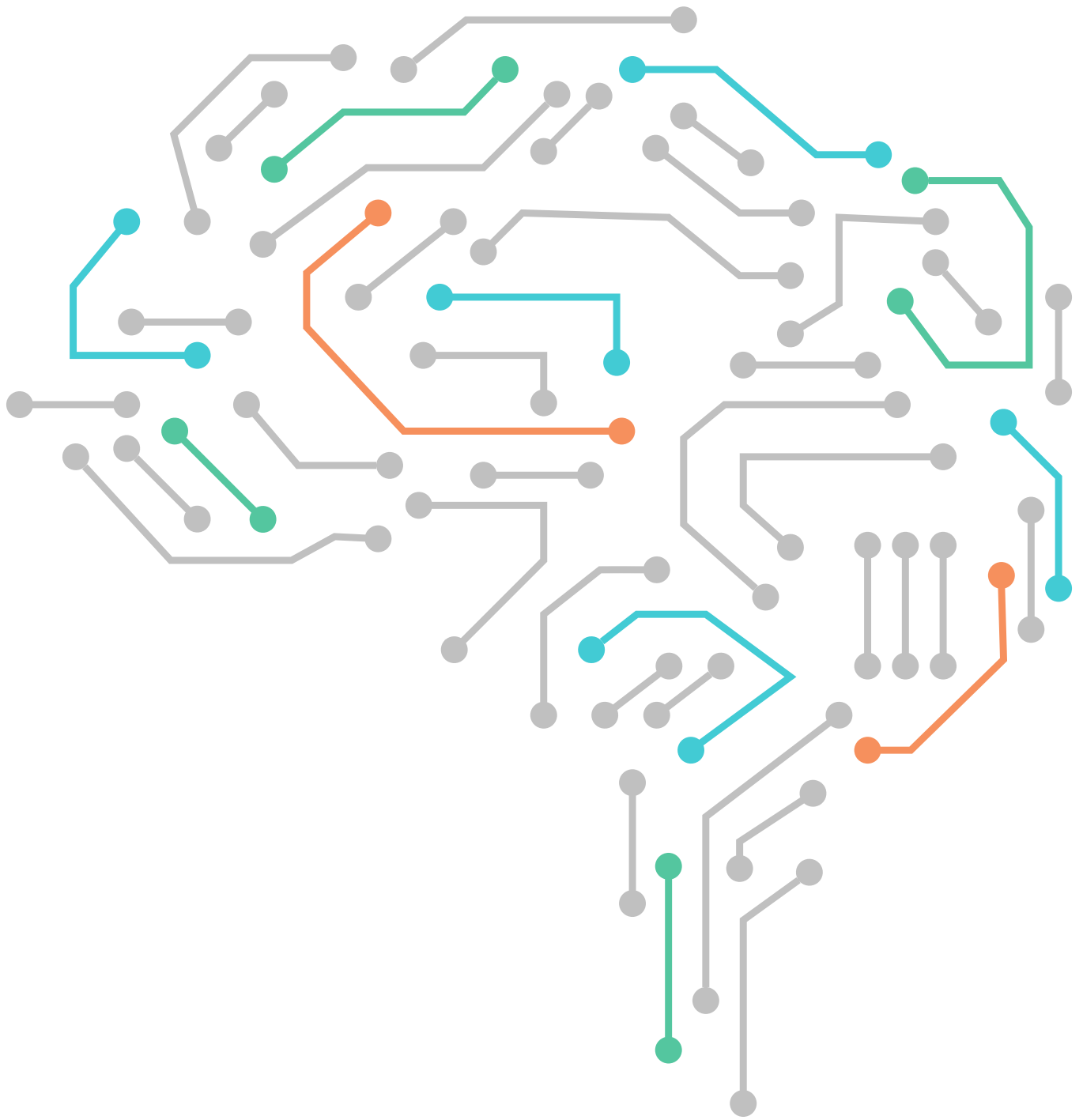


**GENERATING EVIDENCE FOR ARTIFICIAL  
INTELLIGENCE-BASED MEDICAL DEVICES:  
A FRAMEWORK FOR TRAINING,  
VALIDATION AND EVALUATION**





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

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)

© World Health Organization 2021

Some rights reserved. This work is available under the Creative Commons Attribution-NonCommercial-ShareAlike3.0 IGO licence (CC BY-NC-SA3.0 IGO; <https://creativecommons.org/licenses/by-nc-sa/3.0/igo>).

Under the terms of this licence, you may copy, redistribute and adapt the work for non-commercial purposes, provided the work is appropriately cited, as indicated below. In any use of this work, there should be no suggestion that WHO endorses any specific organization, products or services. The use of the WHO logo is not permitted. If you adapt the work, then you must license your work under the same or equivalent Creative Commons licence. If you create a translation of this work, you should add the following disclaimer along with the suggested citation: "This translation was not created by the World Health Organization (WHO). WHO is not responsible for the content or accuracy of this translation. The original English edition shall be the binding and authentic edition".

Any mediation relating to disputes arising under the licence shall be conducted in accordance with the mediation rules of the World Intellectual Property Organization (<http://www.wipo.int/amc/en/mediation/rules/>).

