International Telecommunication Union

ITU-T FG-AI4H Deliverable

TELECOMMUNICATION
STANDARDIZATION SECTOR
OF ITU

15 September 2023

PRE-PUBLISHED VERSION

DEL10.24

FG-AI4H Topic Description Document for the Topic Group on AI-based point-of-care diagnostics (TG-POC)

ITU-T

International Telecommunication Union

Summary

This topic description document (TDD) specifies a standardized benchmarking for AI-based point-of-care. It covers scientific, technical, and administrative aspects relevant for setting up this benchmarking.

Keywords

Artificial intelligence; benchmarking; health; topic groups; overview; ethics; regulations; data quality; data audit; clinical relevance; topic description; point-of-care diagnostics, mobile digital microscopy

Change Log

This document contains Version 1 of the Deliverable DEL10.24 on "*FG-AI4H Topic Description Document for the Topic Group on AI-based point-of-care diagnostics (TG-POC)*" approved on 15 September 2023 via the online approval process for the ITU-T Focus Group on AI for Health (FG-AI4H).

Editor: Nina Linder
TG-POC
University of Helsinki, Finland
E-mail: nina.linder@helsinki.fi

Contributors:

Topic Group Driver:
Nina Linder
Institute for Molecular Medicine Finland,
University of Helsinki, Helsinki, Finland
and
Uppsala University, Sweden
Tel: +358445555407
E-mail: nina.linder@helsinki.fi

Johan Lundin
Institute for Molecular Medicine Finland,
University of Helsinki, Helsinki, Finland
and
Department of Global Public Health,
Karolinska Institutet, Stockholm, Sweden
Tel: +358445009685
E-mail: johan.lundin@ki.se

Andreas Mårtensson
Department of Women's and Children's Health, International Maternal and Child Health, Uppsala University, Uppsala, Sweden

Joar von Baar
Department of Global Public Health,
Karolinska Institutet, Stockholm, Sweden

Antti Suutala
Institute for Molecular Medicine Finland,
University of Helsinki, Helsinki, Finland

Harrison Kaingu
Kinondo Kwetu Health Services Clinic, Kinondo, Kenya

Billy Ngasala
Muhimbili University of Health and Allied Sciences (MUHAS), Tanzania

Mikael Lundin
University of Helsinki, Finland

Ngali Mbuuko
Kinondo Kwetu Health Services Clinic, Kinondo, Kenya

Felix Kinyua
Kinondo Kwetu Health Services Clinic, Kinondo, Kenya

Martin Muinde, Kinondo Kwetu Health Services Clinic, Kinondo, Kenya

Jumaa Mbete
Kinondo Kwetu Health Services Clinic, Kinondo, Kenya

Sara Tönqvist
Department of Global Public Health, Karolinska Institutet, Stockholm, Sweden

Edward Lwidiko
Muhimbili University of Health and Allied Sciences (MUHAS), Tanzania

Edward Lwidiko
Muhimbili University of Health and Allied Sciences (MUHAS), Tanzania

Sara Mårtensson
Department of Women's and Children's Health, International Maternal and Child Health, Uppsala University, Uppsala, Sweden

Hakan Kukucel
Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland

CONTENTS

Page

List of Tables

Page

List of Figures

Page

ITU-T FG-AI4H Deliverable 10.24

FG-AI4H Topic Description Document for the Topic Group on AI-based point-of-care diagnostics (TG-POC)

1 Introduction

The topic group has developed and conducted proof-of-concept studies of a novel method that combines artificial intelligence (AI) and mobile digital microscopy for example for cell-based cervical cancer screening in resource-limited settings. The mobile microscopes are wirelessly connected via mobile networks for AI-based analysis and provide access to diagnostics where there is a lack of medical experts. The method's diagnostic accuracy, technical feasibility, cost and time per test, and acceptance of the AI method is evaluated and compare to conventional diagnostics. Throughout the project, opportunities for larger scale implementation of the diagnostic platform are evaluated, with a strong goal of achieving sustainable solutions for low-resource environments. The methods have great potential as support in cell and tissue-based diagnostics. This means a significant step towards a more equal and sustainable access to high-quality diagnostics in resource-poor countries.

2 About the FG-AI4H topic group on AI for Point-of-care (POC)

To develop this benchmarking framework, FG-AI4H decided to create the TG-POC at the meeting at the virtual meeting 20-21 May 2021

The introduction highlights the potential of a standardized benchmarking of AI systems for POC to help solving important health issues and provide decision-makers with the necessary insight to successfully address these challenges.

The topic group has developed and evaluated a combination of mobile scanners and artificial intelligence (AI) algorithms for point-of-care diagnostics. The mobile scanners are wirelessly connected via mobile networks for AI-based analysis and provide access to diagnostics where there is a lack of medical experts. Within a multicenter prospective study, we validate the diagnostic method for cervical cancer screening at small-to middle-sized hospitals in Kenya. The diagnostic accuracy, technical feasibility, cost and time per test, and acceptance of the method is evaluated in comparison with conventional diagnostics.

Within cancer diagnostics, tasks currently performed visually can be automated to improve precision, accuracy, and consistency. These methods will have a huge societal impact globally where AI can help mitigate a critical shortage of experts. The topic group contributes to the knowledge on AI-based methods for cervical cancer screening, and to the generalizability of AI in medical diagnostics. So far, very few image-based AI-algorithms have been validated in external, independent clinical settings to facilitate the implementation of AI-based diagnostics.

Our AI-based method will provide a cost-effective solution, partly by automation of diagnostics and partly by a decreased need for experts at the point-of-care which can be implemented also in high-recourse settings. Also, the time-to-answer will shorten. Case management by providing access to diagnostics in areas with serious delays in time to diagnosis will improve. Medical images will be created, which have educational value and supports research on AI-based diagnostics. On a public health level, the system will aid the assessment of the disease burden and epidemiology, which is essential to guide adequate control policies.

2.1 Documentation

This document is the TDD for the TG-POC It introduces the health topic including the AI task, outlines its relevance and the potential impact that the benchmarking will have on the health system and patient outcome, and provides an overview of the existing AI solutions for POC. It describes

the existing approaches for assessing the quality of POC systems and provides the details that are likely relevant for setting up a new standardized benchmarking. It specifies the actual benchmarking methods for all subtopics at a level of detail that includes technological and operational implementation. There are individual subsections for all versions of the benchmarking. Finally, it summarizes the results of the topic group's benchmarking initiative and benchmarking runs. In addition, the TDD addresses ethical and regulatory aspects.

The TDD will be developed cooperatively by all members of the topic group over time and updated TDD iterations are expected to be presented at each FG-AI4H meeting.

The final version of this TDD will be released as deliverable "DEL 10.24 (TG-POC)." The topic group is expected to submit input documents reflecting updates to the work on this deliverable **(Table 1)** to each FG-AI4H meeting.

Table 1: Topic Group output documents

Number	Title
FGAI4H-O-029-A01	Latest update of the Topic Description Document of the TG-POC
FGAI4H-M-029-A02	Latest update of the Call for Topic Group Participation (CfTGP)
FGAI4H-M-029-A03	The presentation summarizing the latest update of the Topic Description Document of the TG-POC

The working version of this document can be found in the official topic group SharePoint directory.

- https://extranet.itu.int/sites/itu-t/focusgroups/ai4h/tg/SitePages/TG-POC.aspx

2.2 Status of this topic group

The following subsections describe the update of the collaboration within the TG-POC for the official focus group meetings.

2.2.1 Status update for meeting L

During the meeting in May 2021 the TG-POC topic group was established. Since then, the "Call for topic group participation" document was written and submitted to the secretariat. After that more information has been added to the J-105 document.

The goals for this TG have been discussed with Nina Linder and prof. Johan Lundin who had a meeting with Eva Weicken to clarify the matters regarding updating this document and the process during the meetings.

The topic group leader has discussed and identified which researchers would be of importance to collaborate with regarding this AI for POC.

Regarding studies on cervical cancer at the POC, pare planning a validation study in Kenya and Tanzania. The topic group leaders have discussed potential medical clinics who would be willing to participate that would have enough patients so that the study could be rolled out as fast as possible. The annotation of the cervical sample smears would be done by a cytologist from Finland and pathologists from Tanzania and Kenya.

The topic group members have discussed that any topic groups who would like to collaborate with TG-POC for implementing their software on the point-of-care are welcome to discuss/collaborate with TG-POC.

Regarding the validation study of POC for cervical cancer screening discussions regarding additional clinical sites in Kenya has been started

Inst. for Molecular Medicine, University of Helsinki, Finland,

Karolinska Institute, Sweden

Uppsala University, Sweden

New potential members for the TG-POC will be identified and contacted.

Between Meetings of May and Sept 2021(meeting M), the Topic Group POC onboarded one new member, Johan Lundin. Lundin and Linder had a meeting with Eva Weicken on Sept 14, 2021. During the meeting, TG-POC members and Dr Weicken discussed the processes and upcoming meeting FG-AI4H meeting M. The AI4H principles of action and topic groups in general were discussed.

