JNCI J Natl Cancer Inst (2019) 0(0): djy225

doi: 10.1093/jnci/djy225
Article



ARTICLE

# An Observational Study of Deep Learning and Automated Evaluation of Cervical Images for Cancer Screening

Liming Hu, David Bell, Sameer Antani, Zhiyun Xue, Kai Yu, Matthew P. Horning, Noni Gachuhi, Benjamin Wilson, Mayoore S. Jaiswal, Brian Befano, L. Rodney Long, Rolando Herrero, Mark H. Einstein, Robert D. Burk, Maria Demarco, Julia C. Gage, Ana Cecilia Rodriguez, Nicolas Wentzensen, Mark Schiffman

See the Notes section for the full list of authors' affiliations.
**Correspondence to:** Mark Schiffman, MD, MPH, National Cancer Institute, Room 6E544, 9609 Medical Center Drive, Rockville, MD 20850 (e-mail: schiffmm@mail.nih.gov).

## Abstract

**Background:** Human papillomavirus vaccination and cervical screening are lacking in most lower resource settings, where approximately 80% of more than 500 000 cancer cases occur annually. Visual inspection of the cervix following acetic acid application is practical but not reproducible or accurate. The objective of this study was to develop a "deep learning"-based visual evaluation algorithm that automatically recognizes cervical precancer/cancer.
**Methods:** A population-based longitudinal cohort of 9406 women ages 18–94 years in Guanacaste, Costa Rica was followed for 7 years (1993–2000), incorporating multiple cervical screening methods and histopathologic confirmation of precancers. Tumor registry linkage identified cancers up to 18 years. Archived, digitized cervical images from screening, taken with a fixed-focus camera ("cervicography"), were used for training/validation of the deep learning-based algorithm. The resultant image prediction score (0–1) could be categorized to balance sensitivity and specificity for detection of precancer/cancer. All statistical tests were two-sided.
**Results:** Automated visual evaluation of enrollment cervigrams identified cumulative precancer/cancer cases with greater accuracy (area under the curve [AUC] = 0.91, 95% confidence interval [CI] = 0.89 to 0.93) than original cervigram interpretation (AUC = 0.69, 95% CI = 0.63 to 0.74; $P < .001$) or conventional cytology (AUC = 0.71, 95% CI = 0.65 to 0.77; $P < .001$). A single visual screening round restricted to women at the prime screening ages of 25–49 years could identify 127 (55.7%) of 228 precancers (cervical intraepithelial neoplasia 2/cervical intraepithelial neoplasia 3/adenocarcinoma in situ [AIS]) diagnosed cumulatively in the entire adult population (ages 18–94 years) while referring 11.0% for management.
**Conclusions:** The results support consideration of automated visual evaluation of cervical images from contemporary digital cameras. If achieved, this might permit dissemination of effective point-of-care cervical screening.

Cervical cancer remains a leading cause of cancer mortality and morbidity worldwide (1). Approximately 80% of the half-million cases and 90% of the quarter-million deaths per year occur in low- and middle-income countries, where prevention programs are limited. In some low-resource countries, cervical cancer is the leading female malignancy, with lifetime cumulative incidence exceeding 5%. The number of new cases is projected to increase in the decades ahead as the global population grows and ages (2).

Cervical cancer arises from persistent infection of the cervix with approximately a dozen carcinogenic types of human papillomavirus (HPV) (3). We can prevent it by prophylactic vaccination and screening/treatment of cervical cancer precursor lesions ("precancer"). However, mainstays of cervical cancer

**Received:** August 21, 2018; **Revised:** October 12, 2018; **Accepted:** December 3, 2018

screening programs in high-resource settings, including cervical cytology (Pap tests) and colposcopy, require infrastructure and extensively trained personnel that are lacking in most lower resource settings. Newer, powerful cervical cancer prevention tools like prophylactic HPV vaccination and screening with sensitive HPV tests could be very useful in low-resource settings. However, again, dissemination has been severely limited by lack of resources and organization.

In search of programmatic simplicity and sustainable costs, authorities including the World Health Organization, the US President's Emergency Plan for AIDS Relief, and the Indian government have endorsed screening of the cervix by visual inspection after application (VIA) of acetic acid to highlight precancerous or cancerous abnormalities (4–6) when more advanced methods are not feasible. The health worker performing VIA rates the cervical appearance as normal or abnormal with particular attention to possible invasive cancer. If the appearance is abnormal and if the size and position of the visible lesion permit, the cervical epithelium is destroyed by freezing or heating ("see and treat"). Excision is sometimes required for severe lesions. As a screening test, VIA is simple and inexpensive and has been shown to find some invasive cancers while they are still local and curable by surgery (7). However, the main goal of screening is to prevent cancer by detection and treatment of precancers, and VIA is not accurate in distinguishing precancer from much more common minor abnormalities, leading to both overtreatment and undertreatment (8–10). Increasingly, it is recognized that the visual identification of precancer by health workers, even by experienced nurses and doctors using a colposcope, the reference standard visual tool, is too often unreliable and inaccurate (11,12). Thus, we currently still lack a practical, accurate visual screening approach.

In other similarly subjective medical diagnostic situations, new methods of pattern recognition by computer (referred to as deep or machine learning) have proven useful (13). Machine learning-based approaches to cervical cancer screening have yielded promising early results but have lacked a good measurement of precancer, sufficient sample size, or prospective follow-up (Supplementary Table 1, available online) (14,15).

We applied a deep learning-based object detection method [Faster R-CNN, or faster region-based convolutional neural network (16)] algorithm to cervical images taken during a National Cancer Institute (NCI) prospective epidemiologic study, with long follow-up and rigorously defined precancer endpoints, to develop a detection algorithm that can identify cervical precancer. Here, we demonstrate the proof-of-principle of an "automated visual evaluation" algorithm applied to archived, digitized cervical images.

