Artificial intelligence and visual inspection in cervical cancer screening

Carolyn Nakisige,1,2 Marlieke de Fouw,3 Johnblack Kabukye,4 Marat Sultanov,5 Naheed Nazrui,5 Aminur Rahman,6 Janine de Zeeuw,4 Jaap Koot,7 Arathi P Rao,8 Keerthana Prasad,9 Guruvare Shyamala,9 Premlatha Siddharta,10 Jelle Stekelenburg,4 Jogchum Jan Beltman11

ABSTRACT

Introduction Visual inspection with acetic acid is limited by subjectivity and a lack of skilled human resource. A decision support system based on artificial intelligence could address these limitations. We conducted a diagnostic study to assess the diagnostic performance using visual inspection with acetic acid under magnification of healthcare workers, experts, and an artificial intelligence algorithm.

Methods A total of 22 healthcare workers, 9 gynecologists/experts in visual inspection with acetic acid, and the algorithm assessed a set of 83 images from existing datasets with expert consensus as the reference. Their diagnostic performance was determined by analyzing sensitivity, specificity, and area under the curve, and intra- and inter-observer agreement was measured using Fleiss kappa values.

Results Sensitivity, specificity, and area under the curve were, respectively, 80.4%, 80.5%, and 0.80 (95% CI 0.70 to 0.90) for the healthcare workers, 81.6%, 93.5%, and 0.93 (95% CI 0.87 to 1.00) for the experts, and 80.0%, 83.3%, and 0.84 (95% CI 0.75 to 0.93) for the algorithm. Kappa values for the healthcare workers, experts, and algorithm were 0.45, 0.68, and 0.63, respectively.

Conclusion This study enabled simultaneous assessment and demonstrated that expert consensus can be an alternative to histopathology to establish a reference standard for further training of healthcare workers and the artificial intelligence algorithm to improve diagnostic accuracy.

INTRODUCTION

The WHO global strategy to accelerate the elimination of cervical cancer addressed targets that must be met by 2030. The strategy aims to upscale vaccination, screening, and treatment of cervical pre-cancer and cancer.3 Only 41% of low-income countries administer the human papillomavirus (HPV) vaccine as part of their immunization program, yet these countries suffer from the highest burden of cervical cancer morbidity and mortality.2 Therefore, screening and treatment of pre-cancer remain top priorities. In limited resource settings, visual inspection with acetic acid (VIA) is the most common screening method. It is relatively simple to use, cheap, and allows screening and treatment in one visit.3–5

Screening programs based on VIA need a functioning healthcare infrastructure and well-trained personnel.6,7 The diagnostic accuracy of this method is dependent on the skills of the healthcare worker. Its sensitivity and specificity vary with a sensitivity of 62.5–80% and a specificity of 80–98.8% for the detection of histologically confirmed cervical intra-epithelial neoplasia (CIN) II or more advanced cervical lesions against the reference standard of histology or colposcopy followed by biopsy.5–10 Given its subjectivity, VIA is associated with a problem of over treatment in a single-visit approach and at the same time possibly failing to identify women who are at high risk for cervical cancer.11,12

Recently, artificial intelligence (AI) systems have increasingly been applied in healthcare including cervical cancer screening.13 Studies in both high-income and low-income countries indicate that the application of AI in HPV testing,14–16 cytology,17–21 and colposcopy22–27 has achieved a good detection rate of pre-cancerous lesions with good accuracy. However, few studies investigated the potential of AI to improve the accuracy of VIA and function as a decision support system during screening in the primary care setting.
We investigate an AI-Decision Support System for VIA in three countries as part of a European and Indian funded research project, namely Prevention and Screening Innovation Project Towards Elimination of Cervical cancer (PRESCRIPTEC, www.prescriptec.org).28 Besides the introduction of the AI Decision Support System, the project enhances country-specific cervical cancer prevention programs with focused awareness strategies, community mobilization, and HPV self-testing to increase uptake and quality of screening.

In Uganda, India, and Bangladesh we assessed the baseline quality of VIA of trained healthcare workers, experienced gynecologists, and the AI by measuring their diagnostic accuracy and intra- and inter-observer agreement on magnified VIA images.

METHODS
We conducted a cross-sectional diagnostic study to assess the diagnostic performance in VIA-M (VIA with magnification) by healthcare workers, experts, and the AI. The healthcare workers’ team consisted of 10 members from Uganda, five from Bangladesh, and seven from India. They had been trained in VIA according to their respective national guidelines.

The expert team consisted of nine gynecologists with more than 10 years’ expertise in screening and ablative treatment. Two experts originated from Bangladesh, four from India, one from the Netherlands, and two from Uganda. All were selected by their respective countries (Table 1).