Between Meetings of Sept 2021(meeting M) and February 2022 (meeting N), the Topic Group POC onboarded the following new members: Andreas Mårtensson, Prof, Uppsala University, Ass Prof Billy Ngasala, MUHAS, Tanzania, Prof Andrea Pembe, MUHAS, Tanzania, Deogratias Mzurikwao, IT expert, MUHAS, Tanzania, Prof Jan-Michaél Hirsch, Uppsala University, Sweden, Professor Joakim Lindblad, Professor Bengt Hasséus, University of Gothenburg, Göteborg, Sweden.

TG-POC (Nina Linder and Johan Lundin) had a meeting with Alex Radunsky for TG-Sanitation on Nov 1st, 2021, to discuss possible synergies in our respective research/topics.

On February 10th, 2022, members of TG-POC launched the validation study for an AI -based point-of-care system for detecting pre-cancerous lesions of the cervix at two health care facilities in Kenya. Members of the topic group will evaluate the feasibility of the method for detection of cellular dysplasia by comparing the diagnostic accuracy (sensitivity, specificity, agreement in results) of the novel AI-supported digital method to conventional cytological diagnostics performed by a pathologist as the ground truth in the study. A total minimum number of 720 routinely prepared cervical cytology samples (PAP smears) will be collected from both HIV positive women (target n=227) currently enrolled in the HIV control program and HIV negative women attending the Kinondo Kwetu Hospital and the Diani Health Center (target n=493). If sufficient diagnostic accuracy of the technique is confirmed by the study, a larger controlled study will be indicated to further test the accuracy of the devices. The diagnostic system described has the potential to improve access to screening, and thus treatment and prognosis for cervical cancer on a national level by providing a means to analyse samples at the point-of-care, reducing sample analysis turnover time and facilitating analysis and diagnosis. By utilizing digital image analysis, sample analysis and routine work could potentially be improved even further.

Cost sources related to AI based digital screening for cervical cancer will be calculated from parameters such as: Costs of loss of earnings and/or arranging child/elderly care, Costs of transportation, -Cost of stationary lab equipment, annuitized costs a 10-y lifespan and a 3% interest rate, Cost of stationary lab equipment, annuitized costs a 10-y lifespan and a 3% interest rate, Lab-equipment maintenance and insurance, Consumables costs (reagents), Training of lab technician related to staining, Lab technician time spent staining in minutes, Staining time, Electricity and water supply costs, overhead, Lab-equipment maintenance and insurance, Consumables costs (brushes, glass slides), Training of nurses on counselling and obtaining sample, Nurse time spent counselling patient and obtaining sample, sample preparation time, Electricity and water supply costs, overhead, Training of technician to use the AI, AI cloud platform subscription costs, time spent by technician in uploading, the digital sample, WIFI/Mobile upload costs (airtime), AI processing costs, time spent by pathologist in preparing the report, costs related to sending the report back to the local clinic, (regular mail, WIFI/Mobile download costs, airtime), costs related to potential counselling of local clinic personnel, or consultant doctor related to the results via phone. All SOPs for management of samples, sample retrieval, scanning, AI applications and case report forms are identical with the previous proof-of-concept study.

2.2.2 Status update for meeting R

- Validation of algorithms on a new data set from two health care facilities
- 760 cervical cytology samples collected from HIV positive women attending two hospitals in Kenya.
- TG-POC is currently evaluating the feasibility of the method for detection of cellular dysplasia by comparing the diagnostic accuracy to conventional diagnostics performed by a pathologist.
- Parameters for cost-efficacy collected and to be analysed.
    - Training of technicians
    - AI cloud platform subscription costs, time spent by technician in uploading, the digital sample, WiFi/Mobile upload costs (airtime)
    - AI processing costs, time spent by pathologist in preparing the report, costs related to sending the report back to the local clinic, (regular mail, WiFi/Mobile download costs, airtime)

WHO/ITU Focus Group “AI for Health” Working Group on Clinical Evaluation. During Jan-March members of TG-POC have implemented the Clinical evaluation document together with Eva Weicken.

Members of TG-POC are evaluating the document by using it from the start of a new validation study in Tanzania in May 2023 for cervical screening cancer screening with deep learning.

- Initial contacts with researchers studying breast cancer diagnostics in low resource settings.
- Prepare for publication of cervical screening validation study 07/2023
- Topic Group POC workshop planned for 09/2023

2.2.3 Status update for meeting S

Collaboration project between Fraunhofer and U Helsinki started Q2/2023

- On explainable machine learning "Concept Relevance Propagation" (CRP) on image and clinical data on breast cancer
- Clinical collaborator Prof Heikki Joensuu, UHelsinki

WASP-DDLS funding from the Swedish Wallenberg Foundation
years 2024-2025

- Validation of a new tool for the expert to verify the AI-based cervical cell findings
- Collaboration between Uppsala U and Linköping U, Sweden

Collaboration with the Working Group on Clinical Evaluation

- TG-POC is continuously evaluating and updating he document.
- Using the document from the start of the new validation study in Tanzania for cervical cancer screening with deep learning

New study: Ai for cervical screening for general population in Tanzania

- 1,200 routinely prepared cervical cytology samples from the general female population, regardless of HIV status
- Deep learning for cervical abnormalities
- Human Papilloma Virus (HPV) status

- The study site is St. Benedict's Hospital, a healthcare facility located in Kibamba, Ubungo district of Dar es Salaam, Tanzania.
- A catchment population of >35,000 people
- 1200 patients aimed for recruitment.
- St. Benedict's Hospital has limited access to expert pathologists.

3 Topic description

This section contains a detailed description and background information of the specific health topic for the benchmarking of AI in Point-of-care and how this can help to solve a relevant 'real-world' problem.

Topic Groups summarize related benchmarking AI subjects to reduce redundancy, leverage synergies, and streamline FG-AI4H meetings. However, in some cases different subtopic groups can be established within one topic group to pursue different topic-specific fields of expertise. The TG-POC currently has no subtopics. Future subtopics for different disease groups (e.g. Cervical cancer, helminth diagnostics) might be introduced.

Improved access to lifesaving diagnostics of cancer and infectious diseases is a global health priority. By taking advantage of recent technology innovations this could be an achievable goal. The aim of this project is to assess the feasibility and clinical value of artificial intelligence (AI) combined with mobile digital microscopy scanners (such as MoMic in Figure 1[1]) for improving access to cancer and infectious disease diagnostics (1) (2). The novel methods developed and assessed by our research team can analyze a biological sample with an accuracy comparable to a highly trained expert, but at a fraction of the cost and time. The scanner instruments are connected via mobile networks and samples can be digitized at the point-of-care and rapidly transmitted for remote diagnosis, provided either by AI-supported computer vision or a combination of AI and a human expert. Our novel method enables rapid confirmatory diagnosis of multiple diseases at the point-of-care.

Our proof-of-concept studies show that the novel digital diagnostic method is technically and diagnostically feasible (3-5) and our project has gained international attention in the form of news articles in high impact journals (6, 7). In our most recent study, conducted in rural Kenya, we show that a high diagnostic accuracy can be reached in screening for cervical cancer in a resource-limited setting (3). Also this report gained interested at the time of publication and has been highlighted in healthcare related media (8). A validation study which includes HIV positive patients has been conducted in Kenya during 2022.

[1] The AI at the point-of-care diagnostic system (MoMic) illustrated in Figure 1 includes obtaining a sample, digitizing the sample with a mobile microscope scanner, image transfer over mobile networks, AI-analysis, and verification by remote expert and feed-back of results back to the point-of-care for decision support.

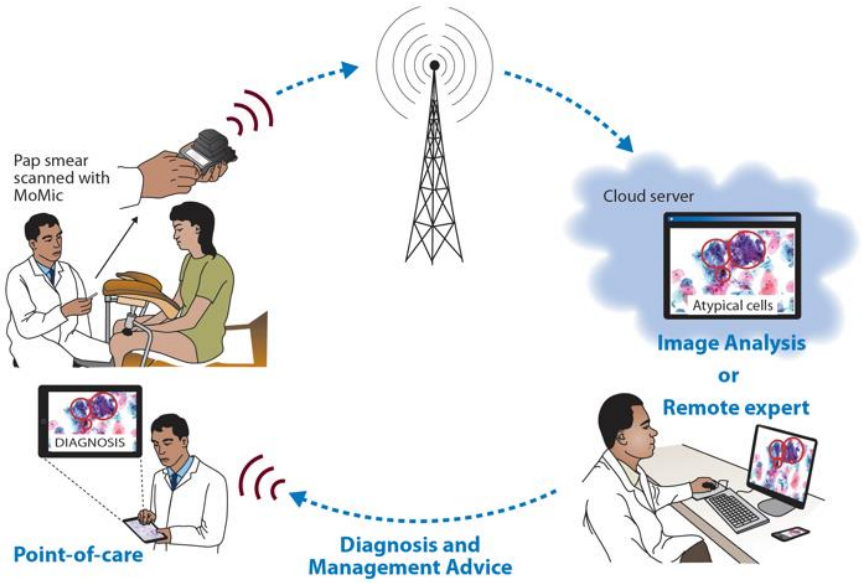

Figure 1 – The AI at the point-of-care diagnostic system (MoMic)

Artificial intelligence for image-based diagnostics

The domain of artificial intelligence (AI) related to image analysis and classification has made huge advances during the last few years. This is much due to the use of convolutional neural networks i.e. deep learning (9). The methods are rapidly being applied to various tasks, such as self-driving cars, industrial robotics, and medical image-based diagnostics (10). Recently, deep learning–based algorithms have been used for many medical image analysis applications, with levels of performance even surpassing human experts in certain tasks. Inspired by excellent results on diabetic retinopathy (11), melanoma (12), tissue biomarkers (13) and our own studies on cancer diagnostics, malaria and parasitic diseases (3-5, 14), TG-POC researchers now aim to implement the AI methods combined with the mobile microscopy system in a clinical validation and implementation study.

