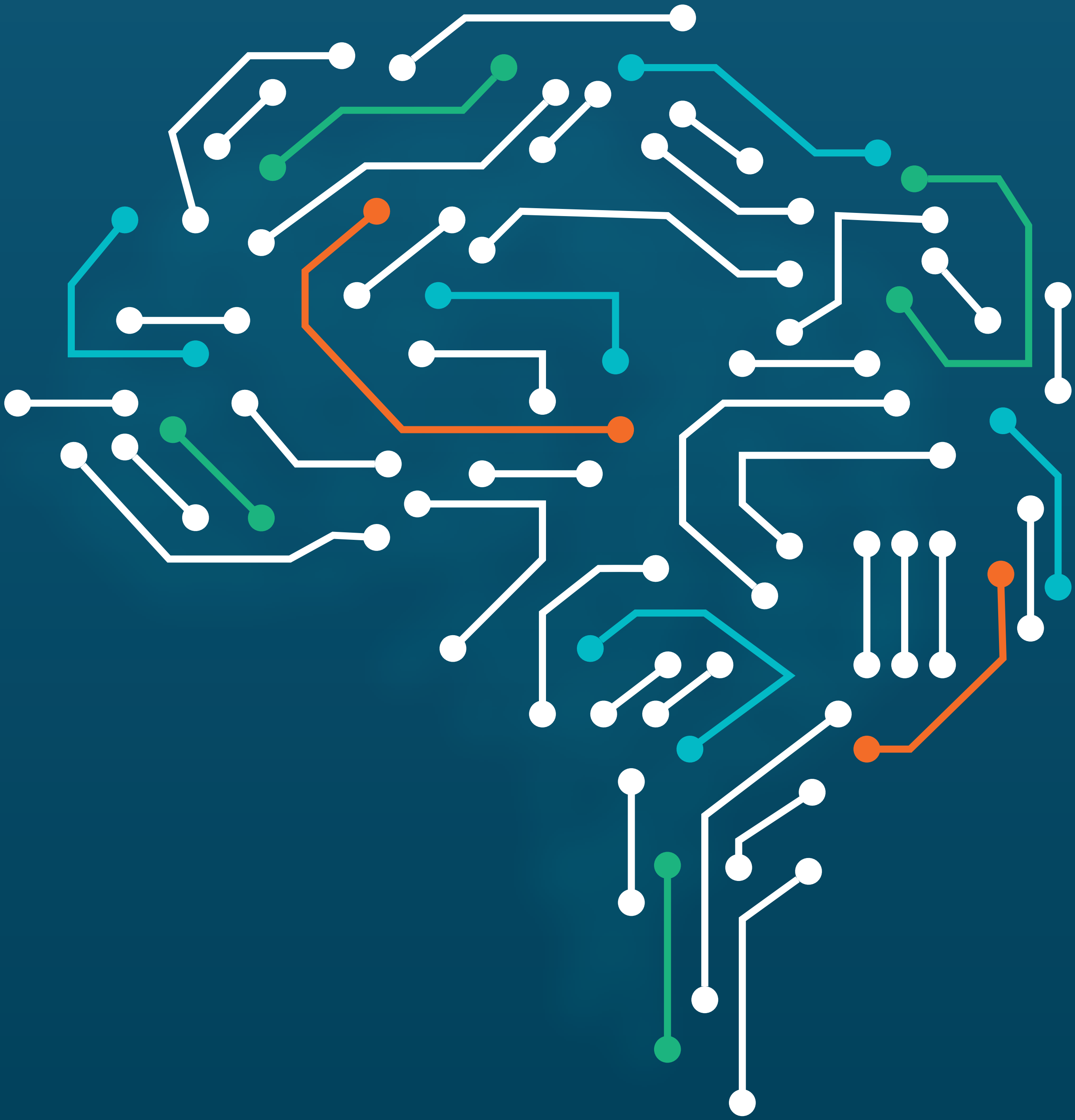
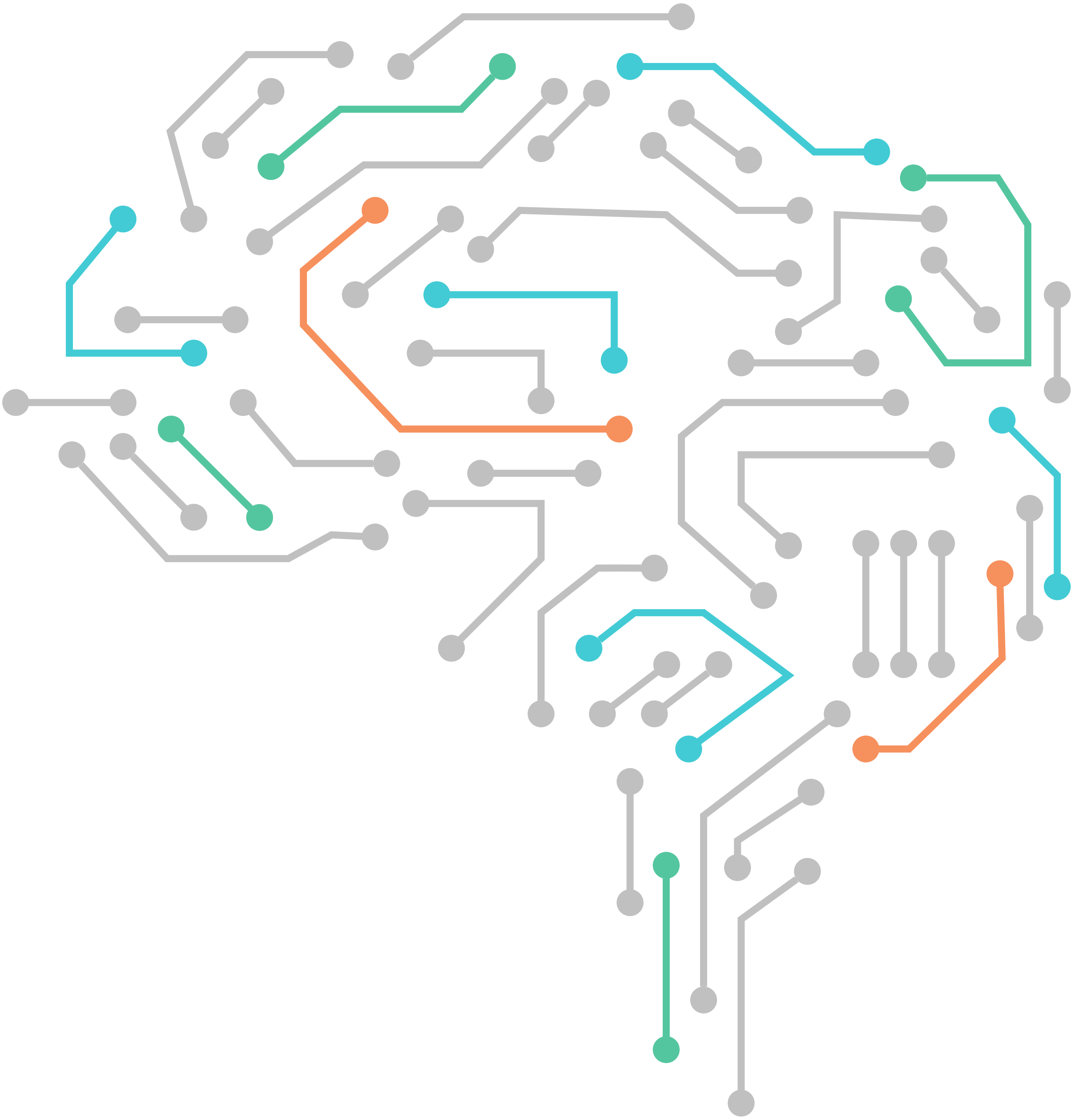


GENERATING EVIDENCE FOR ARTIFICIAL INTELLIGENCE-BASED MEDICAL DEVICES: A FRAMEWORK FOR TRAINING, VALIDATION AND EVALUATION

World Health Organization



GENERATING EVIDENCE FOR ARTIFICIAL INTELLIGENCE-BASED MEDICAL DEVICES: A FRAMEWORK FOR TRAINING, VALIDATION AND EVALUATION

World Health Organization

Generating evidence for artificial intelligence-based medical devices: a framework for training, validation and evaluation

ISBN 978-92-4-003846-2 (electronic version)
ISBN 978-92-4-003847-9 (print version)

Suggested citation. Generating evidence for artificial intelligence-based medical devices: a framework for training, validation and evaluation. Geneva: World Health Organization;2021. Licence: CC BY-NC-SA 3.0 IGO.

Cataloguing-in-Publication (CIP) data. CIP data are available at http://apps.who.int/iris.

Sales, rights and licensing. To purchase WHO publications, see http://apps.who.int/bookorders. To submit requests for commercial use and queries on rights and licensing, see https://www.who.int/copyright.

Third-party materials. If you wish to reuse material from this work that is attributed to a third party, such as tables, figures or images, it is your responsibility to determine whether permission is needed for that reuse and to obtain permission from the copyright holder. The risk of claims resulting from infringement of any third-party-owned component in the work rests solely with the user.

General disclaimers. The designations employed and the presentation of the material in this publication do not imply the expression of any opinion whatsoever on the part of WHO concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. Dotted and dashed lines on maps represent approximate border lines for which there may not yet be full agreement.

The mention of specific companies or of certain manufacturers’ products does not imply that they are endorsed or recommended by WHO in preference to others of a similar nature that are not mentioned. Errors and omissions excepted, the names of proprietary products are distinguished by initial capital letters.

All reasonable precautions have been taken by WHO to verify the information contained in this publication. However, the published material is being distributed without warranty of any kind, either expressed or implied. The responsibility for the interpretation and use of the material lies with the reader. In no event shall WHO be liable for damages arising from its use.

Design and layout by Inis Communication

CONTENTS

List of figures

List of tables

FOREWORD

Artificial intelligence (AI) has potential to optimize the delivery of healthcare and improve outcomes for all. For countries which have yet to achieve universal health coverage, data-driven technology will play a vital role in the next decade. Current AI, machine learning and deep learning applications include the use of clinical decision support tools, diagnostics, and workflow optimisation solutions. AI is also being used to enhance health research and drug development, and in assisting with the deployment of different public health interventions, such as disease surveillance, outbreak response, and health systems management.

AI could greatly benefit low- and middle-income countries, especially in those countries that may have significant gaps in health care delivery and services. AI-based tools and data-driven technology as a whole could help governments extend health care services to underserved populations, improve public health surveillance, and enable healthcare providers to better attend to patients and engage in complex care.

For AI to have a beneficial impact on public health and medicine, ethical considerations must be placed at the centre of the design, development, and deployment of AI technologies for health. The evidence generated from the development and deployment of these devices must be robust and transparent, supporting claims for safety and performance. AI must be generalisable and work to improve outcomes for all populations. Existing biases in healthcare based on race, ethnicity, age, socioeconomic status and gender, that are encoded in data used to train algorithms, must be overcome.

Those same standards for development, deployment and post-market surveillance of AI tools must be applied in the global health context, especially in LMIC populations where governance and regulatory structures for the use of these devices is still evolving. This framework serves as a foundation document and considers minimum requirements for clinical evidence generation in three phases: 1) Software Development, 2) Software Validation and Reporting, and 3) Deployment and Post-Market Surveillance. It uses cervical cancer screening as a use-case to demonstrate the evidence generation considerations. This use-case is appropriate, given the enormous task ahead to eliminate cervical cancer, which remains one of the most common cancers and causes of cancer-related death in women across the globe, even though It is a preventable disease.

As recognised in WHO's *Global strategy to accelerate the elimination of cervical cancer as a public health problem,* high quality diagnostics and research on artificial-intelligence-based technology are key tools for achieving the target to screen at least 70% of all women with a high performance test by 2030. AI-based tools are being applied to address diagnosis of numerous cancers. Where cervical cancer is concerned, we have a chance to show the potential of technology to improve cancer outcomes on a global scale.

Finally, I would like to thank all experts and stakeholders who made essential contributions to the development of this document. I hope that this document will help guide the development of safe and high performing AI for health, with ethical research and evidence generation at its core. This will enable all populations to benefit from the great promise of these technologies in the future.

Soumya

Dr Soumya Swaminathan, Chief Scientist, World Health Organization

ACKNOWLEDGEMENTS

This framework document was developed under the coordination of the Science Division, WHO Department of Digital Health and Innovation. The use-case was chosen as part of the WHO Cervical Cancer Elimination Initiative, Office of the Director-General, and WHO would like to thank Princess Nothemba (Nono) Simelela, Groesbeck Parham, Richard Freeman, Faya El Sheikh, Claire Kimlu and Cherian Verghese. This document was produced after internal consultations with other technical department colleagues: Nathalie Broutet, Elena Fidarova, Raymond Hutubessy, Shona Dalal, Morkor Newman, Lara Vojnov and more. The framework draws together all current standards and frameworks, and references the wide range of academic, regulatory and policy organisations doing work in this area. WHO carefully acknowledges funding support from UNITAID.

Dr Stephanie Kuku (Digital Health, WHO) drafted the document, along with input from Tina Purnat, Sameer Pujari and Bernardo Mariano and other internal contributors Adriana Velasquez, Anita Sands, Yamuna Mundade, Paul Bloem, Denis Falzon, and Marta Seoane Aguilo.

WHO would like to thank the following external experts who contributed to the review of this document: Cameron Po-Hsuan Chen (Technical Lead Manager, AI/ML Researcher, Google Health, San Francisco); Chris Kelly (Senior Clinical Research Scientist, Google Health, London); Hugo Aerts (Director, Artificial Intelligence in Medicine Program (AIM), Harvard University, Boston); Danielle Bitterman (Postdoctoral Researcher, AIM, Boston); Doreen Romogola-Masire, (Professor of Obstetrics and Gynaecology, University of Botswana); Xiao Liu (Clinical AI Researcher, Birmingham University and CONSORT-AI and SPIRIT-AI Consortium); Lynne Denny (Department of Obstetrics and Gynaecology, University of Cape Town,); Anurag Agarwal (Director, CSIR Institute of Genomics and Integrative Biology, Delhi); Rigveda Kadam (Head of Digital Access, FIND, Geneva); Dean Ho (Professor of Biomedical Engineering and Pharmacology, and Director, N.1 Institute for Health, National University of Singapore); Adrian Pacheco-Lopez (Telehealth Director, Ministry of Health, CENETEC-Salud, Mexico City); Nina Schwalbe, Heilbrunn Department of Population and Family Health, Columbia Mailman School of Public Health, New York.)

ABBREVIATIONS

AI	Artificial intelligence
AI-SaMD	Artificial intelligence-based software as a medical device
ASCUS	Atypical squamous cells of undetermined significance
AUC	Area under the receiver operating characteristic curve (also known as AUROC)
AVE	Automatic visual evaluation
CAD	Computer assisted detection
CDS	Clinician decision support
CEAR	Clinical evaluation assessment report
CER	Clinical evaluation report
CHWs	Community health workers
CIN	cervical intraepithelial neoplasia
CONSORT	Consolidated Standards of Reporting Trials
DCNN	Deep convolutional neural network
EQUATOR	Enhancing the Quality and Transparency of Health Research
EU	European Union
FDA	United States Food and Drug Administration
HCPs	Health care professionals
HCW	Healthcare worker
HIC	High-income countries
HPV	Human papillomavirus
HSIL	High grade squamous intraepithelial lesion
HTA	Health technology assessment
HTAs	Health technology assessments
IAS	Image acquisition systems
ICH-GCP	International Conference on Harmonisation – Good Clinical Practice
IMDRF	International Medical Device Regulators Forum
ISO	International Standards Organisation
ITU	International Telecommunication Union
LLETZ	Long loop excision of the transformation zone
LMIC	Low-and-middle income country
LSIL	Low-grade squamous intraepithelial lesion
MDCG	Medical Device Coordination Group (European Union)

MDR	European Union regulations for medical devices
MHRA	Medicines and Healthcare products Regulatory Agency (United Kingdom of Great Britain and Northern Ireland)
ML	machine learning
NHS	National Health Service (United Kingdom)
NICE	National institute of clinical excellence (United Kingdom)
NLP	Natural language processing
NPV	Negative predictive value
NSC	National Screening committee (United Kingdom)
PMCF	Post-market clinical follow-up
PMCF	Post-market follow-up
PMS	Post-market surveillance
PPV	Positive predictive value
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
QMS	Quality management systems
RCT	Randomised controlled trials
ROI	Region of interest
S&T	Screening and testing
SCJ	Squamocolumnar junction
SPIRIT	Standard Protocol Items: Recommendations for Interventional Trials
TPLC	Total Product Life Cycle
TZ	transformation zone
VIA	Visual inspection with acetic acid

EXECUTIVE SUMMARY

The development of artificial intelligence and machine learning based software as a medical device (we will refer to these as AI-SaMD in this document) is rapidly evolving. However, there is currently a lack of globally recognised benchmarking frameworks to assess evidence generated by the use of these devices. International regulatory frameworks for digital health products are also evolving.

The framework provides an overview of considerations used in evaluating clinical evidence regarding AI-SaMD, aiming to help formulate a consensus for guiding validation, evidence generation and reporting across the total product life-cycle within a global health context.

Section I. Development. Chapters 2 to 6 cover evidence generation considerations and minimum standards for AI-SaMD development.

Section II. AI Software Validation and Reporting. Chapters 7 to 10 cover evidence generation during AI-SaMD testing, including data management and evidence reporting.

Section III. AI Software Deployment. Chapters 11 to 14 cover evidence generation considerations for deployment, usability, and post-market surveillance.

Chapter 15 describes evidence generation requirements for procurement of AI-SaMDs in the global health context, in order to ensure that safety, performance within the clinical context, and clinical impact related to intended use are clearly demonstrated.

As well as reviewing the current literature, this framework provides a real-world example in the form of a use-case for AI-SaMD applied to a WHO priority: cervical cancer screening. (Use-case methodology is a systems analysis approach to clarifying product requirements, describing the complete sequence of steps required to reach a specified goal). Several chapters also list minimum standards for different aspects of evidence generation.

1. INTRODUCTION

In July 2020, WHO's Department of Digital Health and Innovation published its draft strategy setting out its long term goals and objectives *(1)*.

Strategic objective 3, *"Strengthen governance for digital health at global, regional and national levels"* aims to improve the assessment and monitoring of research about the application of digital health tools. Evidence generation to demonstrate the health outcomes and impacts of digital health tools is essential to support safe implementation, to establish and promote accountability, and to justify financial investments. It also addresses the need to stimulate the development and testing of technologies, methodologies that allow for comparison, and infrastructure to overcome obstacles to the prioritisation of such technologies for global health.

This objective will be met initially by the publication of frameworks for development and evidence generation, benchmarking, regulating and adoption of digital health tools, including artificial intelligence and data-driven technologies.

In the past decade, the potential for application of artificial intelligence (AI) and data-driven technologies to health care has been investigated extensively. However, for AI to be accepted and implemented in the treatment of patients, proof of safety and performance from well conducted trials is needed.

This document supports the guideline development process both for national health authorities and the WHO. In particular, it describes key considerations for future use of AI-based medical devices in low-and-middle income countries (LMICs), where AI and data-driven technologies could play an important role in addressing global health inequalities at the individual patient, health system and national levels. As such, it is part of the WHO global digital health strategy, and contributes to the strategy's objectives by bringing together international standards and knowledge. It should be considered closely alongside the 2021 WHO guidance document *Ethics and governance of artificial intelligence for health (2)*.

As the regulatory landscape in LMICs evolves, health technology assessment of AI-based medical devices will consolidate evidence from well-conducted clinical trials to ensure safety and performance requirements are met, and ethical standards applied.

Purpose of the framework

The development of artificial intelligence and machine learning based software as a medical device (we will refer to these as AI-SaMD in this document) is rapidly evolving. The evidence base of potential use-cases and validation processes needed to demonstrate safety and performance are also evolving as these tools are being increasingly piloted and incorporated in low resource settings - low- and lower--middle income countries (LMICs) *(3, 4)*.

There is currently a lack of globally recognised benchmarking frameworks to assess evidence generated by the use of these devices. This evidence should span the total product life-cycle, from development to post-implementation, especially in LMIC settings *(5)*.

International regulatory frameworks for digital health products are also evolving. There is still no consensus globally on what current best practice is *(6)*.

Main areas of application of AI-SaMD that feature published evidence for validation include:

- Portable/smartphone based diagnostics
- Clinician decision support (CDS)
- Workflow optimisation
- Population health
- Pre-clinical research and clinical trials

A standardised framework with open benchmarking for the evaluation of AI-SaMDs, including clinician decision support systems is in development by various groups including the ITU/WHO Focus Group on artificial intelligence for health *(7)*.

Current clinical guidance does not go far enough to enable innovators and end-users to know what evidence generation approaches are appropriate, and practical, for all classes of digital health solutions. This is especially true for AI-SaMD, where unforeseen errors at data entry level can lead to catastrophic effects when deployed at scale if performance errors go unchecked.

The framework provides an overview of the considerations used in evaluating clinical evidence regarding AI-SaMD, highlighting the safety and performance requirements that should be considered before and after deployment. The main purpose of this document is to:

1. Provide a global health context by examining existing available frameworks, guidance for clinical study protocols, and evolving evidence from use-cases, in order to recommend a framework for WHO that will underpin evaluation of AI-driven clinical decision support tools in LMICs.
2. Formulate a consensus for guiding validation, evidence generation and reporting across the total product life-cycle from development to post-market surveillance, within a global health context
3. Support the development of guidance for health technology assessment (HTA) related to the use of AI-SaMD for all use-cases, illustrating this with the use-case for the development of AI-SaMD for application in cervical cancer screening.

Structure

This document is divided into three sections:

Section I. AI Software Development. Chapters 2 to 6 cover evidence generation considerations and minimum standards for AI-SaMD Development.

Section II. AI Software Validation and Reporting. Chapters 7 to 10 cover evidence generation considerations during AI-SaMD Testing, and Reporting. The chapters in this section also covers data management and overall evidence reporting, including international standards and existing guidance for evidence generation and reporting.

Section III. AI Software Deployment. Chapters 11 to 14 cover evidence generation considerations for deployment, usability, and post-market surveillance.

Scope of the framework

The following focus areas are covered in the course of the document:

- Clinical evidence generation for AI-SaMD development, validation, testing and post-implementation monitoring
- Contextual considerations for evidence generation in the absence of regulatory standards in some LMICs
- Considerations for developers building AI-SaMD for deployment in LMICs, whether the development is carried within LMICs or external to them
- Building blocks for evaluating evidence, from phased development of diagnostic algorithms, for use until guidance on LMIC-appropriate assessment and evaluation exists
- Provision of references to relevant existing guidance, guidelines, regulatory standards and scientific publications
- Ethical considerations and approaches to fairness and avoiding bias.

This framework will be reviewed periodically and updated in 2022–2023 as technical standards for evaluation of AI-SaMDs evolve.

Out of scope

Guidance on the regulatory landscape for AI in global health is still evolving. Whilst standards for assessing the performance of AI SaMD can and should feed into regulatory decision-making, this framework is not meant to act as a standardised guideline for evidence requirements, or regulatory initiatives.