The AI Decision Support System was developed by Manipal School of Information Sciences, Manipal Academy of Higher Education in India and trained on 100 images after application of acetic acid collected from routine VIA clinics. The images were divided into three sets: training, testing, and validation in the proportion 70:20:10, respectively. There was no overlap between the images in these three sets. The gold standard was a single best expert’s report. The algorithm is built into an Android-based device to enable healthcare workers to introduce the device during screening in field conditions.29 The algorithm will generate an instant report after capturing the image, and can differentiate between a normal cervix (negative) and an abnormal cervix requiring further evaluation (positive).

Databank and Selection of Images
Cervical images from existing databanks of the Uganda Cancer Institute, International Agency for Research on Cancer,30 and Leiden University Medical Center were used to create a new dataset after obtaining consent from each institute. A total of 96 images were selected, in the proportions of 13, 73 and 10, respectively. Four images were purposively duplicated to assess intra-observer agreement making a total of 100 images that were presented for assessment. The expert team of nine gynecologists did not participate in the selection of these images to avoid recall bias.

Data Collection
An online tool to upload cervical images and questionnaires was developed by the Marconi AI laboratory in Makerere University, Uganda.31 The images were uploaded to this annotation platform which the experts and healthcare workers independently accessed on computer monitors. The paired cervical images for each case (before and after application of acetic acid) were presented in a single frame along with a brief questionnaire on the quality, VIA assessment, and if rated ‘positive’ eligibility for ablative therapy (Online supplemental table S1). These images were viewed on a computer monitor which enables VIA-M.

The AI ran on a computer instead of the Android-based device, assessed the images only after application of acetic acid, and provided a binary result (positive or negative).

Procedure
The healthcare workers and experts each created a personal account and provided personal data, including age, country, and

<table>
<thead>
<tr>
<th>Team</th>
<th>Participants n (%)</th>
<th>Age (years) median (IQR)</th>
<th>VIA experience (years) median (IQR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Healthcare workers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bangladesh</td>
<td>5 (23)</td>
<td>37 (30–42)</td>
<td>9 (8.50–11.0)</td>
</tr>
<tr>
<td>India</td>
<td>7 (32)</td>
<td>42 (35–51)</td>
<td>8 (3.0–15.0)</td>
</tr>
<tr>
<td>Uganda</td>
<td>10 (45)</td>
<td>27(25–33)</td>
<td>(0)*</td>
</tr>
<tr>
<td>Experts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bangladesh</td>
<td>2 (22)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>India</td>
<td>4 (44)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Uganda</td>
<td>2 (22)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Netherlands</td>
<td>1 (11)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>9 (100)</td>
<td>52 (38–56)</td>
<td>23 (10.0–26.0)</td>
</tr>
</tbody>
</table>

There were 22 healthcare workers from Bangladesh, India, and Uganda. The median (IQR) were calculated per each country team of healthcare workers.

*Only one healthcare worker out of 10 in Uganda had had VIA experience of 3 years. The team of 9 experts from 4 countries had a median age of 52 years (IQR 38–56) and median years of experience of 23 years (IQR 10.0–26.0).

VIA, visual inspection with acetic acid.
years of experience in screening. Participants were invited to complete the assessment of cervical images individually within a period of 1 week. Before starting they received a video and letter with guiding instructions. The project manager of each country team ensured all healthcare workers filled out the form individually without consulting sources like the internet or the screening manual. After completion, access to the forms was automatically locked.

The images were stored in a folder in a computer and the AI analyzed those images and reported them as being negative or positive.

The gold standard or reference was based on expert consensus due to lack of pathology. An expert consensus meeting was conducted after the images had been assessed individually by all the healthcare workers, experts, and the algorithm. Consensus was reached when at least five out of the nine experts agreed on the VIA assessment. Throughout the consensus meeting, the experts were not aware of the gold standard of the original databanks. In four images the expert consensus was different from the gold standard of the original databanks. All experts collectively agreed that the four images were VIA negative while the original databank stated VIA positive. Therefore, the VIA assessment of these four images was changed to VIA negative. It turned out that the histology of all four images was normal.

In four other images of the final dataset, results were inconclusive based on the individual expert assessment. The cut-off of at least five of nine was not attained to determine their final grades. These were re-evaluated by the consensus group and all experts agreed on grades which turned out to be similar to the original grades.