## Materials and Methods

### Study Population

Conducted from 1993 to 2001, the NCI-funded Proyecto Epidemiologico Guanacaste accumulated approximately 30 000 screening visits of more than 9000 participants (93.3% acceptance) (**Figure 1**). This longitudinal cohort study of HPV and cervical cancer was conducted in a random sample of a moderate-risk, poorly screened Costa Rican province (17,18). Participants aged 18–94 years were screened at baseline and periodically for up to 7 years using multiple methods at intervals determined by their screening results and resultant estimated risk of precancer (mean number of visits = 3.4, SD = 2.5, approximately 90% of women with some follow-up) (17,18). In a subsequent linkage study, the cancer registry was used to extend follow-up for invasive cancer up to 18 years (18,19).

### Cervicography

Each screening visit included multiple kinds of tests: cytology, HPV testing, and cervicography (18). Cervicography, now discontinued, was a visual screening method based on the interpretation of a pair of cervical photographs (called cervigrams) (20,21). Two sequential, duplicate cervigrams of the cervix were taken at each screening visit after acetic acid application using a fixed-focus, ring-lit film camera called a cerviscope. The rolls of film were mailed regularly and developed into projection slides in the United States by National Testing Laboratory Worldwide (Fenton, MO). Each cervigram image was originally projected on a wall screen for magnification and classified by one of two highly experienced National Testing Laboratory Worldwide physician colposcopist evaluators as normal, atypical, positive for minor low-grade HPV-induced changes, or positive for precancer or cancer (the last two combined here). After completion of the field effort, the photographic images were digitized, and the files compressed for storage (17,22).

### Other Screening Tests in the Guanacaste Cohort

Cytology methods included conventional Pap smears, a prototype liquid-based method (23), and a first-generation automated approach incorporating an early version of a neural network-based method (24). The conventional Pap smear performed in Costa Rica best represents the kind typically performed in lower- and middle-resource regions. HPV testing was performed by MY09-MY11 consensus primer PCR (25) but was not used for colposcopy referral in this early epidemiologic demonstration of HPV predictive value. We defined positivity as detection of one of 13 high-risk HPV per the International Agency For Research On Cancer classification (26).

### CIN2+ Cases and Controls

Cases included women who were diagnosed with histologic CIN2 or worse (CIN2+) during enrollment or follow-up. Women found at a cohort screening examination to have abnormal cytology or cervicography screening results were referred to colposcopy performed by a single study gynecologist who biopsied the worst-appearing lesion. Histologic CIN2 or worse (CIN2+) was treated by large loop excision of the transformation zone; all women with definite or even possible CIN2+ (eg, high-grade squamous intraepithelial lesion cytology) were referred for management and censored from further study. The histopathologic reference diagnosis of CIN2+ was determined by majority review of biopsies and excision specimens by a panel of one Costa Rica pathologist and one US consultant pathologist, with disagreements leading to another US pathologist's review.

Collection and use of the visual images were approved by the Costa Rican and NCI ethical committees. The images were collected originally under written informed consent that covered subsequent research use. The specific use for machine learning-based algorithms on study images was also approved by the NCI Institutional Review Board.

ARTICLE

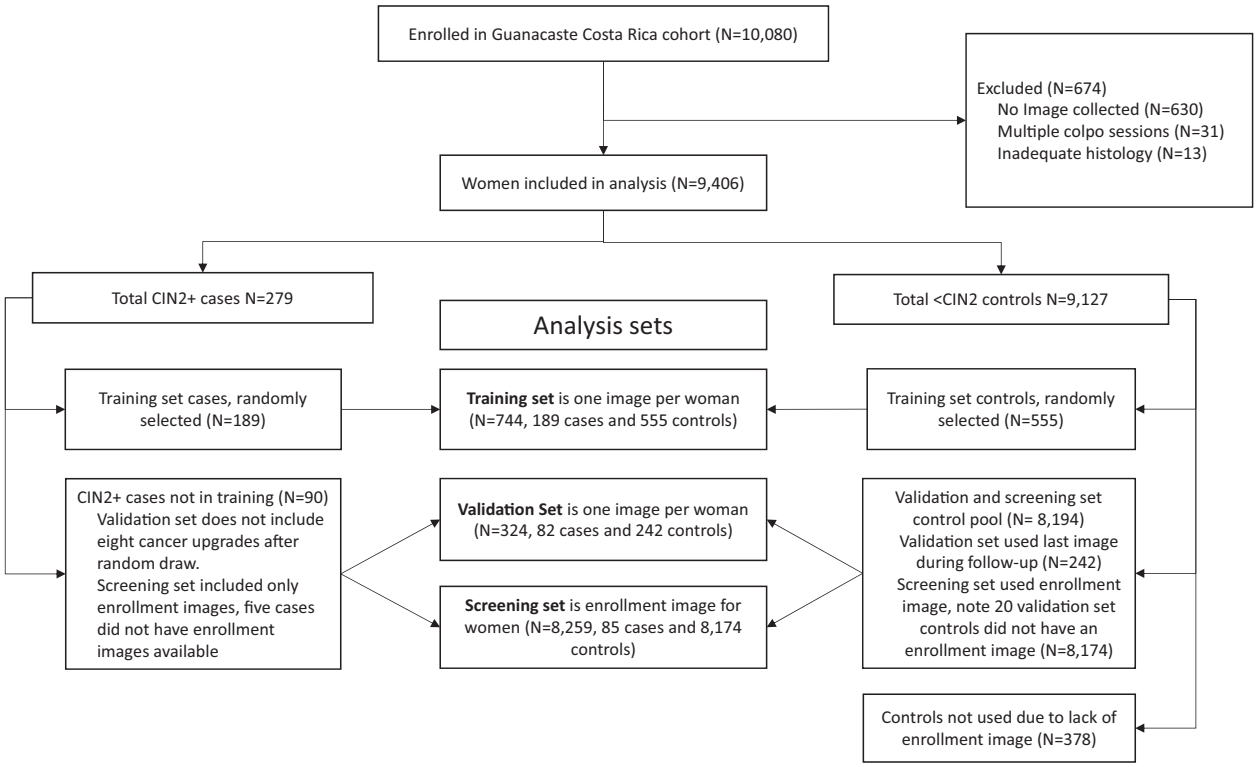


**Figure 1.** Cervical images used for training and validation. The images were drawn from the Proyecto Epidemiologico Guanacaste, a longitudinal cohort study of human papillomavirus infection, other screening tests, and risk of cervical precancer/cancer (1993–2001). The training and initial validation made use of the last images taken prior to case diagnosis and approximately 3:1 controls frequency matched to cases on time of study (enrollment/follow-up). The main analysis focused on images (excluding all images from women in the training set) from cohort enrollment and examined how the automated image analysis of enrollment screening images performed in prediction of cases found over the course of the entire cohort study. In an analysis that counted the absolute numbers of cases detected and controls referred, it was necessary to reweight, that is, to multiply the findings in the randomly selected validation test by the inverse of the 30% sampling fraction to estimate numbers for all cases and matched controls (see Methods). CIN = cervical intraepithelial neoplasia.