The key messages cover evaluation methodologies and considerations, and minimum clinical validation requirements. It should not be considered exhaustive. The framework does not cover requirements related to technical verification and technical files/software and product requirements such as risk management files, version controls, periodic safety update reports etc. Also out of the framework's scope are "adaptive" or "continuous learning' algorithms, an area which is still in early development.

Intended audience

This framework aims to assist global health stakeholders to understand the health technology (HTA) and regulatory (requirements for AI SaMD. These include policy makers with Ministries of Health, industry developers and researchers building AI tools, international stakeholders involved in the implementation of AI tools in global health, and for internal WHO stakeholders. It assumes that readers have at least a basic understanding of the general applications of AI-SaMD.

How this framework was developed

This framework was generated by a lead author selected for expertise in the fields of digital health and artificial intelligence. Content was drawn from internationally recognized standards or sources of guidance, including individuals who made declarations regarding any conflicts of interest. Recommendations are based on evidence-based peer reviewed publications, and reviews with experts and several non-governmental organizations and scientific institutions.

Key takeaways from international multi-disciplinary stakeholder engagement during the creation of this document were as follows:

- Standards for evidence generation of AI-based models and devices can and should inform regulatory decision-making
- There is still some way to go in trying to find a uniform language among stakeholders from different areas of expertise
- Additional input is required from stakeholders in LMICs to make sure their needs are adequately addressed, from data availability to infrastructure and implementation.
- Ethics, fairness and data governance should be a priority in order to prevent further rises in inequalities.

Use-case: cervical cancer screening

WHO's global strategy to accelerate the elimination of cervical cancer as a public health problem makes cervical cancer screening a suitable and justifiable use-case illustration *(2)*. More than 85% of the 311 000 women (2018) who die globally live in LMICs. The WHO states:

> When diagnosed, cervical cancer is one of the most successfully treatable forms of cancer, as long as it is detected early and managed effectively. Cancers diagnosed in late stages can also be controlled with appropriate treatment and palliative care.
>
> With a comprehensive approach to prevent, screen and treat, cervical cancer can be eliminated as a public health problem within a generation.

Relevant WHO publications

WHO has recently published a number of documents that are relevant to his framework document, and to the fast-growing field of artificial intelligence in health.

Global strategy on digital health (1).

2015 global survey on health technology assessment (HTA) by National Authorities (8).

Monitoring and Evaluating Digital Health Interventions (9).

WHO technical guidance and specifications of medical devices for screening and treatment of precancerous lesions in the prevention of cervical cancer (10).

WHO guidance for post-market surveillance and market surveillance of medical devices, including in-vitro-diagnostics (11).

WHO and governance of artificial intelligence for health. Meeting report, 7 March 2021 (12).

SECTION I.
SOFTWARE DEVELOPMENT

2. ARTIFICIAL INTELLIGENCE IN HEALTH

Artificial intelligence (AI) is the broad term used for the capability of machines to perform intelligent tasks; is a subset of AI covering machines capable of learning independently and making accurate predictions *(13)*. In recent years, artificial intelligence has made great progress in the detection, diagnosis, and management of diseases.

Deep learning, a subset of machine learning based on artificial neural networks, has enabled the development of applications that, in controlled settings, have demonstrated performance levels approaching those of trained professionals in tasks including the interpretation of medical images and discovery of drug compounds *(14)*.

Current evidence in health applications

The majority of published evidence to date has consisted of early-phase retrospective validation studies which are in fact *in silico* (i.e. performed by computer, as opposed to *in vivo* and *in vitro*) assessments of datasets used to test performance accuracy of AI algorithms. Currently, most of the regulatory approvals for algorithms by the US Food and Drug Administration (FDA)and the European Commission rely on such preliminary evidence *(14, 15)*. Most recent AI studies have not been adequately reported, and new reporting guidelines identify potential sources of bias specific to AI systems *(16, 17)*.

Prospective studies and randomised controlled trials now evaluating the clinical efficacy and impact of AI-SaMD *(16) (18–24)* have also been met with concerns about evidence generation and reporting *(16) (25–27)*. This has highlighted the need for standardised reporting guidance as well as establishing guidelines for evidence generation.

Geographic distribution of development and implementation

Most AI developments in health care cater to the needs of high-income countries (HICs), where the majority of research is conducted. The majority of AI software for medicine are also developed in HICs, using data from individuals from HICs.

It is acknowledged that AI has the potential to improve medical outcomes in LMICs, where workforce shortages and limited resources constrain access and quality of care. Some evidence is now being generated regarding use of AI algorithms for disease detection in these settings, for example in the detection of TB using chest X-rays *(28)*. AI has the potential to help address complex challenges that underlie poor access to care, quality of care, and poor training of health care workers.

However, deploying models developed in HICs without careful validation in LMICs can introduce and propagate bias *(29)*.

While AI system generalisation to LMIC populations and workflows has so far been limited, this work is absolutely vital to these technologies being adopted safely in these countries. Clinical validation studies showing generalisation of AI-SaMD to new LMIC populations include using AI to detect referable (i.e. high-risk) diabetic retinopathy *(30)*, and applying this new LMIC workflows *(31)*. More work like this is needed as the field evolves.

AI could play an important role in addressing global health care inequities at the individual patient, health system, and population levels *(32)*. However, many challenges must be addressed ahead of wide-spread adoption. Developers may have to be enticed through economic incentives and regulatory action to build or at least validate their solutions in LMIC contexts *(33)*, using data appropriate to local populations. In the field of dermatology, the published validation studies for automated detection of cancer generally use data from mainly white Caucasian patients and are difficult to generalise to black and other minority populations (Fitzpatrick scale V-VI) *(34)*. Evidence generation reporting for all new artificial intelligence tools must therefore include diverse ethnic, racial, age, and sex groups, in order to ensure responsible use of AI in medicine, especially in the global health context *(35)*.

The current dynamic shift in disease burden from communicable to non-communicable diseases underscores the need for high-quality, data-driven care that can rapidly and robustly adapt to these dynamic changes. These emerging challenges have been central to the UN's Sustainable Development Goals, including the aim to reduce, by one-third, premature mortality from NCDs by 2030. AI has the potential to fuel and sustain efforts toward these ambitious goals if used safely and effectively.

Community health workers (CHWs) may eventually use AI to triage patients and identify those requiring close follow-up. Applications also cover laboratory diagnostics, for example, analysing peripheral blood samples to diagnose malaria *(36)*. More applications are expected with the emergence of data-driven pocket diagnostic hardware, including ultrasound probes and microscopes *(32)*.

Cancer screening

Evidence is building for the use of deep learning and neural networks to improve screening, early detection of and overall cancer outcomes. Examples of such applications are wide ranging and include AI systems for screening and triage, diagnosis, prognosis, decision support, and treatment recommendation. Evidence is building in the literature for feasibility of AI systems in screening for skin cancer *(37–39)*, lung cancer *(40)*, breast cancer *(41)*, cervical cancer *(42–45)* and numerous other malignant and premalignant conditions.

Also evolving is evidence for validation of the use of AI and ML-based clinician decision support for screening of cancers and precancerous lesions. Industry-institutional research collaborations and private-public partnerships have also accelerated the rate of scientific publications on validation of AI systems for clinical decision support *(46–48)*.

At the same time, there is concern that AI may worsen disparities in cancer outcomes within populations and between HIC vs LMICs, especially given such challenges as widening access to data and proving generalisability of models and algorithms *(32) (49).*

Medical imaging

The evidence for the use of AI in medical imaging now surpasses that for other applications. Whilst most published evidence is retrospective, ongoing trials of external validation and prospective evaluations are underway. A systematic review of studies published in 2019 found that few studies featured external validation or compared the AI performance with that of health care professionals (HCPs) using the same sample. This review also found that poor reporting of evidence is prevalent in these studies and limits reliable interpretation of performance measures *(50)*.

Within the global health context, there is building evidence demonstrating the potential of deep learning-based automated classification of chest X-ray and CT images to detect tuberculosis and radiological signs of COVID-19 lung disease *(51)*.

Organisations like RAD-Aid are working on infrastructure to enable the incorporation of AI into radiology workflows in LMICs *(52)*. The challenges of integrating AI-SaMDs into workflows, especially in low resource settings, are as challenging as developing high performing devices themselves.

Breast cancer screening

Earlier detection of cancers with population-wide mammography screening has decreased mortality from breast cancer *(53, 54)*. This has not yet been widely introduced in LMICs due to the lack of national screening programmes. In HICs, AI-based computer assisted detection (CAD) is now being widely piloted to aid radiologists in screening, given the global shortage of radiologists. Early use of CAD in breast cancer screening was one of the first examples showing that models that performed well in retrospective studies can have unexpected and even harmful outcomes in the real world *(55)*. Whilst to date there have been no published randomised controlled trials (RCTs), a recent paper reported on external validation of three commercially available AI systems for assessment of screening mammograms *(56)*. Only one of the three AI systems performed adequately, compared to radiologists. This work will need further evidence generation from prospective trials. Similar external validation work needs to be done for LMIC datasets, as well as prospective trials.

Gastrointestinal endoscopy

AI-SaMD for colorectal polyps (benign or precancerous lesions) detection has received regulatory approval in Europe and Japan and is being evaluated by other regulatory bodies. Gastroenterology has leapt ahead of other medical fields due to the several randomised trials of AI interventions that reported clinically meaningful outcomes.

Although most advances in the use of AI in gastrointestinal endoscopy have been in colon polyp detection, several other potential applications exist in which AI could aid gastrointestinal endoscopists in detecting early precancerous and malignancy in the oesophagus and stomach. As of September 2020, here have been only seven RCTs to date evaluating clinical impact of AI-SaMDs - in any medical specialty. Of these, five are in automated visual evaluation in gastrointestinal endoscopy *(15)*. Table 1 summarizes the evidence generated so far in this application, with the randomized trials of AI deep neural networks in endoscopic screening for colorectal polyps. In early 2021, the United States Food and Drug Administration (FDA) approved the first AI-SaMD to assist clinicians in detecting suspicious lesions in the colon in real time during colonoscopy *(57)*.

Table 1. Randomized trials of AI deep neural networks in endoscopic screening

Procedure	Detection	Design	Patients	Trial Sites	Place	Citation
Colonoscopy	Adenomas	Double-bind sham control	1046	1	China	Wang P et al, Lancet Gastro Hep 2020 *(23)*
Colonoscopy	Adenomas	Unmasked	704	1	China	Gong D et al, Lancet Gastro Hep 2020 *(21)*
Colonoscopy	Adenomas	Unmasked	659	1	China	Su et al, Gastro Endoscopy 2020 *(22)*
Esophagogastroduodenoscopy	Blind spots	Unmasked	324	1	China	Wu L et al, Gut 2019 *(19)*
Colonoscopy	Adenomas	Unmasked	1058	1	China	Wang P et al, Gut 2019 *(18)*

Source: Topol EJ. *Welcoming new guidelines for AI clinical research (15)*

Use-case: AI-SaMD in cervical cancer screening

Throughout this framework, the application of AI-SaMD to cervical cancer is featured as a use-case example.

The term “use-case” will be less familiar than “case study” to many readers of this framework, as it is more widespread in the worlds of software development and marketing than in clinical research or public health. Use-case methodology is typically applied in system analysis to explore and clarify product requirements, particularly potential interactions between systems and users with a given environment. The result is usually a report describes the complete sequence of steps required for a user to reach a specified goal *(58)*. Use-cases are typically written by business analysts and are used in various phases of software development, from exploring system requirements to testing applications and preparing user manuals or web-based help. While there is extensive literature on use-cases, a reader-friendly description makes the following distinction *(59)*:

> The difference between Case Studies and Use-cases is the difference between what is real and what is possible. Case Studies are real life, retrospective accounts of real projects that you have delivered for real customers. Use-cases are examples of how a product or service might be deployed. So in that sense Case Studies tell the stories of real customer tried and tested solutions, while Use-cases present examples of solutions to possible problems.

The challenge of cervical cancer

Cervical cancer is the fourth most common cancer among women globally, with an estimated 570 000 new cases and 311 000 deaths in 2018. There are huge global disparities in the burden of cervical cancer. Data for 2018 show that age-standardized cervical cancer incidence rates varied from 75 per 100 000 women in the highest-risk countries to less than 10 per 100 000 women in those with lowest-risk *(60)*.

To achieve elimination this century, the WHO has created a Cervical Cancer Elimination Strategy setting out its “90-70-90” targets to be met by 2030 (see text box).

WHO's global strategy to accelerate the elimination of cervical cancer as a public health problem

Cervical cancer is one cancer the world can actually eliminate: it's time to do it. Following a Call to Action in May 2018 from the World Health Organization (WHO) Director-General, Dr Tedros, 194 countries collectively resolved to end needless suffering from a cancer that is both preventable and curable. The world already has the necessary tools; they just need to be made accessible.

Global strategy to accelerate the elimination of cervical cancer as a public health problem

World Health Organization

In August 2020, the World Health Assembly passed a resolution calling for elimination of cervical cancer and adopting a strategy to make it happen *(61)*. It is a testament to the enthusiasm for this important goal that, even in the context of the COVID-19 pandemic, countries around the world have affirmed their support for this important priority.

This global strategy to eliminate cervical cancer proposes:

- a vision of a world where cervical cancer is eliminated as a public health problem;
- a threshold of 4 per 100 000 women-years for elimination as a public health problem;
- the following 90-70-90 targets that must be met by 2030 for countries to be on the path towards cervical cancer elimination:
- 90% of girls fully vaccinated with human papillomavirus (HPV) vaccine by age 15 years.
- 70% of women are screened with a high-performance test by 35 years of age and again by 45 years of age
- 90% of women identified with cervical disease receive treatment (90% of women with precancer treated, and 90% of women with invasive cancer managed).
- a mathematical model that illustrates the following interim benefits of achieving the 90-70-90 targets by 2030 in low- and lower-middle-income countries:
- median cervical cancer incidence rate will fall by 42% by 2045, and by 97% by 2120, averting more than 74 million new cases of cervical cancer;
- median cumulative number of cervical cancer deaths averted will be 300 000 by 2030, over 14 million by 2070, and over 62 million by 2120.

AI-SaMD use-case in cervical cancer screening

Similar to endoscopy and not unlike many other types of physical examination or imaging, health care professionals must triage, manage, and make diagnostic decisions in real-time (for example, whether to take a biopsy or not) when screening for cervical cancer. This is particularly pertinent in low-resource, infrastructure poor settings. During screening, clinical decisions are usually made immediately and are influenced by visual findings on initial assessment using visual inspection with acetic acid (VIA). The diagnostic accuracy of this procedure is variable and dependent on the training, experience and judgement of the health care professional. This variability id more significant in LMICs due to workforce shortages and a lack of widespread national screening programmes.

Computer vision applied to VIA has the potential to act as clinical decision support and improve visual interpretation. The goal is to provide an accurate, almost instantaneous prediction of whether a high risk precancerous lesion (CIN 3/CIN 2- HSIL) is present in the cervix, or to identify features of low-risk abnormalities (CIN1 - LSIL). (NB. High grade squamous intraepithelial lesion (HSIL) covers the conditions formerly called cervical intraepithelial neoplasia (CIN), including CIN2, CIN3, carcinoma in situ, and moderate and severe dysplasia. Low-grade squamous intraepithelial lesion (LSIL) are also known as mild dysplasia.)

There are ongoing international investigations of methods to automate current methods of cervical assessment (VIA, colposcopy) in order to increase accuracy of detection of high risk precancerous lesions, and streamline the intense monitoring and evaluation required for community health providers *(42) (62)*. Numerous mobile devices and digital colposcopes exist on the market in LMICs for assessing the cervix when screening for precancerous lesions, and the evidence in the literature is evolving. AI-SaMD is also being applied to cytopathology for the assessment of cervical smears by various international groups. To our knowledge, there are currently no validation studies in the LMICs of these applications.

3. FRAMEWORK FOR EVALUATION

Evaluation of digital health interventions AI-SaMD should focus on generating evidence that can be used as a basis for assessing whether observed changes in behaviour, processes or health outcomes can be attributed to the intervention *(9)*.

In general, this involves a combination of types of evaluating evidence that are generated in order to ask questions such as:

Usability

- Is the device usable by the targeted end-users, and does it fit within their workflow?
- What are the requirements of the end-user? Are there any barriers to becoming proficient in using the device as intended? Are instructions for use available as well as a structured "onboarding" (i.e. induction or introductory training) system?
- Is ongoing training available as different versions of the device become available?
- What are the rates of error – in using the system or in workflows – as a result of system use/misuse?
- What are the unintended consequences of using the device?
- What is the relationship between human-AI interaction and how can outcomes (positive or negative) be attributed to the device?

Efficacy

- Has the digital health intervention changed processes or output measures (e.g. time between event X and response Y) in a research setting?
- Has the digital health intervention changed outcomes (e.g. health care worker performance, or patient health outcomes) in a research setting?

Effectiveness

- Has the device changed processes (e.g. time to diagnosis) in the intended setting?
- Has the device changed outcomes (e.g. user accuracy/efficiency, or patient health outcomes) in the intended setting?
- Is the device cost-effective?

Evaluation components

When planning evidence generation, the validation pathway for a given AI-SaMD should aim to outline the process for study design and reporting of clinical data that meet research goals for the device. This will ensure that projects generate a robust evidence base for the components of the device that are of greatest value to stakeholders, and that claims are grounded in the evidence base. Table 2 illustrates a high-level outline of an evaluation framework.

The validation pathway should consider, within the context of the intended use, how the evidence generated will enable stakeholders to make an overall *impact assessment*.

Table 2. Framework for evaluation of an AI-SaMD

Define how the product works (intended use)	Model/flow chart to illustrate device's intended use, desired benefits, operating principle and anticipated risks Define and highlight the hypothesis, assumptions and barriers in LMIC contexts
Design the evaluation	Design of evaluation - clinical trial, clinical study or clinical investigation to illustrate safety and benefit of the AI-SaMD Research ethics - check for bias/fairness in design stage
Choose clinical evaluation methods	Clinical studies: Descriptive: retrospective analysis of routinely collected data/audit/ user feedback Comparative: retrospective, prospective, reader studies, RCTs Qualitative: usability, cost-effectiveness, benefit -risk analysis
Execution of evaluation	Steering group featuring a multi-disciplinary team (all stakeholders) Pre-specified study protocol, data governance, research governance, feasibility checking and piloting, monitoring (e.g. of recruitment)
Analysing data	Analysing qualitative data Preparing quantitative data: (pseudo) anonymization/ outliers/ missing data/sensitivity analysis Analysing quantitative data: descriptive and inferential statistics
Reporting results	Based on reporting guidelines and standards Ensure explainability of AI Peer-review of publications Include recommenwwwwdations on: Anticipated risks (LMIC context) Future evaluation studies Quality improvement Implementation (deployment) Post- deployment monitoring and surveillance

Adapted from Public Health England, *Evaluating digital health products (6)*

Clinical evaluation

Reliable evaluation of AI-SaMD in LMICs is needed before implementation in order to reduce patient and health system risk, and facilitate widespread adoption.

International regulators have proposed frameworks for ensuring the safety and effectiveness of these devices, but these are not exclusively focused on the needs of LMICs. Regulators including the FDA have been guided by the Global Harmonisation Task Force, established in 1993, and superseded in 2011 by the International Medical Device Regulators Forum (IMDRF). The IMDRF recommends that a clinical evaluation must demonstrate "safety, clinical performance and effectiveness of the device" *(63)*. The European Union has codified this requirement for a clinical evaluation report, and also requires manufacturers to prepare and follow a post-market clinical follow-up plan.

The FDA has published a precertification program, a "predetermined change control plan" to monitor changes to an algorithm made after deployment, and has delineated the data to be used in post-release testing and refinement. This clinical evaluation pathway requires an understanding of assessment of clinical data inputs and output from an AI-SaMD, as well as an understanding of the components required to assess safety and performance in both published work in the scientific literature as well as clinical data from devices.

The IMDRF maps these out three phases of evaluation demonstrated in the figure above. Table 3 illustrates the evaluation phases, and the method used to generate evidence.

Table 3. Clinical evaluation methods used to produce desired evidence

Evaluation phase	Question answered	Method used to generate evidence
Valid clinical association	Is there a plausible scientific explanation for your device's use-case?	Scientific literature review
Technical/ analytical validation	Does your device behave correctly in test conditions?	Verification testing of software against technical performance requirements
Clinical validation	Does your device behave correctly in the clinical setting?	Clinical Investigation (Studies) Post Market Clinical Follow up

Adapted (third column added) from IMDRF, *Software as a Medical Device (SaMD) (64)*

Determining the primary outcome (i.e. the result of greatest interest) helps decide what type of clinical evaluation is required. For example, is the primary outcome whether the AI outperforms humans, or is it a more patient-centred outcome (e.g. survival)? Deciding this requires a nuanced discussion regarding intended use. In general, when improving efficiency is not the only goal, patient-centred outcomes should be prioritized.

Use-case: Evaluation for cervical cancer diagnosis

The evaluation of an AI-SaMD to aid in cervical cancer diagnosis should attempt to attribute a range of outcomes to the device over time. This may run from assessing how easily end-users can interact with the system (usability), to measuring the health impacts (efficacy/effectiveness) and calculating the affordability of the system (economic/financial evaluation).

In later stages of development maturity, evaluation questions will arise concerning how the system and its data streams will be integrated within the broader health system architecture and policy environment, with the ultimate goal being to reach and sustain its use on a national scale (implementation research) *(9)*.

Table 4. Evaluation components for AI-SaMD in cervical cancer screening

Formative evaluation	Primary outcome determination AI model development: training/internal validation studies: Feasibility, efficacy study (controlled setting)
Summative evaluation	AI model validation: efficacy/effectiveness studies Effectiveness study (real-world setting, prospective)
Intended use risk categorisation	Does the AI-SaMD assist, inform or drive clinical management during cervical cancer screening? (see Chapter 3: Intended Use)
Population and data	Context (LMIC) and geography (US, EU vs Brazil, India, Africa, etc.) Socio-economic considerations - infrastructure, data governance tools Clinical setting: cervical cancer screening, “screen and treat” clinical pathway *(65)*
Outcome targets	Detection of precancerous lesions of the cervix

4. INTENDED USE

The intended use of an AI-SaMD should define, as clearly as possible, information pertaining to when, where and how it is to be used. This enables evidence generated to be evaluated in the right context for safety and performance requirements.

The manufacturer's documentation that defines the device should include a thorough discussion of intended use which answers the following questions *(64, 66, 67)*:

- What exactly does your AI-SaMD claim to do?
- What is the intended *medical indication(s)*?
- What is the intended use in the *diagnostic process* - e.g. triage vs diagnosis?
- What is the intended *patient population(s)* on which it will be used?
- What is the intended *part of the body/site/tissue* with which it will interact?
- What is the profile of the *intended user(s)*?
- What is the intended *operating environment*?
- What is the intended *operating principle* of the software?
- What is the foreseeable misuse (and mitigation) of the device?