Mobile microscopy

Recent development allows decreasing the size of a microscope to construct miniaturized digital microscopy-imaging devices that can produce high-resolution images. Researchers within TG-POC have invented and studied pioneering methods for imaging that increases the resolution to be comparable with a laboratory level optical microscope and allows scanning of an entire microscope slide (3-5). The instruments are wirelessly connected via mobile networks, widely available also in resource-limited areas, for transfer of data from the POC to remote analysis with AI (the MoMIC system) and verification of AI-findings performed by human experts (3).

3.1.1 Definition of the AI task

The innovation and concept are a combination of a mobile digital microscope scanner and an image-based deep learning algorithm to automatically analyse scanned microscopy samples. The system is developed by partners within in our research collaboration project including University of Helsinki, Finland, University of Uppsala, Sweden, Karolinska Institute, Sweden

Researchers within TG-POC have invented a series of methods that allow any microscope slide with a biological sample (cell, tissue, stool, blood, urine) to be digitally scanned at low cost and at the POC. Digital samples can be viewed both locally (using a smartphone, tablet, laptop, or desktop computer) and transferred to a cloud environment for remote viewing, automated analysis, and archiving. Leaders within the Topic Group have in addition developed AI-algorithms for diagnostic purposes that are based on machine learning with artificial neural networks (e.g., deep learning) and

applied these to diagnostics of soil transmitted helminths in stool and urine samples, and to malaria diagnostics, breast cancer diagnostics and cervical cancer screening.

Members of TG-POC chose cervical cancer as our first most extensive target since this medical problem is immense in many countries. But researchers within TG-POC have also performed studies on neglected tropical diseases such as helminth infections. These are not further discussed in this document which focuses on cervical cancer screening at the POC using deep learning algorithms.

Figure 2 illustrates a workflow for analysis of cervical PAP smears with the AI-based digital diagnostic method compared to conventional diagnostics. The AI algorithm automatically detects atypical cells and counts them for the report.

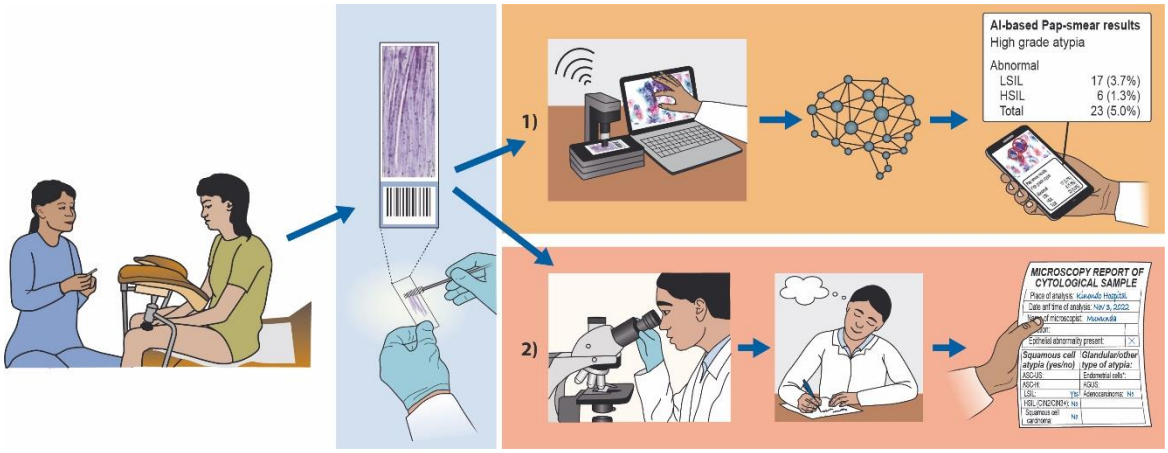

Figure 2 – Analysis of cervical PAP smears with 1) the AI-based digital diagnostic method and 2) conventional diagnostics

The long-term goal is to achieve a digital diagnostic solution that would be widely available and cost less than a smartphone (500-1.000 €) and fit in a carry case bag, orders of magnitude cheaper (typical price 25.000-300.000 €) and smaller (current instruments not suitable to be carried and typical weight is 20-100 kg) than currently available microscope scanners.

To develop a deep learning system for the detection of cervical cell atypia in the digitized Papanicolaou smears, we use a commercially available machine learning and image-analysis platform. Using this platform, an algorithm based on deep convolutional neural networks to detect low-grade squamous intraepithelial lesions (LSILs) and high-grade squamous intraepithelial lesions (HSILs) in the Papanicolaou smear digital whole slides is trained. The samples series is split into a training and tuning set and a validation set. Training is performed by a researcher assisted by a cytotechnologist specialized in cervical cytology screening, using manually defined representative regions of the digitized slides of the training series. Regions (n = 16 133, with cross sections of approximately 25-100 μm) are annotated visually and included areas of both normal cervical cellular morphology and various degrees of atypia. Training of the algorithm uses 30 000 iterations with a predetermined feature size of 30 μm, a weight decay parameter of 0.0001, 20 minibatches, a learning rate of 0.1, and 1000 iterations without progress as the early-stop limit. Training is augmented by using image perturbations. Access to the trained model is possible remotely to analyse samples directly at the POC.

Detection of low- and high-grade squamous intraepithelial lesions in Papanicolaou test whole-slide images. Digitized slides measured approximately 100 000 × 50 000 pixels, corresponding to roughly a standard microscope glass slide (25 mm × 50 mm); i.e., the entire Papanicolaou smear is scanned.

Results of the artificial intelligence (AI) will be compared with conventional visual microscopic assessment of the cervical smears performed by a pathologist (Fig. 2). The conventional microscopy assessment will form the ground truth for diagnosis and the AI results will not influence the

decision-making without expert verification. If the analysed samples, interpreted by the expert shows precancer or cervical cancer, a healthcare provider will be in contact with the patient for referral for appropriate treatment according to local and national guidelines. The trial will be registered at http://www.controlled-trials.com/ and reporting will be done according to the STARD statement (http://www.stard-statement.org/).

3.1.2 Current gold standard

Cervical cancer screening

Access to pathology services is limited in most low-resource countries. For example, in a recent survey (http://www.pathologyinafrica.org/data) in 35 African countries showed that there is in average less than one pathologist per one million people.

Therefore, access to cancer screening and diagnostics is poor and most cancers are detected at a late symptomatic stage. Cervical cancer is the most common cause of cancer death among women in the African countries and 60,000 die each year from a disease that is treatable if detected at an early stage. Improved cervical cancer diagnostics will benefit women both in resource-limited (access to diagnosis) and high-resource countries (faster and more precise diagnosis).

Cervical cancer is a major cause of cancer-related mortality and morbidity globally, and the burden of disease is disproportionally distributed among low-income and middle-income countries (LMICs) and high-income countries (HICs). LMICs account for 80% of cervical cancer cases worldwide and the global age-standardized incidence rate for cervical cancer is 14 per 100 000 women (15), while the incidence rate of cervical cancer is 42.7 per 100 000 women in East Africa (16). During the next decade, the disease incidence is expected to increase, and the yearly mortality is expected to double, with the largest burden of disease occurring in sub-Saharan Africa (17). Major contributing factors include low knowledge of risk factors and how to prevent cervical cancer and lack of organized screening programs. Cervical cancer has in addition to its impact on health, also substantial negative impact on poverty, education, and gender equity – each a separate goal among the Sustainable Development Goals (SDGs) of the United Nations. Many of the women who die of cervical cancer are breadwinners and caretakers of both children and elders.

This section provides a description of the established gold standard of the addressed health topic.

Screening with VIA

The standard screening test in resource-limited settings is visual inspection with acetic acid (VIA) as this can be performed by mid-level health care providers and allows for immediate treatment. However, the result of VIA is subjective resulting in low accuracy, and the utility is questionable in resource-limited settings when the number of screening rounds per women's lifetime is low (18).

Cytology-based screening

Cell-based screening of cervical smears obtained and prepared according to the cytological Papanicolaou method (Pap smears) and analysed visually with conventional microscopy can drastically reduce the incidence and mortality of cervical cancer, but is labour-intensive (19), prone to variations in sensitivity and reproducibility and requires skilled medical experts which makes the process difficult to implement especially in resource-limited settings (20). Pathologist or (usually) cytologist screen samples under a microscope. Conventional cytology screening (Papanicolaou test analysis) can drastically reduce the incidence and mortality of cervical cancer, but the manual analysis of samples is labour intensive, is prone to variations in sensitivity and reproducibility, and requires medical experts to analyse the samples. This makes the process difficult to implement in resource-limited settings.

HPV testing and vaccinations

Human papillomavirus (HPV) infections, which are the causative agent for cervical cancer, can be detected using polymerase chain reaction (PCR) assays with high sensitivity and reproducibility.