## Automated Visual Evaluation Algorithm to Evaluate Cervical Images

Three controls per case were chosen from among women who did not present with CIN2+ during active surveillance, frequency matched to time of diagnosis of case (enrollment/follow-up). A random, approximately 70% (197/279) of cases and frequency-matched controls were chosen for training, and the remaining approximately 30% were reserved as the initial validation test set. We randomly chose a single image from each pair of images available per prediagnostic case visit and control visit. Of note, after the initial validation (Supplementary Methods, available online), the main analysis focused on images taken at enrollment visits.

The system architecture is summarized in Figure 2 and detailed in the Supplementary Methods (available online). The automated visual evaluation detection algorithm performs two main functions: detecting (locating) the cervix within the input image and predicting the probability that the image represents a case of CIN2+. The algorithm based on Faster R-CNN performs object (cervix) detection, feature extraction (calculating the features of the object), and classification (predicting the case probability score). We explored several different techniques, as described in the Supplementary Methods (available online), and Faster R-CNN algorithm provided the best combination of speed and accuracy (27). The inputs to train the model were the cervigram images, cervix location (rectangle encompassing the cervix), and the class ground truth (case of CIN2+ or control <CIN2).

To create the cervix locator function of the automated visual evaluation algorithm, we first trained an independent cervix locator algorithm using Faster R-CNN to separate the cervix from the background. This involved annotating 2000 images (manually drawing a rectangle around the cervix) using a software tool, shown in Supplementary Figure 2 (available online), developed for this task. By incorporating the trained locator function, the only input required for the automated visual evaluation algorithm was the cervigram image. The final model provided both the predicted cervix location and the case probability.

CNN-based methods are data driven and require large quantities of training data to perform well; however, our dataset from a population cohort study was limited to the smaller number of cases that actually occurred. To compensate, we performed transfer learning (28) by first initializing the CNN architecture with pretrained weights from a model trained with the ImageNet dataset (29), a database of millions of (noncervigram) images of all kinds of natural objects. We then retrained with the cervigram images in the training set. We also augmented the image data (30), artificially increasing the number of cervical images, by performing minor distortions to the original images including rotation, mirroring, sheering, and gamma transformation. We used the faster R-CNN end-to-end training configuration of stochastic gradient descent optimization with the parameters in Supplementary Table 2 (available online).

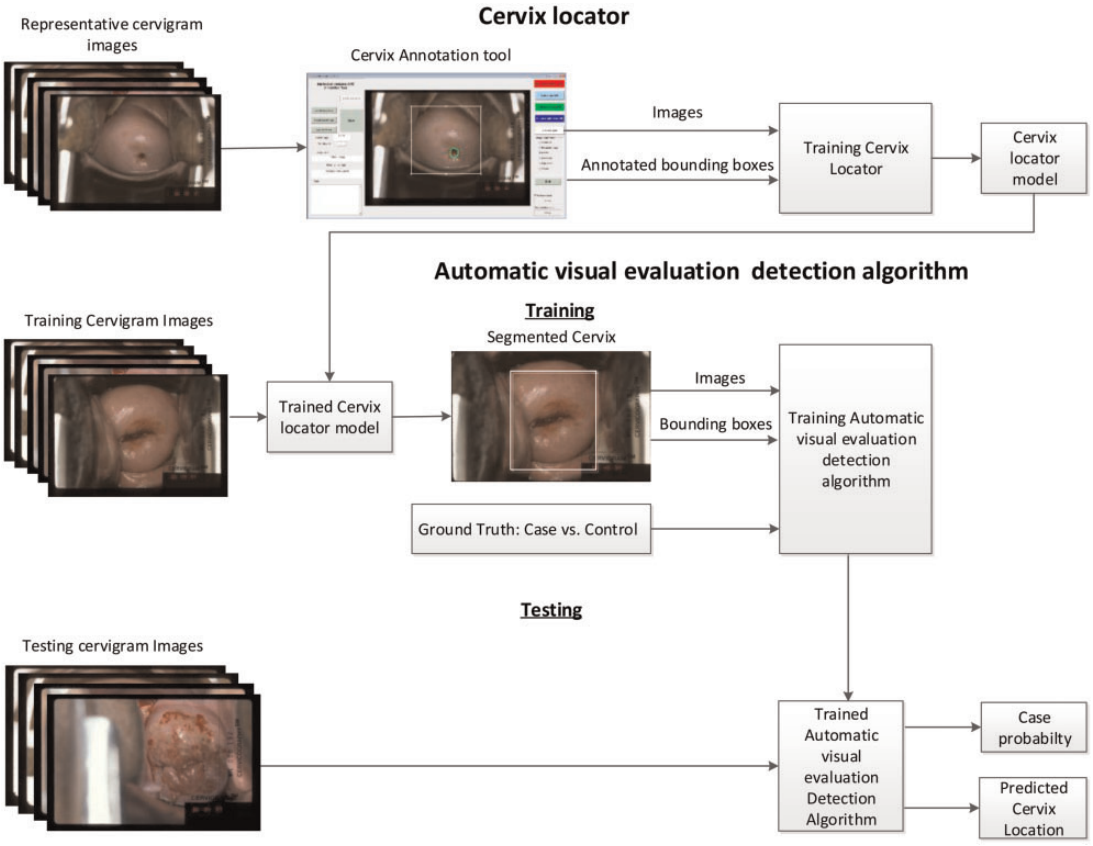


**Figure 2.** The system architecture of the automated visual evaluation algorithm. Two models are trained: a cervix locator (**top**), and the automated visual evaluation detection algorithm (**bottom**). The final validation algorithm incorporated both cervix locator and automated visual evaluation.