The intended use should also state the AI system's model type and architecture, and its use in the context of the clinical pathway. As an AI-SaMD may be replacing, augmenting or assessing components of clinical decision making, it is important to understand the intended use clearly in order to define the purpose for which the AI-SaMD will be evaluated.

Risk classification

A risk based approach is essential for independent review of evidence generated from the evaluation of AI-SaMD. The FDA categorizes medical devices into one of three classes – Class I, II, or III – based on their risks and the regulatory controls necessary to provide a reasonable assurance of safety and effectiveness. The IMDRF has proposed a risk classification schema, shown in Table 5, which classifies increasing potential risk based on the significance of the information (outputs) provided by an SaMD for health care decision making *(64)*. Its four risk categories depend on the health care condition severity (urgent, serious, critical), the health care decision to be made (to inform/drive clinical management, or to treat/diagnose). In the EU this classification is used to determine the IMDRF risk class and thereby the depth of evaluation /assessment procedure.

The Intended Use of a SaMD is the basis for determining its risk. In order to permit comprehensive evaluation, the intended use, operation principle, and foreseeable misuse must be clearly documented in the context of its use and expected impact on health outcomes.

Table 5. IMDRF Risk Categorisation

Health care situation or condition	Significance of information provided by SaMD to assist health care decision. Risk classes I to IV (lowest to highest risk)		
	Treat or diagnose	Drive clinical management	Inform clinical management
Critical	IV	III	II
Serious	III	II	I
Non-serious	II	I	I

Source: IMDRF, *Software as a Medical Device (64)*

Changes to intended use

The intended use of AI-SaMDs may evolve and change during development and evidence generation. The FDA sets out three conditions for which new evidence needs to be reviewed when such modifications are made *(68)*. These types of modifications have the ability to affect the safety and performance of the device, and thus require re-evaluation. These modifications include changes to the intended use of the device, which may or may not accompany changes in performance (clinical and/or analytical), as well as changes in inputs used by the model and their clinical association to the AI-SaMD's output.

Considerations for global health

Detailed analysis has been lacking on how best to deploy and effectively scale up AI solutions in health systems across LMICs. Historically, it has proven very challenging to take disruptive technology innovations from high-income countries and deploy and scale them so that they address the unique needs of populations in low-income environments, and have positive impacts.

There is no current acceptable performance standards, accuracy rates, or patient health outcome benchmarks against which to measure and compare AI-SaMDs in LMICs. Further, perspectives vary widely on what standards AI tools should adhere to in order to be used across LMICs, and for patients and doctors to have trust in them. There is also debate over variability in accuracy standards across different types of AI tools.

Acceptable accuracy rates for algorithms are likely to vary depending on clinical area, given that physician performance in diagnosis and treatment is itself highly variable. The Hippocratic oath– *do no harm* – must be the guiding principle for all efforts to scale AI tools. Yet, how to operationalize this principle is unclear since there is still insufficient evidence from the use of AI-SaMD in LMICs to know which tools might have potential to cause harm, including if misused *(31)*.

Clearly defining the intended use of the AI-SaMD being tested and deployed in LMICs, and doing so in the context of the complete clinical pathway and the infrastructure available, will ensure robust evidence is generated to evaluate safety and performance metrics. Wahl et al discuss these challenges and describe some early AI algorithms deployed in LMICs, such as those to assist predicting birth asphyxia and estimate the spread of dengue fever *(3) (69)*.

Minimum standards for defining intended use

- What is the medical indication for use of the AI-SaMD?
- What part of the body/ system is being investigated?
- What use environment will the device be used in?
- What are the ways in which the device can be misused?
- For what patient population?
- What is the specific user profile and expertise?
- What is the exact operating principle of the device?
- What are the possible unintended consequences of the device and how will these risks be mitigated?

Use-case: Defining intended use for AI-SaMD in cervical cancer screening

This part of the use-case examines what must be included when developing intended use descriptions for AI-SaMD in cervical cancer screening, with particular emphasis on device description and architecture.

Description of the AI-SaMD software and hardware:

- Image capture hardware
- Data storage and image quality classification algorithm
- Mobile phone with built-in AI-SaMD application
- Application with integrated software housing an AI-based algorithm trained to detect (high grade) precancerous lesions (CIN3/HSIL) of the cervix
- Output display- Probability Score, normal vs abnormal, etc.

Considerations for defining intended use:

- Expectations of how the user will interact with the device or algorithm output (must be clearly specified as this can vastly affect the performance of the model).
- Intended users (nurses, doctors, gynaecologists, etc. or levels of expertise and training)
- Intended use context (e.g. clinic, community, hospital, etc.)
- Principles and operating procedures of the device (illustrated for users in instruction manuals)
- Foreseeable misuse, (e.g. if the device is used as a sole reader without input of the technical staff or health care worker's clinical judgement, risking subsequent mis-management of a patient).

Considerations for defining intended use within the cervical cancer screen and treat (S&T) Pathway:

- Triage (or detection), as an assist to visual assessment by a health care worker to improve detection of precancerous lesions of the cervix
- Indications for use:
- Primary screening of women in the general population
- Secondary screening for triage of HPV-screened positive women (high risk group)
- Will the device be used as a clinical decision support in combination with other tests (VIA, HPV Testing) or as an isolated test?
- How will the performance of the AI-SaMD be assessed as a combination modality therapy in terms of intended use?

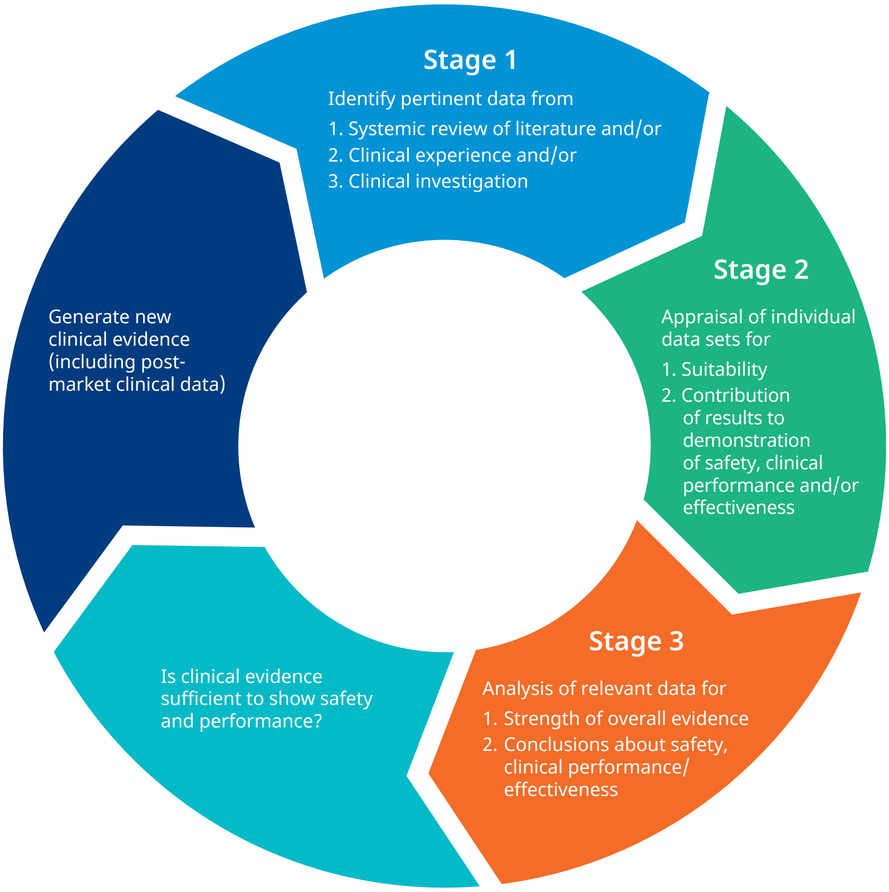
Table 6. IMDRF risk categorisation for AI-SaMD use in cervical cancer screening

State/Stage of healthcare condition **Screening of pre-cancer/ cervical cancer detection**	**Significance of information provided by SaMD to health care decision**		
	Treat or diagnose (Minimal/no clinical input)	Drive clinical management (guides subsequent management)	Inform clinical management (partial assessment only, with clinical assessment overrides)
Critical (e.g. cancer)	IV	III	II
Serious (e.g. early in-situ disease/precancer)	III	II	I
Non-serious (e.g. cervicitis/ other abnormalities)	II	I	I

5. MODEL DEVELOPMENT AND TRAINING FOR CLINICAL EVALUATION

The evidence generation pathway for AI-SaMD begins from model development, i.e. training of the algorithm that is tasked with predicting an output/outcome. Inconsistent and incomplete reporting of evidence remains one of the barriers to the assessment of impact of all digital health interventions, especially in the LMIC context. Using frameworks to set out the objectives of the evidence generation process enables robust collection of data throughout the product life cycle of an AI-SaMD. Figure 1 below illustrates the stages of clinical evaluation identified by IMDRF and the corresponding evidence generation activity or question.

Figure 1. Evidence generation and stages of clinical evaluation



Stage 1

Identify pertinent data from

1. Systemic review of literature and/or
2. Clinical experience and/or
3. Clinical investigation

Stage 2

Appraisal of individual data sets for

1. Suitability
2. Contribution of results to demonstration of safety, clinical performance and/or effectiveness

Stage 3

Analysis of relevant data for

1. Strength of overall evidence
2. Conclusions about safety, clinical performance/ effectiveness

Is clinical evidence sufficient to show safety and performance?

Generate new clinical evidence (including post-market clinical data)

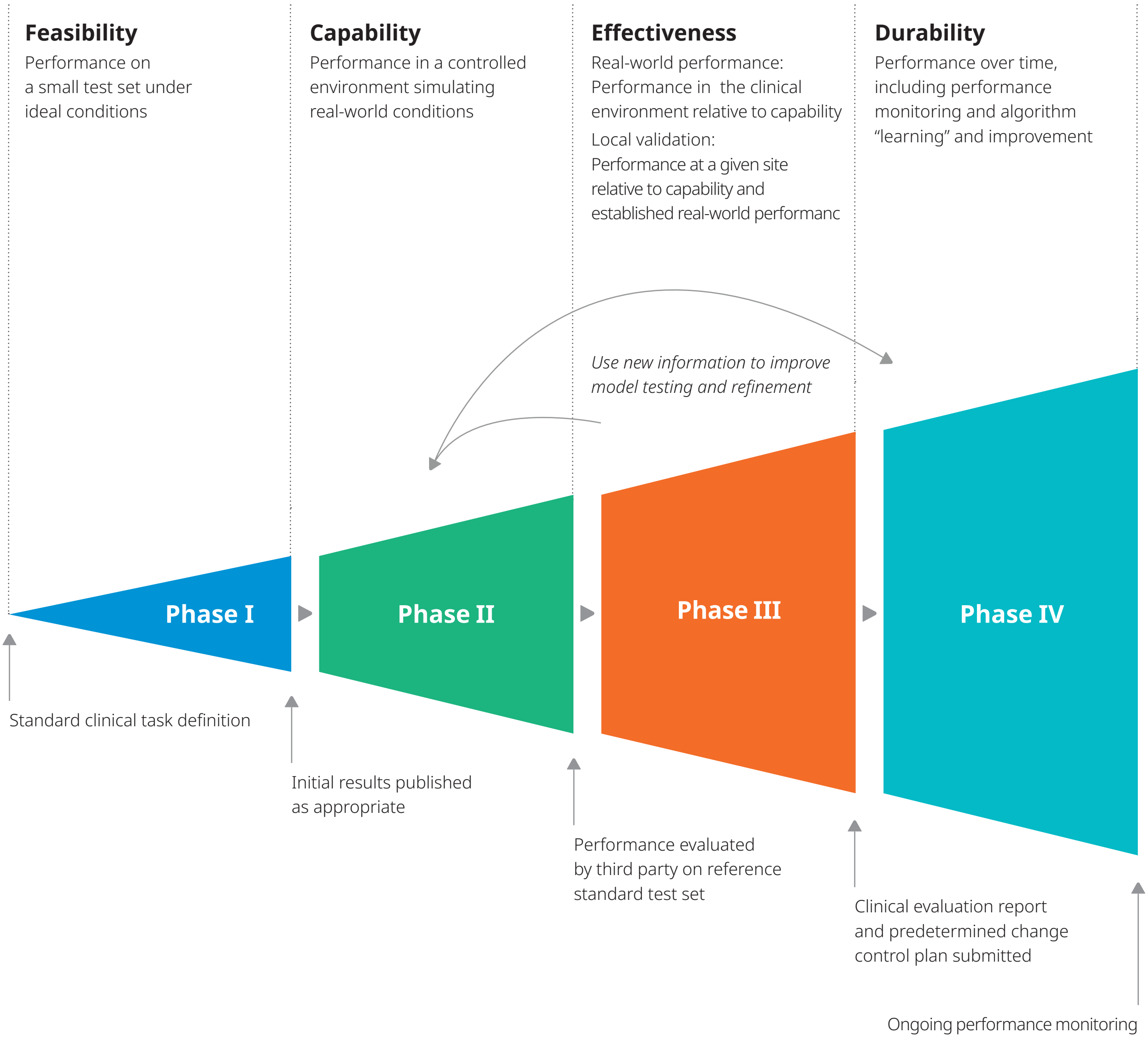
Source: IMDRF. Clinical Evaluation, 2019 *(63)*

Designing an AI-based model

Once the intended use has been defined, development of an AI model requires the "training" of the algorithm using input data that is representative of the population for which the AI-SaMD is to be deployed.

Figure 2 illustrates the phases of clinical trials required to demonstrate feasibility, capability, effectiveness, and durability in post-market surveillance.

Figure 2. Phases of development and evaluation for AI-SaMD diagnostic algorithms



Feasibility
Performance on a small test set under ideal conditions

Capability
Performance in a controlled environment simulating real-world conditions

Effectiveness
Real-world performance: Performance in the clinical environment relative to capability
Local validation: Performance at a given site relative to capability and established real-world performanc

Durability
Performance over time, including performance monitoring and algorithm "learning" and improvement

Use new information to improve model testing and refinement

Standard clinical task definition

Initial results published as appropriate

Performance evaluated by third party on reference standard test set

Clinical evaluation report and predetermined change control plan submitted

Ongoing performance monitoring

Adapted from Larson et al. *Regulatory frameworks for development and evaluation of Artificial Intelligence-Based Diagnostic Imaging Algorithms: Summary and Recommendations (70)*

Feasibility studies

Feasibility studies aim to assess whether the AI-SaMD works as intended in a controlled setting, and involve training and internal validation of the algorithm. These studies normally utilise retrospective data to prove the research hypothesis and demonstrate technical validity. Algorithms do not need to be fully robust at this stage, as the goal is simply to demonstrate feasibility. The resulting findings may be worthy of publication even if the algorithm is not ready for clinical application at this stage.

Capability studies

The next phase involves testing the accuracy of the model in a controlled environment that simulates real world conditions, and applying it to a dataset independent of that used for training the model. Such studies aim to demonstrate that the algorithm performs as intended, and measure its accuracy, reliability and safety.

- "Accuracy" refers to how closely the algorithm output matches the ground truth, including sensitivity, specificity and positive predictive value. (NB The term ground truth typically refers to information acquired from direct observation such as biopsy or laboratory results. It is sometimes used interchangeably with "gold standard" *(71)*, though the terms are not perfectly equivalent).
- "Reliability" refers to the algorithm's ability to consistently perform accurately in all conditions under which it may be used.
- "Safety" refers to the algorithm's ability to minimize the risk of harm when deployed, including when subjected to unanticipated situations.

Before proceeding to deployment in the clinical setting, the algorithm should be evaluated by a third party on a reference standard test set.

Effectiveness studies

These studies aim to confirm that the real-world performance of the algorithm matches its performance in the test environment. Local validation can be performed by clinical/industry researchers at each site before or at the time of clinical implementation. This may also include usability testing, for example, the deployment of an AI-SaMD in clinics in Thailand (see below) aimed at detecting diabetic eye disease *(4) (31)*.

Real-world deployment may reveal quality control problems in local environments.

Evaluating a clinical deep learning system

In the study by Beede et al described in their paper *A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy,* examined the much-repeated expectation that the deployment of AI-SaMDs will lead to improvements in clinician workflows and patient outcomes. Hoping to document this in a real world clinical setting, they examined a deep learning system used in eleven clinics across Thailand for the detection of diabetic eye disease. The study was carried out through interviews and observation, and examined items such as current eye-screening workflows, user expectations for an AI-assisted screening process, and post-deployment experiences. The findings showed the impact of a number of socio-environmental factors on device performance, nursing workflows, and the patient experience, suggesting a valuable role for human-centred evaluative research alongside prospective evaluations of model accuracy.

Source: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (31)*

Durability (post-market clinical follow-up)

This stage comprises generation of clinical data to track ongoing performance, for use in evaluation and monitoring. The IMDRF recommends that manufacturers embed monitoring or auditing systems within their product to automatically detect, recover from, and report errors. They should also seek less-structured sources of feedback, including customer inquiries, complaints, market studies, focus groups, and field service reports. (Note that this is particularly important for continual learning systems, which are beyond the scope of this publication and for which international regulatory standards are still unclear).

During the course of these studies, data can also be generated for evaluating the SaMD's usability and integration into the intended clinical setting and workflow.

Clinical study design

Whilst from a regulatory perspective, clinical investigations collect data in order to provide evidence of a medical device's *compliance* (e.g. safety, performance, benefit), for device developers the aim of clinical studies is to gain *new, real world data* about safety and effectiveness.

Methodologies for designing clinical studies to generate new evidence of clinical effectiveness should be planned immediately after the intended use is defined. This is to ensure that the evidence generated meets the claims of the researcher or developer, and to facilitate complete reporting for independent review and assessment of clinical impact.

Better protocols and reporting of clinical trials

In the past two decades, major international efforts have been made to improve the quality, transparency and impact of clinical trials, and the ways that they are reviewed and evaluated. One was the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) initiative, developed by an international group of scientific, industrial and regulatory stakeholders. The *SPIRIT 2013 Statement* was one of its major outputs, presenting a checklist of items to include in clinical trial protocols, including a rationale, detailed description, model example from an actual protocol, and supporting references *(72)*.

A similar initiative called CONSORT (Consolidated Standards of Reporting Trials) was developed to improve reporting of randomised controlled trials (RCTs), allowing readers to judge the reliability and validity of trial findings and extract information for systematic reviews. The first CONSORT statement was published by a group of scientists and editors in 1996 and updated in 2001 and 2007. It consists of a checklist and flow diagram for researchers to use when reporting an RCT, and has been endorsed by leading international medical journals and editorial groups have endorsed the CONSORT statement *(73)*.

June 2008 saw the launch of Enhancing the Quality and Transparency of Health Research (EQUATOR), a network focused on improving the value, transparency and reliability of published health research by promoting high-quality, robust reporting guidelines. The EQUATOR Network's website offers a database of reporting guidelines, both completed and under development, as well as a variety of other information and educative activities *(74)*.

SPIRIT-AI extension

The huge amount of progress and activity in health-related artificial intelligence in recent years and led to the creation of sub-initiatives such as SPIRIT-AI and CONSORT-AI. In September 2020, the first guidance was published to help develop clinical study protocols for AI-based interventions *(75)*. Called the SPIRIT-AI extension, and supported by CONSORT and EQUATOR, it aims to extend or elaborate on the existing SPIRIT 2013 statement and help develop consensus-based AI-specific protocols. The guidance is not prescriptive about methodological approaches to AI trials. Instead, it aims to promote transparency in reporting the design and methods of a clinical trial and thus to facilitate understanding, interpretation, and peer review.

The SPIRIT-AI extension includes 15 new items to be included clinical trial protocols of AI interventions (see Table 7). It recommends that investigators and developers provide clear descriptions of new AI interventions, instructions and skills required for use, the setting in which the interventions will be integrated, handling of input and output data, the human/AI interaction, and analysis of error cases. As well as researchers, SPIRIT-AI will be used to assist editors and peer reviewers, as well as the general readership, to understand, interpret, and critically appraise the design and risk of bias of a planned clinical study.

Table 7. SPIRIT-AI checklist items and explanations

Section	Item	SPIRIT 2013 Item	SPIRIT-AI item	
Title	1	Descriptive title identifying the study design, population, interventions, and, if applicable, trial acronym	SPIRIT-AI 1(i) Elaboration	Indicate that the intervention involves artificial intelligence/machine learning and specify the type of model.
			SPIRIT-AI 1(ii) Elaboration	Specify the intended use of the AI intervention.
Background and rationale	6a	Description of research question and justification for undertaking the trial, including summary of relevant studies (published and unpublished) examining benefits and harms for each intervention	SPIRIT-AI 6a (i) Extension	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g. healthcare professionals, patients, public).
			SPIRIT-AI 6a (ii) Extension	Describe any pre-existing evidence for the AI intervention.
Study setting	9	Description of study settings (e.g., community clinic, academic hospital) and list of countries where data will be collected. Reference to where list of study sites can be obtained	SPIRIT-AI 9 Extension	Describe the onsite and offsite requirements needed to integrate the AI intervention into the trial setting.
Eligibility criteria	10	Inclusion and exclusion criteria for participants. If applicable, eligibility criteria for study centres and individuals who will perform the interventions (e.g., surgeons, psychotherapists)	SPIRIT-AI 10 (i) Elaboration	State the inclusion and exclusion criteria at the level of participants.
			SPIRIT-AI 10 (ii) Extension	State the inclusion and exclusion criteria at the level of the input data.

Section	Item	SPIRIT 2013 Item	SPIRIT-AI item	
Interventions	11a	Interventions for each group with sufficient detail to allow replication, including how and when they will be administered	SPIRIT-AI 11a (i) Extension	State which version of the AI algorithm will be used.
			SPIRIT-AI 11a (ii) Extension	Specify the procedure for acquiring and selecting the input data for the AI intervention.
			SPIRIT-AI 11a (iii) Extension	Specify the procedure for assessing and handling poor quality or unavailable input data.
			SPIRIT-AI 11a (iv) Extension	Specify whether there is human-AI interaction in the handling of the input data, and what level of expertise is required for users.
			SPIRIT-AI 11a (v) Extension	Specify the output of the AI intervention.
			SPIRIT-AI 11a (vi) Extension	Explain the procedure for how the AI intervention's output will contribute to decision-making or other elements of clinical practice.
Harms	22	Plans for collecting, assessing, reporting, and managing solicited and spontaneously reported adverse events and other unintended effects of trial interventions or trial conduct	SPIRIT-AI 22 Extension	Specify any plans to identify and analyse performance errors. If there are no plans for this, justify why not.
Access to data	29	Statement of who will have access to the final trial dataset, and disclosure of contractual agreements that limit such access for investigators	SPIRIT-AI 29 Extension	State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use.