However, because most HPV infections are transient, the specificity for precancerous lesions is low (21). In high-resource areas, both molecular and cytology-based screening methods are commonly used and are often combined (i.e., co-testing) to improve the diagnostic accuracy (22). In low-resource settings HPV screening is still rarely available, due to costs and requirement of highly trained personnel.

Ultimately, vaccinations against human papillomavirus (HPV) have the potential to significantly reduce the disease incidence but given that the full benefits of even the most efficient vaccination programs will take decades to be fully realized, millions of women remain at risk (23). Therefore, cytology-based screening tests remain essential, and innovative POC diagnostic solutions like the one presented in the current proposal are needed (24).

3.1.3 Relevance and impact of an AI solution

This section addresses the relevance and impact of the AI solution (e.g., on the health system or the patient outcome) and describes how solving the task with AI improves a health issue.

Overall, our AI based diagnostic solution can: scale productivity, increase diagnostic accuracy, reduce costs, enhance staff satisfaction, improve patient outcomes.

Deep learning AI for image-based diagnostics

The domain of artificial intelligence (AI) related to image recognition and classification has during the last five years made huge advances. This is much due to the use of convolutional neural networks or so-called deep learning (9). These methods have rapidly been applied to tasks where interpretation of visual scenes is crucial.

Deep learning has recently been successfully applied to medical image-based diagnostics and inspired by excellent results in our recent studies, researchers within TG-POC now aim to validate the method based on deep learning AI and mobile digital microscopy systems that our group has developed. To our knowledge, our proof-of-concept study is the first one where digital whole slide images of entire Pap smears were captured and analyzed with AI at the point-of-care in a rural, resource-limited environment (3). The proposed external validation in a prospective study and in multiple laboratories represent a further scientific novelty and addresses the request for assessment of AI-performance in an independent setting outside the institution where data for the original algorithm was collected the AI trained (Kinondo Hospital, Kenya). According to the literature only less than 10% of image-based AI-algorithms have so far been validated in external, independent settings and this has been highlighted as something that needs to be addressed to facilitate the implementation of AI-based diagnostics (25).

Mobile microscope scanners used for the studies represent novel technology and are constructed of cameras, optical elements and microelectronics typically used in smartphone systems. The instruments include software that improves the quality of captured images and performs image compression for rapid and cost-effective data transfer and AI analysis. A commercially available scanner developed in Finland (Ocus, Grundium) will be used for sample digitization (3).

Clinical relevance

The project contributes to improved equitable access to cell-based screening for cervical cancer in a in a country that currently has one of the highest incidences of the disease globally. The digital methods will enable accurate, efficient, and accessible diagnostics, partly by automation and partly by a decreased need for expertise at the point-of-care, and thereby contribute to the Sustainable Development Goals (SDGs) of the United Nations to ensure health and well-being for all. Improved prevention also indirectly has significance for the SDGs related to poverty, gender equality and quality education through prevention of unnecessary morbidity and mortality in relatively young women.

The patients. The methods will improve case management by providing access to diagnostics in areas where access to high quality microscopy services previously has been limited or resulted in delays in time to diagnosis.

The health care system and professionals. The methods enable remote consultation and thus directly addresses the need for task shifting and a more efficient use of available experts.

The scientific community. Within the project high-quality digitized medical images will be created, which will support research on medical diagnostics and applied AI.

The medical and technical educational system. The project will give teachers and students access to state-of-the art instruments for sample digitization, enable the establishment of digital sample archives for educational purposes, as well as build capacity for development of AI algorithms for a wide variety of diagnostic purposes.

Impact

The digital methods will enable more accurate and personalized, efficient, and accessible diagnostics, partly by automation and partly by a decreased need for expertise at the POC. Only less than 10% of image-based AI-algorithms have so far been validated in external, independent settings and this has been highlighted as something that needs to be addressed to facilitate the implementation of AI-based diagnostics. The methods will improve patient management by providing access to diagnostics in areas were with previously limited or delays in time to diagnosis. The project aims are not only to conduct research on novel digital diagnostics, but also to enable capacity building and to create sustainable solutions for mobile and connected health in general.

The project results will be highly relevant for industry, not the least for Small and Medium Enterprises. Commercialization of the diagnostic system will be considered to support implementation and achieve sustainability. The technology transfer and capacity building proposed will inspire entrepreneurs and nurture local spin-off companies both in the Nordics and in East Africa. The targeted computer science advances are broadly applicable for AI use in clinical settings and tackle a key challenge for adoption of AI solutions outside of academia, thus having tremendous potential for industrial and societal impact.

3.1.4 Existing AI solutions

The currently known AI system for cervical cell diagnostics developed by our research group include the following inputs, outputs, key features, target user groups: input digital image of a cervical cell smear, classification of cells into cell s according to the Bethesda classification (see below). Target user groups may be any laboratory that does cervical diagnostics from small field-based laboratories to high end pathology and diagnostic laboratories.

The common features found in AI solutions that might be benchmarked in this field are the ones in the Bethesda classification, i.e.:

1. Atypical squamous cells of undetermined significance (ASC-US)
2. Atypical squamous cells for which a high-grade lesion cannot be excluded (ASC-H)
3. Low-grade squamous intraepithelial lesion (LSIL) encompassing HPV infection or mild dysplasia (CIN 1)
4. High-grade squamous intraepithelial lesion (HSIL) encompassing moderate (CIN 2) and severe dysplasia (CIN 3/CIS), noting whether the lesion has features suggesting invasion
5. Squamous cell carcinoma

The following are several relevant metadata dimensions characterizing the AI systems in this field:

- System Name: The AI system's commercial, project, or codename.
- AI Type: The type of AI model or system, such as Supervised Learning, Unsupervised Learning, Reinforcement Learning, or a hybrid approach.
- Architecture: Specific architectural choice, like a Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), Transformer, or GPT-based model.
- Training Data: The datasets used to train the AI, including the size, source, content, and any preprocessing techniques used.
- Evaluation Metrics: Metrics used to assess the system's performance, such as accuracy, precision, recall, F1 score, Area Under Curve (AUC), etc., depending on the task.
- Performance: Results on benchmark datasets or in specific application contexts, including any limitations or edge cases.
- Parameters: The number of parameters in the model, which often relates to the model's complexity and capacity.
- Training Methodology: The approach to training the model, such as transfer learning, fine-tuning, or from scratch.
- Hardware Requirements: The computational resources needed to train and run the AI, including memory, processing power, and energy consumption.
- Software Dependencies: The software libraries or frameworks required to run the AI, such as TensorFlow, PyTorch, or custom libraries.
- Release Date: When the AI was released or updated.
- Operational Environment: The context in which the AI operates, such as a cloud server, local machine, mobile device, or embedded system.
- Usage Constraints: Any limitations or constraints in the use of the AI, including terms of service, licensing, and privacy considerations.
- Bias and Fairness: Information about potential biases in the AI's decisions, impact on different demographic groups, and any strategies employed to mitigate these biases.
- Interpretability and Explainability: The degree to which the AI's decision-making process can be understood or explained.
- Security and Privacy: How the AI handles user data, the steps taken to protect data privacy and ensure the security of the system.
- Sustainability: The environmental impact of training and running the AI system, such as its carbon footprint.

Although significant advances have been made in digital microscopy diagnostics at the point of care (POC), their clinical implementation has been slow.

A previously developed deep learning based algorithm developed by our research group will be used for detection of premalignant lesions in the digitized Pap smears. The algorithm has been trained on 16,133 manually annotated regions from 350 samples, including areas of both normal cervical cellular morphology and various degrees of atypia as previously described. The AI-algorithm was validated on 390 samples. Access to the trained model is possible remotely to analyse samples directly at the POC on a cloud-based platform via upload over mobile or landline networks. The AI results are reviewed and verified by a pathologist.

Studies on deep learning algorithms for analysis of cervical cytology smears have mainly analysed only small areas of samples with instruments not suitable for POC usage. To our knowledge, no

research has been conducted on the analysis of digital whole-slide images of entire Papanicolaou tests, captured in more challenging real-world clinical environments. Thus, this technology has not yet been applied in basic laboratories that are able to perform simple staining procedures but lack access to molecular testing, where the need for improved diagnostics is highest.

Previous studies have reported results with the deep learning–based analysis of smaller cropped images from Papanicolaou smears (Bora et al, William et al, Zhang et al) were digitized with conventional slide scanners, but clinical application requires the examination of substantially larger sample areas.

– Bora K, Chowdhury M, Mahanta LB, Kundu MK, Das AK. Automated classification of Pap smear images to detect cervical dysplasia. Comput Methods Programs Biomed. 2017; 138:31-47. doi:10.1016/j.cmpb.2016.10.001

– William W, Ware A, Basaza-Ejiri AH, Obungoloch J. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. Comput Methods Programs Biomed. 2018; 164:15-22. doi:10.1016/j.cmpb.2018.05.034

– Zhang L, Le Lu, Nogues I, Summers RM, Liu S, Yao J. DeepPap: deep convolutional networks for cervical cell classification. IEEE J Biomed Health Inform. 2017;21(6):1633-1643. doi:10.1109/JBHI.2017.2705583

Data sharing

The project fully embraces an open science approach. All the data collected in the project will be shared according to FAIR principles. TG-POC will employ the sharing services that has been ranked by EOSC-Nordic to be among the top 12% in FAIR readiness among tested repositories in the Nordics. The services of AIDA Data Hub are well established and mature in technical, legal, and ethical aspects of data sharing. Members of TG-POC are committed to *Open Access* publication and prioritize gold open access journals whenever possible.