These techniques are further described in the Supplementary Methods (available online).

National Library of Medicine (National Institutes of Health, USA) collaborators independently checked the technique. This series of confirmatory experiments was not designed to stand alone; rather, it was meant to increase credibility of the original findings by replication outside of the inventing group. They first followed the architecture of the proposed method to evaluate the performance of the cervix region locator and of the visual evaluation algorithm, independently evaluating different subsets of the Guanacaste cohort images for model training than the original. They also tested the trained network with test images that were altered to represent a different camera zoom and moderate image resizing.

To visualize what the algorithm bases its prediction on, we also implemented a system that generates a heat map showing which regions of the image most strongly influence the prediction. This system is described in Supplementary Methods (available online).

## Statistical Analyses

The automated visual evaluation detection algorithm yields a score (ranging from 0.0 to 1.0) that predicts whether the image represents a case of histologic CIN2+. Examining first the reproducibility of the scores, we compared 322 replicate pairs of images included for this purpose in the validation set; the Pearson correlation within pairs was 0.97 (95% confidence interval [CI] = 0.97 to 0.98); the very high agreement justified using only one image per pair for subsequent analyses.

We conducted the analyses presented here using cohort enrollment images from the 30% of women in the initial validation set plus enrollment images from the "leftover" women in the Guanacaste cohort with <CIN2 not chosen for either the training or initial validation set (Figure 1). The main validation analysis, focused on images taken at cohort enrollment visits, evaluated accuracy of the automated visual evaluation algorithm compared with the originally performed baseline screening tests (cervicography, cytology) and HPV testing introduced for validation but not used as a screening test at that time.

We evaluated the accuracy of the automated visual evaluation detection algorithm for identification of CIN2+ cases in the validation set using Receiver Operating Characteristic (ROC) curves and its summary statistic, area under the curve (AUC). We created ROC-like curves for the ordinal categories of the original screening tests and a ROC curve of the continuous automated visual evaluation empirically by the trapezoid method. These AUC values were tested for statistical significance against the AUC for automated visual evaluation by two-sided chi-squared tests (31). For analyses requiring a categorical positive/negative result for automated visual evaluation, we chose the cutpoint in the continuous score distribution, specific to that age group in age-stratified analyses, that maximized Youden's index (sensitivity + specificity – 1) (32).

In a separate analysis to estimate the impact of a single screening round for the full Guanacaste cohort, we filled in for images from women used in the training set and therefore

excluded from the validation of the algorithm. To do so, we reweighted; specifically, because the initial validation set was a random sample of the training validation group, we could multiply the results for women in the validation set by the inverse of the sampling fraction (30%) to represent all women used for training and validation. We then added the remaining noncases to reconstitute the full cohort.

Cases that resulted in differences between automated visual evaluation algorithm results and original evaluator interpretation of the same cervigram images were rereviewed without masking by an expert gynecologic oncologist and colposcopist (ME) to note any subjective patterns that might explain discrepancies.

## Results

The cohort study population enrolled in 1993–1994 (Table 1) was a random sample of adult women in Guanacaste Province, ranging in age from 18 to 94 years (median 35 years). The province was mixed rural and small town. Most women were married (76.1%) and parity was high (92.3). Smoking was unusual (10.7% were current smokers), but oral contraceptive use (present or past) was common (77.7% had ever used oral contraceptives). The prevalence of carcinogenic HPV types was less than 20% and declined sharply with age.

There were 241 histologically confirmed cases of precancer (CIN2/CIN3) and 38 cases of cancer observed among 9406 women followed for up to 7 years in the population-based Guanacaste cohort. Glandular lesions were very uncommon and categorized with squamous lesions of equivalent severity (eg, rare AIS was included with CIN3). The mean automated visual evaluation algorithm result scores at enrollment were equivalent for CIN2 vs CIN3 (0.70 and 0.69, $P = .51$) and statistically nonsignificantly higher for the uncommon cancers (0.80, $P = .42$) (Figure 3). In subsequent analyses, we combined CIN2, CIN3, and cancer in a single case group.

The AUC for automated visual evaluation of the enrollment cervical images, for women of all ages, was 0.91 (95% CI = 0.89 to 0.93) (Figure 4). Automated visual evaluation was statistically significantly more accurate than cervigram result (AUC = 0.69, 95% CI = 0.63 to 0.74; $P < .001$).

As shown in Figure 4, automated visual evaluation performance was also statistically significantly more accurate than conventional Pap smears (AUC = 0.71, 95% CI = 0.65 to 0.77; $P < .001$), liquid-based cytology (AUC = 0.79, 95% CI = 0.73 to 0.84; $P < .001$), first-generation neural network-based cytology (AUC = 0.70; 95% CI = 0.63 to 0.76; $P < .001$), and HPV testing (AUC = 0.82, 95% CI = 0.77 to 0.87; $P < .001$).

We examined all 16 discrepant findings that were algorithm positive/HPV negative. The additional positives by automated visual evaluation tended to occur among younger women (median = 26.5, $P = .06$ by Wilcoxon test compared with the rest of the cases in the cohort whose median age was 35 years). They were likely to be diagnosed with CIN2 (12/16 [75.0%] compared with 20/61 [32.8%] of other cases, $P = .008$).

In anticipation of designing screening programs, we considered three distinct age ranges: 18–24 years, 25–49 years, and 50+ years, with the intermediate age group of primary interest because it coincides with the majority (130/228) of CIN2-CIN3 cases and higher sensitivity (127/130, $P < .001$ compared with younger women) (Table 2). We estimated that a single automated visual evaluation screening round targeting women at the prime screening ages of 25–49 years could identify 55.7% (127/228) of precancers (CIN2/CIN3/AIS) diagnosed cumulatively in the entire adult population. To achieve this level of sensitivity in the entire population by a single round of screening women aged 25–49 years would require referring 11.0% (982/8906) of the entire population (and 18.0% [982/5460] of those aged 25–49 years) for treatment.

