Computers in Biology and Medicine 186 (2025) 109710

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

A multimodal deep learning model for cervical pre-cancers and cancers prediction: Development and internal validation study

Sreenath Madathil [a,b], Mohamed Dhouib [c], Quitterie Lelong [c], Ahmed Bourassine [c], Joseph Monsonego [d,*]

[a] Faculty of Dental Medicine and Oral Health Sciences, McGill University, Montreal, Canada
[b] Gerald Bronfman Department of Oncology, Faculty of Medicine, McGill University, Montreal, Canada
[c] École Polytechnique, Institut Polytechnique de Paris, Palaiseau, France
[d] Institute of the Cervix, Paris, France

ARTICLE INFO

Keywords:
Artificial intelligence
CIN
Cervical cancer
Colposcopy
Risk prediction
Deep learning

ABSTRACT

Background: The current cervical cancer screening and diagnosis have limitations due to their subjectivity and lack of reproducibility. We describe the development of a deep learning (DL)-based diagnostic risk prediction model and evaluate its potential for clinical impact.

Method: We developed and internally validated a DL model which accommodates both clinical data and colposcopy images in predicting the patients CIN2+ status using a retrospective cohort of 6356 cases of LEEP-conization/cone-biopsy (gold-standard diagnosis) following an abnormal screening result. The overall performance, discrimination, and calibration of the model were compared to expert clinician's colposcopic impression. The potential for clinical impact was assessed with rate of unnecessary conizations that could be avoided by using our model.

Results: The model combining clinical history and colposcopy images demonstrated superior performance prediction of CIN2+(AUC-ROC = 95.3 %, accuracy = 90.8 %, PPV = 94.1 %, NPV = 87.9 %) and better calibration compared to models that used image or clinical history data alone and outperformed clinician's colposcopic impressions. Moreover, if a decision threshold of 10 % is applied to the predicted probability from this model to recommend conization, up to 35 % of conizations could be avoided without missing any true CIN2+ cases.

Conclusion: We present a novel DL model to predict cervical neoplasia with potential for reducing unnecessary conization. External validation studies are warranted for assessing generalizability.

1. Introduction

Cervical cancer remains a major contributor to female cancer incidence and morbidity worldwide [1–4]. Despite the advances in screening and diagnostic techniques, many cases still evade early detection and appropriate management [5–7].

Current recommendations for screening and diagnosis of cervical precancers and cancers include several procedures including Human Papillomavirus (HPV) testing, cytology triage, colposcopy, punch biopsy, endocervical curettage, cone biopsy and histopathological examination [8]. Nevertheless, despite the historical success of cytology, colposcopy, and histology in screening and diagnosis, these methods are not devoid of limitations [6,9,10]. The reliance on human visual assessment results in significant inter- and intra-observer variability, leading to suboptimal sensitivity and specificity [11–14]. For example, approximately 14 % of invasive cervical cancers are observed after inappropriate colposcopy management, and 9 % after a delay in care of more than 3 months following an abnormal colposcopy screening [6]. Moreover, sensitivity of colposcopy for high grade cervical intraepithelial neoplasia (HG-CIN) lesions is estimated to be between 50 % and 70 % and specificity lower than 50 % [14–16].

The reliability of punch biopsies also raises concern, with challenges in recognizing and localizing severe lesions, particularly when the squamocolumnar junction (SCJ) is not visible [9,16]. Histological analysis suffers similar fates of poor reproducibility and variability, especially when distinguishing between metaplasia and different grades of CIN [17,18]. The concordance between colposcopic impression and histological findings is less than optimal, estimated at under 65 % [9,19,20].

* Corresponding author. Institute of the Cervix, 174 rue de Courcelles, 75017, Paris, France.
E-mail address: jmonsonego@orange.fr (J. Monsonego).

https://doi.org/10.1016/j.compbiomed.2025.109710
Received 10 June 2024; Received in revised form 10 October 2024; Accepted 15 January 2025

List of abbreviations:

LEEP	Loop Electrosurgical Excision Procedure
PPV	Positive Predictive Value
NPV	Negative Predictive Value
AUC-ROC	Area Under Curve of the Receiver Operating Characteristic
HPV	Human Papillomavirus
HR-HPV	High Risk Human Papillomavirus
HG-CIN	High Grade Cervical Intraepithelial Neoplasia
CIN	Cervical Intraepithelial Neoplasia
DL	Deep Learning
AI	Artificial Intelligence
SCJ	Squamocolumnar Junction
VaIN	Vaginal Intraepithelial Neoplasia
LSIL	Low-Grade Squamous Intraepithelial Lesion
HSIL	High-Grade Squamous Intraepithelial Lesion
ACS-US	Atypical Squamous Cells of Undetermined Significance
ASC-H	Atypical Squamous Cells, HSIL cannot be excluded
AGC	Atypical Glandular Cells
TZ	Transformation Zone
TPR	True Positive Rate
FPR	False Positive Rate

Clinical management of cervical lesions is heavily dependent on the colposcopic impression, which is influenced by lesion characteristics, including the volume and visibility of the SCJ [16]. Discrepancies in assessment can lead to the underestimation of HG-CIN in the presence of large lesions or overestimation when the endocervical junction is not visible, resulting in potentially harmful overtreatment, including excessive conization with its deleterious effects [16,21].

The shift to HPV screening accompanied by cytological triage has led to an increase in colposcopy referrals and a consequent rise in the detection of HG-CIN [22–24]. In addition, this has also resulted in a surge in normal colposcopies and a heightened risk of overdiagnosis and overtreatment, given the reduced specificity of an HPV based screening protocol [22,24,25]. Such a trend demands critically evaluating current practices and an exploration of more reliable alternatives.

The expertise required for colposcopy is directly related to the volume of abnormal cases examined, making high proficiency less common in everyday clinical practice [14]. Training programs have attempted to address these deficiencies, yet gaps remain, especially in rural areas of high income countries and in low-resource settings, where the scarcity of skilled practitioners is most acute [26–29]. These disparities in skill and resource allocation contribute to inequities in care and can lead to considerable healthcare costs.

In light of these challenges, there is a pressing need for more reliable, reproducible, and objective tools that can enhance the sensitivity and specificity of colposcopy. Artificial Intelligence (AI) based prediction models stand out as a promising solution in this regard, potentially transforming cervical cancer screening and management into a more accurate, less subjective, and accessible procedure across diverse healthcare settings [30–34].

A recent systematic review highlighted the development of eleven deep learning models designed for cervical cancer screening [35]. Out of these, four models used colposcopy images, and only one compared the performance of the model to that of human experts. The review estimated a pooled sensitivity (83 %) and specificity (80 %) of these models, indicating their potential effectiveness in identifying pre-cancerous and cancerous lesions. However, the review also pointed out critical shortcomings in these models, such as their limited ability to generalize across diverse patient populations and imaging conditions.

Additionally, most of these studies failed to report key evaluation metrics such as calibration and potential clinical impact. For instance, metrics such as expected calibration error and net clinical benefit of using a prediction model can aid in understanding how well these models would perform in a real-world clinical setting. Furthermore, except for one study, none of the existing models used clinical history data in addition to the colposcopy images. This limitation is significant because cervical neoplasia diagnosis often involves integrating information from various sources, such as patient epidemiological profile, clinical history, screening findings, and laboratory results. A model that only focuses on colposcopy images may not fully capture the complexity of clinical decision-making in this context.

These shortcomings are particularly problematic when considering the application of these models in resource-limited settings. In such environments, where healthcare resources are scarce and access to expert pathologists may be limited, the need for reliable and calibrated tools is even more critical. A model that fails to incorporate the full spectrum of clinical data or that lacks robust validation may not be suitable for these settings.

In this manuscript, we report on the development and internal validation of a deep learning model designed to predict cervical precancerous and cancerous lesions from clinical history data and colposcopy images. Furthermore, we present a comprehensive suite of evaluations, including the potential for clinical impact.