Source: Rivera et al., *Guidelines for Clinical Trial Protocols for Interventions Involving Artificial Intelligence (75)*

Minimum standards for model development

Full description of AI model and architecture including associated hardware:

Training set	Dataset description
Tuning set	Dataset description
Internal validation set	Dataset description

Use-case: Model development and training for AI-SaMDs in cervical cancer screening

When it comes to evaluate an AI-SaMD for use in cervical cancer screening, a key task is the device's training and tuning. As a first step, the training and validation study methodologies should be pre-specified, including:

- Technical specifications for digital cervical images and analysis
- Study cohort and contextual Information, particularly for LMIC settings
- Input data management (including relevant clinical data such as HPV and HIV status)
- Approach to validation, with primary and secondary endpoints clearly defined.

It is essential that the AI-SaMD be evaluated in its intended stage in the cervical cancer screening screen and treat (S&T) pathway.

If the AI-SaMD is to be evaluated as a clinical decision support or "assist" to the current standard approach (e.g. VIA) in lower and middle income countries), researchers should aim to define how the information arising from the AI-SaMD (output) will fit in within the overall diagnostic function. This should be used as a guide when selecting the optimal study design for efficacy/effectiveness trials *(75, 76)*.

Table 8 below uses the SPIRIT-AI checklist to illustrate some considerations that might be used to generate evidence of the safe and effective use of AI-SaMDs in cervical cancer screening.

Table 8. SPIRIT-AI items used in evaluation of AI-SaMD for cervical cancer screening

SPIRIT-AI ITEM	Example 1	Example 2
Specify the **Intended Use** of the AI Intervention in the study title	Detection of presence of precancerous cells (normal or abnormal)	Detection of high or low grade precancerous lesions of the cervix (e.g. CIN1/2 vs CIN3 and above)
Specify the **intended role** of the AI intervention in the context of the clinical pathway	Use as assist to VIA in primary screening of the cervix	Use as assist to VIA in triaging high risk (HPV positive) women in cervical cancer screening
State the inclusion and exclusion criteria at the level of the input data State the Inclusion and Exclusion Criteria at the level of the participant	Inclusion: full cervical image with visible transformation zone (TZ) in region of interest (ROI) Exclusion: Image artefact e.g. speculum Inclusion: age range specific Exclusion: Previous ablation to the cervix	Inclusion: visible TZ in ROI Exclusion: inadequate image resolution Inclusion: Visible squamocolumnar junction (SCJ) (or TZ types) Exclusion: HPV or HIV positive women
Specify the **output** of the Intervention	Normal vs abnormal	Probability score with thresholds for different actions in pathway
Specify whether there is **human-AI *Interaction*** and the required level of expertise of the users	Use of healthcare worker decision as input data into the AI-SaMD system for comparison	Healthcare worker overrides output of the AI-SaMD if disagreement

6. DATASET MANAGEMENT

This chapter covers the clinical data components used in the training and validation of an algorithm for AI-SaMD. Machine learning engineers who work closely with multi-disciplinary clinical and technical development teams generate evidence to observe the performance of a model and fine-tune its parameters.

Note that software verification and product requirements, including documentation and reporting requirements for technical files, are not covered. However, they can be found in guidance for regulatory approval such as the US FDA's Artificial intelligence and machine learning in software as a medical device *(69)*, EU Regulations for medical devices (MDR) *(66)*, and the ITU AI Guidelines for AI based medical devices *(77)*.

Terminology

Dataset construction and evaluation between studies can be difficult to discern and compare between studies. The following terms for different sets of data are widely used:

The *Development set* is prominent in the clinical literature to refer to the dataset used for ML models. It is the overall set used to generate and then test the hypothesis in exploratory analyses. Within this set:

- *Training data* is used to fit the model in an iterative fashion for majority of the "learning'
- *Tuning data* provides an opportunity for ML engineers to observe the performance of, and fine-tune the model weights. (Model parameters and hyperparameters can be adjusted based on performance of the model on the tuning set)

The *Test set* is used to validate the hypothesis and check robustness of the model after model methods are locked following tuning. This may also be referred to as the validation set.

Clear reporting is essential for evaluation of clinical data generated from the use of AI-SaMD for demonstration of efficacy, effectiveness and usability. Table 9 below describes the terminology and different datasets reported in published clinical literature on evidence generation for AI-SaMD used in (1) triage of diabetic retinopathy, and (2) in identifying metastatic breast cancer (pathology).

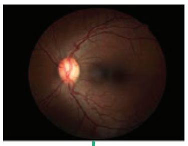
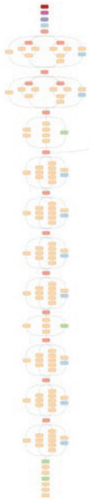
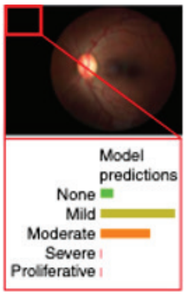
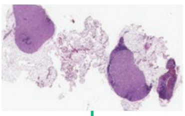
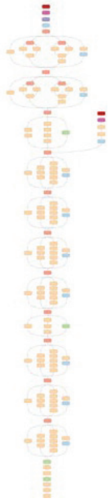
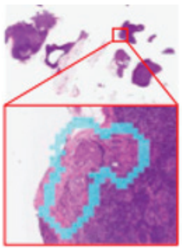
Table 9. Dataset naming in Clinical and ML Studies

Terminology in clinical literature	Development set		Validation set
Terminology in the ML literature	Training set	Tuning set (also called validation set)	Test set (holdout set)
Use-case: referable diabetic retinopathy Gulshan et al, (2016) Bellemo et al, (2019)b	102 540 images	25 635 images	11 711 images
Use-case: metastatic breast cancer Bejnordi et al, (2017) Liu et al, (2018)	216 images	54 images	129 images

Source: Chen et al. *How to develop machine learning models for health care (78)*

Table 10 illustrates these two use-cases, which were developed and validated in HICs, showing the different phases of evidence generation in the translational process, moving from validation AI models to implementation and demonstration of clinical impact. The same evidence generation considerations should be applied in LMICs:

Table 10. Datasets in training, validating and implementing AI models for healthcare

Problem selection	Data collection	ML development	Validation	Assessment of impact	Deployment and monitoring
 Referable diabetic retinopathy? Gulshan et al. (2016) Ting et al. (2017)	Developement dataset 128 175 images Validation dataset 11 711 images Gulshan et al. (2016)	 Gulshan et al. (2016)	Retrospective Sensitivity: 0.90 Specificity: 0.98 Gulshan et al. (2016) Prospective Sensitivity: 0.87 Specificity: 0.91 Abramoff et al. (2018)	 Model predictions None Mild Moderate Severe Proliferative 40% reduction in false negatives Sayres et al. (2018)	Future work
 Metastatic breast cancer? Bejnordi et al. (2017)	Developement dataset 270 images (millions of patches) Validation dataset 129 images Bejnordi et al. (2017)	 Bejnordi et al. (2017) Liu et al. (2018)	Retrospective AUC: 0.99 Sen@1*: 0.77 Bejnordi et al. (2017) Retrospective AUC: 0.99 Sen@1*: 0.86 Liu et al. (2018) *: tumour detection sensitivity at one false positive per slide	 2x review speed ½ false negatives Steiner et al. (2018)	Future work

Source: Chen et al. *How to develop machine learning models for health care (78)*

Model training

Evidence generated from AI model training and validation is essential to evaluate data and research management and research reporting of the inputs into the system, to ensure steps have been taken to mitigate bias, and to demonstrate generalizability of the model. The data used to train the model iteratively is "seen" by the developers/investigators whilst a smaller "unseen" tuning set is used to optimise the predictive/diagnostic accuracy of the model.

Published and unpublished evidence should include:

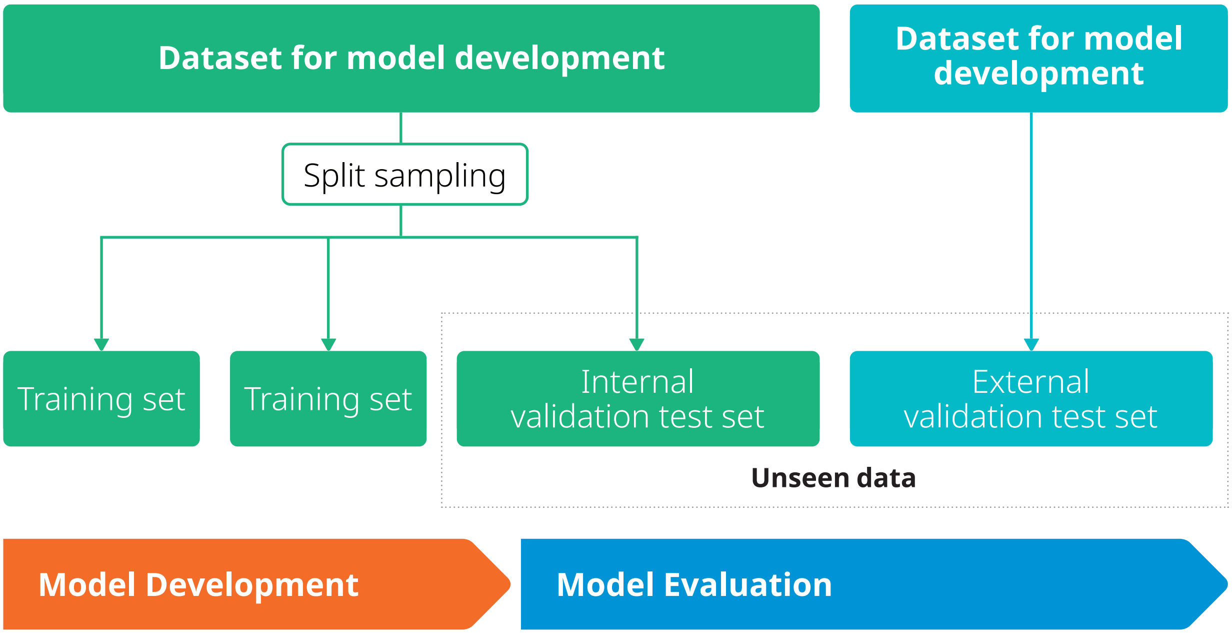
- An overview of the AI system
- Detailed description of AI system and architecture (e.g. deep learning, deep convolutional neural network for image classification tasks)
- Description of the model training datasets
- Technical requirements for software verification

Model validation

Internal validation of the model, an important part of the model development process, should use data "unseen" by the algorithm during training of the model. Methods should be pre-specified and locked (i.e. set) before validation. A clear description of how data (including patient-level data) are divided into training, tuning and internal validation test sets should be documented to demonstrate the absence of overlap. Data from the same patient must not appear in both the training and test sets.

External validation should be performed with an independent test set from an external source; this is to demonstrate generalisability of the model and should be performed by independent evaluators. Faes et al, describe the model development process and illustrate the dataset evaluation components for evidence generation (Figure 3)

Figure 3. Overview of dataset evaluation components


Dataset for model development
Dataset for model development
Split sampling
Training set
Training set
Internal validation test set
External validation test set
Unseen data
Model Development
Model Evaluation

Source: Faes et al, *A Clinician's Guide to Artificial Intelligence: How to Critically Appraise Machine Learning Studies (76)*

Use-case: Dataset construction in AI-SaMDs for cervical cancer screening

During the process of evaluating dataset construction, a variety of consideration must be explored. An opening question to be answered is whether the developers/investigators provide evidence of a pre-specified AI algorithm training, tuning, and testing methodologies

Second, it is essential to ascertain to determine whether the Dataset used in training the AI-SaMD's Algorithm reflective of the clinical setting in which the model will be applied?

A final consideration is selection bias, where thought must be given to whether the data represents the complete spectrum of disease (precancerous and malignant lesions of the cervix) for the target population? Issues to be carefully investigated regarding possible spectrum bias and class imbalance include:

- Normal controls /ASCUS
- Inflammatory changes/Cervicitis (Benign)
- CIN1 (Low grade abnormality - LSIL)
- CIN2 (High grade abnormality - HSIL)
- CIN3 (High grade abnormality - HSIL)
- Atypical glandular cells
- AIS (adenocarcinoma in-situ)
- Invasive disease (malignant)

Table 11. Training dataset considerations for cervical cancer screening

Training dataset (participants & input data)	Use-case training dataset considerations (non-exhaustive)
Patient demographic	Age, race
Digital cervical image selection criteria	Image acquisition system, image quality, risk group
Clinical setting	Community, primary care, gynaecology, specialist oncology/cancer screening centre
Data collection time period	Pre-curated (e.g. data-set from previous study/investigation), duration of collection of cervical images
Geographic location	Rural, local, country (relative to target population as defined in intended use statement)
Clinical demographics	HPV or HIV status, Transformation Zone Category, previous pathology, previous ablation or long loop excision of the transformation zone (LLETZ), etc.

7. INTERNAL VALIDATION AND DATA MANAGEMENT

The piloting and use of AI-SaMD is relatively recent in LMICs and there are few robust evaluations of evidence generated in these settings. AI-based models require a large amount of high-quality data which is generally very difficult to collect, and more so in LMICs.

When evaluating evidence, care should be taken to assess the risk of developing biased AI models or "defective" AI-SaMDs. The bias or defects to be avoided are those stemming from the model or device having been trained on datasets that differ to the local/target populations where these devices are intended to be used *(49)*.

International datasets currently used to train algorithms do not necessarily reflect the diversity of patients or health conditions of lower-resource settings *(32)*. Evaluation of data management and sources are therefore essential components in evaluating the evidence generated from clinical trials and studies.

Data handling

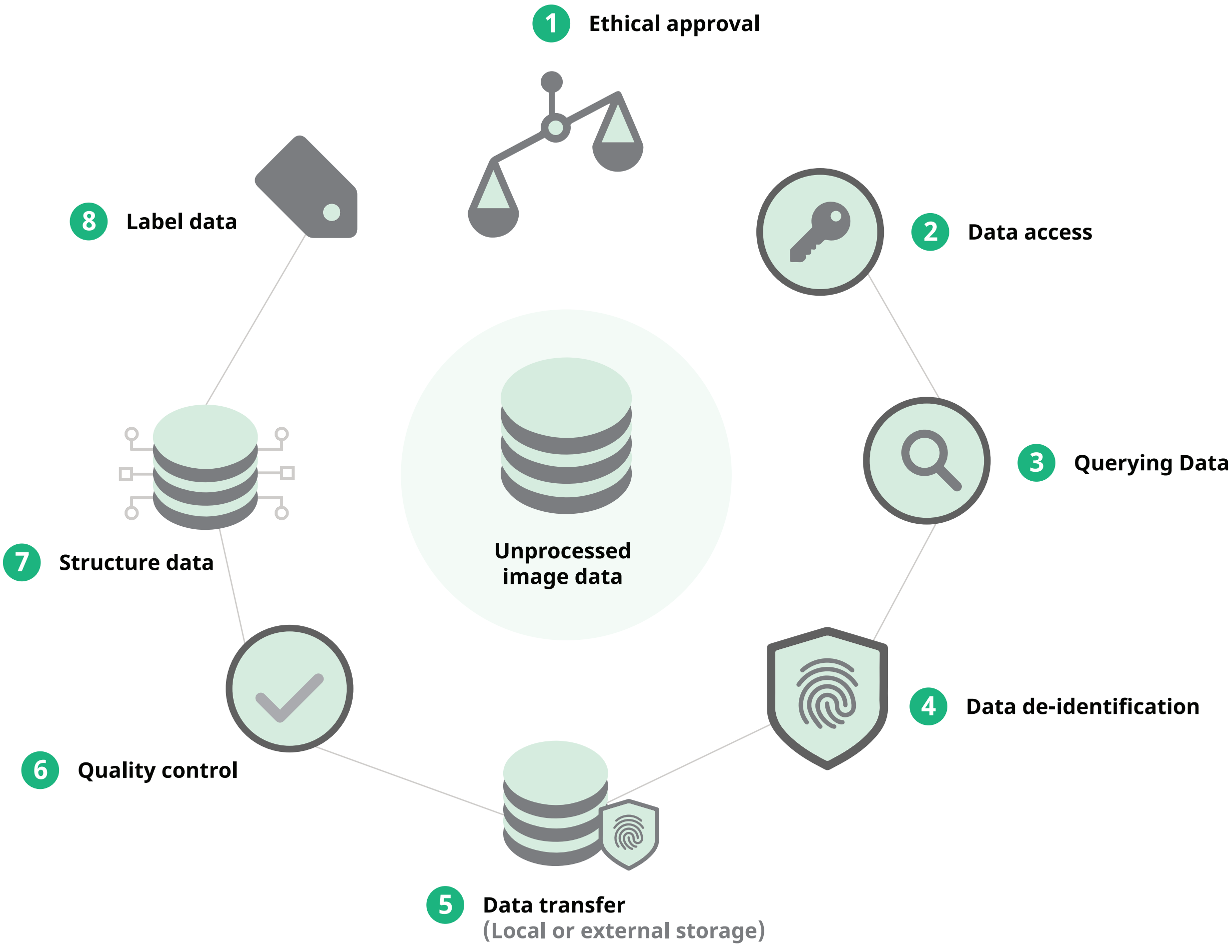
Evidence generated for validation should document the procedure used to acquire and select input data, since the performance of an AI-SaMD may be critically dependent on the nature and quality of the input data. This is particularly important for AI-based devices that have been developed in HICs specifically for use in LMICs.

The completeness and transparency of this evidence is integral to the feasibility assessment, and to successful replication of safety and performance metrics beyond the study being evaluated.

At a minimum, evidence for data management should be provided on:

1. Access to and handling of data (Figure 4)
2. Acquisition of data, including data sources
3. Data selection, including inclusion and exclusion criteria
4. Data de-identification (anonymization), pre-processing and augmentation
5. Data analysis, including for missing and poor-quality data
6. Data labelling

Figure 4. Data governance: the process of handling medical image data



Source: Willemink et al, *Preparing Medical Imaging Data for Machine Learning (79)*

Ground truth confidence

Before an AI algorithm can be trained and tested, the "ground truth" needs to be defined and linked to the image (chest x-ray, digital cervigram etc.) *(76)*.

Image labels are annotations performed by medical specialists such as radiologists. These annotations can be considered ground truth if imaging is the reference standard (e.g., pneumothorax). However annotations alone are insufficient as ground truth where a biopsy or pathologic investigation is needed to confirm the prediction/diagnosis. Since manual labelling processes cannot efficiently deal with the large datasets that are required to train models, natural language processing (NLP) is used to label annotations using free text data from radiology (and/ electronic health records) *(80)*.

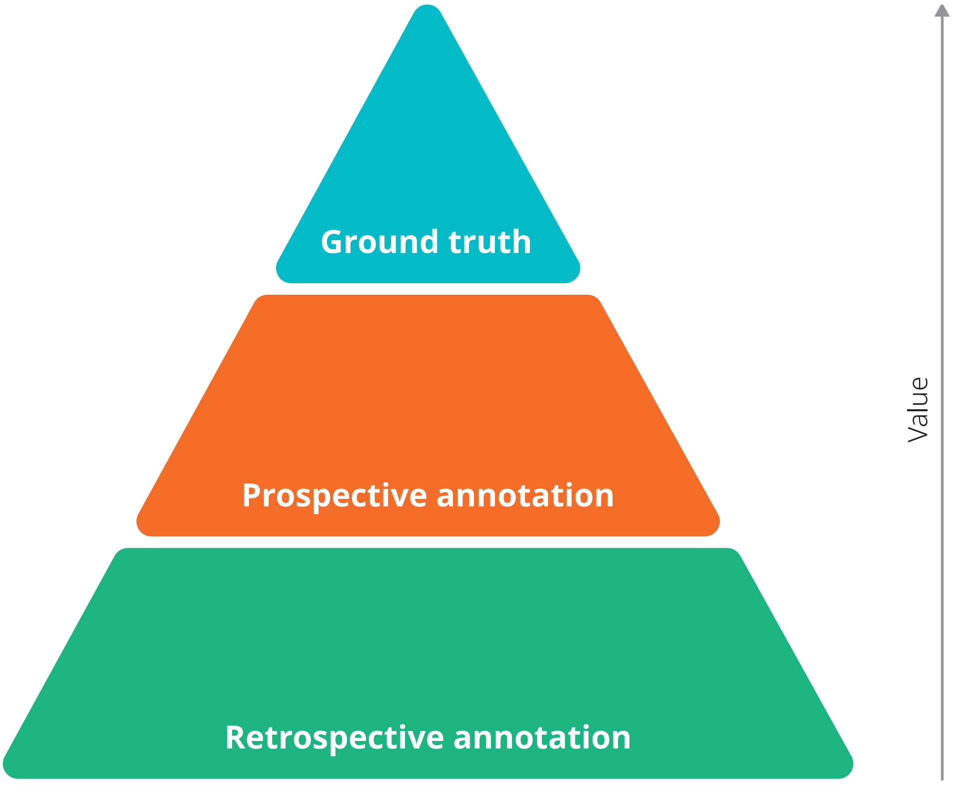
In summary then, assessing the accuracy of a "supervised learning" AI model requires the ground truth – provides the essential reference point for comparison. A useful measure of reliability of ground truth is inter-observer agreement between the labellers, and it is good practice for a threshold for inclusion of cases to be prespecified *(76)*.

A list of ground truth labels includes:

- Surgical findings
- Histopathological data
- Genomic / other laboratory diagnostic data
- Clinical outcome data (short and long-term follow-up)

Evidence generated by an AI-SaMD should be assessed in terms of the data annotation (labelling) methods used to evaluate the accuracy and robustness of the model. Figure 5 illustrates the value hierarchy of data annotation (labelling). Most useful but least abundant is ground truth data, including pathologic information from biopsies, genomic data, or clinical outcome data.

Figure 5. Value hierarchy of data annotation



Source: Willemink et al, *Preparing Medical Imaging Data for Machine Learning (79)*

Use-case: internal validation for AI-SaMDs in cervical cancer screening

The internal validation phase for an AI-SaMD's deep-learning algorithm follows training and tuning. In the use-case, this will involve testing the technical performance of the model *on a dataset of digital cervical images different to that used in training the model.*

There are several steps that are crucial to validation. One is to adopt a set of pre-defined primary and secondary outcomes for the training, for example:

- Target lesion identification - CIN3 or HSIL/LSIL (i.e. 'abnormal' vs 'normal')
- Identification of normal controls
- Prediction threshold for risk of CIN3/HSIL (high grade lesion).

A second step is to adopt a methodology for internal validation, for example:

- Cross-validation
- Bootstrapping
- Split-sample validation (Internal validity not generalisability).

A number of important questions must answered in preparation for acquisition, care and documentation of datasets, often known as "curation" *(81)*:

- Do the developers/investigators provide sufficient clarity on how the datasets were split?
- Is disease prevalence of target lesion (e.g. CIN3 or HSIL) in the internal validation test dataset representative of prevalence in the target population in a real world setting? Dataset split justification should take into account the prevalence of the disease subtypes/severity.
- When validating models based on a pre-curated dataset, for what purpose was this dataset originally curated? Does the disease probability within this cohort differ according to the setting in which the model will be deployed?
- Are there any under-represented or over-represented subgroups within the training dataset? For example: HPV- or HIV-positive women; high BMI or TZ III/IV (i.e. more difficult to visualise the cervix), etc.
- Do the exclusion criteria include any elements which create a selection bias?
- Has a sampling method been used to reduce the risk of spectrum bias (i.e. performance is affected by the specific set of patients in the test compared another set)?
- Are image labels likely to reflect the true disease state? This will be used to assess accuracy of the model.