4 Ethical considerations

The studies are conducted in accordance with the Declaration of Helsinki and International Conference on Harmonization Good Clinical Practice (IC H-GC P) guidelines. Microscopy samples are digitized (‘scanned’) using digital microscope scanners deployed at each of the research sites. The digital samples are pseudonymized, meaning that the digital sample will contain only the study number and no clinical information, study subject information or other personal identifiers. The digital image of the slide (i.e., ‘digital sample’) will be stored on local hard drives at the research sites, in locked rooms accessible only to study personnel. Digital images (without personal identifiers) will be uploaded to a cloud-server, based in Helsinki (Primed 6, Meilahti Campus Library Terkko, University of Helsinki, Helsinki, Finland), from where they can be accessed remotely. Servers are stored in locked rooms, not accessible to the public. Remote access to the server for sample viewing is established with secured SSL encryption and a password-protected web-based interface. Identification of individuals based on digital samples is not possible.

Informed consent will be sought from all participants in their native language. Sample and corresponding clinical data will be pseudonymized, data will be stored on computers with electronically controlled access or on secured local servers. Backups are acquired of all data and stored in physically separate spaces from the main server, but secured, locked and only accessible by authorized personnel.

In cases of abnormal cervical tests, treatment expenses are covered by study funding, and treatment was arranged by a gynaecologist in accordance with national guidelines.

Serious health issues when precancerous or cancer is missed or found too late. Cervical cancer remains a common and deadly cancer in areas without screening programs. During the next decade,

the disease incidence is expected to increase, and the yearly mortality is expected to double, with the largest burden of disease occurring in sub-Saharan Africa.

5 Existing work on benchmarking

This section focuses on the existing benchmarking processes in the context of AI and TG-POC for quality assessment. It addresses different aspects of the existing work on benchmarking of AI systems (e.g., relevant scientific publications, benchmarking frameworks, scores and metrics, and clinical evaluation attempts). The goal is to collect all relevant learnings from previous benchmarking that could help to implement the benchmarking process in this topic group. The work done by our research group are the first bench marking studies until there is a consensus. The results could be further developed towards a more common consensus classification.

5.1.1 Publications on benchmarking systems

For cervical cancer screening:

1. Holmström O, Linder N, Kaingu H, Mbuuko N, Mbete J, Kinyua F, Törnquist S, Muinde M, Krogerus L, Lundin M, Diwan V, Lundin J. Point-of-Care Digital Cytology With Artificial Intelligence for Cervical Cancer Screening in a Resource-Limited Setting. JAMA network open. 2021;4(3):e211740-e.

The data management plan will include descriptions on the types of data generated, what standards will be used within the project, what database (at this point TG-POC intend to use RedCap, Vanderbilt University, Nashville, TN) and how data and knowledge will be curated, stored, managed, shared, and preserved. The plan will also include information data ownership, how data will be shared and made available for verification and re-use. Aspects related to how study participants will be protected and how costs related to data curation and preservation will be covered. The Data Protection Directive (95/46/EC), regarding protecting and handling personal data will be followed.

Data will be collected through common data standards such as .csv, .tiff, .jpg and .jp2. The scanned whole-slide images will also be converted to the novel standard developed by the working group 26 of the Digital Imaging and Communications in Medicine (DICOM) organization (https://www.dicomstandard.org/wgs/wg-26) For terminology related to laboratory procedures and findings our researchers will use SNOMED CT (http://www.snomed.org/) and for diagnoses ICD coding.

Inclusion criteria for the latest prospective study (unpublished):

– Female sex, age between 18 and 64

– Women accessing health services at Kinondo Kwetu Hospital (n=500), Diani Health Center (n=500) and Oasis Health Center (n=500).

– Written informed consent from the patient

Exclusion criteria:

– Symptoms and signs of acute, severe disease, requiring immediate referral

– Patient is currently pregnant

– Patient refuses consent

– Analysis of samples fails due to unsatisfactory smear sample

Study endpoints:

– Diagnostic accuracy of the digital diagnostic method for detection of low grade (LSIL) and high grade cellular dysplasia (HSIL) by remote analysis of the digital, cervical samples, as compared to conventional pathology light microscopy of the physical samples

Secondary outcome measures:

– Assessing accuracy for detection of other types of cellular atypia (e.g. ASC-US, ASC-H)

– Detection of infectious pathogens in the sample

– Determination of sample evaluation time (turnover time), evaluation of user experience ("ease-of-use")

Outcome measures: Diagnostic accuracy (sensitivity, specificity, positive and negative predictive value) for pre-cancerous lesions. Cost-efficiency (e.g., time spent for analysis, training, patient waiting time) is calculated for all survey and diagnostic activities, cost per patient tested, and cost per case detected.

The metrics used are as following:

Diagnostic accuracy (sensitivity, specificity, positive and negative predictive value) for pre-cancerous lesions. Cost-efficiency will be calculated for all survey and diagnostic activities, and cost per patient tested, cost per case detected.

Cost parameters studied for the POC system:

Women attending the health facility:

– Costs of loss of earnings and/or arranging child/elderly care,

– Costs of transportation

Cervical smear obtained:

– Cost of stationary lab equipment

– Lab-equipment maintenance and insurance

– Consumables costs (reagents)

– Training of lab technician related to staining

– Lab technician time spent staining in minutes

– Staining time

– Electricity and water supply costs, overhead

Cervical smear stained with Papanicolaou stain in the local lab:

– Cost of stationary lab equipment, annuitized costs a 10-y lifespan and a 3% interest rate

– Lab-equipment maintenance and insurance

– Consumables costs (reagents)

– Training of lab technician related to staining

– Lab technician time spent staining in minutes

– Staining time

– Electricity and water supply costs, overhead

Cervical smear stained with Papanicolaou stain digitally scanned with a microscope scanner at the local hospital:

– Cost of scanner and computer equipment, annuitized costs a 10-y life-span and a 3% interest rate
– Scanner maintenance and insurance
– Training of technician related to scanning
– Technician time spent scanning
– Scanning time
– Electricity costs
– WiFi/mobile network subscription costs
– Local data storage costs

The digitally scanned cervical smear stained with Papanicolaou stain uploaded to the cloud for analysis with artificial intelligence (AI):

– Training of technician to use the AI
– AI cloud platform subscription costs
– Time spent by technician in uploading the digital sample
– WiFi/Mobile upload costs (airtime)
– AI processing costs

The results of the AI applied to the digitally scanned cervical smear stained with papanicolau stain sent to the pathologist for verification.

– Cost of computer equipment costs (computer, display, mobile/wifi router)
– Training of pathologist to use the AI
– Time spent by pathologist in verifying the AI results
– WiFi/Mobile download costs (airtime)

Report on the cervical smear stained with Papanicolaou stain created by a pathologist at the central pathology lab:

– Time spent by pathologist in preparing the report
– Costs related to sending the report back to the local clinic (regular mail, WiFi/Mobile download costs, airtime)
– Costs related to potential counselling of local clinic personnel or consultant doctor related to the results via phone

Patient recruitment

Individuals presenting to the health facility wishing to take part in the study will initially be screened for eligibility by a nurse, and informed about the study, including how it will be carried

out and the potential risk and benefits from taking part in the study. Clients meeting the inclusion criteria and willing to participate will be scheduled an appointment date upon which they will return to the health facility (Fig. 3). Study participants will be reimbursed for the public transportation cost for their scheduled appointment at the facility. Clients presenting for their scheduled appointments will first be counselled in Swahili on the procedure and its benefits, the study design and follow up.

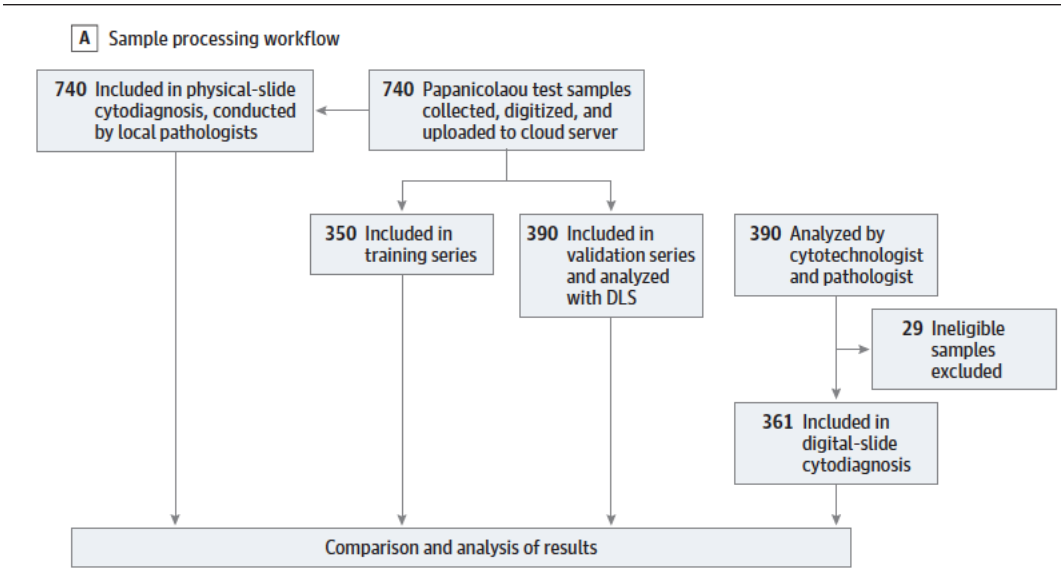

Figure 3 – Sample processing workflow for the benchmarking study (Holmström et al)

After giving written informed consent the client will be taken to the examination room and a cervical sample will be obtained using a brush, and a Pap smear prepared and fixed on a glass slide, followed by staining according to the Papanicolaou method, checked for initial quality using a conventional microscope, scanned with the microscopy scanner, uploaded to the cloud platform (Aiforia Hub, Aiforia Technologies, Helsinki, Finland) and analysed using AI-based algorithm trained to specifically find cervical cell abnormalities.