2. Methods

Clinical decision point of interest: Our focus is on the clinical decision-making process following an abnormal cervical cancer screening test result using cytology alone or HR-HPV positivity combined with reflex cytology triage. Our fundamental aim was to evaluate the potential of deep learning-based diagnostic risk prediction models to aid clinicians in the decision to perform cone biopsy or LEEP-conization.

Data source: The data for this retrospective cohort study originates from clinical records obtained from a gynecology and cervical pathology specialist clinic in Paris, France. The database is meticulously compiled, cleaned and prepared by an expert (JM) who brings decades of expertise in the field of cervical pathology. The patient cohort in this database consists of women referred for colposcopy following abnormal screening results. The database was established initially to support clinical documentation and decision-making and facilitate effective communication with referral clinicians. All colposcopic examinations and cone biopsy or LEEP-conizations were performed by a single clinician- JM using a Carl Zeiss Jenna Colposcope. And images were recorded using a Toshiba 3CCD camera.

At each consultation, a detailed clinical history is collected including age, type of contraception; the number of pregnancies, smoking status; HPV vaccination status; history of abnormal cytology; history of high-risk human papillomavirus (HR-HPV) positivity; history of untreated vaginal (VaIN) or cervical intraepithelial neoplasia (CIN); history of treatment for CIN or VaIN; history of conization, loop electrosurgical excision procedure (LEEP) or endocervical curettage. In addition, colposcopy images and the clinical impression of satisfactory colposcopy for assessment (Normal, doubtful, LSIL, HSIL, or cancer) or non-satisfactory (SCJ not visible in the endocervix) are recorded for each colposcopy visit. This information was recorded using a clinical chart template; an example of which is provided in the supplementary materials. We used a document parsing algorithm to extract clinical information (Fig. S1).

Cone-biopsy and endocervical curettage were performed for women with persistent abnormal screening and non-satisfactory colposcopy (Transformation Zone (TZ) not visible on the endocervix). The following classification of Transformation Zone were used for this purpose: TZ1: Transformation Zone type 1 – the whole transformation zone is ectocervical and is fully visible; TZ2: Transformation Zone type 2: the upper limit of TZ is partially or entirely observable within the canal and is fully visible in a 360-degree circumference. TZ3: Transformation Zone type 3:

Either a portion or the whole upper limit of the TZ is not visible within the canal. In TZ3, the outer edge may be observable on the ectocervix, within the canal, or it may also be not visible at all. Finally, a cone biopsy or LEEP-conization was performed for women with biopsy(s) result of high-grade CIN.

Sample size: The reference database included 20,693 unique participants with more than 64,000 consultations of colposcopy between January 2010 and May 2023. Out of which 5009 women met our inclusion-exclusion criteria and contributed a total of 6356 data points (Normal/CIN1 = 3534; CIN2+/Cancer = 2822). However, most women [3992 (79.7 %)] contributed only one LEEP-conization or cone-biopsy. We kept 400 data points (Normal/CIN1 = 200; CIN2+/Cancer = 200) aside as the test dataset. To avoid data leakage between training and test datasets, we restricted data from the same patient to either of those sets.

Population: Majority of women who report to the recruiting clinic is between the ages of 32–48 and approximately 2 in 3 of women (74.6 %) are referred there after their primary cervical cancer screening with cytology (before 2019) or with HR-HPV testing + reflex cytology (after 2019). A minority of women (25.4 %) are referred with one of the following: i) a history of HR-HPV positivity (including persistent infections after one year) or a history of abnormal cytology screening (ACS-US+) or history of untreated or treated CIN1 and or Vaginal Intraepithelial Neoplasia (VaIN) (10.9 %); ii) history of treatment including cone biopsy or LEEP-conization for CIN2+(14.5 %). All participants signed an informed consent form, and the research protocol was approved by the institutional research ethics board of McGill University, Montreal, Canada.

Inclusion criteria: Our inclusion criteria were women who: i) had undergone colposcopy following a response to either an abnormal screening result or a persistent HR-HPV positivity after one year; ii) had at least one colposcopy image, iii) cone biopsy or LEEP conization was available.

For this study, an abnormal screening result that recommend performing a colposcopy was defined as any of the following:

1) For data before 2019, an abnormal screening result was defined based on cytology result such as i) Atypical Squamous Cells of Undetermined Significance (ASC-US) and HPV positive; ii) Atypical Squamous Cells, HSIL cannot be excluded (ASC-H); iii) Low-grade Squamous Intraepithelial Lesion (LSIL); iv) High-grade Squamous Intraepithelial Lesion (HSIL); v) Atypical glandular cells (AGC).
2) For data after 2019, an abnormal screening result was defined as positive for HR-HPV and a reflex cytology result of ASC-US+ and for women with normal cytology a persistent HR-HPV positive one year after an initial HPV positive test.

The difference in the definition is due to the shift in the first-line cervical cancer screening recommendation for women aged 30–65 years in France [36].

Exclusion criteria: In order to have a homogenous quality of dataset we excluded data for women who had undergone prior colposcopy with intervention (e.g, LEEP-conization) in a different clinic, than to the recruiting clinic; as the colposcopy images prior to any intervention for those women were not available.

All participants followed the standard of care management following the colposcopy.

Outcome: The binary ground truth outcome was chosen as the results from cone biopsy or LEEP-conization (Normal/CIN1 vs CIN2+/Cancer including Adenocarcinoma in situ) performed within the subsequent six months of the colposcopy consultation. If both punch biopsy and LEEP-conization results were available, and a discordance existed between them, the more severe of the two results was chosen as the ground truth.

Model development: Motivated by the need to add incremental value to the AI-based diagnostic prediction, we developed a multimodal deep learning model (*CerVital Predict)* which can utilize clinical history data in text form and colposcopy images as input. Moreover, to adapt to different clinical settings this model was trained to take inputs as i) clinical history data extracted from clinical notes alone (*CerVital-history*), ii) colposcopy images alone (*CerVital-colpo*) and iii) a combination of both (*CerVital-combo*). The data extracted from clinical notes included age, smoking, parity, menopause status, HR-HPV test results with or without genotyping, HPV vaccine status and prior cytology - histology results or treatment for CIN. Multiple colposcopy images (on average, 3.8 images per consultation) from the same consultation were used as input for the latter two models. Technical details of the architectures, training, validation and testing procedures are provided in the supplementary materials. Briefly, we used ConvNeXt architecture [37] to extract visual features from the images (Fig. 1). An encoder-only transformer [38] was used to extract clinical feature vector from the clinical history information (Supplementary Fig. S1). These extracted features are then used by three distinct classifiers to predict CIN2+/Cancer status.

3. Model evaluations

Performance on CIN2+ prediction: Most often, AI-based prediction models are evaluated using classification-based metrics (e.g., Sensitivity, Specificity). Although these metrics are easy to interpret, they are less rigorous than predicted probability-based metrics and combine a decision to essentially a prediction problem. Following several recommendations for comprehensive and robust evaluation of predictions models [39,40] we compare our three models in four dimensions: i) overall performance (Brier score, Scaled Brier score, Brier skill score); ii) discrimination (AUC-ROC, Discrimination slope); iii) calibration (calibration-in-the large, calibration slope, and calibration plot); iv) classification-based metrics (Sensitivity, Specificity, Accuracy) [40]. See the legend of Table 3 for detailed description of each metric and their interpretation.

Potential for reducing unnecessary interventions: Several recent reports suggest the need to evaluate prediction of models beyond the usual performance metrics and the importance of assessing potential for clinical impact [39,41,42]. We evaluated our model's potential clinical impact (net benefit) by calculating unnecessary LEEP-conization or cone-biopsy that could be avoided using decision curve analysis [43]. The net benefit of the three models was also compared to a scenario where all participants who reported for colposcopy were followed up with a conization or biopsy.