When assessing ground truth, the following should be considered:

- Are the labels manually or automatically generated from associated records?
- Are any ground truths missing?
- Were the images labelled prospectively or retrospectively? How confident are researchers that these labels are indeed ground truth?
- What biases might be present? For example, is there a gatekeeper bias where a biopsy is only performed if a suspicious lesion is identified and referred specifically for biopsy, resulting in false negatives in missed cases.
- Have there been reports of inter-observer agreement/disagreements between labellers?

SECTION II.

SOFTWARE VALIDATION AND REPORTING

8. EXTERNAL VALIDATION

The design, execution and reporting of external validation studies focuses how well the diagnostic accuracy of the model translates to clinical accuracy in a real world setting.

External validation is a continuum rather than a single event to ensure performance accuracy is maintained over time. It should be conducted prior to product release and during post-market surveillance after product release, and applies even to "locked" algorithms, which will need regulatory review for changes that go beyond the original market authorization *(82)*.

External validation datasets can adopt a variety of characteristics *(76)*:

- Independent but different in setting and population
- Independent but differ in geographic location
- Independent in same/new population over time, to test for degradation of algorithm performance as the population evolves
- Independent and with the use of different image capture devices

Validation can also be temporal (i.e. focused on generalisability) or geographic.

Published case studies

There is a growing literature documenting external validation of AI-based clinical decision support systems. Two recent published studies are indicative of the work being done, one for breast cancer screening and the other for assessing tuberculosis.

Assessment of screening mammograms

The first, a case-control study in Stockholm, Sweden evaluated the performance of three commercially available AI-SaMDs, seeking to establish whether any of them performed as well as or above the level of radiologists in mammography screening *(56)*. The external validation dataset derived from Swedish Cohort of Screen-Aged Women, included 8805 women, and was geographically independent from all three commercial systems.

The study found that only one of the three algorithms was more accurate than first-reader radiologists in assessing screening mammograms, with sufficient diagnostic performance to act as an independent reader in prospective clinical studies. However, the greatest number of positive cases was detected by combining this best algorithm with first-reader radiologists.

Tuberculosis detection from chest X-ray

The second study compared the performance of five AI-SaMDs on an "unseen" dataset of chest X-rays collected in three TB screening centres in Dhaka, Bangladesh, a high TB-burden setting *(83)*. The external validation dataset was derived from a TB screening centre in Bangladesh, and was independent from the training and development set; the study was also geographically independent of the five AI systems tested, which originated in China, India, and South Korea.

A total of 23 566 individuals took part in the study, all receiving a chest X-ray which was read by three Bangladeshi radiologists. A sample of these were re-read by US radiologists. Xpert was utilized as the reference standard. Of the five algorithms, all significantly outperformed the human readers. The performance of the algorithms across the subgroups of age, use-cases, and prior TB history was also assessed, showing that threshold scores performed differently across different subgroups. It was concluded that these AI-SaMDs offer effective screening and triage tools for active case finding in regions with high TB-burdens.

Minimum standards to be met in external validation

- Dataset management should feature out-of-sample "unseen" (i.e. protected from developers/ investigators) test sets of input data or images
- Piloting and monitoring of data collection should be carried out to ensure diagnostic accuracy is maintained
- Independent (peer) review should be carried out on output data
- The algorithm should be retrained if performance of AI-SaMD does not meet pre-specified performance target
- The algorithm version should be updated and re-tested on prospective independent test set

Use-case: External validation for AI-SaMDs in cervical cancer screening

The external validation of an AI-SaMD in cervical cancer screening must take in a variety of considerations when choosing datasets. They may be:

- Independent but different in setting and population, e.g. general screening population vs high-risk population)
- Independent but different in geographic location, e.g. rural vs urban areas; India vs Africa, etc.
- Independent in same/new population but over time, e.g. addition of other screening tests into the pathway e.g. HPV DNA, E6/7 oncoprotein, DNA methylation)
- Independent and with the use of different image capture devices, e.g. digital cervicography, colposcopy images, digital colposcopes, pocket colposcope images, etc.

The independent review of the AI-SaMD should cover:

- Performance accuracy and sub-group analysis (age, women with HIV, HPV status, etc.)
- Comparison to readers of varying levels of expertise (nurses, general physicians, gynaecologists, gynaecological oncologists.)

9. DATA MANAGEMENT

As part of the evidence generation effort, data management must encompass a variety of activities, notably regarding data splitting, curation, and selection. Measures to ensure data quality and permit data augmentation must also be considered.

Dataset splitting management issues include:

- Dataset splitting must be "clean" at the level of patients/participants, e.g. all images from the same patient should be in the same set
- Image similarity detection in order to identify duplicate lesions, taking into account that merging datasets from different sources might involve patient-level overlap
- Dataset sample sizes.

Dataset curation issues covering the acquisition, maintenance and documentation of datasets:

- Training datasets must be representative in order to avoid class imbalance. Underrepresentation of important diagnostic features may limit performance of the model (sub-group analysis of results may reveal this)
- Datasets must be reflective of the setting in which the model will be applied. A lack of diverse data (age, race, geographic areas) will limit the generalisability and accuracy of developed AI-SaMDs
- Ensure the dataset represents the spectrum of disease manifestation (severity, stage, distribution of alternative diagnoses) to mitigate spectrum bias.

Dataset selection issues:

- Ensure that inclusion and exclusion criteria at the patient level and input data level do not create a selection bias
- Ensure data is representative of the image acquisition types that the model will be applied to in the target population
- Ensure that the data selected is representative of the data quality that will be encountered in the real world. For example, performance may be overestimated or safety overlooked in a highly curated/cleaned research dataset

Dataset quality and augmentation issues (including handling of expanded training dataset and addition of supplementary datasets):

- Ensure ground truth labels of the training dataset are of high quality. Subjective labelling and variability between labellers can introduce systematic and random errors
- Ensure transparency of methods to ensure data quality. Quality issues may be resolved by training a model only with images which are robust to the classification task, or by using a separate "image-quality" ML model *(84)*.

Minimum standards for data management

- Data sources and selection (including missing data)
- Data curation, processing and augmentation
- Data quality and demographic distribution
- Dataset split methodology and any overlaps in use of data

Use-case: Data management for AI-SaMDs in cervical cancer screening

When evaluating an AI-SaMD for use in cervical cancer screening, a clear data management plan should be pre-specified. This will cover the following considerations:

- Data collection (retrospective/prospective); selection, inclusion and exclusion criteria (at patient level and input data (image) level)
- Digital image capture specifications, including image acquisition types (e.g. digital cervigrams, colposcopic images, mobile phone images, pocket colposcopes, etc.)
- Data de-identification, including anonymization given other sensitive clinical data (e.g. HPV and HIV status)
- Data storage format
- Data quality assessment and machine language model features
- Reference standard determination, including methodology for annotation / ground truth labelling of images
- Dataset split management and sample size determination
- Details of an expanded training dataset and addition of supplementary datasets
- Data augmentation strategy (addition of independent training and test dataset, control access to both training and test dataset as additional data are being included and revised algorithm is being tuned, retrained and tested)

10. EVIDENCE GENERATION STANDARDS

The objective of a clinical investigation or any systematic study involving AI-SaMDs is to assess the safety, clinical performance and effectiveness of a medical device for a particular indication or intended use.

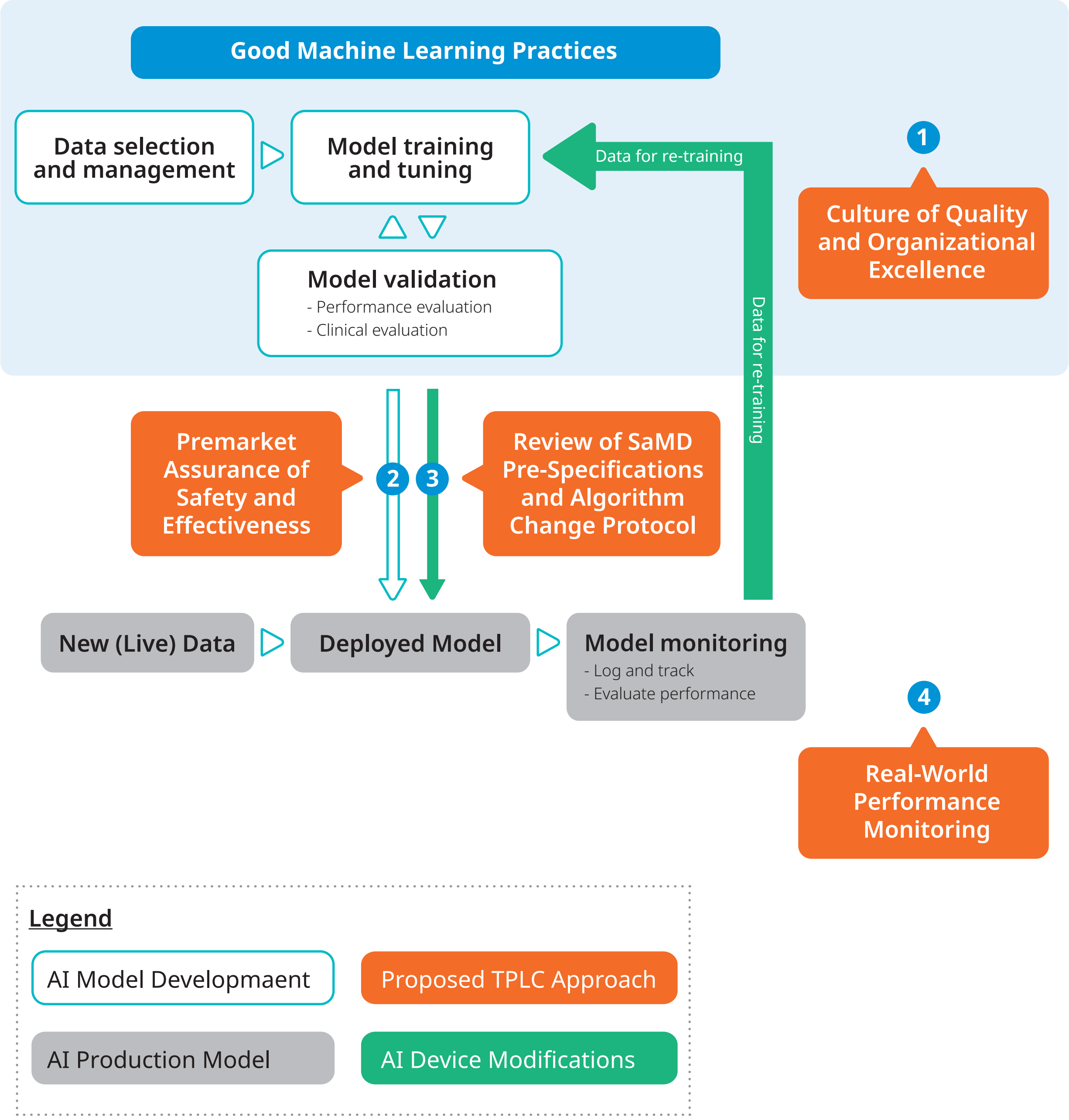
Over the last decade there have been numerous guidance documents issued by academics, regulatory agencies, and technical organisations like IMDRF and ITU. However, there is still no international consensus on how to evaluate and compare evidence generated from the development and implementation of AI-SaMDs.

Good clinical trials practice

Clinical studies and investigations evaluating AI-SaMD must be conducted following International Conference on Harmonisation - Good Clinical Practice (ICH-GCP) principles to ensure safety and transparency of reporting *(85)*.

In the context of AI-SaMD, these steps also align with the FDA's Good Machine Learning Principles, which were set out in a 2019 discussion paper, which was updated in 2021 *(82)*. Figure 6 illustrates the paper's "total product life cycle" approach to development and validation of such devices, whereby the essential requirements for evidence generation are achieved at each stage of the life cycle of the AI-SaMD. The approach highlights the relationship between post-deployment monitoring, real-world performance monitoring, and re-training of the algorithm in the event that safety and performance targets are not met. Underlying this relationship between the algorithm development, the device development, and modifications is a culture of good practices which allows for clear and transparent reporting of evidence from these devices.

Figure 6. Good machine learning practices: total product life cycle approach



Source: FDA. *Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) (82)*

Ethics of deployment

AI-enabled tools have considerable potential for bringing new health care solutions to some LMICs. For example, they offer the possibility of clinical decision support to compensate for a lack of radiologists, thus making access to breast cancer screening more widespread.

A discussion of the ethics of introducing such technologies in the LMIC context is required, covering questions such as: Should the implementer consider prior health care provision as the baseline (which may be no provision), or should HIC standards be set as the minimum acceptable standard?

How does one weigh up the economic and practical compromises that may be required to deploy an AI-SaMD into the clinical pathway in low-resource settings?

In all cases, the basic ethical principles of screening still apply, such as the need to ensure that downstream treatment must be available and affordable *(2)*. For example, even if technically possible, it would not be ethically acceptable for breast cancer detection using an expensive AI-driven second reader system to be deployed in a country with no screening programme.

User studies and user experience research

The interaction between the human being and the computer is a key area of research at present, with many surprising (and occasionally unintended) consequences of AI-enabled devices. However, these studies are largely done in HICs, and there may be very different findings if performed with LMIC user groups. This is covered more fully in Chapter 12.

Hazards and safety

Post-deployment monitoring and surveillance is crucial to put in place mechanisms for hazard identification and mitigations, with a particular emphasis on the specific challenges of a LMIC setting. For example, data drift may occur if device calibration unless regularly checked than in HICs. Poor quality data (both input and output data) are particular issues that should be anticipated and reported.

Technical infrastructure

As health systems evolve and health technology assessment tools specific for LMICs are established, there will be more discussion amongst international stakeholders about ensuring that appropriate technical infrastructure is available to support AI-enabled tools. For example, if a system requires cloud-based analysis, it will be essential for a reliable and fast internet connection be provided to the relevant clinical location.

Bias and fairness

Algorithms developed in HICs are very likely to have blind spots with underperforming subgroups when used in LMICs. These need to be rigorously explored and mitigated. Safety reporting in evidence generation at all stages of development and deployment will be particularly important for LMIC settings. This is covered in more detail below in Annex B.

International standards

Some of the key guidance relevant to evidence generation during the life cycle of an AI-SaMD is provided in Table 12 as reference guide to international standards. The table is non-exhaustive, and Annex A provides greater detail, including links and references).

Table 12. Evidence generation standards: selected guidance

Title	Description	Date	Organisation
Study protocols and reporting			
DECIDE-AI	Guidelines for developmental and exploratory clinical investigations for decision support systems driven by AI (human factors and early clinical evaluation) *(86)*	In development	EQUATOR
STARD-AI	Reporting guidelines for diagnostic accuracy studies assessing AI Interventions	In development	EQUATOR
TRIPOD-ML	Reporting standards for ML based predictive models	In development	EQUATOR
CONSORT-AI	Reporting standards for studies incorporating AI-based Interventions	2020	EQUATOR
SPIRIT-AI	Study protocol standards for AI-based Interventions	2020	EQUATOR
International Medical Device Regulators Forum			
SaMD: Clinical Evidence (N55)	Guidance to all those involved in the generation, compilation and review of clinical evidence sufficient to support the marketing of medical devices	2019	IMDRF
SaMD: Clinical Investigation (N57)	Guidance focusing on the activities needed to clinically evaluate SaMD	2019	IMDRF
SaMD: Clinical Evaluation (N56)	Guidance outlining general principles of clinical evaluation; how to identify relevant clinical data to be used in a clinical evaluation; how to appraise and integrate clinical data into a summary; how to document a clinical evaluation in a clinical evaluation report.	2017	IMDRF
Regulatory Frameworks and Guidance			
MDCG 2020-5	Clinical Evaluation – Equivalence: A guide for manufacturers and notified bodies	2020	MDCG (EU)
MDCG 2020-1	Guidance on Clinical Evaluation (MDR) / Performance Evaluation (IVDR) of Medical Device Software	2020	MDCG (EU)
Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan *(69)*	Based on IMDRF's risk categorisation, FDA's TPLC (Total Product Life Cycle) approach and Pre-certification Programs	2021	FDA
MEDDEV 2.7/1. Revision 4. Clinical Evaluation	Guidelines for clinical evaluation of medical devices and evidence generation. Analysis and appraisal of clinical data generated from medical devices to demonstrate safety and performance	2016	European Commission
EU 2017/745 Medical Device Regulation	Guidance with sections pertaining to clinical evaluation and post-market clinical follow-up	2017	European Commission

Title	Description	Date	Organisation
International Standards and Frameworks (Regulatory)			
ISO/IEC CD 23053 Framework for Artificial Intelligence (AI) Systems using Machine Learning (ML)	Guidelines for developing Artificial Intelligence applications	In development	International Standards Organisation (ISO)
ITU-T. FG-AI4H-I-036. Guidelines for AI based medical device: regulatory requirements *(77)*	Defines a set of guidelines intended to serve AI developers/manufacturers on how to conduct a comprehensive requirements analysis and to streamline conformity assessment procedures to ensure regulatory compliance for the AI-SaMDs	2020 (Draft)	ITU-T Focus Group on AI for Health
EN ISO 14971 Application of Risk Management to Medical Devices	Guidelines for risk management: analysis of risks, benefit-risks analysis, evidence generation and reporting for pre- and post-market risk management	2019	International Standards Organisation (ISO)
IEC 62366 Application of Usability Engineering to Medical Devices	Guidelines for usability engineering and validation	2014	International Standards Organisation (ISO)
ISO 13485 Quality Management Systems for Medical Devices	Current guidelines for quality management systems (QMS) which incorporates	2016	International Standards Organisation (ISO)
ISO 14155:2011 Clinical Investigation of Medical devices	Guideline covering good clinical practice, clinical investigation, planning, design, reporting and monitoring of risks, quality assurance and documentation	2011	International Standards Organisation (ISO)
National Standards, Guidance and Frameworks			
Interim guidance for those wishing to incorporate artificial intelligence into the National Breast Screening Programme	Draft Guidance to start discussions on evidence requirements for AI in Breast Cancer Screening Programme, includes incorporating and piloting and research governance submission committee	2019	National Screening committee NSC (UK)
NICE Evidence Standards Framework	Evidence generation for effectiveness standards. Gold standard for evaluating effectiveness, high-quality intervention trial or RCT	2019	National institute of clinical excellence (UK)
Human factors and usability engineering guidance for medical devices	Standards for usability evaluation, post-market surveillance and monitoring, summative testing. Adapted from the FDA's *Applying human factors and usability engineering to medical devices 2016*	2017	MHRA (UK)

Use-case: Applying international standards to AI-SaMDs in cervical cancer screening

Table 13 provides an example of how international standards – in this case the SPIRIT-AI guidelines – can inform the use of AI-SaMDs for use in cervical cancer screening.

Table 13. Applying SPIRIT-AI checklist to cervical cancer screening

Components of Evaluation	SPIRIT-AI Item	Considerations for AI-SaMD for use in cervical cancer screening
Clear statement of objectives	Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g., health care professionals, patients, public)	Primary screening: General population Secondary Screening: Triage for HPV positive women, HIV
Appropriate subject population(s)	State the inclusion and exclusion criteria at the level of i. the participant ii. input data	Screening age group Screening setting and image acquisition type: community, clinic, colposcopy, etc.
Choice of appropriate controls	Specify the procedure for acquiring and selecting the input data for the AI intervention	Normal controls Normal +/- (HPV negative, HIV negative)
Design configuration	Describe the onsite and offsite requirements needed to integrate the AI intervention into the trial setting	Usability and integration into clinical workflow: Well-designed observational cohort study (AVE + VIA) vs randomised trial of VIA +/- AVE
Type of comparison and comparators	Specify whether there is human-AI interaction in the handling of the input data, and what level of expertise is required for users	Superiority of (AVE + VIA) vs VIA alone Non-inferiority of AVE vs VIA Equivalence but more cost-effective (e.g. cost-savings on biopsies of CIN1) Level of expertise for users - generalists, gynaecologists, nurses, nurse colposcopists, etc.)
Study endpoints	Specify the output of the AI intervention	Primary: identification of CIN3 Secondary: identification of CIN2/3 in HIV and HPV positive women
Minimization of bias	Explain the procedure for how the AI intervention's output will contribute to decision-making or other elements of clinical practice	Phase II: Efficacy/effectiveness study design Randomization? VIA vs VIA + AVE Identification of confounding factors (e.g. concurrent therapies, comorbidities)
Follow-up duration and monitoring	Specify the procedure for assessing and handling poor quality or unavailable input data	Follow-up duration to confirm: Ground truth Outcome of S&T pathway Interval until next screening test (define)
Adverse event definitions and reporting	Specify any plans to identify and analyse performance errors. If there are no plans for this, explain why not	Missed/ interval cervical cancers Data mismanagement Post management clinical follow-up and reporting

11. EVIDENCE REPORTING

Optimal reporting of studies is crucial to ensure the results can be used to inform policy decisions and Health technology assessments (HTAs). As global health standards evolve, reporting of evidence to enable the evaluation of a given AI-SaMD and comparison to other equivalent devices will be essential to ensure reliable impact assessments for safety, performance and cost-benefit analyses.

Data Sources

The IMDRF sets out clear guidance for data sources for evaluation of clinical data related to safety and performance claims of a SaMD. Three main data sources are useful for evaluation of evidence: published data, data from clinical experience, and data from clinical investigations (or clinical trials).

1. Published peer-reviewed data gathered through literature searches

A systematic review of the literature will be required to find relevant evidence related to the AI SaMD being evaluated. This must include studies with clinical data related to the same intended use. It is recommended a methodological approach to the literature search – e.g. PRISMA *(87)* – is utilised to identify and appraise only those publications which can demonstrate reference standards for safety and performance.

2. Data from clinical experience

Clinical experience data is most useful for:

- Identifying less common but serious adverse events
- Providing long term data about safety, clinical performances +/-effectiveness (including durability data and information about failure modes
- Elucidating the end-user “learning curve”.

3. Data from clinical investigations or trials

Clinical trials with pre-specified methodology and results of clinical investigations, along with a clinical investigation plan/protocol should be reported using available reporting guidelines (Table 10) to allow for evaluation and comparison of evidence generated from the validation and implementation of AI-SaMD.

Study reports and appraisals

Analysis of clinical data in reported studies, investigations and from clinical experience data is required to evaluate AI-SaMD to mitigate the risk of unintended consequences when these devices are deployed at scale. Clinical data needs to be assessed for:

- Quality and relevance
- Significance with respect to safety and clinical performance
- Contribution of each dataset
- Methods used to generate/collect the data (to avoid bias and confounder effects)
- Selection of cases form available datasets
- Generalisability and applicability to target population.