The Pap smears will also be examined by a pathologist at Muhimbili National Hospital and the pathologist's diagnosis is the gold standard, and the basis for clinical decision making. The clients will be informed of the results, via phone call or a scheduled appointment. In case of abnormal findings, the patient will be referred to Ocean Road Cancer Institute for further management according to national guidelines. However, the study organisers will not financially cover the costs of further treatment but will see to that the patient is referred accordingly. Clients who test positive for high risk HVP strains with normal cervical smear will be referred to Ocean Road Cancer Institute's cancer screening programme within a year for follow up screening. If a client presents with any other medical concern or if found to have a medical condition when taking part in the study be it during consultation, follow up or examination the patient will be referred for further management as per the national guidelines for the presenting complaint, symptoms or findings and the treatment will not be financially covered by the study organisers.

Sample series

The samples series was split with a 50–50 distribution of the target number of samples into the training series (n = 350), used for training and tuning of the model, and external validation series (n = 390). Individual digitized slides measured approximately 100,000 × 50,000 pixels. Training was performed by a researcher, assisted by a cytotechnologist specialized in cervical-cytology screening, using manually defined representative regions of the digitized slides of the training series. Regions (n = 16,133, with cross-sections of ~25–100 µm) were selected visually and included areas of both normal cervical cellular morphology and various degrees of atypia; visible atypia (low-grade and high-grade) was manually annotated.

Performance of the deep learning algorithms

Performance of the AI system compared the current gold standard (see Figures 4 and 5): The deep learning system achieved high sensitivities (95·7%; 95% CI 85·5—99·5%, and 100%; 95% CI 82·4—100·0%) and AUCs (0·94–0·96) for detection of cervical-cellular atypia. Specificity was higher for high-grade atypia (98·5%; 95% CI 96·5—99·5%, and 93·3%; 95% CI 90·1—95·6%), than for low-grade atypia (86·0%; 95% CI 81·8—89·5%, and 82·4%; 95% CI 78·0—86·3%). Negative predictive values were high (99·3–100%), and no samples classified as high grade by manual sample analysis had false-negative assessment by the deep learning system.

In Figure 4, low grade (green) and high-grade (red) dysplastic cell indicative of premalignant changes in a cervical smear detected and results visualized using colour overlays.

In Figure 5, areas under the receiver operating characteristic curves (AUCs) for the detection of general atypia, high-grade atypia, and low-grade atypia with the deep learning system compared with manual assessment of digital slides by a cytotechnologist and a pathologist (A) and physical slides by a local pathologist (B).

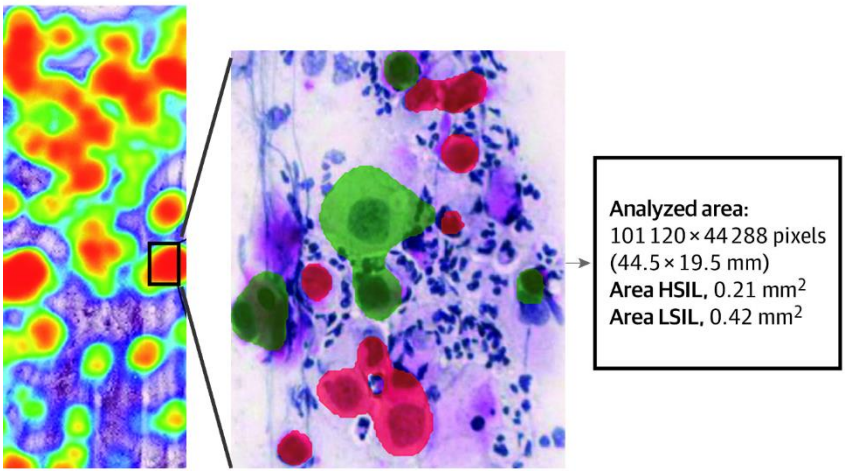

Figure 4 – A cervical Pap smear scanned and analysed with the AI-based method

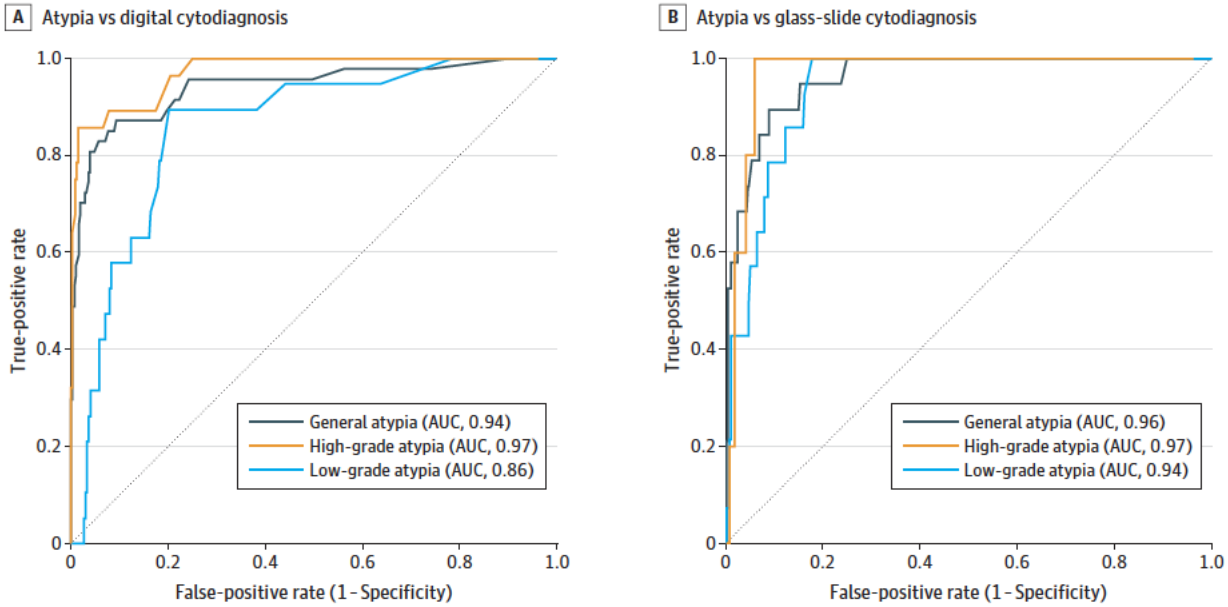

Figure 5 – Detection of general, high-grade and low-grade atypia with a) deep learning versus b) manual assessment of digital slides

Researchers within TG-POC have ongoing and clinical validation studies in Kenya and Tanzania regarding AI solution at the POC for cervical cancer screening and soil transmitted helminths. During 2023-2026 the topic group will further validate the AI systems in prospective clinical studies in East Africa in collaboration with Universities in Sweden and Finland.

Furthermore, population related shifts may include rare subtypes of cellular atypia which are often not sufficiently represented in the data to perform a reliable AI classification. Our assumption is that the uncertainty predictions will allow the human expert to focus on the challenging cases, whereas the AI predictions with high confidence will need minimal human intervention and can be verified by the human experts in seconds (Figure 6). This will be further studied by our research group in collaboration with Joakim Lundström (Univ of Linköping, Sweden), where AI results are presented in a verification panel and uncertainty estimates will be integrated to allow for effective human expert review.