Added value of colposcopy images for prediction of CIN2+: We also evaluated the incremental value in the predictive power of the model, added by utilizing the colposcopy images in addition to the clinical history of the participants. This metric is essential to consider in settings where access and resources are limited (e.g., skilled colposcopist is not easily accessible) or colposcopy images are not available.

Comparison to colposcopy impression of the expert colposcopist: We compared the sensitivity, specificity, PPV, and NPV between the models and the expert's impression of the colposcopy before the conization/cone-biopsy was performed.

4. Results

The description of the database is presented in Table 1. The median age of the participants was 38 years, with the majority having HPV-positive screening results (59 %), 30 % were smokers, 88 % are non-menopausal, approximately 60 % did not use any contraceptives and 5.5 % have been vaccinated against HPV with the quadrivalent vaccine.

As discussed previously, data from multiple colposcopy-conization cone biopsy pairs from the same women are considered. Details of the clinical history for colposcopy-conization cone-biopsy pairs included in the dataset is presented in Table 2. The majority of these datapoints had a history of abnormal cytology [ACS-US+](73 %) or were positive (59 %) for HPV at the colposcopy consultation visit. Approximately one

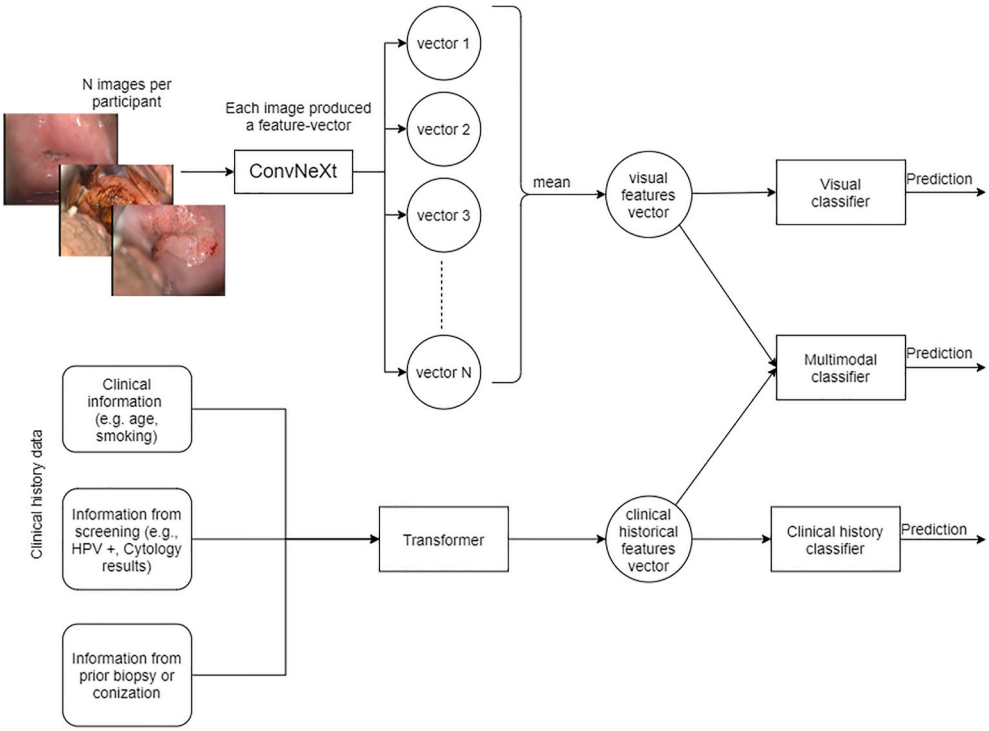

Fig. 1. Illustration of data flow and models. ConvneXt model is used to process colposcopy images. Embedding models are used to process clinical history data. Models were trained to predict the probability of CIN2+ using colposcopy image alone, clinical data alone and combination of both.

Table 1
Baseline (at recruitment) characteristics of the women included in the study.

Characteristic	N = 5,009[a]
Age at recruitment	
Median (IQR)	38 (32, 48)
Range	17, 87
Menopausal status	
Menopausal	613 (12 %)
Non-Menopausal	4396 (88 %)
Type of contraception	
No Contraception	2690 (54 %)
Mechanical	538 (11 %)
Hormonal	1781 (36 %)
Number of pregnancies	
None	2314 (46 %)
One	1022 (20 %)
Two	1162 (23 %)
More than Two	511 (10 %)
Number of birth	
None	2519 (50 %)
One	927 (19 %)
Two	1123 (22 %)
More than Two	440 (8.8 %)
Tobacco smoking	
Non-smoker	3475 (69 %)
<5	507 (10 %)
5 to 10	574 (11 %)
>10	453 (9.0 %)
HPV vaccination status	
Un-vaccinated	4734 (95 %)
Vaccinated	275 (5.5 %)

[a] n (%).

fourth of the datapoints had a history of abnormal screening (21 %) or HPV positivity (11 %), and 17 % had a history of CIN2+ confirmed cone-biopsy or LEEP-conization.

Table 2
Clinical history of datapoints included in the study.

Characteristic	N = 6,356[a]
Cytology result at recruitment	
Normal	1704 (27 %)
ASC-US+	4652 (73 %)
HPV status at recruitment	
Not Performed	2140 (34 %)
Negative	469 (7.4 %)
Positive	3747 (59 %)
Prior cytology result	
Normal	4997 (79 %)
ACS-US+	1359 (21 %)
Prior HPV test result	
Not Performed	5532 (87 %)
Negative	128 (2.0 %)
Positive	696 (11 %)
Prior Conization/Cone-Biopsy result	
Not performed	5062 (80 %)
Normal/CIN1	191 (3.0 %)
CIN2+	1103 (17 %)
Year of recruitment	
[2010,2019]	4358 (69 %)
(2019,2023]	1998 (31 %)

[a] n (%).

4.1. Performance of deep learning models improving detection of CIN2+

Overall, the *CerVital-combo* model, which combined both clinical history data and colposcopy images, performed the best, achieving an AUC-ROC of 95.3 %, Accuracy of 90.8 %, Specificity of 94.5 %, Positive Predictive Value (PPV) of 94.1 % and Negative Predictive Value (NPV) of 87.9 % (Table 3 & Supplementary Fig. S2). This model also showed overall better calibration than the other two models. Interestingly, the *CerVital-history* model had acceptable performance with 88 % AUC-ROC, 82 % accuracy and 90 % Specificity and was calibrated well, even though the only input was clinical history data (Table 3 & Supplementary Fig. S3).

The CerVital-colpo model demonstrated a 28 % improvement in predicting the probabilities of CIN2+/Cancer compared to the clinical history data-only model, as measured by the Brier Skill Score. Similarly, the CerVital-combo model showed an almost 50 % increase in performance compared to the CerVital-history model based on the same metric (Table 3).

Table 3
Performance of deep learning models[a] for CIN2+/Cancer prediction.