Review of enrollment images from cases with discrepant human vs automated visual evaluations (and a random tenth of discrepant noncases) revealed that many of the automated visual evaluation algorithm result-positive/cervigram-negative cases of CIN2+ (additional true positives) had suboptimal images (eg, poor focus or washed-out color during scanning of film), or there were obstructing vaginal folds or blood. No clear patterns emerged from review of images with discrepant results among women with <CIN2.

**Table 1.** Selected demographic features of the 9406 women in this analysis of the Guanacaste cohort, and their enrollment screening results

| Feature or test result | No. % |
|---|---|
| Enrollment age, y | |
| 18–29 | 2598 (27.6) |
| 30–49 | 4357 (46.3) |
| 50–94 | 2451 (26.1) |
| Marital status | |
| Married | 7157 (76.1) |
| Separated/widowed | 508 (10.2) |
| Single | 1290 (13.7) |
| Age at first intercourse, y | |
| ≤16 | 3269 (34.8) |
| 17–18 | 2579 (27.4) |
| 19+ | 3550 (37.8) |
| Lifetime No. of partners | |
| 1 | 4934 (52.5) |
| 2–3 | 1947 (33.1) |
| 4+ | 1358 (14.4) |
| Lifetime No. of pregnancies | |
| 0 | 723 (7.7) |
| 1–2 | 2588 (27.5) |
| 3–4 | 2475 (26.3) |
| 5+ | 3620 (38.5) |
| Current smoker | |
| No | 8398 (89.3) |
| Yes | 1003 (10.7) |
| Ever use oral contraception | |
| Yes | 5677 (77.7) |
| No | 1634 (22.4) |
| Enrollment Pap result* | |
| Normal | 7906 (84.1) |
| ASC-US | 4227 (8.4) |
| LSIL | 471 (5.0) |
| High-grade | 232 (2.5) |
| Enrollment cervicography* | |
| Negative | 8176 (86.9) |
| Atypical | 868 (9.2) |
| Low-grade | 316 (3.4) |
| High-grade/cancer | 46 (0.5) |
| Enrollment HPV result* | |
| Negative for carcinogenic types | 6683 (71.1) |
| Positive (not HPV16) | 2314 (24.6) |
| HPV16 | 393 (4.2) |

*Missing test results not shown but counted in totals such that percentages do not sum to 100%. ASC-US = atypical squamous cells of undetermined significance; HPV = human papillomavirus; LSIL = low-grade squamous intraepithelial lesion.

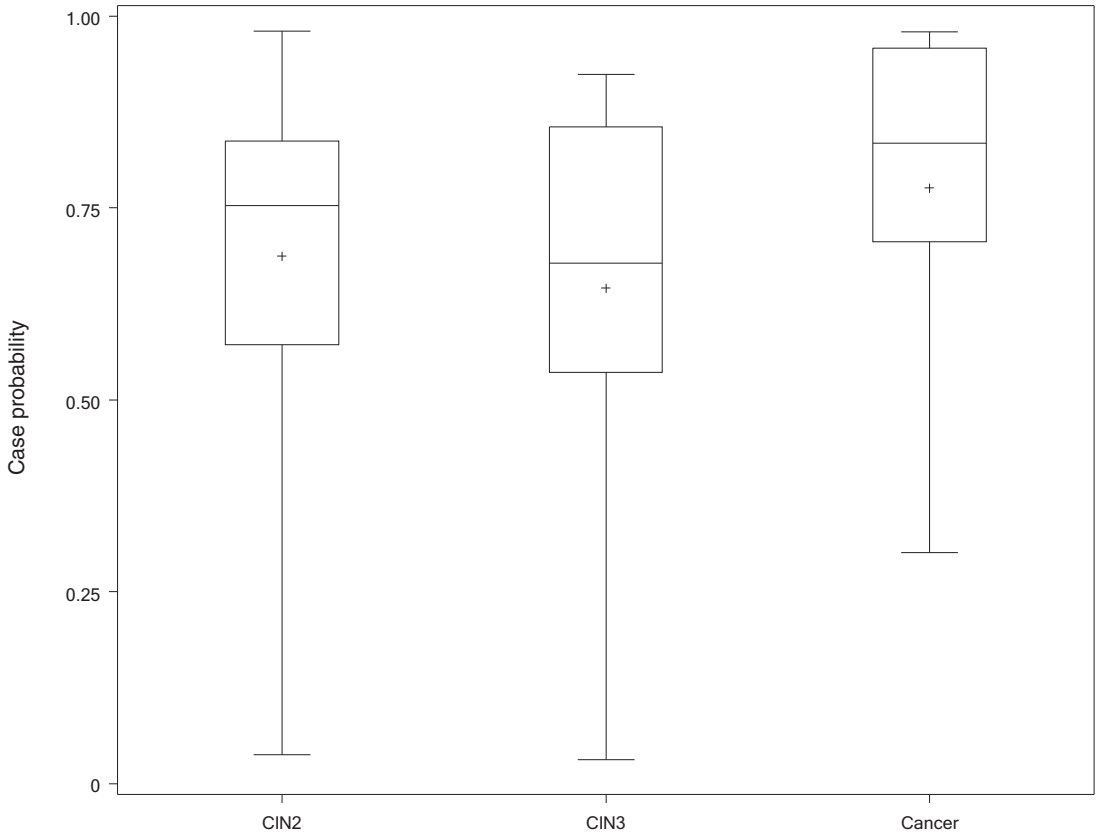


**Figure 3.** Comparison of the mean scores obtained by comparing enrollment images from women whose worst diagnosis within the cohort was cervical intraepithelial neoplasia (CIN) 2, CIN3, or cancer. These subsets of the cases were not statistically significantly different with regard to severity scores generated by the algorithm. Therefore, we combined all into a single case group. We show a box and whisker plot of 32 CIN2, 38 CIN3, and seven cancers, giving quartiles, means, and outliers of case probability for each diagnosis within the cohort.