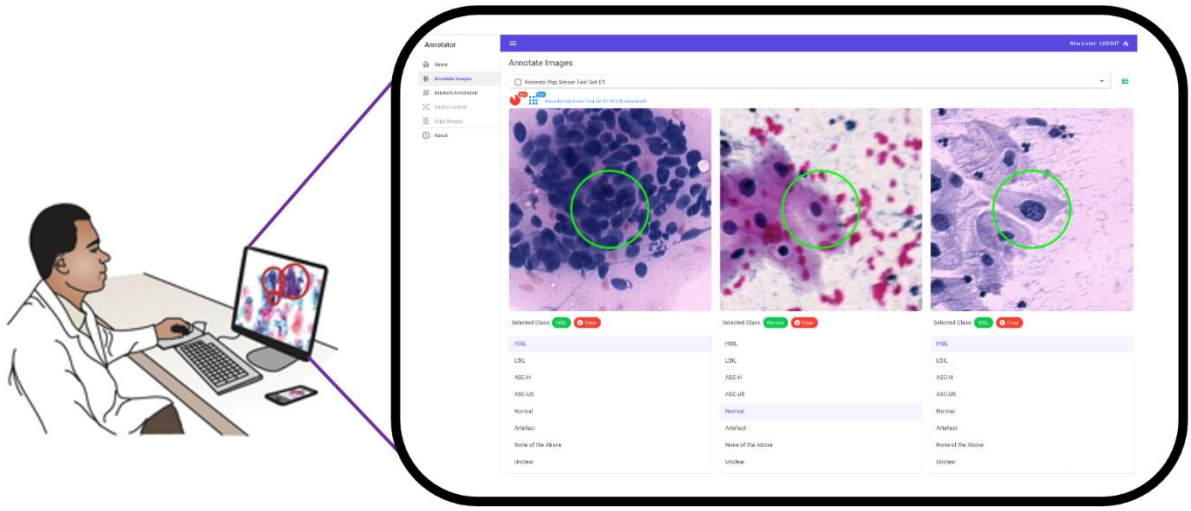

Figure 6 – Tentative interface for the remote expert

Benchmarking by AI developers

All developers of AI solutions for TG-POC implemented internal benchmarking systems for assessing the performance. This section will outline the insights and learnings from this work of relevance for benchmarking in this topic group. TG-POC will accomplish the AI STARD initiative is to improve the completeness and transparency of reporting diagnostic accuracy, to allow to assess the potential for bias in the studies (internal validity) and to evaluate generalisability (external validity) regarding the cervical cancer and other target studies using AI at the POC.

A classification consensus is lacking as well as data sets for annotations for training the algorithms, the goal is to work towards a consensus on classification of cervical atypia's using AI. The scores and metrics used are Sensitivity, specificity, positive predictive value, negative predictive value, AUC, F1 score.

The test data set was acquired by AI developers within the TG-POC from the above studies.

5.1.2 Relevant existing benchmarking frameworks

Out topic group will arrange large-scale classification competitions between existing platforms, also those endorsed by FG-AI4H regarding AI for cervical cancer detection. The variability in interobserver classification on a cellular level is of particular importance.

6 Benchmarking by the topic group

This section describes all technical and operational details regarding the benchmarking process for the point-of-care diagnostics AI task including subsections for each version of the benchmarking that is iteratively improved over time.

Within DEL107: *Clinical Evaluation of AI for health"* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health) topic group TG-POC has participated when it comes to implementation of the document to our clinical study regarding cervical cancer screening using AI at the POC in Tanzania.

It reflects the considerations of various deliverables: DEL5 *"Data specification"* (introduction to deliverables 5.1-5.6), DEL5.1*"Data requirements"* (which lists acceptance criteria for data submitted to FG-AI4H and states the governing principles and rules), DEL5.2 *"Data acquisition"*, DEL5.3 *"Data annotation specification"*, DEL5.4 *"Training and test data specification"* (which provides a systematic way of preparing technical requirement specifications for datasets used in training and testing of AI models), DEL5.5 *"Data handling"* (which outlines how data will be handled once they are accepted), DEL5.6 *"Data sharing practices"* (which provides an overview of the existing best practices for sharing health-related data based on distributed and federated environments, including the requirement to enable secure data sharing and addressing issues of data governance), DEL06 *"AI training best practices specification"* (which reviews best practices for proper AI model training and guidelines for model reporting), DEL7*"AI for health evaluation considerations"* (which discusses the validation and evaluation of AI for health models, and considers requirements for a benchmarking platform), DEL7.1 *"AI4H evaluation process description"* (which provides an overview of the state of the art of AI evaluation principles and methods and serves as an initiator for the evaluation process of AI for health), DEL7.2 *"AI technical test specification"* (which specifies how an AI can and should be tested *in silico*), DEL7.3 *"Data and artificial intelligence assessment methods (DAISAM)"* (which provides the reference collection of WG-DAISAM on assessment methods of data and AI quality evaluation), DEL7.4*"Clinical Evaluation of AI for health"* (which outlines the current best practices and outstanding issues related to clinical evaluation of AI models for health), DEL7.5 *"FG-AI4H assessment platform"* (which explores assessment platform options that can be used to evaluate AI for health for the different topic groups), DEL9 *"AI for health applications and platforms"* (which introduces specific considerations of the benchmarking of mobile- and cloud-based AI applications in health), DEL9.1 *"Mobile based AI applications,"* and DEL9.2 *"Cloud-based AI applications"* (which describe specific requirements for the development, testing and benchmarking of mobile- and cloud-based AI applications).

6.1 Subtopic [A]

For further study

The benchmarking of point-of-care diagnostics is going to be developed and improved continuously to reflect new features of AI systems or changed requirements for benchmarking. This section outlines all benchmarking versions that have been implemented thus far and the rationale behind them. It serves as an introduction to the subsequent sections, where the actual benchmarking methodology for each version will be described.

6.1.1 Benchmarking version [Y]

This section includes all technological and operational details of the benchmarking process for the benchmarking version [Y] (latest version, chronologically reversed order).

6.1.1.1 Overview

This section provides an overview of the key aspects of this benchmarking iteration, version [Y].

6.1.1.2 Benchmarking methods

This section provides details about the methods of the benchmarking version [Y]. It contains detailed information about the benchmarking system architecture, the dataflow and the software for the benchmarking process (e.g., test scenarios, data sources, and legalities).

6.1.1.2.1 Benchmarking system architecture

This section covers the architecture of the benchmarking system. For well-known systems, an overview and reference to the manufacturer of the platform is sufficient. If the platform was developed by the topic group, a more detailed description of the system architecture is required.

6.1.1.2.2 Benchmarking system dataflow

This section describes the dataflow throughout the benchmarking architecture.

6.1.1.2.3 Safe and secure system operation and hosting

This section addresses security considerations about the storage and hosting of data (benchmarking results and reports) and safety precautions for data manipulation, data leakage, or data loss.

In the case of a manufactured data source (vs. self-generated data), it is possible to refer to the manufacturer's prescriptions.

6.1.1.2.4 Benchmarking process

This section describes how the benchmarking looks from the registration of participants, through the execution and resolution of conflicts, to the final publication of the results.

6.1.1.3 AI input data structure for the benchmarking

This section describes the input data provided to the AI solutions as part of the benchmarking of point-of-care diagnostics. It covers the details of the data format and coding at the level of detail needed to submit an AI for benchmarking. This is the only TDD section addressing this topic. Therefore, the description needs to be complete and precise. This section does *not* contain the encoding of the labels for the expected outcomes. It is only about the data the AI system will see as part of the benchmarking.

6.1.1.4 AI output data structure

Similar to the input data structure for the benchmarking, this section describes the output data the AI systems are expected to generate in response to the input data. It covers the details of the data format, coding, and error handling at the level of detail needed for an AI to participate in the benchmarking.

- What are the general data output types returned by the AI and what is the nature of the output (e.g., classification, detection, segmentation, or prediction)?
    - How exactly are they encoded? Discuss points like:
        - The exact data format with all fields and metadata (including examples or links to examples)
        - Ontologies and terminologies
- What types of errors should the AI generate if something is defective?

6.1.1.5 Test data label/annotation structure

While the AI systems can only receive the input data described in the previous sections, the benchmarking system needs to know the expected correct answer (sometimes called 'labels') for

each element of the input data so that it can compare the expected AI output with the actual one. Since this is only needed for benchmarking, it is encoded separately. The details are described in the following section.

Scores and metrics are at the core of the benchmarking. This section describes the scores and metrics used to measure the performance, robustness, and general characteristics of the submitted AI systems.

6.1.1.6 Test dataset acquisition

Test dataset acquisition includes a detailed description of the test dataset for the AI model and, in particular, its benchmarking procedure including quality control of the dataset, control mechanisms, data sources, and storage.

6.1.1.7 Data sharing policies

This section provides details about legalities in the context of benchmarking. Each dataset that is shared should be protected by special agreements or contracts that cover, for instance, the data sharing period, patient consent, and update procedure (see also DEL5.5 on *data handling* and DEL5.6 on *data sharing practices*).

6.1.1.8 Baseline acquisition

The main purpose of benchmarking is to provide stakeholders with the numbers they need to decide whether AI models provide a viable solution for a given health problem in a designated context. To achieve this, the performance of the AI models needs to be compared with available options achieving the same clinically meaningful endpoint. This, in turn, requires data on the performance of the alternatives, ideally using the same benchmarking data. As the current alternatives typically involve doctors, it might make sense to combine the test data acquisition and labelling with additional tasks that allow the performance of the different types of health workers to be assessed.

6.1.1.9 Reporting methodology

This section discusses how the results of the benchmarking runs will be shared with the participants, stakeholders, and general public.

6.1.1.10 Result

This section gives an overview of the results from runs of this benchmarking version of your topic. Even if your topic group prefers an interactive drill-down rather than a leader board, pick some context of common interest to give some examples.

6.1.1.11 Discussion of the benchmarking

This section discusses insights of this benchmarking iterations and provides details about the ‘outcome’ of the benchmarking process (e.g., giving an overview of the benchmark results and process).

6.1.1.12 Retirement

This section addresses what happens to the AI system and data after the benchmarking activity is completed. It might be desirable to keep the database for traceability and future use. Alternatively, there may be security or privacy reasons for deleting the data. Further details can be found in the reference document of this section DEL4 “*AI software lifecycle specification”* (identification of standards and best practices that are relevant for the AI for health software life cycle).

7 Overall discussion of the benchmarking

For further study.