	Model with clinical history only *(CerVital-history)*	Model with colposcopy images only *(CerVital-colpo)*	Model with both images and clinical history *(CerVital-combo)*	Expert clinician's colposcopy impression
Overall performance[b]				
Brier score	0.14	0.1	**0.071**	
Scaled Brier score	0.45	0.6	**0.71**	
Brier skill score	ref	28 %	**49 %**	
Discrimination[c]				
AUC-ROC	88.1	92.4	**95.3**	
Discrimination slope	0.46	0.59	**0.71**	
Calibration[d]				
Calibration-in-large	0.24	−0.23	**−0.1**	
Calibration slope	0.93	1.13	**1.05**	
Expected calibration error	0.04	**0.01**	0.02	
Classification				
Sensitivity	**90.0 %**	79.5 %	87.0 %	80.0 %
Specificity	74.0 %	93.5 %	**94.5 %**	85.0 %
Positive predictive value	77.6 %	92.4 %	**94.1 %**	84.2 %
Negative predictive value	88.1 %	82.0 %	87.9 %	81.0 %
Accuracy	82.0 %	86.5 %	**90.8 %**	82.5 %

[a] Best metrics in each type are presented in bold font.
[b] *Overall performance:* metrics that combine the ability of the model to predict the probabilities of precisely and how reliable the model's predictions are: i) *Brier score*: measures the accuracy of probabilistic predictions(lower is better); ii) *scaled brier score*: compares the model's ability to predict the probabilities precisely compared to a random guess(higher values are better); iii) *Brier skill score*: compares the brier score of two models and aids in relative comparison of models (higher values are better).
[c] *Discrimination:* is the ability of the model to discriminate between Normal/CIN1 vs CIN2+. *AUC-ROC:* area under the receiver operator curve (higher values are better); *Discrimination slope:* asessess how well a model can differentiate between two groups on probability scale. It is the difference between the average predicted probability CIN2+ between the two groups(higher values are better).
[d] *Calibration:* is a measure of how well the predicted probabilities of an event match the actual occurrence of that event. In fields of medicine, calibration is crucial for assessing how well a model's predictions reflect reality. A well-calibrated model means that for all instances where the model predicts a certain probability of an event, that event actually occurs with that frequency. *Calibration-in-large:* is a measure of the systematic bias in the predicted probabilities of a model, indicating whether the model tends to overpredict or underpredict the probability of an event on average. (values closer to zero are better). *Calibration slope:* assesses how well the predicted probabilities correlate with the actual outcomes, particularly focusing on the consistency of this relationship across different levels of predicted risk (values closer to 1 are better). *Expected calibration error:* quantifies the average difference between predicted probabilities and actual outcomes, providing a clear indication of a model's calibration accuracy (lower values are better).

4.2. Performance of deep learning models compared to expert clinician's impression

We also compared the colposcopic impression of the expert clinician (JM) with the histopathological results from the LEEP-conization or biopsy. The *CerVital-combo* model outperformed expert clinician in all classification metrics [sensitivity (87%vs80 %), specificity (94%vs85 %), PPV (94%vs84 %), NPV (88%vs81 %) and accuracy (91%vs82 %)]. Additionally, the *CerVital-colpo* outperformed expert clinician's impressions in PPV (92%vs84 %), NPV(82%vs81 %), Accuracy(86%vs82 %) and Specificity(93%vs85 %) (Table 3).

4.3. Added predictive value of colposcopy images over clinical history-based prediction of CIN2+/Cancer

The added predictive value analysis aims to estimate to what extent the image data adds to the model's predictive power in addition to the clinical history data. The analysis compares the variance in the predicted probabilities explained by a reference model and estimates the fraction of information added into the model by the new input data [44–46]. Our results show that adding colposcopy images can add 33 % additional information into the prediction model with clinical history alone (Supplementary materials Table S1).

4.4. Clinical utility of using deep learning-based prediction models

The decision curve analysis is designed to assess the clinical value of prediction models by considering the trade-offs between the benefits of an intervention (LEEP-conization or cone biopsy) for women who have a high risk of CIN2+ and harms of the intervention for women who have low risk of CIN2+ across a range of decision thresholds. The decision thresholds are probability values at which a clinician might decide to do cone biopsy or LEEP-conization for a patient. A novice clinician might choose a very low decision threshold for cone biopsy or do LEEP-conization for everyone after colposcopy (Intervention for all strategy). The decision curve analysis can also be used to compare such strategies verses a scenario where clinician uses a prediction model to aid in their decision making.

We present the standardized-net-benefit calculated as a function of true positive rate(TPR), false positive rate(FPR) and the prevalence of CIN2+/Cancer in the test data. A standardized-net-benefit is the measure of relative clinical utility of a prediction model. Where the maximum possible clinical utility is 1 (TPR = 1 and FPR = 0).

Our results show that using our prediction models, by clinicians to decide for whom, to recommend cone biopsy or LEEP-conization will yield better clinical utility compared to all other strategies of decision making, across a broad range of threshold probabilities (Fig. 2). Interestingly, the *CerVital-history* model that uses clinical history data alone showed higher net benefit for high-risk groups (>65 % probability of CIN2+/Cancer), compared to the expert's impression.

4.5. Reduction in unnecessary cone biopsy or LEEP-conization

Following the abnormal screening result, a default strategy may be to perform a punch biopsy for all colposcopy consultations and cone or biopsy LEEP-conization for non-fully visible disease. However, this would lead to unnecessary intervention for patients with a low probability of CIN2+/Cancer. Fig. 3 presents the evaluation of the extent to which use of our models can reduce unnecessary cone biopsy or LEEP-conization without missing diagnosis of any CIN2+/Cancer.

For example, a strategy where LEEP-conization or cone biopsy is recommended for all women for whom the *CerVital-combo model* predicts at least 10 % probability of CIN2+/Cancer, can avoid 35 unnecessary conizations for every 100 participants without missing any diagnosis of CIN2+Cancer among this group. The corresponding estimates for the *CerVital-colpo* and *CerVital-history* models were 24 per 100

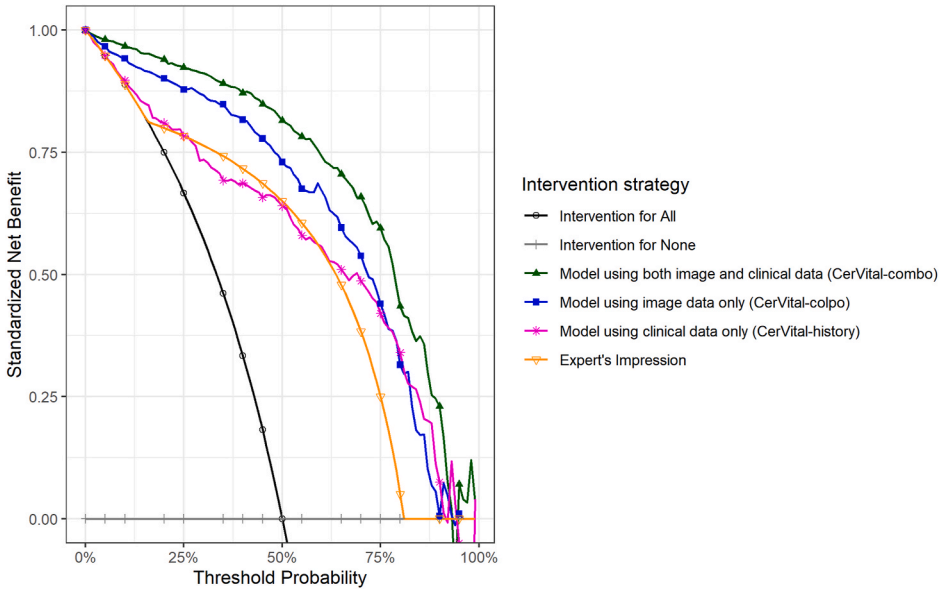

Fig. 2. The standardized-net-benefit curve plot compares different deep learning models to conization for all strategies. * In advising LEEP-conization or cone-biopsy for women surpassing a specified risk threshold, the assumption is that such intervention proves beneficial for those with CIN2+/Cancer and potentially harmful for those without. The nature of benefits (e.g., monetary gain, survival odds improvement, early treatment) and harms (e.g., added costs, stress, missed workdays, pain, complications) varies, yet decision curve analysis doesn't necessitate quantifying these values. Our results indicate that recommending LEEP-conization based on predicted risk from our AI-based model almost always leads to better net-benefit. It is further important to note that prevalence of CIN2+/Cancer affects net-benefit. For example, a strategy of recommending LEEP-conization to all women may be more favorable in high-prevalence populations.