Clinical evaluation reporting

The requirements for clinical evaluation apply to all classes of AI-SaMD. The evaluation should be appropriate to the device under evaluation, its specific properties, and its intended use.

Benefits and risks should be specified as to their nature, probability, extent, duration and frequency. Core issues are (a) the proper determination of the benefit-risk profile in the intended target groups and medical indications, and (b) demonstration of acceptability of that profile based on current knowledge/ state of the art in the medical fields concerned.

In the European Union (EU), clinical evaluation is a responsibility of the AI-SaMD developer. As part of the European Commission's regulatory requirements that permit a product to display CE marking, the clinical evaluation report (CER) is required part of a medical device's technical documentation *(88, 89)*. A clinical evaluation assessment report (CEAR) is used by a Notified Body (the responsible entity for assessing medical devices and diagnostics within the EU) to document its conclusions about the clinical evidence presented by the manufacturer in the CER and about the related clinical evaluation that was conducted; this is a core requirement of the European Union's Medical Device Regulation *(66).*

A clinical evaluation should be a part of the developer's quality management documentation which must be available for AI-SaMD to be deployed in low resource settings. It should also be aligned with and reflected in the rest of the technical documentation.

Reporting standards

SPIRIT-AI and CONSORT-AI extensions

There is as yet no international consensus about whether it is feasible for all interventions and devices involving artificial intelligence to undergo rigorous prospective evaluation to demonstrate impact on health outcomes. However, the SPIRIT-AI extension provides a new reporting guideline for clinical trial protocols that are evaluating interventions with an AI component *(35)*. Developed in parallel with CONSORT-AI, its companion statement for trial reports, the guidance highlights the following:

- Minimum technical evidence reporting requirements
- Full details on development of the AI algorithm (including intended use, subject populations, training and testing data, and public accessibility of the code)
- Technical information regarding on-site application of the AI technology
- Details about human–AI interactions, including required expertise of the user and how the AI output contributed to clinical decision making
- Specificity with regards to what version of an AI algorithm was used, given that performance of some algorithms can change iteratively, or in some cases, continuously.

Minimum standards for reporting technical evidence

- Full details on development of the AI algorithm including intended use, subject populations, training and testing data, and public accessibility of the code
- Technical information regarding on-site application of the AI technology
- Details about Human–AI interactions, including required expertise of the user and how the AI output contributed to clinical decision making
- Specificity with regards to what version of an AI algorithm was used, given that performance of some algorithms can change iteratively, or in some cases, continuously

Use-case: Reporting for AI-SaMDs in cervical cancer screening

The following considerations apply to reporting on trials of a specific SaMD aimed at cervical cancer screening, namely automatic visual evaluation (AVE) of cervical images using a deep-learning algorithm.

In line with the SPIRIT-AI and CONSORT-AI extensions, the report should provide a detailed device description including:

1. The hardware (and/or image acquisition device) that will house the algorithm (e.g. mobile phone, pocket colposcope)
2. The image capture device(s) for cervical images after acetic acid staining (digital cervicography); a list of the different image acquisition systems (IAS), with pre-specified acceptance criteria for image acquisition characteristics intended for future compatibility with the algorithm
3. The software application that will house the algorithm
4. The architecture of the algorithm - e.g. Deep convolutional neural network (DCNN)

Results reporting and analysis should include:

1. Performance accuracy: ability to detect high grade precancerous lesions of the cervix or low grade abnormality, ability to correctly identify a normal cervix (the absence of precancerous lesions)
2. Explainability of the AI: how the model derived its predictions
3. Clinical outcomes: definition of endpoints for clinical impact assessment; confusion matrices and statistical analysis e.g. precision, positive predictive value (PPV), negative predictive value (NPV), recall rates
4. Usability: impact on clinical workflow, e.g. What was the experience of using the device? Did it operate as expected? Did it negatively affect existing workflows by being too slow, unreliable, etc?
5. Peer-review publication of studies - evidence generated should be published for appraisal by evaluators

Statistical analysis should be prospectively specified and based on sound scientific principles and methodology:

1. Clinically relevant endpoints
2. Analysis of performance in target population and sub-group analysis for generalizability (e.g. age, geographic location, HIV and HPV status)
3. Statistical significance levels, power sample size calculation and justification analysis methodology
4. Management of potential confounding factors
5. Measurement of human-AI Interaction and comparison of performance of health care workers of varying expertise with and without the use of the AI-SaMD
6. Statistical methods for demonstrating accuracy, e.g. AUROC, Youden’s Index

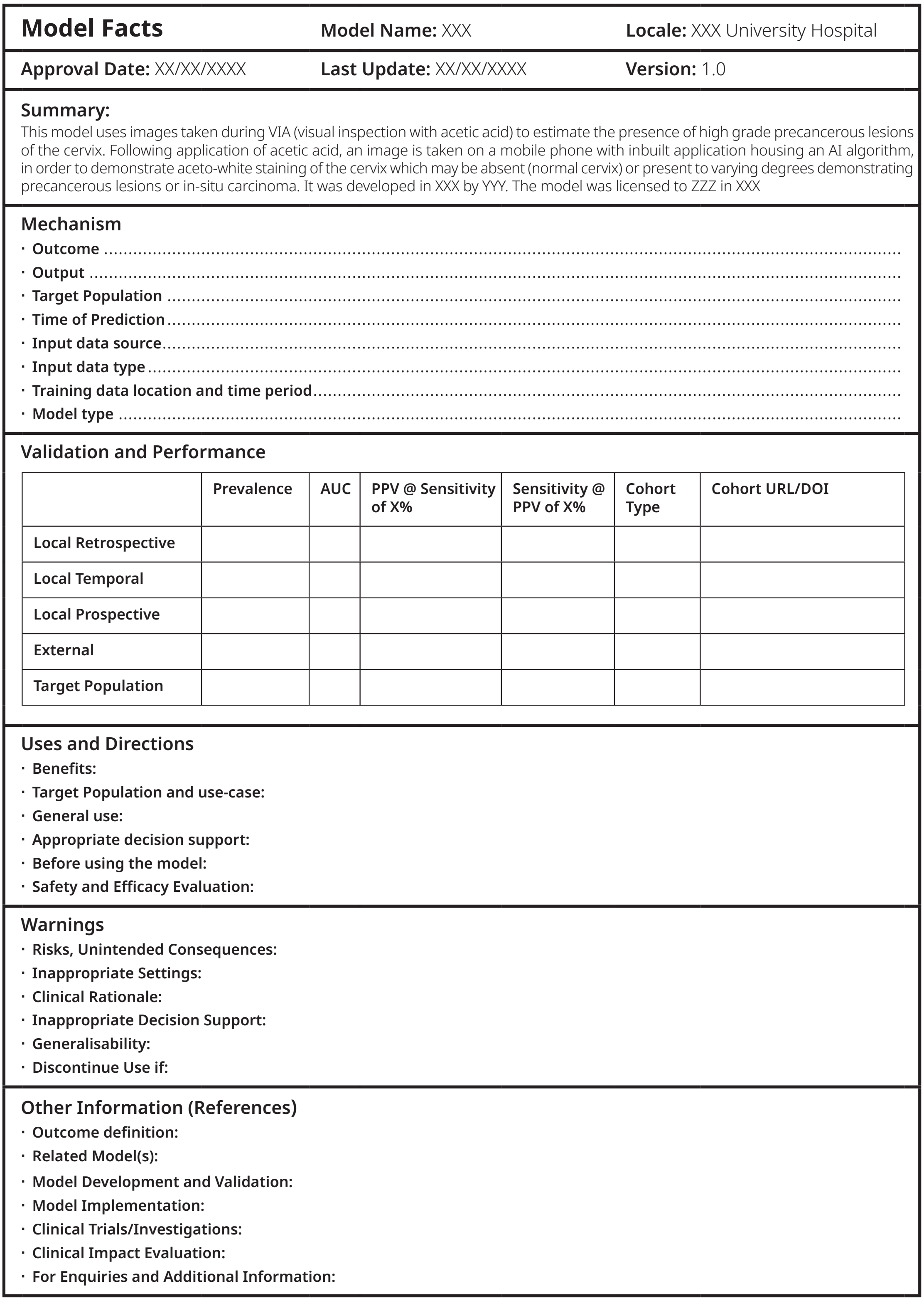
Model Facts label

Designed by an interdisciplinary team including developers, clinicians, and regulatory experts, the Model Facts label provides a reporting template aimed at assisting "clinicians who make decisions supported by a machine learning model", i.e. an AI-SaMD *(90)*. The major sections of the label include the model name, locale, and version, summary of the model, mechanism of risk score calculation, validation and performance, uses and directions, warnings, and other information. The structure is meant to mirror product information for food, drugs and devices. Its authors propose the Model Facts labels be used for all AI-SaMD to help improve the current level of reporting and transparency of AI and ML evaluations in the published literature. It is based on an earlier study on reporting of machine learning studies *(91).* The authors comment:

> The purpose is to collate relevant, actionable information in 1-page to ensure that front-line clinicians know how, when, how not, and when not to incorporate model output into clinical decisions. It is not meant to be comprehensive and individual sections may need to be populated over time as information about the model becomes available. For example, a model may be used in a local setting before it has been externally validated in a distinct geographical setting. There is also important information about the model, such as the demographic representation of training and evaluation data, that may need to be immediately available to an end user preceding full publication of a model.

Figure 7. below illustrates a partially populated sample "Model Facts" label which has been applied to the use-case for a cervical precancer prediction model.

Figure 7. Partially populated sample "Model Facts" label for cervical precancer prediction



Model Facts	**Model Name:** XXX	**Locale:** XXX University Hospital
Approval Date: XX/XX/XXXX	**Last Update:** XX/XX/XXXX	**Version:** 1.0

Summary:

This model uses images taken during VIA (visual inspection with acetic acid) to estimate the presence of high grade precancerous lesions of the cervix. Following application of acetic acid, an image is taken on a mobile phone with inbuilt application housing an AI algorithm, in order to demonstrate aceto-white staining of the cervix which may be absent (normal cervix) or present to varying degrees demonstrating precancerous lesions or in-situ carcinoma. It was developed in XXX by YYY. The model was licensed to ZZZ in XXX

Mechanism

- **Outcome** ....
- **Output** ....
- **Target Population** ....
- **Time of Prediction** ....
- **Input data source** ....
- **Input data type** ....
- **Training data location and time period** ....
- **Model type** ....

Validation and Performance

	Prevalence	AUC	PPV @ Sensitivity of X%	Sensitivity @ PPV of X%	Cohort Type	Cohort URL/DOI
Local Retrospective						
Local Temporal						
Local Prospective						
External						
Target Population						

Uses and Directions

- **Benefits:**
- **Target Population and use-case:**
- **General use:**
- **Appropriate decision support:**
- **Before using the model:**
- **Safety and Efficacy Evaluation:**

Warnings

- **Risks, Unintended Consequences:**
- **Inappropriate Settings:**
- **Clinical Rationale:**
- **Inappropriate Decision Support:**
- **Generalisability:**
- **Discontinue Use if:**

Other Information (References)

- **Outcome definition:**
- **Related Model(s):**
- **Model Development and Validation:**
- **Model Implementation:**
- **Clinical Trials/Investigations:**
- **Clinical Impact Evaluation:**
- **For Enquiries and Additional Information:**

Adapted from Sendak et al. *Presenting machine learning model information to clinical end users with model facts labels (90)*

SECTION III.
DEPLOYMENT AND POST-MARKET SURVEILLANCE

12. EVALUATION OF USABILITY

Education and training are required to "grow" an AI-literate workforce, able to take full advantage of AI-SaMDs and other innovative interventions. Full understanding of these new tools, of product labels and instructions for use, will reduce foreseeable misuse and improve compliance and clinical reporting for evaluation and evidence generation.

User interface and data interpretation

Multidisciplinary teams involved in building AI-SaMDs must carry out thorough usability testing before deployment in order to ensure adoption within clinical workflow and clinical setting.

Considerations for user interface which may affect output interpretation and performance data include:

1. Specifications of user interface in case of:
   - Errors of system in reading input data
   - Incomplete datasets
   - Internal errors in the output display (including warnings/alerts/output failure
2. Instructions for Use. This is essential to ensure AI-SaMD are used as intended to ensure safety and performance thresholds are met

Guidance for usability evaluation

There are numerous examples of international guidance covering evidence requirements for usability technical and human factors (see Table 12 above). There is general agreement that evidence generated during clinical validation should include the evaluation of clinical risks pertaining to usability, including:

1. **Clinical experience data** showing evidence arising from the output of the AI-SaMD being misunderstood, overlooked, or ignored (e.g. due to disagreement with the user)
2. **Outcome data** arising from foreseeable misuse by users (e.g. blindly trusting the AI-SaMD without engaging in the clinical decision making). This is of particular concern in the global health context where devices may be used by inexperienced HCWs

Minimum standards for evidence in evaluating usability

- Evidence of integration into clinical workflow with sustained overall benefit
- Infrastructure and conditions to allow for use of device as intended
- Effects of adding AI-SaMD to current standard
- Effects of disagreements between output of AI-SaMD and clinical decision of health care worker
- Users' interaction with output of AI-SaMD. Is the output interpretable? Error rates? Is the image readable?

Case-Study

A recent study from Thailand illustrates the evaluation of a deep-learning system's usability for the detection of diabetic retinopathy in rural clinics *(31) (84)*. Interviews and observation were carried out in 11 clinics in Thailand, investigating workflows, user expectations, and post-deployment experiences.

Pre-deployment findings included: high variation of the eye-screening process across the clinics in the study (including image capture and workflow); clinic screening conditions varied across teams (infrastructure, rooms, lighting etc); volume (i.e. variation in number of patients to screen per clinic, time allocation and effect on workflow, image capture etc.). Users (nurses) saw both advantages and disadvantages to adding the AI-SaMD to their workflow for decision support

Post-deployment findings included detecting an effect of AI-SaMD deployment on the patient consent process. Clinical factors affected performance accuracy of the AI-SaMD: 21% of images were too poor in quality to permit gradability by the AI; staff in some clinics developed workarounds to the study protocol (especially if output was ungradable); poor internet and connectivity impeded workflow and thus affected patient experience.

Evaluating the AI-SaMD within an LMIC context highlighted several socio-environmental factors which impacted model performance, nursing workflows, and the patient experience. The findings supported the value of conducting human-centred evaluative research within real-world settings alongside prospective evaluations of model accuracy.

Use-case: Usability for AI-SaMDs in cervical cancer screening

When evaluating an AI-SaMD's usability for augmenting visual inspection of the cervix during cervical cancer screening in an LMIC setting, the following should be considered:

- Evaluate cervical cancer screening clinical workflow before and after deployment of the AI-based system, through observational research and interviews
- Evaluate and pre-determine contextual challenges that may affect usability, output score, and decision-making depending on output score, e.g. poor lighting, expertise of user, time constraints in clinics
- Evaluate clinical factors that may affect performance accuracy, e.g. HIV or HPV status, BMI, age, etc

13. EVALUATION OF CLINICAL IMPACT

A well performing machine learning algorithm of an AI-SaMD is rarely sufficient alone to demonstrate clinical impact. Whilst no algorithm will be 100% efficient in real-world scenarios, it is important to recognise the challenges to translating performance metrics seen in efficacy and effectiveness studies into the intended clinical pathways.

The feasibility of designing both retrospective and prospective studies to generate evidence of clinical impact has to be considered when the intended use of the AI-SaMD is defined.

Crucial to large scale adoption and scalability are user trust, user and patient experience, and integration into the actual clinical workflow, with appropriate safeguards for patient safety.

User trust, and thus successful evidence generation during prospective real-world studies, can be improved with:

- Clear instructions for use (including labelling)
- A well-designed user interface
- Training and experience in using the AI-SaMD
- Prospectively conducted studies
- Completely reported validation studies

Evaluation Metrics

To evaluate an AI-SaMD model, evaluation metrics have to be consistent with metrics in the relevant community/research setting.

There are two main categories of evaluation metrics: discrimination metrics, and calibration metrics *(92)*.

Discrimination metrics measure the ability to correctly rank or distinguish two classes. The most common threshold-free discriminative metric is the area under the receiver operating characteristic curve (also called AUROC, AUC or c-statistic). Threshold-dependent metrics include sensitivity (recall), specificity and precision (positive predictive value). Thresholds tend to play a much larger role in health care relative to foundational ML papers because clinical applications commonly involve binary decisions, such as applying or withholding treatment. Threshold selection depends on the clinical use-case (e.g. high sensitivity for screening and high specificity for diagnosis) and resource constraints (e.g. only a certain percentage of patients can be screened based on time, manpower or monetary limitations).

Calibration metrics evaluate how well the predicted probabilities match the actual probabilities. Some ML models do not output a probability by default and may require post-training calibration. Although under-reported, calibration metrics (for example, the Hosmer–Lemeshow statistic) are crucial for real-world use because these probabilities are used for expected cost–benefit analysis.

AI-SaMD developers and ML engineers should report widely used metrics for the specific field to facilitate comparisons across studies. If no standard metrics exist, then care should be taken to report clinically relevant metrics based on the expected use-case. A performant (i.e. well-functioning) model should demonstrate both good discriminative performance and, where applicable, also generate well-calibrated probabilities. AI-SaMD validation should be done using large, heterogeneous datasets to ensure generalization to diverse patient populations.

Sub-group analyses and population adjustment

Evaluation of model performance in subgroups can be relevant in determining the clinical use-case. For example, the prevalence and presentation of precancerous lesions of the cervix differs in women with HIV, or in those who test positive for HPV subtypes. Subgroup analysis can also be based on non-patient factors, such as imaging hardware models, or the site where the data was collected. In addition, evaluation can also be affected by factors such as inclusion or exclusion of specific subgroups for the analysis. In these situations, sensitivity analysis might be prudent to ensure that these choices did not meaningfully affect the evaluation.

Evaluation of dataset augmentation to reduce class imbalance should be performed, as the validation set collected may have a different distribution of disease subtypes relative to real-world populations. In this situation, the evaluation should be adjusted according to realistic prevalence distributions *(76)*. This can facilitate the comparison of evaluation across the evidence in the scientific literature because the metrics would have been corrected for bias attributable to differences in prevalence between studies.

Human performance comparators

Evaluation of model performance in efficacy/effectiveness studies usually requires comparisons with a “human baseline” for context. For example, AI-SaMDs for diagnostic tasks may benefit from comparing model accuracy with that of human graders. In these situations, care should be taken to ensure a fair comparison: the experience level of the humans comprising the baseline should be representative of those in the real world, and the baseline comparators should be given a reasonable amount of time relative to real-world constraints. The comparators should be provided additional data such as patient history and results of other tests where relevant *(93)*.

For AI-SaMDs that predict a previously unknown association, comparison with a baseline model (e.g. logistic regression) based on variables that are readily available in the clinic such as demographics, may be useful to evaluate the added value of the proposed novel association. It is also possible to evaluate how well an output predicts a clinically relevant outcome relative to human performance to add potential value.

Comparing the AI-SaMD’s performance relative to human clinicians’ performance can be performed as “reader studies” in external validation whereby the same image is shown to the AI algorithm and health care workers. The latter are normally specialists such as radiologists, of varying levels of expertise reflecting the intended users in their real world clinical setting.

Real world performance testing

When a study is being designed to test real-word performance, the following issues should be taken into consideration:

- Adequacy of the sample size and power calculation
- Adequacy and relevance of endpoints (including validity of surrogate endpoints, if used)
- Adequacy of applied controls (including choice of the study type and of comparators, if applicable). For example, what is the performance of the AI-SaMD being compared to? If clinical decision support, will this be evaluated with and without the use of the AI-SaMD?
- Prospective randomisation of patients in case of multiple treatment arms
- Adequacy of inclusion and exclusion criteria, and of stratification of patients in respect to age, medical indication, severity of the condition, gender, other prognostic factors, etc.
- Distribution of prognostic factors. In the use-case of multiple groups, were the groups comparable for these factors? Sub-group analysis of performance accuracy should be carried out to determine if the impact measured is across all groups
- Blinding of patients, including use of sham devices or sham surgery, professional users, and outcome assessors (blinded endpoints)
- Adequacy of the follow-up period, including if follow-up was long enough for outcomes to occur, and if follow-up was frequent enough to detect temporary side effects and complications
- Reliability of the methods used for quantifying symptoms and outcomes, including validation of the methods
- Adequate recording and reporting of unintended consequences, serious adverse events, and device deficiencies
- Adequacy of procedures for retrieving complete information (e.g. procedures to be applied when contacts with patients are lost, disclosure of reasons for patients leaving the study, conduct of sensitivity analysis for determining if missing data affect conclusions)

The evaluators should verify whether clinical investigations have been defined in such a way as to confirm or refute the developer's claims for the AI-SaMD.

In summary, the:

- Comparison to gold standard
- Measures of improvement in patient outcomes, clinical process, or time efficiency
- Measures of acceptable unintended consequences and absence of harm to patients
- Changes in experience of patient or user (i.e. health care worker).

Minimum standards for clinical impact evaluation

- Comparison to gold standard
- Measures of improvement in patient outcomes, clinical process, or time efficiency
- Measures of acceptable unintended consequences and absence of harm to patients
- Changes in experience of patient or user (i.e. health care worker)

Use-case: Clinical impact for AI-SaMDs in cervical cancer screening

In evaluating an AI-SaMD for cervical cancer screening, the following measures of clinical impact should be considered.

Comparison to gold standard

- Clinician /health care worker actions: clinical judgement after visual assessment by the worker, and with the addition of the AI-SaMD: Behaviour/decisions altered? short and long term outcomes of adding the AI-SaMD to clinical decision-making process?

Effectiveness

- Patient care, including measures of improvement and meaningful outcome measures:
  - Increased detection rate of precancerous lesions or early cancers
  - Increased overall detection rate of screen-detected early cancers
  - Reduction in prevalence of symptom-detected cervical cancers
  - Reduction in prevalence of missed/interval cervical cancers
  - Reduction of false positive biopsies
  - Increased follow-up rate
  - Overall reduction of short, medium, and long term cervical cancer incidence
  - Improved survival
- Healthcare Process:
  - Efficiency of screening process
  - improvement in detection rates by LMIC health care professionals
- Cost-Effectiveness Analysis in comparison to other screening modalities e.g. primary HPV testing accessibility and cost

Safety

- Unintended consequences, performance errors
- Negative Patient outcomes - missed cancers, interval cancers, etc

Usability

- User experience and effect on clinician workflow: objective measures of effectiveness from health care worker
- User interface: usability, trust, acceptance and adoptability
- Health system use: implementation and sustainability
- Limitations of AI-SaMD technology

Patient experience

- Impact of adding device to VIA (or other gold standard) on consultation time, patient satisfaction, engagement, compliance with follow-up.

14. EVIDENCE ON IMPLEMENTATION

International guidance has recognized the inherent incremental software changes which could impact safety performance of AI-SaMDs. This highlights a need for rigorous AI software version management and post-deployment surveillance to ensure that safety and performance metrics are maintained over time.