**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

---

<b>Foreword</b>	<b>vi</b>
<b>Acknowledgements</b>	<b>vii</b>
<b>Abbreviations and acronyms</b>	<b>viii</b>
<b>Executive summary</b>	<b>x</b>
<b>1. Introduction</b>	<b>1</b>
<b>SECTION I. SOFTWARE DEVELOPMENT</b>	<b>5</b>
<b>2. Artificial intelligence in health</b>	<b>7</b>
Current evidence in health applications	7
Use-case: AI-SaMD in cervical cancer screening	10
<b>3. Framework for evaluation</b>	<b>13</b>
Evaluation components	13
Clinical evaluation	14
Use-case: Evaluation for cervical cancer diagnosis	15
<b>4. Intended use</b>	<b>17</b>
Risk classification	17
Changes to intended use	18
Considerations for global health	18
<i>Minimum standards for defining intended use</i>	19
Use-case: Defining intended use for AI-SaMD in cervical cancer screening	19
<b>5. Model development and training for clinical evaluation</b>	<b>21</b>
Designing an AI-based model	22
Clinical study design	24
Better protocols and reporting of clinical trials	24
<i>Minimum standards for model development</i>	26
Use-case: Model development and training for AI-SaMDs in cervical cancer screening	27
<b>6. Dataset management</b>	<b>29</b>
Terminology	29
Model training	31
Model validation	31
Use-case: Dataset construction in AI-SaMDs for cervical cancer screening	32
<b>7. Internal validation and data management</b>	<b>33</b>
Data handling	33
Ground truth confidence	34
Use-case: internal validation for AI-SaMDs in cervical cancer screening	36

<b>SECTION II. SOFTWARE VALIDATION AND REPORTING</b>	<b>37</b>
<b>8. External validation</b>	<b>39</b>
Published case studies	39
<i>Minimum standards to be met in external validation</i>	40
Use-case: External validation for AI-SaMDs in cervical cancer screening	40
<b>9. Data management</b>	<b>41</b>
<i>Minimum standards for data management</i>	42
Use-case: Data management for AI-SaMDs in cervical cancer screening	42
<b>10. Evidence generation standards</b>	<b>43</b>
International standards	45
Use-case: Applying international standards to AI-SaMDs in cervical cancer screening	48
<b>11. Evidence reporting</b>	<b>49</b>
Data Sources	49
Reporting standards	50
<i>Minimum standards for reporting technical evidence</i>	51
Use-case: Reporting for AI-SaMDs in cervical cancer screening	51
<b>SECTION III. DEPLOYMENT AND POST-MARKET SURVEILLANCE</b>	<b>55</b>
<b>12. Evaluation of usability</b>	<b>57</b>
Guidance for usability evaluation	57
<i>Minimum standards for evidence in evaluating usability</i>	57
Use-case: Usability for AI-SaMDs in cervical cancer screening	58
<b>13. Evaluation of clinical impact</b>	<b>59</b>
Real world performance testing	61
<i>Minimum standards for clinical impact evaluation</i>	61
Use-case: Clinical impact for AI-SaMDs in cervical cancer screening	62
<b>14. Evidence on implementation</b>	<b>63</b>
Software development	63
Product development risk analysis	63
Post-market surveillance and monitoring	64
Post-market clinical follow-up	64
<i>Minimum standards for post-market clinical follow-up</i>	65
Use-case: Post-market follow-up for AI-SaMDs in cervical cancer screening	65
<b>15. Evidence on procurement</b>	<b>67</b>
Guidance for procurement	68
<b>References</b>	<b>69</b>
<b>Glossary</b>	<b>77</b>
<b>Annexes</b>	<b>83</b>
<b>Annex 1. Summary of guidance and regulations</b>	<b>85</b>
<b>Annex 2. Evidence generation checklists</b>	<b>87</b>
<b>Annex 3. Minimum standards summary</b>	<b>88</b>

## List of figures

Figure 1. Evidence generation and stages of clinical evaluation.....	21
Figure 2. Phases of development and evaluation for AI-SaMD diagnostic algorithms.....	22
Figure 3. Overview of dataset evaluation components.....	31
Figure 4. Data governance: the process of handling medical image data.....	34
Figure 5. Value hierarchy of data annotation.....	35
Figure 6. Good machine learning practices: total product life cycle approach.....	44
Figure 7. Partially populated sample “Model Facts” label for cervical precancer prediction.....	53
Figure 8. Procurement checklist.....	68

## List of tables

Table 1. Randomized trials of AI deep neural networks in endoscopic screening.....	10
Table 2. Framework for evaluation of an AI-SaMD.....	14
Table 3. Clinical evaluation methods used to produce desired evidence.....	15
Table 4. Evaluation components for AI-SaMD in cervical cancer screening.....	16
Table 5. IMDRF Risk Categorisation.....	18
Table 6. IMDRF risk categorisation for AI-SaMD use in cervical cancer screening.....	20
Table 7. SPIRIT-AI checklist items and explanations.....	25
Table 8. SPIRIT-AI items used in evaluation of AI-SaMD for cervical cancer screening.....	27
Table 9. Dataset naming in Clinical and ML Studies.....	29
Table 10. Datasets in training, validating and implementing AI models for healthcare.....	30
Table 11. Training dataset considerations for cervical cancer screening.....	32
Table 12. Evidence generation standards: selected guidance.....	46
Table 13. Applying SPIRIT-AI checklist to cervical cancer screening.....	48
Table 14. Evaluating PMCF of AI-SaMDs for cervical cancer screening.....	65
Table 15. Procurement guidance: evidence requirements.....	68

## 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*,

预览已结束，完整报告链接和二维码如下：

[https://www.yunbaogao.cn/report/index/report?reportId=5\\_23495](https://www.yunbaogao.cn/report/index/report?reportId=5_23495)