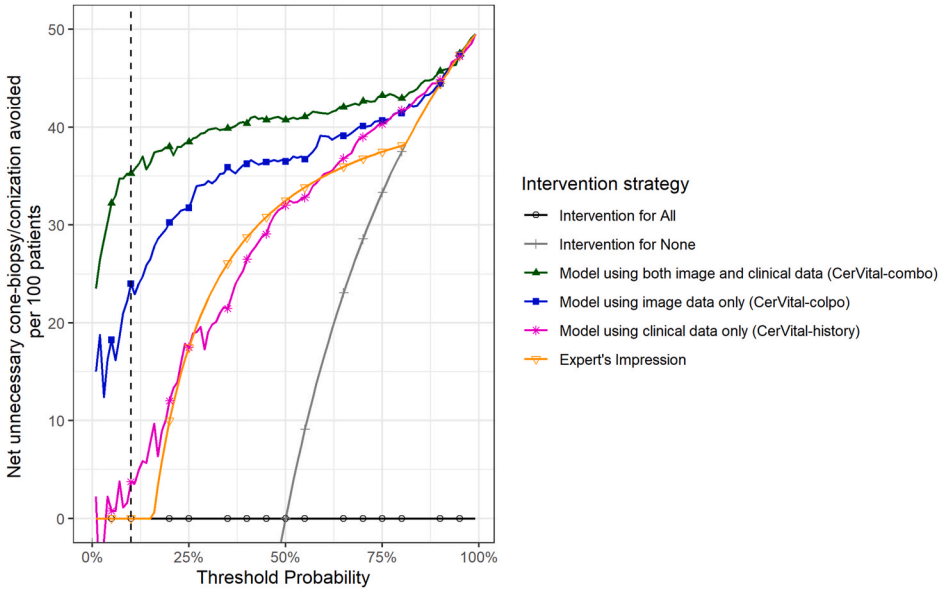

Fig. 3. Net unnecessary conization avoided per 100 patients reporting for colposcopy for different intervention strategy. Intervention refers to the act of performing conization or cone-biopsy. Black line represents the strategy where conization or cone-biopsy is performed for all women reporting for a colposcopy. Green, Blue and Pink lines represents the strategies where conization or cone-biopsy is performed for woman who are above a threshold of predicted probability for CIN2+ based on the AI-model that uses image and clinical data, image data alone, and clinical data alone respectively. Yellow line represents the strategy where conization or cone-biopsy is performed for those women for whom the expert clinician in the study assigned a probability of CIN2+ above a threshold. Grey line represents the strategy where none of the women undergo conization or cone-biopsy. The plot shows the net unnecessary intervention avoided for a range of thresholds. The vertical dashed line represents a threshold probability of 10 %. A strategy suggesting cone biopsy or LEEP-conization for women with over 10 % CIN2+/Cancer risk could reduce unnecessary procedures by 35 per 100 participants using our model, which combines image and clinical data. In comparison, models using only images or clinical data resulted in 24 and 4 avoidable procedures per 100 participants, respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

participants and 4 per 100 participants, respectively. Moreover, our prediction models show clear potential for clinical impact in reducing unnecessary conization across a wide range of risk thresholds. Interestingly, the *CerVital-history* model which used clinical history data alone showed comparable potential for impact as the expert clinician's impression at colposcopy consultation (Supplementary Table S3).

4.6. Performance of deep learning models improving detection of CIN2+ among participants with different transformation zone types (TZ)

We assessed the performance of our models among subcategories of participants with differing transition zone types. The majority of datapoints in the test dataset were of TZ1 type (297, 74.3 %) followed by TZ2 (82, 20.5 %) and TZ3(13, 3.2 %). As expected, *CerVital-combo* and *CerVital-colpo* models showed the best performance across different TZ types, and the performance of all models are lower among TZ3 group compared to TZ1 and TZ2. However, even among participants with TZ3 type, the *CerVital-combo* model showed acceptable levels of classification performance with a PPV of 60 %, NPV of 75 % and discrimination [AUC-ROC of 76 %] (Supplementary Table S2).

Model performance in out of distribution samples: Although some previous studies report external validation, those datasets are not openly available for testing excluding the possibility of a direction comparisons to other DL models for CIN2+ classification. Moreover, currently, there are no open access benchmark data sets for cervical lesion diagnosis. Hence indirect comparisons of performance of models are also not possible. A welcoming move in this direction is the International Agency for Research on Cancer (IARC) Cervical Cancer Image Bank [47]. This image bank provides access to 913 colposcopy images from 200 cases. However, LEEP conization results, gold standard ground truth, were only available for 55 participants (CIN1 = 2, CIN2+ = 53). Thus, not suitable as an external validation dataset. Nonetheless, this dataset could inform on how the model performs on out of distribution samples (samples that may have been captured by other types of colposcopy devices). Hence, we calculated the sensitivity of our CerVital-colpo model among the 53 CIN2+ participants from this dataset. Forty-seven (88.6 %) out of 53 cases were correctly identified as CIN2+ by the model. In the interest of reproducibility, we provide the code to implement our models and the predictions for the 55 cases from IARC Cervical Image Bank dataset at https://github.com/MadathilSA/CerVital.

5. Discussion

Precision in colposcopy-based impression and management of cervical precancers and cancers remains a challenge. AI-based decision support systems have been proposed to aid clinicians in this process [6,9,14,34,48]. A recent scoping review identified 20 publications that developed deep learning models using colposcopy images for this purpose [49]. The majority of these publications used clinical opinion as the ground truth label, whereas only seven studies used a comparator of biopsy [49]. However, among those that used biopsy results as the ground truth almost all of them used punch biopsy results instead of the gold-standard cone-biopsy or LEEP-conization. To the best of our knowledge, none of the previous studies used the gold-standard diagnostic test (LEEP-conization/cone-biopsy) as the ground truth label. Moreover, this scoping review highlighted the need to investigate the clinical utility of deep learning models.

Motivated by the limitations of the previous studies, a recent report [50] put forth five guiding principles for developing and evaluating AI models for clinical decision support in cervical cancer screening and management. These fundamental principles include 1) identifying the clinical decision point where the model is intended to be used (e.g., screening, diagnosis, prognosis); 2) focusing on clinically important errors; 3) evaluating models using clinical epidemiologic criteria (e.g. calibration); 4) evaluate algorithms on absolute risk scale which are clinically meaningful; 5) risk-based scenarios of clinical use which can match to local resources limitations and priorities. In alignment with these principles, here we describe the development and internal validation of a multimodal deep learning-based diagnostic risk prediction model that use the cone biopsy or LEEP-conization gold-standard histopathological results as ground truth. Moreover, we trained our model to be adaptable under different application scenarios where only colposcopy image or only clinical history data is available. We further evaluated the model using clinically important metrics of overall performance, calibration that uses absolute risk predicted by the model, and further assessment of the potential for clinical impact via decision curve analysis. We focused on the model's clinical utility in reducing an important error of unnecessary cone-biopsy or conization. Finally, we demonstrate the added value of using colposcopy images in making diagnostic risk predictions.

Our *CertVital-history* model showed an acceptable accuracy of 82 %, a specificity of 90 % which showed similar performance, in predicting CIN2+, compared to expert colposcopist's impressions before biopsy or LEEP-conization without using colposcopy images as input. However, it is important to acknowledge that our results do not suggest replacing the current diagnostic methods. Rather the results highlight the potential of our deep learning-based model to aid clinicians in making decisions such as situations where a discrepancy exists between histology and colposcopy. Importantly, the *CerVital-history* model may have significant benefit in resource-limited settings or in under screened populations where access to colposcopy images or a non-expert colposcopist would like to use it to support clinicians in their patient assessment and decision. Our model, after external prospective validation, has also the potential to be used as immediate triage after HR-HPV positive screening avoiding the high risk of loss in the follow up (up to 50 %) [51,52].