Some trends emerged based on analysis of the heat maps highlighting regions of the image that most influence the algorithm prediction (Supplementary Figures 3–7, available online). The model usually based its prediction primarily on the region surrounding the os, and performed best when the cervix was oriented directly towards the camera. It was more affected by regions with visible texture than by smooth regions. In general, the model was not excited by glare from the camera's light or other distractors, although in some cases the presence of cotton swabs or other noncervix objects did distract the model.

Analysis of the long-term follow-up of the cohort including linkage to the Costa Rican Tumor Registry revealed 39 cancers. Most were in the training set; 17 were not and had enrollment severity scores (Table 3). Women with cancer missed by automated visual evaluation tended to be diagnosed by the tumor registry merge (mainly older women diagnosed after 7 years of active follow-up).

The automated visual evaluation detection algorithm could theoretically be used to triage women testing positive for HPV rather than for primary screening. Training another automated visual evaluation model restricted to HPV-positive women (data not shown), a sensitivity of 100% (13/13) was achieved with a specificity of 57.5% (103/179). Therefore, primary HPV testing using self-sampling, if it matched the HPV test performance of the early assays we used in 1993–1994, followed by the automated visual evaluation algorithm restricted to positives could achieve the same aggregate sensitivity while more than halving the number of women requiring cervical examinations.

## Discussion

Using a deep learning approach called Faster R-CNN with extensive image augmentation based on a pretrained model, we trained and validated an image analyzer that performs "automated visual evaluation" of the cervix. As a primary screening method, the algorithm, trained using digitized cervigrams from the Guanacaste Cohort, achieved excellent sensitivity for detection of CIN2+ in the age group at highest risk of precancers. The performance surpassed colposcopist evaluator interpretations of the same images (cervicography) and compared favorably to conventional Pap smears (and alternative kinds of cytology) while matching the screening accuracy of an early version of PCR-based HPV testing. The additional positive evaluations among women that were negative for HPV need further study before they can be believed; logically, it is hard to accept and there was an indication that the cases were an unusually young group with incident CIN2, suggesting some possibility of misclassification of both visual appearance and outcome.

Restricted to the age group at which risk of precancer peaks, to achieve nearly perfect sensitivity for cases occurring up to 7 years after examination generated a large number of false positives among screened noncases. More balanced cutpoints for positivity might be chosen to limit excessive treatments, although sensitivity would drop. Another possibility for improving specificity while retaining high sensitivity might be combination screening (commonly called "cotesting") with