During the meeting, three images were excluded from analysis; one image was of insufficient quality for evaluation due to poor lighting, and two images were taken at follow-up after initial treatment with the Loop Electro-surgical Excision Procedure. These two images were excluded because women with previous cervical treatment will be excluded from the implementation study. Additionally, the algorithm could not assess 10 out of the 93 images due to file format issues. Therefore, a total of 83 unique images were used for final analysis (Figure 1). Of the 83 images, 24 had histopathological diagnoses: six normal, four CIN I, two CIN II, three CIN III, two with features of HPV infection, and seven squamous cell carcinomas. The expert consensus ended up with 48 VIA negative images, 26 VIA positive images, and nine suspected cancers. All cancer images were tagged positive to enable comparison with the binary outcome of the AI, resulting in 35 VIA positive images.

Analysis
Analysis was done using Statistical Package for Social Science version 26 where diagnostic performance and intra/inter-observer agreement were analyzed using sensitivity, specificity, area under the curve (AUC), false positives and false negatives as well as Fleiss kappa (κ) values.

Diagnostic Performance
Diagnostic performance was assessed by comparing the individual assessment, the majority vote of the assessment within the teams of healthcare workers, and the assessment of the algorithm to that of the expert consensus.

We evaluated the diagnostic accuracy of healthcare workers, individual experts, and the AI using sensitivity, specificity, receiver operating characteristics (ROC), and its summary statistic AUC. True positives, true negatives, false positives, and false negatives of the individual country teams were also reported. Feedback on their performance was given and the healthcare workers re-trained on the same images in their settings.

Intra-/Inter-Observer Agreement
We determined the intra-observer and inter-observer agreement within teams of healthcare workers, experts, and the algorithm using Fleiss κ values.32 Activities under the PRESCRIP-TEC were approved by the institutional review boards of Bangladesh, India, and Uganda and written informed consent obtained from both the healthcare workers and experts.

RESULTS
Characteristics of the Participants
A total of 31 individuals and the AI participated in this study. Table 1 illustrates their country of work, age, and VIA experience. The experience varied among the teams, from median 8 years (IQR 3.0–15.0) in India to 9 years (IQR 8.50–11.0) in Bangladesh, while Uganda had only one healthcare worker with 3 years of experience. The rest of the Uganda team had no experience prior to the training for the project.

Diagnostic Performance
Table 2 shows that the sensitivity of healthcare workers assessment was 80.4% and specificity 80.5%, and the sensitivity of expert assessment was 81.6% and specificity 93.5%. The algorithm demonstrated a sensitivity of 80.0% and a specificity of 83.3%.

Online supplemental figure S1 demonstrates the AUC, which was 0.80 (95% CI 0.70 to 0.90) for the healthcare workers, 0.93 (95% CI 0.87 to 1.00) for the experts, and 0.84 (95% CI 0.75 to 0.93) for the AI.

Table 3 illustrates the relationship between the AUC and diagnostic accuracy.33 Individual performance is presented in Online supplemental table S2.

Intra-/Inter-Observer Agreement
There was 100% intra-observer agreement of all healthcare workers and experts concerning the four duplicate images. Healthcare worker agreement within teams was moderate with κ values of 0.43, 0.44, and 0.48 in Bangladesh, India, and Uganda, respectively. Overall κ values for all healthcare workers, experts, and the algorithm were 0.45, 0.68, and 0.63, respectively. When excluding outliers in the team of healthcare workers, defined as having κ scores <0.5, agreement among healthcare workers was substantial with κ=0.63 (Table 4).

DISCUSSION
Summary of Main Results
We found that the diagnostic performance of healthcare workers was adequate and enabled sufficient quality of screening. The algorithm performed better than the healthcare workers but lower than the experts. Agreement between teams was comparable.
Figure 1  Flow chart of the VIA dataset assessed, comprising 96 unique images and four duplicates. Three were excluded at an expert consensus meeting, leaving 93. Ten were excluded after AI failed to assess them, leaving 83 images for final assessment. These comprised 48 VIA negative images, 28 VIA positive images, and nine suspected cancers. The nine suspected cancers were added to the VIA positives to make 35 VIA positive images and 48 VIA negatives to enable comparison with the binary outcome of the AI. AI, artificial intelligence; VIA, visual inspection with acetic acid.
Results in the Context of Published Literature

The diagnostic performance of the healthcare workers compared well to the sensitivity and specificity described in two reviews which mentioned sensitivities of 80% and 73.2% and specificities of 92% and 86.7%.9 10 In clinical practice the diagnostic performance might be less comparable to research settings. It was remarkable that the Ugandan team had the highest specificity, leading to fewer false positive results and lower risk of over treatment, while the sensitivity was low compared with the team in India and Bangladesh. The high sensitivity of the Bangladesh and Indian teams translates to few false negative results at the cost of possible over diagnosis and over treatment. This could be attributed to the higher prevalence of HPV infections and higher VIA positivity rate in Uganda compared with India and Bangladesh (unpublished project data). Furthermore, the overall performance measured by the AUC reflected the number of years’ experience with VIA which was highest in Bangladesh and lowest in Uganda.