On the other hand, *CerVital-colpo* model has the potential to be used in settings where colposcopy images are available but several important clinical history information is missing or as a second opinion during a colposcopy visit and in combination with telemedicine. Moreover, in resource limited settings the current strategy of HR-HPV testing followed by reflex cytology-based screening approach requires multiple appointments that in turn may lead to lost to follow-up and higher economic burden. Validated and well calibrated prediction models have the potential to reduce this loss-to follow-up by flagging those women, in the real-time, with a high risk of CIN2+ at the primary HR-HPV screening (e.g., our model that uses clinical data only) or at colposcopy visit (e.g., our model that uses both images and clinical data). Subsequently, a targeted approach to streamline the screening process for these high-risk women can be pursued.

The CerVital-hitory model showed slightly higher Sensitivity and PPV compared to the CerVital-combo models. One possible explanation for this phenomenon is the difference in the classification ability of historical data and colposcopy image data. For example, HPV testing results have shown to have higher sensitivity and negative predictive value, compared to colposcopy, for detecting high-grade CIN [53–55]. However, the CerVital-combo model has shown higher overall, discrimination and calibration performances and hence comparatively better model for probabilistic prediction of CIN2+. Moreover, our target population are those women who are referred to colposcopy after an abnormal screening results (referral population), and not at primary screening stage. In this context, the goal is to enhance the positive predictive value (PPV) and specificity, which the CerVital-combo model achieves, rather than to increase sensitivity. For primary screening, which is beyond the scope of our work, the emphasis would be on improving sensitivity and negative predictive value (NPV).

Reducing unnecessary cone biopsy and LEEP-conization is of crucial importance, especially since the transition to HPV screening was followed by cytological triage (non-blind). For example, we have observed a significant increase in colposcopies and, consequently, the number of normal colposcopies compared to when screening was performed by cytology alone [22]. While this new approach contributes to detecting more high grade CIN, the reduced specificity may lead to overdiagnosis and overtreatment. Moreover, after the abnormal screening, a non-visible endocervical squamo-columnar junction (TZ3) may lead to an overestimation of the lesion, resulting in excessive cone biopsy or LEEP-conization and morbidity. All our models show clear potential for clinical impact when measured using the net avoided LEEP-conization or cone biopsy per 100 patients without losing the ability to capture

all true positives among them correctly. To the best of our knowledge, none of the previous studies have investigated this potential, which is widely recognized as critical to clinical decision-making [49]. Furthermore, AI models have the ability to improve the current decision making based on HPV and risk management categories [56–58].

Only a minority of patients (5.5 %) in the training dataset were vaccinated. This limitation might reduce the generalizability of our models in a population where vaccination against cervical HPV uptake is high. However, recent data from WHO/UNICEF Joint Reporting Form on Immunization shows that, in 2022, the vaccination program coverage is low globally (15 %), with the highest reported in European region and Region of Americas (52 %) [59]. This report highlight the need for increasing screening coverage and advanced tools for screening for women who are still unvaccinated. Moreover, among vaccinated women, as the prevalence of the disease decreases, the performance (PPV and sensitivity) of cytology and colposcopy is expected to decrease. The model based on historical data should be assessed for the residual risk of CIN2+ and may have the potential to triage for colposcopy women screened with persistent non-HPV vaccine types positive.

Additional methodologic strengths of our study include: i) homogeneous dataset with quality control and reduced variability in colposcopy images and biopsy procedures, increasing the internal validity of the models; ii) unlike most of the previous studies, we use the gold-standard diagnostic tool of cone biopsy or LEEP- conization specimen and pathological examination as our ground truth; iii) we also integrate multimodal data (clinical history and colposcopy images); iv) majority(74.6 %) of the data is for woman who undergoing their abnormal primary cervical cancer screening (with no prior history of HPV positive testing, abnormal cytology, LSIL or prior treatment for low and high grade lesions).

We recognized the need for external prospective validation of the model in different settings (e.g., images from different colposcopes) and data sources (e.g., diverse populations). It is recommended that external validation is done with a dataset that was not available at the time of model and by a different research group [60]. Moreover, recent reports highlight the illusion of a truly externally validated prediction model [61]. The authors highlighted that due to shift in patient population, measurement procedures heterogeneity in model performance is expected, and prediction models are never truly externally validated. Instead, any model must be updated and recalibrated to each use case setting. Furthermore, importance of reporting development of a prediction model was further highlighted [62].

Nevertheless, estimating the external performance of a model, from the available data is important; this process is known as internal-external validation [60]. Here we report the results of internal-external validation using 5 fold cross validation approach, an essential step in the development of a model to estimate how the model would perform in a new sample which is not too far from the training data.

Our data set only included participants who had a cone-biopsy or LEEP-conization after a punch biopsy of high-grade CIN mainly with TZ1 or TZ2 accessible to punch biopsy, and for inaccessible TZ2 and TZ3 cases as a diagnostic procedure. Thus, we may have missed a minority of patients for whom a cone-biopsy or LEEP conization was not performed after an abnormal screening result. This group consists of women with no prior history of abnormal screening, with the latest screening visit being normal and the TZ zone being fully visible and normal in colposcopy. In our experience, missing significant disease in such a situation is a very rare. Moreover, the likelihood of missing a hidden lesion, such as a deep glandular lesion not yet visible on colposcopy, is also very low. Furthermore, our model incorporates several years of patient history, so these sporadic cases, which are usually detected in subsequent screenings, are captured in the follow-up screening visits and thus would have been included in our dataset. It should be noted that to develop high-grade CIN or a glandular lesion, HPV must have been positive for at least 8–10 years prior, and in such cases, historical data would take this into account. However, our care point of interest is for women who had an abnormal screening visit and are referred to colposcopy, hence the above scenario is out of scope of our target population.

Due to limited event rates, we were not able to train models for more fine-grained categories (Normal vs CIN1 vs CIN2 vs CIN3+) —the same limitation applied to the diagnosis of adenocarcinoma in-situ and invasive cervical cancer due to the low number rarity of these events cases in our dataset.

While the performance of our models in predicting CIN2+ was lower among sub-group of women with TZ3 compared to women with TZ1 or TZ2; this behavior of the model is expected as the prevalence of TZ3 type in our dataset is less than 5 %. It is important to note that for women with TZ3 type a clinician may not be able to predict or exclude CIN2+ and may opt for endocervical curettage and/or cone-biopsy, following a persistent abnormal screening test. Interestingly, even among these women our model showed acceptable PPV and NPV, hence highlighting the strong potential for clinical utility.

Although we compared the performance of the model against the clinical expert impression, the model must be compared to a representative group of clinicians in different settings.

6. Conclusion

In conclusion, this study contributes significantly to the growing field of AI-assisted diagnostic aids for cervical precancers and cancer. We demonstrate the potential clinical utility of deep learning models in reducing unnecessary conization or biopsy, a critical concern in cervical neoplasia management. Our model, utilizing gold-standard histopathological results as ground truth, displayed commendable accuracy and specificity, offering valuable support for CIN2+ prediction in real time and in clinical decision-making, particularly in resource-limited settings. However, external prospective validation across diverse clinical settings and populations is warranted. Future research should focus on expanding the models' diagnostic scope, particularly for finer categorization of cervical precancers and invasive cancers and comparing their performance against a wider range of clinical practitioners.

CRediT authorship contribution statement

Sreenath Madathil: Writing – review & editing, Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Mohamed Dhouib:** Writing – review & editing, Validation, Software, Resources, Methodology. **Quitterie Lelong:** Writing – review & editing, Validation, Software, Resources, Methodology. **Ahmed Bourassine:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation. **Joseph Monsonego:** Writing – review & editing, Supervision, Project administration, Methodology, Data curation, Conceptualization.

Ethical approval and consent to participate

All participants signed an informed consent form, and the research protocol was approved by the institutional research ethics board of McGill University, Montreal, Canada[A01-E05-24B]. Furthermore, the study was performed in accordance with the Declaration of Helsinki.

Data availability statement

Due to the possibility of potentially identifiable personal medical data used for the project the training or testing dataset is not made available at this stage.

Funding information

The author(s) received no specific funding for this work.