ARTICLE

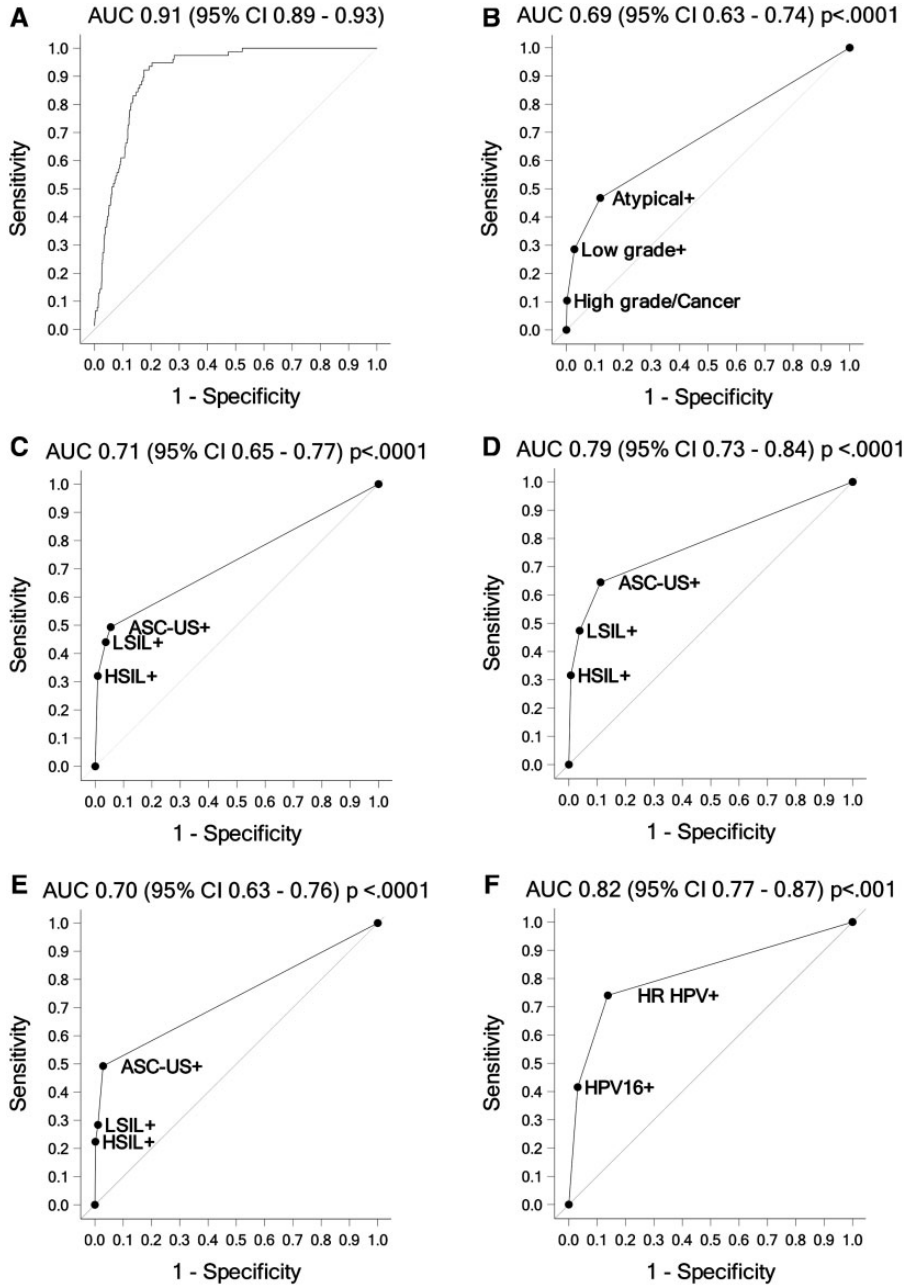


**Figure 4.** Receiver operating characteristic (ROC) curve of automated visual evaluation of cervical images and comparison of performance in identification of cervical intraepithelial neoplasia 2+. ROC-like curves are shown for the categorical variables for simple visual and statistical comparison with automated visual evaluation (two-sided chi-squared tests). The thresholds are listed on each curve, showing the sensitivity and 1-specificity applicable to that cutpoint. Automated visual evaluation was as accurate or more than all of the screening tests used in the cohort study, including: **A**) automated visual evaluation; **B**) cervicography: area under the curve (AUC); **C**) conventional Pap smear; **D**) liquid-based cytology; **E**) first-generation neural network-based cytology; and **F**) MY09-MY11 PCR-based human papillomavirus (HPV) testing. ASC-US = atypical squamous cells of undetermined significance; HSIL = high-grade squamous intraepithelial lesion; LSIL = low-grade squamous intraepithelial lesion.

automated visual evaluation and HPV testing, when such tests become more affordable and more widely available than they currently are.

To permit widespread use of automated visual evaluation detection algorithms, the method would need to be transferred from digitized cervigrams to contemporary digital cameras, because cervicography is obsolete and discontinued. If the transfer is achieved, health workers would still need to visualize the cervix and take well-lit, in-focus images for analysis with the cervix in full view. The minimal required equipment (in addition to the algorithm software) would be acetic acid (vinegar), disposable specula (or sterilization equipment), and the imaging system, such as a dedicated smart phone or digital camera.

Images that resulted in discrepancies between the automated visual evaluation algorithm and clinician interpretation revealed that the performance was affected by image quality and obstructions as well as traditional discordant evaluations due to human subjectivity. So, rather than focusing on training

**Table 2.** Estimated comparison of automated visual evaluation performance by age groups

| Automated visual evaluation by age | CIN2+, No. | <CIN2, No. | Total No. | Age-specific sensitivity, % | Age-specific specificity, % |
|---|---|---|---|---|---|
| <25 y | | | | | |
| + | 46 | 226 | 272 | 82.1 | 77.2 |
| − | 10 | 765 | 775 | | |
| Age-specific total | 56 | 991 | 1047 | | |
| 25–49 y | | | | | |
| + | 127 | 855 | 982 | 97.7 | 84.0 |
| − | 3 | 4475 | 4478 | | |
| Age-specific total | 130 | 5330 | 5460 | | |
| 50+ y | | | | | |
| + | 39 | 399 | 438 | 92.9 | 83.2 |
| − | 3 | 1969 | 1972 | | |
| Age-specific total | 42 | 2368 | 2410 | | |
| Total | 228 | 8689 | 8917 | — | — |

**Table 3.** Severity scores from automatic visual evaluation algorithm and screening results, from enrollment visit of the Guanacaste cohort study, for cases of invasive cancer*

| Years to diagnosis | Enrollment age, y | HPV type result | Pap smear result | Cervigram result | Algorithm severity score |
|---|---|---|---|---|---|
| Enrollment | 21 | 16, 52 | Cancer | Cancer | Training |
| Enrollment | 26 | 16, 18, 51 | Normal | High-grade | Training |
| Enrollment | 29 | 18, 31 | HSIL (CIN2) | High-grade | Training |
| Enrollment | 34 | 16 | Microinvasive | High-grade | Training |
| Enrollment | 35 | 31, 45 | Microinvasive | Cancer | 0.98 |
| Enrollment | 38 | 16 | HSIL (CIN2) | Cancer | Training |
| Enrollment | 41 | 16 | ASC-US | Cancer | 0.78 |
| Enrollment | 47 | 18 | Microinvasive | Cancer? | Training |
| Enrollment | 54 | 16 | Microinvasive | Ccancer? | Training |
| Enrollment | 71 | 53, 82v | Microinvasive | Negative | 0.13 |
| Enrollment | 73 | 33 | Microinvasive | Cancer | 0.96 |
| Enrollment | 74 | 35 | HSIL (CIN3) | Negative | 0.35 |
| Enrollment | 42 | Equivocal | Microinvasive | Cancer? | 0.84 |
| 1 y | 37 | 18 | Normal | Negative | Training |
| 1 y | 61 | Negative | Normal | Negative | Training |
| 2 y | 23 | 16 | ASC-US | Low-grade | 0.87 |
| 2 y | 64 | 45, 51, 58 | Normal | Negative | 0.30 |
| 3 y | 49 | 16 | ASC-US | Negative | Training |
| 5 y | 29 | 18 | Normal | Negative | Training |
| 5 y | 36 | 16 | Normal | Negative | 0.71 |
| 6 y | 38 | Negative | HSIL (CIN2) | Negative | Training |
| 6 y | 45 | 31 | Normal | Missing | Missing |
| 6 y | 48 | Negative | Normal | Negative | Training |
| 6 y | 66 | 56 | Microinvasive | Negative | 0.38 |
| 7 y | 50 | Negative | Normal | Negative | Training |
| 7 y | 63 | Negative | Normal | Negative | Training |
| 8 y | 35 | Negative | Normal | Negative | 0.02 |
| 8 y | 52 | 16 | Normal | Negative | 0.70 |
| 8 y | 63 | 16 | Normal | Negative | Training |
| 8 y | 81 | Negative | Normal | Negative | 0.34 |
| 9 y | 28 | 18 | Normal | Negative | Training |
| 10 y | 65 | 16 | HSIL (CIN3) | Negative | Training |
| 10 y | 75 | 85 | Microinvasive | Negative | Training |
| 11 y | 75 | Negative | Normal | Negative | 0.16 |
| 12 y | 68 | Negative | Normal | Negative | 0.02 |
| 13 y | 56 | Negative | Normal | Negative | 0.75 |
| 15 y | 78 | Missing | Missing | Atypical | Training |
| 16 y | 52 | Negative | Normal | Negative | Training |
| 17 y | 21 | Negative | Inadequate | Atypical | 0.04 |

*ASC-US = atypical squamous cells of undetermined significance; CIN2 = cervical intraepithelial neoplasia 2; CIN3 = cervical intraepithelial neoplasia 3; HSIL = high-grade squamous intraepithelial lesion.