Product and software requirements, whilst part of the technical documentation (NB not clinical evidence files), contain some evidence for safety and performance requirements that contribute to the overall assessment of clinical impact.

Software development

When software is in development, data management design should specify how the AI-SaMD will deal with:

- Incomplete datasets
- Paucity of datasets
- Wrong data format
- Data outside of specified value ranges
- Wrong temporal sequence of data.

Given the likelihood of AI-SaMD version updates, it is worth noting that incremental software changes – whether continuous or iterative, intentional or unintentional – could have serious consequences on safety performance after deployment. It is therefore vitally important that such changes are documented and identified by software version, and that a robust post-deployment surveillance plan is in place

Product development risk analysis

Risk analysis considers threats to data management and the quality of evidence that might be caused by internal errors and monitoring data.

A benefit-risk analysis of the AI-SaMD is required after review of both clinical and non-clinical evidence generated from post-market surveillance.

In the European Union, risk management files include *(94):*

- risks and instructions for their mitigation by users
- a list of scenarios where users can deal with any issues related to usability or evaluation of outputs.

Post-market surveillance and monitoring

Post-deployment, developers in the EU are required to implement and maintain a system that routinely monitors output from the AI-SaMD and downstream clinical outcomes for clinical performance and clinical safety *(95)*. The scope and nature of such post-market surveillance (PMS) should be appropriate to the device and its intended purpose.

Post-market surveillance and monitoring regularly generates new data such as safety reports, results from published literature, registries, post-market clinical follow-up (PMCF) studies, and other data about the use of AI-SaMD. This data needs to be checked for information that has the potential to change the evaluation of the risk/benefit analysis, and the clinical performance and clinical safety of the device.

Such data are required to be fed into the clinical evaluation process in a timely manner to avoid negative consequences. Data sources to be monitored for adverse events and unintended consequences include: scientific literature; clinical data and reports from users/health care professionals; customer communications; IT security databases; "bug reports"; databases of regulatory and governance bodies such as the UK's MHRA and the US FDA.

When analysing PMS data, the following should be carefully examined:

1. Quality of metrics: precision and accuracy, sensitivity/specificity
2. Selection of operating points for thresholds: the validation set is usually employed to set operating points as this better simulates prospective deployment; clinical outcome data should be monitored to ensure expected performance metrics are being observed
3. Variance of performance metrics over time
4. Is the data in the field (real-world performance data) consistent with expected data
5. Threshold values to trigger actions and modifications:
   - Re-evaluation of the benefits-risk analysis
   - Re-training of the algorithm (unlock, re-train, version update)
   - Product recall.

Post-market clinical follow-up

The European Union's MDCG 2020-8 document provides a template for evaluation reporting on post-market clinical follow-up *(89)*. It informs manufacturers of the activities that must be undertaken to generate evidence for evaluating PMCF. These include:

- Analysis of clinical data collected in the PMCF study
- Deviations from the initial plan for collecting PMCF data, if any, and their impact
- Discussion of results – impact on the risk-benefit analysis
- Conclusions relating back to intent of original post market surveillance (PMS) plan
- Identification and implementation of any corrective or preventive actions
- Evaluation of clinical data relating to similar AI-SaMD (same intended use)
- Impact of the results on clinical performance

Minimum standards for post-market clinical follow-up

Post market clinical follow-up	Considerations for evidence generation
Analysis of clinical data from a PMCF Study	Non-inferior to gold standard
Deviations from pre-specified PMCF Plan	Lack of adequate post-market reporting
Result and impact on Benefits-Risk Analysis	Evidence from Safety and Usability
Conclusions relating to initial PMS plan	Data management adherence in post-market monitoring and surveillance
Identification and Implementation of corrective actions	Re-training or re-tuning of the algorithm with additional datasets
Evaluation of clinical data relating to similar or equivalent device	Compare similar devices for effectiveness
Impact of results on clinical performance	Performance accuracy maintained over time

Use-case: Post-market follow-up for AI-SaMDs in cervical cancer screening

Table 14 below lists some considerations for evidence generated from post-market surveillance. Whilst not exhaustive, it should be noted that safety and performance monitoring post-implementation is required to show sustained clinical impact.

Table 14. Evaluating PMCF of AI-SaMDs for cervical cancer screening

PMCF activity	Considerations for evidence generation in cervical cancer screening
Analysis of clinical data from a PMCF study	Non-inferior to gold standard - VIA or Primary HPV screening
Deviations from pre-specified PMCF plan	Lack of adequate post-market reporting - follow-up of outcomes and data collection must be planned before deployment
Result and impact on benefits-risk analysis	Evidence from safety and usability - reports and logging of device use, acceptability of cervical images, barriers to adequate image collection, error reports, etc
Conclusions relating to initial post-market surveillance (PMS) plan	Data management adherence in post-market monitoring and surveillance - reports of data collected from screening programme and audit of effect of adding AI-SaMD to workflow
Identification and Implementation of corrective actions	Re-training or re-tuning of the algorithm with additional datasets - Identification of changes to performance accuracy over time and any groups or sub-populations of women in which the algorithm may show poor generalisability
Evaluation of clinical data relating to similar or equivalent device	Compare similar devices for effectiveness - comparison to other modalities for assisting screening and other alternatives to identifying women at high risk - et HPV screening, digital pathology, other available devices
Impact of results on clinical performance	Performance accuracy maintained over time - adequate detection of high grade abnormalities of the cervix, low false positive rates

15. EVIDENCE ON PROCUREMENT

As more studies of AI-SaMDs generate real-world evidence illustrating clinical impact, there remain many challenges to translating such evidence into clinical practice. Kelly et al *(96)* describe a number of these key challenges:

1. Lack of peer-reviewed RCTs as an evidence gold standard
2. Metrics do not always reflect clinical applicability
3. Difficulty comparing algorithms due to non-standardised reporting
4. Challenges related to machine learning science:
   - Algorithmic bias - the risk of increasing existing health inequalities
   - Dataset shift - due to shifting patient populations and where clinical and operational practices evolve over time
   - Accidentally overfitting confounders versus true signal
   - Challenges in generalisations to new populations and settings
   - Susceptibility to adversarial attack or manipulation
5. Logistical difficulties in Implementing AI systems

These challenges are compounded in LMICs *(97)* by additional factors:

- Lack of adequate supporting infrastructure for evidence generation
- Lack of regulatory structures and compliance
- Capacity for monitoring, surveillance and post-market evidence generation
- Resources capacity for training and sufficient expertise
- Adequate data storage/analytics capabilities

According to Mehta et al, the ability of AI to fulfil its promise to improve global health will depend on at least three key challenges being addressed *(98)*. Each of these challenges has implications for evidence generation:

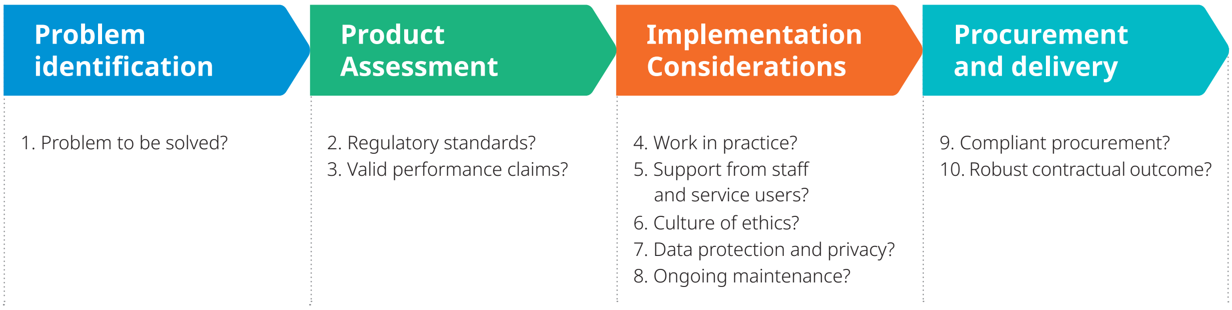
1. **Reliability and availability of data.** The limited availability or even absence of well-curated, high-fidelity, applicable clinical data sets in LMICs is described as a “foundational challenge”. Using algorithms built with input data from high-income countries can create biases in the AI system’s training and hence its responses. Due to the cost and complexities of conducting validation studies, many AI-based solutions are being adopted without a full understanding of their local applicability. Machine-driven decision making must be validated using data that are relevant to the context in which it will be deployed.
2. **Applying AI tools in health systems.** Matching the right AI-based tools to the right providers can be difficult, and that health systems face a daunting challenge in motivating untrained or poorly trained providers to use them. Adequate resources and infrastructure for training in the use of AI-SaMDs must be central to usability and post-market monitoring evaluations to ensure safety and maximum clinical benefit is abstained from implementing these devices as clinical decision support. Compliance is usually difficult to facilitate in these settings without adequate training and monitoring.
3. **Regulatory capacity.** Few LMIC health systems have the regulatory capacity to oversee and manage rapidly changing technologies. This poses challenges to effectively scaling AI-SaMDs to the health system level.

Guidance for procurement

To meet these various challenges, and to facilitate the implementation of safe and highly-performing AI-SaMDs, several organisations including the UK's NHSX AI Lab have published guidance for buyers and implementers *(99)*.

Figure 8 below illustrates the NHS's procurement pathway whilst Table 15 describes the considerations in the procurement guidance with respect for evidence requirements. These considerations should be applied to the global health context in order to ensure procurement only of AI-SaMDs that demonstrate evidence of **safety, performance within the clinical context,** and **clinical impact** related to its **intended use.**

Figure 8. Procurement checklist



Problem identification

1. Problem to be solved?

Product Assessment

2. Regulatory standards?
3. Valid performance claims?

Implementation Considerations

4. Work in practice?
5. Support from staff and service users?
6. Culture of ethics?
7. Data protection and privacy?
8. Ongoing maintenance?

Procurement and delivery

9. Compliant procurement?
10. Robust contractual outcome?

Source: NHSX, *A Buyer's Guide to AI for Health and Care (99)*

Table 15. Procurement guidance: evidence requirements

Questions and considerations	Evidence requirements
What problem are you trying to solve and is **AI the right solution?**	Literature review and appraisals
Does this product meet **regulatory standards?**	National or international standards
Does this product perform in line with the vendor's claims?	Evidence from a well-designed prospective study
Will this product **work in practice?**	Evidence from real-world performance studies (effectiveness, usability)
Can you secure the support you need from staff and service users?	Evidence of usability and integration into clinical workflows
Can you build and maintain a culture of **ethical responsibility** around this project?	Transparent reporting to international standards
What **data protection protocols** do you need to safeguard privacy and comply with the law?	National protocols, where they exist
Can you manage and maintain this product after you adopt it?	Evidence from post-market surveillance data (clinical and device related)
Is your procurement process fair, transparent and competitive?	Local and country level procurement processes should be established for the global health context
Can you ensure a commercially and legally robust **contractual outcome** for your organisation?	Relevant global health context - local/hospital level

Source: NHSX, *A Buyer's Guide to AI for Health and Care (99)*

REFERENCES

1. WHO. Global strategy on digital health 2020-2025 [Internet]. World Health Organization; [cited 2021 Jun 13]. Available from: https://cdn.who.int/media/docs/default-source/documents/gs4dhdaa2a9f352b0445bafbc79ca799dce4d.pdf?sfvrsn=f112ede5_75
2. WHO. Ethics and governance of artificial intelligence for health: WHO Guidance [Internet]. World Health Organization; [cited 2021 Jul 19]. Available from: https://www.who.int/publications-detail-redirect/9789240029200
3. Wahl B, Cossy-Gantner A, Germann S, Schwalbe NR. Artificial intelligence (AI) and global health: how can AI contribute to health in resource-poor settings? BMJ Glob Health. 2018 Aug 1;3(4):e000798.
4. USAID, Rockefeller Foundation. AI in Global Health: Defining a Collective Path Forward. [Internet]. USAID; 2019 Apr [cited 2021 Jul 29]. Available from: https://www.usaid.gov/sites/default/files/documents/1864/AI-in-Global-Health_webFinal_508.pdf
5. Buston, O, Chowdhury, Pick A, P. Digital health in low- and lower-middle-income countries. 2019 Sep [cited 2021 Jul 29]; Available from: https://pathwayscommission.bsg.ox.ac.uk/Digital-health-paper
6. Public Health England. Evaluating digital health products. In: GOVUK [Internet]. Public Health England; 2021 [cited 2021 Jul 29]. Available from: https://www.gov.uk/government/collections/evaluating-digital-health-products
7. Wiegand T, Krishnamurthy R, Kuglitsch M, Lee N, Pujari S, Salathé M, et al. WHO and ITU establish benchmarking process for artificial intelligence in health. The Lancet. 2019 Jul 6;394(10192):9–11.
8. WHO. 2015 Global Survey on Health Technology Assessment by National Authorities [Internet]. World Health Organization; 2016 [cited 2021 Jun 19]. Available from: https://www.who.int/publications-detail-redirect/9789241509749
9. WHO. Monitoring and evaluating digital health interventions. A practical guide to conducting research and assessment [Internet]. World Health Organization; 2016 [cited 2021 Jun 19]. Available from: http://www.who.int/reproductivehealth/publications/mhealth/digital-health-interventions/en/
10. WHO. WHO technical guidance and specifications of medical devices for screening and treatment of precancerous lesions in the prevention of cervical cancer [Internet]. World Health Organization; 2020 [cited 2021 Jun 19]. Available from: https://apps.who.int/iris/handle/10665/331698
11. WHO. Guidance for post-market surveillance and market surveillance of medical devices, including in vitro diagnostics [Internet]. 2020 [cited 2021 Jun 19]. Available from: https://www.who.int/publications-detail-redirect/guidance-for-post-market-surveillance-and-market-surveillance-of-medical-devices-including-in-vitro-diagnostics
12. WHO. WHO Consultation towards the development of guidance on ethics and governance of artificial intelligence for health [Internet]. World Health Organization; 2021 Mar [cited 2021 Jun 19]. Available from: https://www.who.int/publications-detail-redirect/who-consultation-towards-the-development-of-guidance-on-ethics-and-governance-of-artificial-intelligence-for-health
13. Kilic A. Artificial Intelligence and Machine Learning in Cardiovascular Health Care. Ann Thorac Surg. 2020 May 1;109(5):1323–9.
14. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019 Jan;25(1):44–56.

15. Topol EJ. Welcoming new guidelines for AI clinical research. Nat Med. 2020 Sep;26(9):1318–20.

16. Liu X, Faes L, Calvert MJ, Denniston AK, CONSORT/SPIRIT-AI Extension Group. Extension of the CONSORT and SPIRIT statements. Lancet Lond Engl. 2019 Oct 5;394(10205):1225.

17. Haibe-Kains B, Adam GA, Hosny A, Khodakarami F, Waldron L, Wang B, et al. Transparency and reproducibility in artificial intelligence. Nature. 2020 Oct;586(7829):E14–6.

18. Wang P, Berzin TM, Brown JRG, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. 2019 Oct 1;68(10):1813–9.

19. Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. Gut. 2019 Dec;68(12):2161–9.

20. Wijnberge M, Geerts BF, Hol L, Lemmers N, Mulder MP, Berge P, et al. Effect of a Machine Learning-Derived Early Warning System for Intraoperative Hypotension vs Standard Care on Depth and Duration of Intraoperative Hypotension During Elective Noncardiac Surgery: The HYPE Randomized Clinical Trial. JAMA. 2020 Mar 17;323(11):1052–60.

21. Gong D, Wu L, Zhang J, Mu G, Shen L, Liu J, et al. Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. Lancet Gastroenterol Hepatol. 2020 Apr 1;5(4):352–61.

22. Su J-R, Li Z, Shao X-J, Ji C-R, Ji R, Zhou R-C, et al. Impact of a real-time automatic quality control system on colorectal polyp and adenoma detection: a prospective randomized controlled study (with videos). Gastrointest Endosc. 2020 Feb;91(2):415-424.e4.

23. Wang P, Liu X, Berzin TM, Brown JRG, Liu P, Zhou C, et al. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol Hepatol. 2020 Apr 1;5(4):343–51.

24. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, et al. Diagnostic Efficacy and Therapeutic Decision-making Capacity of an Artificial Intelligence Platform for Childhood Cataracts in Eye Clinics: A Multicentre Randomized Controlled Trial. EClinicalMedicine. 2019 Mar 1;9:52–9.

25. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet Lond Engl. 2019 Apr 20;393(10181):1577–9.

26. Gregory J, Welliver S, Chong J. Top 10 Reviewer Critiques of Radiology Artificial Intelligence (AI) Articles: Qualitative Thematic Analysis of Reviewer Critiques of Machine Learning/Deep Learning Manuscripts Submitted to JMRI. J Magn Reson Imaging. 2020;52(1):248–54.

27. 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 Mar 25;368:m689.

28. Qin ZZ, Sander MS, Rai B, Titahong CN, Sudrungrot S, Laah SN, et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. Sci Rep. 2019 Oct 18;9(1):15000.

29. Panch T, Mattie H, Atun R. Artificial intelligence and algorithmic bias: implications for health systems. J Glob Health. 9(2):020318.

30. Bellemo V, Lim ZW, Lim G, Nguyen QD, Xie Y, Yip MYT, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. Lancet Digit Health. 2019 May 1;1(1):e35–44.

31. Beede E, Baylor E, Hersch F, Iurchenko A, Wilcox L, Ruamviboonsuk P, et al. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems [Internet]. New York, NY, USA: Association for Computing Machinery; 2020 [cited 2021 Jun 28]. p. 1–12. (CHI ’20). Available from: https://doi.org/10.1145/3313831.3376718

32. Hosny A, Aerts HJWL. Artificial intelligence for global health. Science. 2019 Nov 22;366(6468):955–6.

33. Kamulegeya LH, Okello M, Bwanika JM, Musinguzi D, Lubega W, Rusoke D, et al. Using artificial intelligence on dermatology conditions in Uganda: A case for diversity in training data sets for machine learning. bioRxiv. 2019 Oct 31;826057.

34. Adamson AS, Smith A. Machine Learning and Health Care Disparities in Dermatology. JAMA Dermatol. 2018 Nov 1;154(11):1247–8.

35. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, 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 Sep;26(9):1364–74.

36. Oliveira AD, Prats C, Espasa M, Zarzuela Serrat F, Montañola Sales C, Silgado A, et al. The Malaria System MicroApp: A New, Mobile Device-Based Tool for Malaria Diagnosis. JMIR Res Protoc. 2017 Apr 25;6(4):e70.

37. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, et al. Human–computer collaboration for skin cancer recognition. Nat Med. 2020 Aug;26(8):1229–34.

38. Brinker TJ, Hekler A, Enk AH, Berking C, Haferkamp S, Hauschild A, et al. Deep neural networks are superior to dermatologists in melanoma image classification. Eur J Cancer. 2019 Sep 1;119:11–7.

39. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb;542(7639):115–8.

40. Kim H, Goo JM, Lee KH, Kim YT, Park CM. Preoperative CT-based Deep Learning Model for Predicting Disease-Free Survival in Patients with Lung Adenocarcinomas. Radiology. 2020 Jul;296(1):216–24.

41. Geras KJ, Mann RM, Moy L. Artificial Intelligence for Mammography and Digital Breast Tomosynthesis: Current Concepts and Future Perspectives. Radiology. 2019 Nov;293(2):246–59.

42. Hu L, Bell D, Antani S, Xue Z, Yu K, Horning MP, et al. An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening. JNCI J Natl Cancer Inst. 2019 Sep 1;111(9):923–32.

43. Miyagi Y, Takehara K, Nagayasu Y, Miyake T. Application of deep learning to the classification of uterine cervical squamous epithelial lesion from colposcopy images combined with HPV types. Oncol Lett. 2020 Feb;19(2):1602–10.

44. Yuan C, Yao Y, Cheng B, Cheng Y, Li Y, Li Y, et al. The application of deep learning based diagnostic system to cervical squamous intraepithelial lesions recognition in colposcopy images. Sci Rep. 2020 Jul 15;10(1):11639.

45. Zhang Y, Li L, Gu J, Wen T, Xu Q. Cervical Precancerous Lesion Detection Based on Deep Learning of Colposcopy Images. J Med Imaging Health Inform. 2020 May 1;10(5):1234–41.

46. McKinney SM, Sieniek M, Godbole V, Godwin J, Antropova N, Ashrafian H, et al. International evaluation of an AI system for breast cancer screening. Nature. 2020 Jan;577(7788):89–94.

47. Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, et al. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019 Sep 1;111(9):916–22.

48. Rodríguez-Ruiz A, Krupinski E, Mordang J-J, Schilling K, Heywang-Köbrunner SH, Sechopoulos I, et al. Detection of Breast Cancer with Mammography: Effect of an Artificial Intelligence Support System. Radiology. 2019 Feb;290(2):305–14.

49. Alami H, Rivard L, Lehoux P, Hoffman SJ, Cadeddu SBM, Savoldelli M, et al. Artificial intelligence in health care: laying the Foundation for Responsible, sustainable, and inclusive innovation in low- and middle-income countries. Glob Health. 2020 Jun 24;16(1):52.

50. 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 Oct;1(6):e271–97.

51. Zhu J, Shen B, Abbasi A, Hoshmand-Kochi M, Li H, Duong TQ. Deep transfer learning artificial intelligence accurately stages COVID-19 lung disease severity on portable chest radiographs. PLOS ONE. 2020 Jul 28;15(7):e0236621.

52. Mollura DJ, Culp MP, Pollack E, Battino G, Scheel JR, Mango VL, et al. Artificial Intelligence in Low- and Middle-Income Countries: Innovating Global Health Radiology. Radiology. 2020 Dec;297(3):513–20.

53. Marmot MG, Altman DG, Cameron DA, Dewar JA, Thompson SG, Wilcox M. The benefits and harms of breast cancer screening: an independent review. Br J Cancer. 2013 Jun 11;108(11):2205–40.

54. Seely JM, Alhassan T. Screening for breast cancer in 2018—what should we be doing today? Curr Oncol. 2018 Jun;25(Suppl 1):S115–24.

55. Fenton JJ, Taplin SH, Carney PA, Abraham L, Sickles EA, D'Orsi C, et al. Influence of computer-aided detection on performance of screening mammography. N Engl J Med. 2007 Apr 5;356(14):1399–409.

56. Salim M, Wåhlin E, Dembrower K, Azavedo E, Foukakis T, Liu Y, et al. External Evaluation of 3 Commercial Artificial Intelligence Algorithms for Independent Assessment of Screening Mammograms. JAMA Oncol. 2020 Oct 1;6(10):1581–8.