Declaration of competing interest

The authors declare no conflicts of interest.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2025.109710.

References

[1] H. Sung, J. Ferlay, R.L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, et al., Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA Cancer J Clin 71 (2021) 209–249.

[2] M. Arbyn, E. Weiderpass, L. Bruni, S. de Sanjosé, M. Saraiya, J. Ferlay, et al., Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis, Lancet Global Health 8 (2020) e191–e203.

[3] S. Lin, K. Gao, S. Gu, L. You, S. Qian, M. Tang, et al., Worldwide trends in cervical cancer incidence and mortality, with predictions for the next 15 years, Cancer 127 (2021) 4030–4039.

[4] W.Q. He, C. Li, Recent global burden of cervical cancer incidence and mortality, predictors, and temporal trends, Gynecol. Oncol. 163 (2021) 583–592.

[5] A. Shiraz, R. Crawford, N. Egawa, H. Griffin, J. Doorbar, The early detection of cervical cancer. The current and changing landscape of cervical disease detection, Cytopathology 31 (2020) 258–270.

[6] C.R. Chao, J. Chubak, E.F. Beaber, A. Kamineni, C. Mao, M.J. Silverberg, et al., Gaps in the screening process for women diagnosed with cervical cancer in four diverse US health care settings, Cancer Med. 12 (2022) 3705–3717.

[7] J. Wang, H. Edvardsson, B. Strander, B. Andrae, P. Sparén, J. Dillner, Long-term follow-up of cervical cancer incidence after normal cytological findings, Int. J. Cancer 154 (2024) 448–453.

[8] D. Saslow, D. Solomon, H.W. Lawson, M. Killackey, S.L. Kulasingam, J.M. Cain, et al., American cancer society, American society for colposcopy and cervical pathology, and American society for clinical pathology screening guidelines for the prevention and early detection of cervical cancer, J. Low. Genit. Tract Dis. 16 (2012) 175–204.

[9] J. Jeronimo, M. Schiffman, Colposcopy at a crossroads, Am. J. Obstet. Gynecol. 195 (2006) 349–353.

[10] R.G. Pretorius, W.H. Zhang, J.L. Belinson, M.N. Huang, L.Y. Wu, X. Zhang, et al., Colposcopically directed biopsy, random cervical biopsy, and endocervical curettage in the diagnosis of cervical intraepithelial neoplasia II or worse, Am. J. Obstet. Gynecol. 191 (2004) 430–434.

[11] D.G. Ferris, M.S. Litaker, Prediction of cervical histologic results using an abbreviated Reid Colposcopic Index during ALTS, Am. J. Obstet. Gynecol. 194 (2006) 704–710.

[12] W.P. Soutter, Advances in the imaging and detection of cervical intra-epithelial neoplasia, Future Oncol. 5 (2009) 371–378.

[13] L.S. Massad, J. Jeronimo, H.A. Katki, M. Schiffman, National Institutes of Health/American Society for Colposcopy and Cervical Pathology Research Group, The accuracy of colposcopic grading for detection of high-grade cervical intraepithelial neoplasia, J. Low. Genit. Tract Dis. 13 (2009) 137–144.

[14] B.H. Brown, J.A. Tidy, The diagnostic accuracy of colposcopy – a review of research methodology and impact on the outcomes of quality assurance, Eur. J. Obstet. Gynecol. Reprod. Biol. 240 (2019) 182–186.

[15] M. Sideri, P. Garutti, S. Costa, P. Cristiani, P. Schincaglia, P. Sassoli de Bianchi, et al., Accuracy of colposcopically directed biopsy: results from an online quality assurance programme for colposcopy in a population-based cervical screening setting in Italy, BioMed Res. Int. 2015 (2015) 614035.

[16] M. Underwood, M. Arbyn, W. Parry-Smith, S. De Bellis-Ayres, R. Todd, C. Redman, et al., Accuracy of colposcopy-directed punch biopsies: a systematic review and meta-analysis, BJOG An Int. J. Obstet. Gynaecol. 119 (2012) 1293–1301.

[17] M.H. Stoler, M. Schiffman, Atypical squamous cells of undetermined significance-low-grade squamous intraepithelial lesion triage study (ALTS) group. Interobserver reproducibility of cervical cytologic and histologic interpretations: realistic estimates from the ASCUS-LSIL triage study, JAMA 285 (2001) 1500–1505.

[18] P.E. Castle, M.H. Stoler, D. Solomon, M. Schiffman, The relationship of community biopsy-diagnosed cervical intraepithelial neoplasia grade 2 to the quality control pathology-reviewed diagnoses: an ALTS report, Am. J. Clin. Pathol. 127 (2007) 805–815.

[19] D.T. Howe, A.C. Vincenti, Is large loop excision of the transformation zone (LLETZ) more accurate than colposcopically directed punch biopsy in the diagnosis of cervical intraepithelial neoplasia? Br. J. Obstet. Gynaecol. 98 (1991) 588–591.

[20] C. Zuchna, M. Hager, B. Tringler, A. Georgoulopoulos, A. Ciresa-Koenig, B. Volgger, et al., Diagnostic accuracy of guided cervical biopsies: a prospective multicenter study comparing the histopathology of simultaneous biopsy and cone specimen, Am. J. Obstet. Gynecol. 203 (2010) 321.e1–321.e6.

[21] M. Cárdenas-Turanzas, M. Follen, J.L. Benedet, S.B. Cantor, See-and-treat strategy for diagnosis and management of cervical squamous intraepithelial lesions, Lancet Oncol. 6 (2005) 43–50.

[22] M. Rebolj, J. Rimmer, K. Denton, J. Tidy, C. Mathews, K. Ellis, et al., Primary cervical screening with high risk human papillomavirus testing: observational study, BMJ 364 (2019) l240.

[23] M.A. Smith, M. Sherrah, F. Sultana, P.E. Castle, M. Arbyn, D. Gertig, et al., National experience in the first two years of primary human papillomavirus (HPV) cervical screening in an HPV vaccinated population in Australia: observational study, BMJ 376 (2022) e068582.

[24] V.M. Partanen, J. Dillner, A. Tropé, Á.I. Ágústsson, S. Lönnberg, S. Heinävaara, et al., Divergent effects of switching from cytology to HPV-based screening in the Nordic countries, Eur. J. Publ. Health (2024) ckad225.

[25] J. Wang, M. Elfström, J. Dillner, A randomized healthcare policy trial of human papillomavirus-based cervical screening [Internet]. Rochester, NY; Available from: https://papers.ssrn.com/abstract=4845172, 2024. (Accessed 9 June 2024).

[26] N. Phoolcharoen, M.L. Varon, E. Baker, S. Parra, J. Carns, K. Cherry, et al., Hands-on training courses for cervical cancer screening, diagnosis, and treatment procedures in low- and middle-income countries, JCO Glob Oncol 8 (2022) e2100214.

[27] N. Phoolcharoen, M. Kremzier, V. Eaton, V. Sarchet, S.C. Acharya, E. Shrestha, et al., American society of clinical oncology (asco) cervical cancer prevention program: a hands-on training course in Nepal, JCO Glob Oncol 7 (2021) 204–209.

[28] B.A. Grema, I. Aliyu, G.C. Michael, M.B. Mafala, Diagnosing premalignant lesions of uterine cervix in A ResourceConstraint setting: a narrative review, W. Afr. J. Med. 36 (2019) 48–53.

[29] A.H. Rossman, H.W. Reid, M.M. Pieters, C. Mizelle, M. von Isenburg, N. Ramanujam, et al., Digital health strategies for cervical cancer control in low- and middle-income countries: systematic review of current implementations and gaps in research, J. Med. Internet Res. 23 (2021) e23350.

[30] X. Hou, G. Shen, L. Zhou, Y. Li, T. Wang, X. Ma, Artificial intelligence in cervical cancer screening and diagnosis, Front. Oncol. 12 (2022) 851367.