8 Regulatory considerations

For further study

For AI-based technologies in healthcare, regulation is not only crucial to ensure the safety of patients and users, but also to accomplish market acceptance of these devices. This is challenging because there is a lack of universally accepted regulatory policies and guidelines for AI-based medical devices. To ensure that the benchmarking procedures and validation principles of FG-AI4H are secure and relevant for regulators and other stakeholders, the working group on *"Regulatory considerations on AI for health" (WG-RC)* compiled the requirements that consider these challenges.

The deliverables with relevance for regulatory considerations are DEL2 *"AI4H regulatory considerations"* (which provides an educational overview of some key regulatory considerations), DEL2.1 *"Mapping of IMDRF essential principles to AI for health software",* and DEL2.2 *"Guidelines for AI based medical device (AI-MD): Regulatory requirements"* (which provides a checklist to understand expectations of regulators, promotes step-by-step implementation of safety and effectiveness of AI-based medical devices, and compensates for the lack of a harmonized standard). DEL4 identifies standards and best practices that are relevant for the "*AI software lifecycle specification."* The following sections discuss how the different regulatory aspects relate to the TG-POC.

8.1 Existing applicable regulatory frameworks

For further study

Most of the AI systems that are part of the FG-AI4H benchmarking process can be classified as *software as medical device* (SaMD) and eligible for a multitude of regulatory frameworks that are already in place. In addition, these AI systems often process sensitive personal health information that is controlled by another set of regulatory frameworks. The following section summarizes the most important aspects that AI manufacturers need to address if they are developing AI systems for point-of-care diagnostics.

8.2 Regulatory features to be reported by benchmarking participants

For further study

In most countries, benchmarked AI solutions can only be used legally if they comply with the respective regulatory frameworks for the application context. This section outlines the compliance features and certifications that the benchmarking participants need to provide as part of the metadata. It facilitates a screening of the AI benchmarking results for special requirements (e.g., the prediction of prediabetes in a certain subpopulation in a country compliant to the particular regional regulatory requirements).

8.3 Regulatory requirements for the benchmarking systems

For further study

The benchmarking system itself needs to comply with regulatory frameworks (e.g., some regulatory frameworks explicitly require that all tools in the quality management are also implemented with a quality management system in place). This section outlines the regulatory requirements for software used for benchmarking in this topic group.

8.4 Regulatory approach for the topic group

For further study

Building on the outlined regulatory requirements, this section describes how the topic group plans to address the relevant points in order to be compliant. The discussion here focuses on the guidance and best practice provided by the DEL2 *"AI4H regulatory considerations."*

References

1. Vaisman A, Linder N, Lundin J, Orchanian-Cheff A, Coulibaly JT, Ephraim RK, Bogoch, II. Artificial intelligence, diagnostic imaging and neglected tropical diseases: ethical implications. Bull World Health Organ. 2020;98(4):288-9.
2. Bogoch II, Lundin J, Lo NC, Andrews JR. Mobile phone and handheld microscopes for public health applications. The Lancet Public Health. 2017;2(8):e355.
3. Holmström O, Linder N, Kaingu H, Mbuuko N, Mbete J, Kinyua F, Törnquist S, Muinde M, Krogerus L, Lundin M, Diwan V, Lundin J. Point-of-Care Digital Cytology With Artificial Intelligence for Cervical Cancer Screening in a Resource-Limited Setting. JAMA network open. 2021;4(3):e211740-e.
4. Holmström O, Linder N, Ngasala B, Mårtensson A, Linder E, Lundin M, Moilanen H, Suutala A, Diwan V, Lundin J. Point-of-care mobile digital microscopy and deep learning for the detection of soil-transmitted helminths and Schistosoma haematobium. Global Health Action. 2017;10(sup3):1337325.
5. Holmström O, Stenman S, Suutala A, Moilanen H, Kücükel H, Ngasala B, Mårtensson A, Mhamilawa L, Aydin-Schmidt B, Lundin M, Diwan V, Linder N, Lundin J. A novel deep learning-based point-of-care diagnostic method for detecting Plasmodium falciparum with fluorescence digital microscopy. PLOS ONE. 2020;15(11):e0242355.
6. Perkel JM. Pocket laboratories. Nature. 2017;545:119.
7. Wetsman N. Artificial intelligence aims to improve cancer screenings in Kenya. Nature Medicine. 2019;25(11):1630-1.
8. Nicholls M. AI delivers cervical cancer screening to rural areas of Kenya. Healtcare in Europe, [Internet]. 2021 07.04.2021. Available from: https://healthcare-in-europe.com/en/news/ai-delivers-cervical-cancer-screening-to-rural-areas-of-kenya.html.
9. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.
10. Beam AL, Kohane IS. Translating artificial intelligence into clinical care. JAMA. 2016;316(22):2368-9.
11. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402-10.
12. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115-8.
13. Lancellotti C, Cancian P, Savevski V, Kotha SR, Fraggetta F, Graziano P, Di Tommaso L. Artificial Intelligence & Tissue Biomarkers: Advantages, Risks and Perspectives for Pathology. Cells. 2021;10(4).
14. Linder N, Turkki R, Walliander M, Martensson A, Diwan V, Rahtu E, Pietikainen M, Lundin M, Lundin J. A malaria diagnostic tool based on computer vision screening and visualization of Plasmodium falciparum candidate areas in digitized blood smears. PLOS ONE. 2014;9(8):e104855.
15. Serrano B, Brotons M, Bosch FX, Bruni L. Epidemiology and burden of HPV-related disease. Best Practice & Research Clinical Obstetrics & Gynaecology. 2018;47:14-26.
16. Torre LA, Islami F, Siegel RL, Ward EM, Jemal A. Global Cancer in Women: Burden and Trends. Cancer Epidemiology Biomarkers & Prevention. 2017;26(4):444.

17. Bouassa RSM, Prazuck T, Lethu T, Meye JF, Bélec L. Cancer du col de l'utérus en Afrique subsaharienne: une maladie associée aux papillomavirus humains oncogènes, émergente et évitable. Médecine et Santé Tropicales. 2017;27(1):16-22.

18. Fokom-Domgue J, Combescure C, Fokom-Defo V, Tebeu PM, Vassilakos P, Kengne AP, Petignat P. Performance of alternative strategies for primary cervical cancer screening in sub-Saharan Africa: systematic review and meta-analysis of diagnostic test accuracy studies. BMJ : British Medical Journal. 2015;351:h3084.

19. Elsheikh TM, Austin RM, Chhieng DF, Miller FS, Moriarty AT, Renshaw AA. American society of cytopathology workload recommendations for automated pap test screening: Developed by the productivity and quality assurance in the era of automated screening task force. Diagnostic Cytopathology. 2013;41(2):174-8.

20. Wilson ML, Fleming KA, Kuti MA, Looi LM, Lago N, Ru K. Access to pathology and laboratory medicine services: a crucial gap. The Lancet. 2018;391(10133):1927-38.

21. Mapanga W, Girdler-Brown B, Feresu SA, Chipato T, Singh E. Prevention of cervical cancer in HIV-seropositive women from developing countries through cervical cancer screening: a systematic review. Systematic reviews. 2018;7(1):198.

22. Katki HA, Kinney WK, Fetterman B, Lorey T, Poitras NE, Cheung L, Demuth F, Schiffman M, Wacholder S, Castle PE. Cervical cancer risk for women undergoing concurrent testing for human papillomavirus and cervical cytology: a population-based study in routine clinical practice. The Lancet Oncology. 2011;12(7):663-72.

23. Randall TC, Ghebre R. Challenges in Prevention and Care Delivery for Women with Cervical Cancer in Sub-Saharan Africa. Frontiers in Oncology. 2016;6(160).

24. Sayed S, Chung M, Temmerman M. Point-of-care HPV molecular diagnostics for a test-and-treat model in high-risk HIV populations. The Lancet Global Health. 2020;8(2):e171-e2.

25. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, Mahendiran T, Moraes G, Shamdas M, Kern C, Ledsam JR, Schmid MK, Balaskas K, Topol EJ, Bachmann LM, Keane PA, Denniston AK. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The Lancet Digital Health. 2019;1(6):e271-e97.

Annex A:
Glossary

This section lists all the relevant abbreviations, acronyms and uncommon terms used in the document.

Acronym/Term	Expansion	Comment
AI	Artificial intelligence	
AI4H	Artificial intelligence for health	
AI-MD	AI based medical device	
API	Application programming interface	
CfTGP	Call for topic group participation	
DEL	Deliverable	
FDA	Food and Drug administration	
FGAI4H	Focus Group on AI for Health	
GDP	Gross domestic product	
GDPR	General Data Protection Regulation	
IMDRF	International Medical Device Regulators Forum	
IP	Intellectual property	
ISO	International Standardization Organization	
ITU	International Telecommunication Union	
LMIC	Low-and middle-income countries	
MDR	Medical Device Regulation	
PII	Personal identifiable information	
SaMD	Software as a medical device	
TDD	Topic Description Document	Document specifying the standardized benchmarking for a topic on which the FG AI4H Topic Group works. This document is the TDD for the Topic Group POC
TG	Topic Group	
WG	Working Group	
WHO	World Health Organization	

Annex B:
Declaration of conflict of interests

No conflicts of interest were declared by the TG participants.