The overall performance of the experts compared favorably with previous studies.8 9 The wide range of diagnostic performance was similar to what Vidya et al reported among a group of gynecologists and residents where sensitivity was 57.1–92.9% and specificity was 54.3–94.5%. The wide variation could be due to varying experiences of the experts and the subjectivity of VIA assessment.29 The sensitivity and specificity of the AI was lower than its previous performance after initial training.29 The lesser performance could be due to being tested on images from various sources with expert consensus as reference compared with the assessment of a single expert who was the reference during the training of the AI. To mitigate the risk of over fitting during the development phase of the algorithm, images were divided into three sets of training, validation and testing without overlap however, the total number was small.

Interestingly, the AI identified all images with suspected or invasive cancer as positive, although it was not trained on such images. It is trained only on images after application of acetic acid and this procedure is not performed when cancer is suspected. Further training of the AI including images of suspected cervical cancer will probably improve its diagnostic accuracy and applicability in the field.

The intra-observer agreement of the four images for all the participants was 100%, which is very rare but could be attributed
to the very small number of duplicate images. The inter-observer agreement of all healthcare workers was moderate, with an overall weighted \( \kappa \) value of 0.454. The moderate level of agreement could be due to the varying levels of experience between and within country teams, the different exposure to refresher training and supervision, and the very low specificity and sensitivity of a few individuals. After excluding the five outliers, overall weighted \( \kappa \) values improved from 0.454 to 0.625, thus from moderate to substantial agreement. Among the outliers with lower sensitivity and specificity were two healthcare workers without previous VIA experience. Therefore, we will conduct ongoing supervision and regular performance evaluations to continuously assess the quality of screening.

**Strength and Weaknesses**

This study is unique in evaluating the diagnostic performance of VIA-M among both experts and healthcare workers from different low resource settings simultaneously. The use of images from publicly available databases could have introduced recall bias. We mitigated this risk by adding pictures from different datasets, presenting the pictures in random order, and checking for intra-observer agreement with duplicate images, although the number was small. This presented varying gold standards, making comparison difficult. We developed a common expert consensus gold standard applicable to all the datasets.

We based our gold standard on expert consensus, given that pathology is not readily available in the study settings, while the desired reference in cervical cancer screening studies is histopathology. We used a team of nine experts to ensure the quality of the expert consensus; however, the bigger the number the more difficult it is to reach consensus. Screening by healthcare professionals will provide three possible outcomes, namely positive, negative, or suspected cancer. The AI provides only two possible outcomes, positive and negative. Given that detection of suspected cancers has great clinical importance, in future studies we hope to introduce a trinary outcome for the algorithm.

**Implications for Practice and Future Research**

During project implementation, the outcome of screening with VIA will be based on the assessment of healthcare workers with supervision provided by experts. At the same time, images will be captured on the Android-based device with the algorithm to further train the AI and validate its performance in field conditions.

Over time, the diagnostic performance of the AI and healthcare workers will be re-evaluated to assess the effect of training, regular feedback, and supervision. The data generated will confirm feasibility of the use of the AI as a decision support system in the field for cervical cancer screening. The AI could lead to task shifting of screening to less-trained healthcare providers, after rigorous validation in field conditions. Task shifting will increase the opportunities to provide quality screening to women, especially in resource constrained settings, and will potentially translate into increased uptake of screening and adherence to follow-up.

We recommend for future research to assess the feasibility of detection of suspected cancer and the eligibility for direct treatment with thermal ablation by the AI, and evaluate the perception of screened women and healthcare providers about the implementation of AI in screening programs. In addition, we recommend to evaluate the use of expert consensus as an alternative to histology as the gold standard for training the AI, especially in settings where pathology is not readily accessible.

**CONCLUSION**

The diagnostic accuracy in VIA-M of the healthcare workers and the AI was good, though lower than the experts. Agreement within teams was moderate to substantial. This signifies need for further training of the healthcare workers and the algorithm to improve diagnostic accuracy and agreement, with strict validation of the AI to maximize its performance in the field and allow task shifting of screening to less trained healthcare workers. This study showed that in the absence of pathology, expert consensus can be used as a reference to train healthcare workers and the AI.
REFERENCES

1 World Health Organization. Global strategy to accelerate the elimination of cervical cancer as a public health problem. 17 Nov 2020/global strategy;


30 International Agency of Research on Cancer. Cervical cancer image bank to aid artificial intelligence cancer detection–IARC screening group;

31 The Marconi Society machine learning laboratory, College of Engineering, Design Art and Technology. Makerere University, Kampala, Uganda.