[31] S. Kim, H. Lee, S. Lee, J.Y. Song, J.K. Lee, N.W. Lee, Role of artificial intelligence interpretation of colposcopic images in cervical cancer screening, Healthc Basel Switz 10 (2022) 468.

[32] N. Nazir, B.S. Saini, A. Sarwar, Early diagnosis of cervical cancer using AI: a review, in: Y. Singh, P.K. Singh, M.H. Kolekar, A.K. Kar, P.J.S. Gonçalves (Eds.), Proceedings of International Conference on Recent Innovations in Computing, Springer Nature, Singapore, 2023, pp. 105–116 (Lecture Notes in Electrical Engineering).

[33] A.A. Swanson, L. Pantanowitz, The evolution of cervical cancer screening, J. Am. Soc. Cytopathol. 13 (1) (2024 Jan-Feb) 10–15, https://doi.org/10.1016/j.jasc.2023.09.007.

[34] J.J.M. Kowsigan, A comprehensive assessment of recent advances in cervical cancer detection for automated screening, in: Image Processing and Intelligent Computing Systems, CRC Press, 2023.

[35] P. Xue, J. Wang, D. Qin, H. Yan, Y. Qu, S. Seery, et al., Deep learning in image-based breast and cervical cancer detection: a systematic review and meta-analysis, Npj Digit Med 5 (2022) 1–15.

[36] Évaluation de la recherche des papillomavirus humains (HPV) en dépistage primaire des lésions précancéreuses et cancéreuses du col de l'utérus et de la place du double immuno-marquage p16/Ki67 [Internet]. Haute Autorité de Santé. [cited 2023 December 2]. Available from: https://www.has-sante.fr/jcms/c_2806160/fr/evaluation-de-la-recherche-des-papillomavirus-humains-hpv-en-depistage-primaire-des-lesions-precancereuses-et-cancereuses-du-col-de-l-uterus-et-de-la-place-du-double-immuno-marquage-p16/ki67..

[37] Z. Liu, H. Mao, C.Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A ConvNet for the 2020s, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11966–11976.

[38] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, et al., Attention is all you need, in: Advances in Neural Infor- Mation Processing Systems, 2017, pp. 6000–6010.

[39] A.R. Localio, S. Goodman, Beyond the usual prediction accuracy metrics: reporting results for clinical decision making, Ann. Intern. Med. 157 (2012) 294–295.

[40] E.W. Steyerberg, A.J. Vickers, N.R. Cook, T. Gerds, M. Gonen, N. Obuchowski, Assessing the performance of prediction models: a framework for traditional and novel measures, Epidemiol. Camb Mass 21 (2010) 128–138.

[41] M. Fitzgerald, B.R. Saville, R.J. Lewis, Decision curve analysis, JAMA 313 (2015) 409–410.

[42] K.F. Kerr, M.D. Brown, K. Zhu, H. Janes, Assessing the clinical impact of risk prediction models with decision curves: guidance for correct interpretation and appropriate use, J. Clin. Oncol. 34 (2016) 2534–2540.

[43] A.J. Vickers, B. van Calster, E.W. Steyerberg, A simple, step-by-step guide to interpreting decision curve analysis, Diagn Progn Res. 3 (2019) 18.

[44] H. Fe, C. Rm, P. Db, L. Kl, R. Ra, Evaluating the yield of medical tests, JAMA (1982) 247 [Internet]. Available from: https://pubmed.ncbi.nlm.nih.gov/7069920/. (Accessed 4 October 2024).

[45] F. Harrell, Statistical thinking - statistically efficient ways to quantify added predictive value of new measurements [Internet]. Available from: https://www.fharrell.com/post/addvalue/, 2018. (Accessed 21 September 2023).

[46] Jr.FE. Harrell, K.L. Lee, R.M. Califf, D.B. Pryor, R.A. Rosati, Regression modelling strategies for improved prognostic prediction, Stat. Med. 3 (1984) 143–152.

[47] IARC cervical cancer Image Bank [Internet]. Available from: https://screening.iarc.fr/cervicalimagebank.php. (Accessed 9 June 2024).

[48] M. Bokil, B. Lim, Colposcopy: a closer look into its past, present and future, BJOG An Int. J. Obstet. Gynaecol. 126 (2019), 543–543.

[49] H.D. Vargas-Cardona, M. Rodriguez-Lopez, M. Arrivillaga, C. Vergara-Sanchez, J. P. García-Cifuentes, P.C. Bermúdez, et al., Artificial intelligence for cervical cancer screening: scoping review, 2009-2022, Int. J. Gynaecol. Obstet Off Organ Int. Fed . Gynaecol. Obstet. 165 (2) (2024) 566–578.

[50] D. Egemen, R.B. Perkins, L.C. Cheung, B. Befano, A.C. Rodriguez, K. Desai, et al., Artificial intelligence–based image analysis in clinical testing: Lessons from cervical cancer screening 116, JNCI J Natl Cancer Inst, 2024, pp. 26–33.
[51] P. Sasieni, P.E. Castle, J. Cuzick, Further analysis of the ARTISTIC trial, Lancet Oncol. 10 (2009) 841–842.
[52] H.C. Kitchener, M. Almonte, C. Thomson, P. Wheeler, A. Sargent, B. Stoykova, et al., HPV testing in combination with liquid-based cytology in primary cervical screening (ARTISTIC): a randomised controlled trial, Lancet Oncol. 10 (2009) 672–682.
[53] G. Ronco, J. Dillner, K.M. Elfström, S. Tunesi, P.J.F. Snijders, M. Arbyn, et al., Efficacy of HPV-based screening for prevention of invasive cervical cancer: follow-up of four European randomised controlled trials, Lancet Lond. Engl. 383 (2014) 524–532.
[54] E. Alfonzo, C. Zhang, F. Daneshpip, B. Strander, Accuracy of colposcopy in the Swedish screening program, Acta Obstet. Gynecol. Scand. 102 (2023) 549–555.
[55] M.F. Mitchell, D. Schottenfeld, G. Tortolero-Luna, S.B. Cantor, R. Richards-Kortum, Colposcopy for the diagnosis of squamous intraepithelial lesions: a meta-analysis, Obstet. Gynecol. 91 (1998) 626–631.
[56] J. Monsonego, L. Zerat, K. Syrjänen, J.C. Zerat, J.S. Smith, P. Halfon, Prevalence of type-specific human papillomavirus infection among women in France: implications for screening, vaccination, and a future generation of multivalent HPV vaccines, Vaccine 30 (2012) 5215–5221.
[57] J. Monsonego, J.T. Cox, C. Behrens, M. Sandri, E.L. Franco, P.S. Yap, et al., Prevalence of high-risk human papilloma virus genotypes and associated risk of cervical precancerous lesions in a large U.S. screening population: data from the ATHENA trial, Gynecol. Oncol. 137 (2015) 47–54.
[58] D. Egemen, L.C. Cheung, X. Chen, M. Demarco, R.B. Perkins, W. Kinney, et al., Risk estimates supporting the 2019 ASCCP risk-based management consensus guidelines, J. Low. Genit. Tract Dis. 24 (2020) 132–143.
[59] WHO immunization data portal [Internet]. Available from: https://immunizationdata.who.int/index.html. (Accessed 11 January 2024).
[60] E.W. Steyerberg, F.E. Harrell, Prediction models need appropriate internal, internal–external, and external validation, J. Clin. Epidemiol. 69 (2016) 245–247.
[61] B. Van Calster, E.W. Steyerberg, L. Wynants, M. van Smeden, There is no such thing as a validated prediction model, BMC Med. 21 (2023) 70.
[62] H.M. la Roi-Teeuw, F.S. van Royen, A. de Hond, A. Zahra, S. de Vries, R. Bartels, et al., Don't be misled: three misconceptions about external validation of clinical prediction models, J. Clin. Epidemiol. (2024) 111387.