57. Office of the Commissioner. FDA Authorizes Marketing of First Device that Uses Artificial Intelligence to Help Detect Potential Signs of Colon Cancer [Internet]. FDA. FDA; 2021 [cited 2021 Jul 29]. Available from: https://www.fda.gov/news-events/press-announcements/fda-authorizes-marketing-first-device-uses-artificial-intelligence-help-detect-potential-signs-colon

58. Kate Brush. Use Case [Internet]. SearchSoftwareQuality. 2020 [cited 2021 Jun 22]. Available from: https://searchsoftwarequality.techtarget.com/definition/use-case

59. Julie Platt. Case Studies or Use Cases? [Internet]. Case Study Writer. 2020 [cited 2021 Jun 22]. Available from: https://casestudywriter.co.uk/whats-the-difference-between-case-studies-and-use-cases/

60. IARC. Global Cancer Observatory [Internet]. Global Cancer Observatory. [cited 2021 Jul 29]. Available from: https://gco.iarc.fr/

61. WHO. Global strategy to accelerate the elimination of cervical cancer as a public health problem [Internet]. World Health Organization. 2020 [cited 2021 Jun 13]. Available from: https://www.who.int/publications-detail-redirect/9789240014107

62. Xue Z, Novetsky AP, Einstein MH, Marcus JZ, Befano B, Guo P, et al. A demonstration of automated visual evaluation of cervical images taken with a smartphone camera. Int J Cancer. 2020 Nov 1;147(9):2416–23.

63. IMDRF. Clinical Evaluation (Final Document) [Internet]. 2019. Available from: http://www.imdrf.org/docs/imdrf/final/technical/imdrf-tech-191010-mdce-n56.pdf

64. IMDRF. Software as a Medical Device (SaMD). IMDRF; 2017.

65. World Health Organization, editor. WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention. Geneva: World Health Organization; 2013. 40 p.

66. European Union. REGULATION (EU) 2017/745 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 5 April 2017 on medical devices, amending Directive 2001/83/EC, regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC [Internet]. European Union; 2017 [cited 2021 Jun 23]. Available from: http://www.bloomsburycollections.com/book/fundamental-texts-on-european-private-law-1

67. ISO. ISO 14971:2019. Medical devices — Application of risk management to medical devices [Internet]. International Standards Organization. 2019 [cited 2021 Jun 23]. Available from: https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/07/27/72704.html

68. Health C for D and R. Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. FDA [Internet]. 2021 Jan 11 [cited 2021 Jul 27]; Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

69. Center for Devices and Radiological Health. Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan. FDA [Internet]. 2021 Jan 11 [cited 2021 Jul 29]; Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

70. Larson DB, Harvey H, Rubin DL, Irani N, Tse JR, Langlotz CP. Regulatory Frameworks for Development and Evaluation of Artificial Intelligence–Based Diagnostic Imaging Algorithms: Summary and Recommendations. J Am Coll Radiol. 2021 Mar;18(3):413–24.

71. Cardoso JR, Pereira LM, Iversen MD, Ramos AL. What is gold standard and what is ground truth? Dent Press J Orthod. 2014;19(5):27–30.

72. Chan A-W, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ. 2013 Jan 9;346:e7586.

73. 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 Mar 24;340:c869.

74. Catalá-López F, Alonso-Arroyo A, Page MJ, Hutton B, Ridao M, Tabarés-Seisdedos R, et al. Reporting guidelines for health research: protocol for a cross-sectional analysis of the EQUATOR Network Library. BMJ Open. 2019 Mar 1;9(3):e022769.

75. Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI Extension. BMJ. 2020 Sep 9;370:m3210.

76. Faes L, Liu X, Wagner SK, Fu DJ, Balaskas K, Sim DA, et al. A Clinician’s Guide to Artificial Intelligence: How to Critically Appraise Machine Learning Studies. Transl Vis Sci Technol. 2020 Jan 28;9(2):7–7.

77. ITU-T Focus Group on AI for Health. Updated DEL2.2: Guidelines for AI based medical device: Regulatory requirements (Draft: April 2020) [Internet]. ITU; 2020 [cited 2021 Jul 18]. Available from: https://www.itu.int

78. Chen P-HC, Liu Y, Peng L. How to develop machine learning models for healthcare. Nat Mater. 2019 May;18(5):410–4.

79. Willemink MJ, Koszek WA, Hardell C, Wu J, Fleischmann D, Harvey H, et al. Preparing Medical Imaging Data for Machine Learning. Radiology. 2020 Apr;295(1):4–15.

80. Zech J, Pain M, Titano J, Badgeley M, Schefflein J, Su A, et al. Natural Language-based Machine Learning Models for the Annotation of Clinical Radiology Reports. Radiology. 2018 May;287(2):570–80.

81. Miller RJ. Big Data Curation. In: COMAD. 2014. p. 4.

82. FDA. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback [Internet]. US Food & Drug Administration; 2019 [cited 2021 Jul 23]. Available from: https://www.fda.gov/files/medical%20 devices/published/US-FDA-Artificial-Intelligence-and-Machine-Learning-Discussion-Paper.pdf

83. Qin ZZ, Ahmed S, Sarker MS, Paul K, Adel ASS, Naheyan T, et al. Can Artificial Intelligence (AI) Be Used to Accurately Detect Tuberculosis (TB) from Chest X-Rays? An Evaluation of Five AI Products for TB Triaging in a High TB Burden Setting [Internet]. Rochester, NY: Social Science Research Network; 2020 Oct [cited 2021 Jul 26]. Report No.: ID 3702975. Available from: https://papers.ssrn.com/abstract=3702975

84. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016 Dec 13;316(22):2402–10.

85. Vijayananthan A, Nawawi O. The importance of Good Clinical Practice guidelines and its role in clinical trials. Biomed Imaging Interv J. 2008 Jan 1;4(1):e5.

86. DECIDE-AI Steering Group. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. Nat Med. 2021 Feb;27(2):186–7.

87. Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009 Jul 21;6(7):e1000097.

88. MEDDEV Guidance List [Internet]. Medical Device Regulation. [cited 2021 Jul 30]. Available from: https://www.medical-device-regulation.eu/meddev-guidance-list-download/

89. Medical Device Coordination Group. MDCG 2020-8 Post-market clinical follow-up (PMCF) Evaluation Report Template. A guide for manufacturers and notified bodies [Internet]. European Union; 2020. Available from: https://ec.europa.eu/docsroom/documents/40906/attachments/1/translations/en/renditions/native

90. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. Npj Digit Med. 2020 Mar 23;3(1):1–4.

91. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. Model Cards for Model Reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency [Internet]. New York, NY, USA: Association for Computing Machinery; 2019 [cited 2021 Jul 28]. p. 220–9. (FAT* ’19). Available from: https://doi.org/10.1145/3287560.3287596

92. Alba AC, Agoritsas T, Walsh M, Hanna S, Iorio A, Devereaux PJ, et al. Discrimination and Calibration of Clinical Prediction Models: Users’ Guides to the Medical Literature. JAMA. 2017 Oct 10;318(14):1377–84.

93. van Smeden M, Van Calster B, Groenwold RHH. Machine Learning Compared With Pathologist Assessment. JAMA. 2018 Apr 24;319(16):1725–6.

94. Chan T, Tong RKY. ISO 14971: Application of Risk Management to Medical Devices. In: Wong J, Tong R, editors. Handbook of Medical Device Regulatory Affairs in Asia [Internet]. 2nd ed. Jenny Stanford Publishing; 2018 [cited 2021 Jul 29]. p. 175–91. Available from: https://www.taylorfrancis.com/books/9780429996771/chapters/10.1201/9780429504396-16

95. Ecker W, Labek G, Mittermayr T. Clinical Evaluation and Investigation of Medical Devices under the new EU-Regulation. Books On Demand; 2020. 282 p.

96. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Med. 2019 Oct 29;17(1):195.

97. Paul AK, Schaefer M. Safeguards for the use of artificial intelligence and machine learning in global health. Bull World Health Organ. 2020 Apr 1;98(4):282–4.

98. Mehta MC, Katz IT, Jha AK. Transforming Global Health with AI. N Engl J Med [Internet]. 2020 Feb 26 [cited 2021 Jul 29]; Available from: https://www.nejm.org/doi/10.1056/NEJMp1912079

99. Indra Joshi, Dominic Cushnan. A Buyer's Guide to AI in Health and Care [Internet]. NHS; [cited 2021 Jul 29]. Available from: https://www.nhsx.nhs.uk/ai-lab/explore-all-resources/adopt-ai/a-buyers-guide-to-ai-in-health-and-care/

GLOSSARY

Note on sources and references: Definitions and terminology as described by the International Medical Device Regulators Forum (IMDRF) have been used in cases where multiple related definitions of the term exist.

Technical terms

Algorithm	A final set of instructions (or rules) that defines a sequence of operations for solving a particular computational problem for all the problem instances for a problem set
Artificial Intelligence (AI)	The most general of computer reasoning terms - it includes any system that aims to mimic human intelligence by learning from data and/or by applying manually defined decision rules.
Class-activation map	Class-activation maps are particularly relevant to image classification AI interventions. Class-activation maps are visualisations of the pixels that had the greatest influence on predicted class, by displaying the gradient of the predicted outcome from the model with respect to the input. They are also referred to as “saliency maps” or “heat maps”.
Convolutional Neural Networks (CNN)	The type of deep neural networks most frequently applied in medical image analysis
Computer Vision	A scientific field that deals with how computers gain a high-level understanding from digital images or videos. From the perspective of engineering, it aims to automate tasks that the human visual system can do.
Deep Learning	Among neural networks, deep learning, which involves the study of neural networks consisting of many layers, is currently the most successful in practical applications and the subject of most intense research.
Fine-tuning	Modifications or additional training performed on the AI intervention model, done with the intention of improving its performance.
Locked Algorithm	An algorithm that provides the same result each time the same input is applied to it and does not change with use.
Machine Learning (ML)	A field of computer science concerned with the development of models/ algorithms that can solve specific tasks by learning patterns from data, rather than by following explicit rules. It is seen as an approach within the field of AI.
Model	Output from a predictive algorithm using training data
Neural Networks	Simplified from Artificial Neural Networks (ANN). An ANN is based on a collection of connected units or nodes called artificial neurons which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons.

Statistical terms and abbreviations

AUC (also AUROC)	The Area Under a Receiver Operating Characteristic curve is a measure of the usefulness of a test in general, where a greater area means a more useful test, the areas under ROC curves are used to compare the usefulness of tests
Confusion Matrix	A table used to describe the performance of a statistical classification model (or classifier) on a set of test data for which the true values are known.
Negative Predictive Value (NPV)	Describes the accuracy and precision of a performance test. NPV refers to the proportion of negative results that are true negative. NPV = True Negatives / (True Negatives + False Negatives)
Precision	Precision, also called positive predictive value, is the fraction of relevant instances among the retrieved instances
Positive Predictive Value (PPV)	Describes the accuracy and precision of a performance test. PPV refers to the proportion of positive results that are true positives. PPV = True Positives/ (True Positives + False Positives)
Recall	Recall (also known as sensitivity) is the fraction of relevant instances that were retrieved.
Receiver Operating Characteristic	A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied
Sensitivity	Measures the proportion of positives that are correctly identified. Sensitivity = True Positives/ True Positives + False Negatives.
Specificity	Measures the proportion of negatives that are correctly identified. Specificity = True Negatives/True Negatives + False Positives.
Youden’s J Index	Youden's J statistic (also called Youden's index) is a single statistic that captures the performance of a dichotomous diagnostic test. J = (Sensitivity + Specificity) - 1

Clinical and scientific terms

Colposcopy	Visualisation of the cervix under magnification
Digital Cervicography	The process of digital image capture of the cervix, usually after staining with acetic acid
Interval Cancers	Cancers presenting in the interval following a negative screening. This may represent missed cancers, occult or new cancers arising in the interim)
Screen-detected Cancers	Cancers diagnosed at screening before symptoms occur.
Symptomatic detected cancers	Cancers diagnosed outside of routine screening when symptoms present

Product terms

Application ("app")	A program designed for end users
Hardware	Physical parts of a computer such as CPU (central processing unit), monitor, keyboard, computer data storage, graphic cards, etc
Software	A collection of data or instructions that tell a computer how to work
Software as a Medical Device (SaMD)	Software intended to be used for one or more medical purposes that perform these purposes without being part of a hardware medical device.
Total Product Life-cycle (TPLC)	Framework for assessing a device from development (pre-market) to post-market and eventual demise.

Evaluation terms

Clinical data	Safety, clinical performance, and/or effectiveness information that is generated from the clinical use of a medical device
Clinical evaluation	A set of ongoing activities that use scientifically sound methods for the assessment and analysis of clinical data to verify the safety, clinical performance and/or effectiveness of the device when used as intended by the manufacturer.
Clinical evidence	The clinical data and its evaluation pertaining to a medical device.
Clinical investigation	Any systematic investigation or study in or on one or more human subjects, undertaken to assess the safety, clinical performance and/or effectiveness of a medical device
Clinical outcome	Measured variables in a clinical trial that are used to assess the effects of an intervention
Clinical outcome assessment	The FDA defines a clinical outcome assessment as "a measure that describes or reflects how a patient feels, functions, or survives." They then go into four different types of clinical outcome that you could describe.
Clinical performance	The ability of a medical device to achieve its intended clinical purpose as claimed by the manufacturer.
Clinical trials	A properly conducted clinical investigation, including compliance to the clinical investigation plan and local laws and regulations, ensures the protection of human subjects, the integrity of the data and that the data obtained is acceptable for the purpose of demonstrating the SaMD's conformity to the Essential Principles
Clinical validation	The ability of a SaMD to yield a clinically meaningful output associated to the target use of SaMD output in with the target health care situation or condition identified in the SaMD definition statement

Development environment	The clinical, and operational settings from which the data used for training the model are generated. This includes all aspects of the physical setting (such as geographical location, physical environment), operational setting (such as integration with an electronic record system, installation on a physical device) and clinical setting (such as primary, secondary and/or tertiary care, patient disease spectrum)
Effectiveness	The ability of a medical device to achieve clinically meaningful outcome(s) in its intended use as claimed by the manufacturer.
Human–computer interaction	Human-computer interaction is a multidisciplinary field that focuses on the design of computer technology and, in particular, the interaction between humans (the users) and computers.
Input data	The data that needs to be presented to the AI system to allow it to serve its purpose.
Intended use	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.
Internal validation	
Locked algorithm	
Operational environment	The environment (technical, physical or clinical) in which the AI system will be deployed, including the infrastructure required to enable the AI system to function.
Output data	The predicted output given by the AI system based on processing of the input data. The output data can be presented in different forms, including a classification (including diagnosis, disease severity or stage, or recommendation such as referability), a probability, a class-activation map, etc.
Performance error	Instances in which the AI system fails to perform as expected. This term can describe different types of failures, and it is up to the investigator to specify what should be considered a performance error, preferably based on prior evidence. This can range from small decreases in accuracy (compared to expected accuracy) to erroneous predictions or the inability to produce an output, in certain cases.
Post-market clinical follow-up study (PMCF-study)	Study carried out following marketing approval intended to answer specific questions relating to clinical safety or performance (i.e. Residual risks) of a medical device when used in accordance with its approved labelling (ISO 20416)
Post-market surveillance (PMS)	Systematic process to collect and analyse the performance of medical devices that have been placed on the market (ISO 13485)
Real world data	Data generated after a product has been released to the market that can provide insight into the performance of the product used in actual clinical settings, in routine medical practice, and by regular use by consumers
Real world evidence	Evidence derived from aggregation and analysis of real world data

Real world performance	Information on real-world device use and performance from a wider patient population than a more controlled study or pertinent literature, and thus provide information that cannot be obtained through such studies
Safety	Acceptability of risks as weighed against benefits, when using the medical device according to the manufacturer's labelling.
Scientific validity (valid clinical association)	The extent to which the SaMD's output (concept, conclusion, measurements) is clinically accepted or well founded (based on an established scientific framework or body of evidence) and corresponds accurately in the real world to the health care situation and condition identified in the SaMD definition statement (corresponds to the level of clinical acceptance of the SaMD's output)
Verification	Objective evidence that the specific requirements have been fulfilled
Validation	Objective evidence that the requirements for a specific intended use have been fulfilled

ANNEXES

ANNEX 1. SUMMARY OF GUIDANCE AND REGULATIONS

Evidence reporting guidance

SPIRIT-AI and CONSORT-AI	https://www.clinical-trials.AI/
EQUATOR NETWORK	https://www.equator-network.org/reporting-guidelines/
STARD-AI	Sounderajah V, Ashrafian H, Aggarwal R, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: The STARD-AI Steering Group. Nat Med 2020; 26: 807–08
TRIPOD-ML	Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393:1577–1579. doi:10.1016/s0140-6736(19) 30037-6

International medical device regulators forum (IMDRF)

http://www.imdrf.org/documents/documents.asp

- SaMD: Key Definitions (N10)
- SaMD: Possible Framework for Risk Categorisation and considerations (N12) 2014
- SaMD: Application of Quality Management System (QMS) (N23) 2015
- SaMD: Clinical Evaluation (N41) 2017
- SaMD: Clinical Evidence (N55) 2019
- SaMD: Clinical Evaluation (N56) 2019
- SaMD: Clinical Investigation (N57) 2019

WHO guidance

WHO	Monitoring and Evaluating Digital Health Interventions, 2016 https://www.who.int/reproductivehealth/publications/mhealth/digital-health-interventions/en/
WHO DHI	Digital Health Strategy. Draft, July 2020 https://www.who.int/health-topics/digital-health#tab=tab_1

International organisation for standardization (ISO)

ISO/IEC CD 23053 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)

https://www.iso.org/standard/74438.html
https://www.iso.org/standards.html

International regulatory guidance

US-FDA	Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning [AI/ML]- Based Software as a Medical Device (SaMD) 2019
ITU-T. FG-AI4H-I-036.	Guidelines for AI based medical device: Regulatory requirements (Draft: April 2020)
EU -European Union Medical Device Regulation EU 2017/745	MEDDEV 2.7/1 revision 4 (June 2016) CLINICAL EVALUATION: A Guide for Manufacturers and Notified Bodies Under Directives 93/42/EEC and 90/385/EEC MDCG 2020-5. Medical Device Coordinating Group. Clinical Evaluation – Equivalence: a guide for manufacturers and notified bodies, April 2020

White papers and reports

USAID, Rockefeller Foundation, Gates Foundation	Artificial Intelligence in Global Health. Defining a collective path forward.; 2019 https://www.usaid.gov/cii/AI-in-global-health
Digital Health in LLMICs - Pathway Commission Report 2019	Chowdhury, A. & Pick, A. (2019) Digital Health in LLMICs: Current and future technological developments with the potential to improve health outcomes in low- and lower-middle-income countries Pathways for Prosperity Commission Background Paper Series; no. 28. Oxford, United Kingdom. https://pathwayscommission.bsg.ox.ac.uk/
United Kingdom - National Health Service (NHS)	Artificial Intelligence: How to get it right. NHSX, 2019 A Buyer's Guide to AI for Health and Care NHSX, 2020 https://www.nhsx.nhs.uk/key-tools-and-info/
United Kingdom - NICE	Evidence Standards Framework for Digital Health Technologies 2019 https://www.nice.org.uk/about/what-we-do/our-programmes/evidence-standards-framework-for-digital-health-technologies
Public Health England.	Guide to Evaluating Digital Health Products, 2020. https://www.gov.uk/guidance/get-started-evaluating-digital-health-products
United Kingdom - National Screening committee:	Interim Guidance for those wishing to incorporate artificial intelligence into the National Breast Cancer Screening Programme. Gov.uk, 2020

ANNEX 2. EVIDENCE GENERATION CHECKLISTS

Validation roadmap: evidence generation and evaluation components

Components of evaluation	AI algorithm training and tuning	AI algorithm internal validation	AI algorithm external validation	Post implementation clinical follow-up	Evidence generation
Input data	**Image quality** **Image labelling** **Training dataset**	Input data, dataset split management	External test set management	Prospective real world data in target clinical setting	
Research standards	**Ground truth confidence**	Internal validation methods	Study design Methodology Comparators (healthcare workers vs AI)	Piloting & monitoring Independent (peer) review	
Clinical trial and investigation	**Pre-specified hypothesis, intended use,** **Methodology**	Feasibility study	Efficacy and effectiveness studies	Prospective real world performance study in clinical pathway	
Reporting Standards	**Outcome data and AI explainability**	Model accuracy	Clinical performance metrics	Measures of clinical impact, usability	Scientific review of literature - AI intervention studies

ANNEX 3. MINIMUM STANDARDS SUMMARY

Minimum standards for defining intended use

- What is the medical indication for use of the AI-SaMD?
- What part of the body/ system is being investigated?
- What use environment will the device be used in?
- What are the ways in which the device can be misused?
- For what patient population?
- What is the specific user profile and expertise?
- What is the exact operating principle of the device?
- What are the possible unintended consequences of the device and how will these risks be mitigated?

Minimum standards for model development

Full description of AI model and architecture including associated hardware:

Training set	Dataset description
Tuning set	Dataset description
Internal validation set	Dataset description

Minimum standards to be met in external validation

- Dataset management should feature out-of-sample “unseen” (i.e. protected from developers/ investigators) test sets of input data or images
- Piloting and monitoring of data collection should be carried out to ensure diagnostic accuracy is maintained
- Independent (peer) review should be carried out on output data
- The algorithm should be retrained if performance of AI-SaMD does not meet pre-specified performance target
- The algorithm version should be updated and re-tested on prospective independent test set

Minimum standards for data management

- Data sources and selection (including missing data)
- Data curation, processing and augmentation
- Data quality and demographic distribution
- Dataset split methodology and any overlaps in use of data

Minimum standards for reporting technical evidence

- Full details on development of the AI algorithm including intended use, subject populations, training and testing data, and public accessibility of the code
- Technical information regarding on-site application of the AI technology
- Details about Human–AI interactions, including required expertise of the user and how the AI output contributed to clinical decision making
- Specificity with regards to what version of an AI algorithm was used, given that performance of some algorithms can change iteratively, or in some cases, continuously

Minimum standards for evidence in evaluating usability

- Evidence of integration into clinical workflow with sustained overall benefit
- Infrastructure and conditions to allow for use of device as intended
- Effects of adding AI-SaMD to current standard
- Effects of disagreements between output of AI-SaMD and clinical decision of health care worker
- Users' interaction with output of AI-SaMD. Is the output interpretable? Error rates? Is the image readable?

Minimum standards for clinical impact evaluation

- Comparison to gold standard
- Measures of improvement in patient outcomes, clinical process, or time efficiency
- Measures of acceptable unintended consequences and absence of harm to patients
- Changes in experience of patient or user (i.e. health care worker)

Minimum standards for post-market clinical follow-up

Post market clinical follow-up	Considerations for evidence generation
Analysis of clinical data from a PMCF Study	Non-inferior to gold standard
Deviations from pre-specified PMCF Plan	Lack of adequate post-market reporting
Result and impact on Benefits-Risk Analysis	Evidence from Safety and Usability
Conclusions relating to initial PMS plan	Data management adherence in post-market monitoring and surveillance
Identification and Implementation of corrective actions	Re-training or re-tuning of the algorithm with additional datasets
Evaluation of clinical data relating to similar or equivalent device	Compare similar devices for effectiveness
Impact of results on clinical performance	Performance accuracy maintained over time

Digital Health and Innovation
Division of the Chief Scientist

World Health Organization
Avenue Appia 20
1121 Geneva 27
Switzerland

9789240038462