ARTICLE

for the subtleties of visual appearance, which is a difficult skill, the training for automated visual evaluation could highlight more easily acquired skills of improving image quality and lighting and removing obstructions.

We also anticipated a second approach to using the automated visual evaluation detection algorithm, specifically as a triage method restricted to women testing HPV positive, if HPV testing is the primary screen. HPV testing offers the possibility of cervicovaginal self-sampling (33), with only a small minority of women testing HPV positive and requiring gynecologic examinations. Moreover, HPV testing has proven very long-term negative predictive value, that is, reassurance when negative (34,35). The combination of self-sampled HPV screening and triage use of automated visual evaluation detection could greatly reduce the need for speculum examinations. Choice of the automated visual evaluation algorithm screening strategy (general screening vs triage of HPV positive women) will depend on setting and cost-effectiveness analyses. Currently, HPV tests still take several hours and cost too much for many settings, but low-cost point-of-care and batch testing methods applicable to self-sampling are in advanced research and development and likely will be available within a very few years.

It is a strength of this study that it was conducted in a random population sample of a previously poorly screened population that resembles HPV-positive women in other low- and middle-income country (LMIC) target screening populations. Definition of precancer was very good based on repeated screening during extended follow-up and panel review of histopathology.

The major limitations are the following. We studied a relatively small numbers of cases from a single cohort study. We included CIN2 cases in the case group, whereas it would be ideal to train on more definite cases of precancer (ie, CIN3 and AIS caused by types of HPV prevalent in invasive cervical cancers). The images were captured by a small team of highly trained nurses and not a wide variety of health workers. The work relied on images taken with a discontinued film camera technique, rather than contemporary digital image technology.

Nonetheless, the proof-of-principle strongly supports further evaluation of automated visual evaluation. Rather than resurrecting obsolete film camera technology to achieve the observed results, we are currently working to transfer automated visual evaluation to images from contemporary phone cameras and other digital image capture devices to create an accurate and affordable point-of-care screening method that would support the recently announced World Health Organization initiative to accelerate cervical cancer control.

Our findings extend and improve upon the results obtained in more preliminary reports listed in the Supplementary Table 1 (available online). Internationally, a sizable number of commercial and academic research groups are now exploring deep learning algorithms for cervical screening. In collaborative work moving ahead, we will be considering the following topics: assessing the international variability in cervical appearance (eg, inflammatory changes) and need for region-specific training; examining which camera characteristics affect the algorithm and which devices are adequate for image capture; training the camera to take an image automatically when a focused, adequately lit and visualized cervix is visible; training an algorithm by segmentation studies focused on the squamocolumnar junction to assess when ablational therapy is possible or whether (mainly endocervical) extent of lesion suggests need for excision; and evaluating how best to combine automated visual evaluation with HPV testing, and eventually with vaccination, to accelerate control of cervical cancer.

## Funding

This work was supported by the National Institutes of Health, NCI and National Library of Medicine Intramural Research Programs; and the Global Good Fund.

## Notes

Affiliations of authors: Intellectual Ventures Global Good Fund, Bellevue, WA (LH, DB, MPH, NG, BW, MSJ); National Library of Medicine, NIH, Bethesda, MD (SA, ZX, LRL); Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, MD (KY, MD, JCG, NW, MS); Information Management Services, Calverton, MD (BB); Early Detection and Prevention Section, International Agency for Research on Cancer, Lyon, France (RH); Rutgers New Jersey Medical School, Newark, NJ (MHE); Albert Einstein College of Medicine, Bronx, NY (RDB); Consultant to National Cancer Institute, NIH, Bethesda, MD (ACR).

The funding sources had no role in the design of the study, interpretation of the data, the collection, analysis, and interpretation of the data, or the writing of the manuscript.

No author reports a conflict of interest with regard to this research.

We acknowledge the expert advice of Drs Emily Wu and Corey Casper who helped define the clinical outcomes in the initial stage of the project.

## References

1. de Martel C, Plummer M, Vignat J, Franceschi S. Worldwide burden of cancer attributable to HPV by site, country and HPV type. *Int J Cancer*. 2017;141(4):664–670.
2. Bray F, Jemal A, Grey N, Ferlay J, Forman D. Global cancer transitions according to the Human Development Index (2008-2030): a population-based study. *Lancet Oncol*. 2012;13(8):790–801.
3. Schiffman M, Doorbar J, Wentzensen N, et al. Carcinogenic human papillomavirus infection. *Nat Rev Dis Primers*. 2016;2(2):16086.
4. Bagcchi S. India launches plan for national cancer screening programme. *BMJ*. 2016;355;i5574.
5. WHO Guidelines Approved by the Guidelines Review Committee. *WHO Guidelines for Screening and Treatment of Precancerous Lesions for Cervical Cancer Prevention*. Geneva: World Health Organization; 2013.
6. PEPFAR 2019 Country Operational Plan Guidance for all PEPFAR Countries (Draft). Renewed Partnership to Help End AIDS and Cervical Cancer in Africa. https://www.pepfar.gov/documents/organization/288160.pdf#page=318
7. Shastri SS, Mittra I, Mishra GA, et al. Effect of VIA screening by primary health workers: randomized controlled study in Mumbai, India. *J Natl Cancer Inst*. 2014;106(3):dju009.
8. Denny L, Kuhn L, De Souza M, Pollack AE, Dupree W, Wright TC Jr. Screen-and-treat approaches for cervical cancer prevention in low-resource settings: a randomized controlled trial. *JAMA*. 2005;294(17):2173–2181.
9. Sankaranarayanan R, Nene BM, Shastri SS, et al. HPV screening for cervical cancer in rural India. *N Engl J Med*. 2009;360(14):1385–1394.
10. Catarino R, Schafer S, Vassilakos P, Petignat P, Arbyn M. Accuracy of combinations of visual inspection using acetic acid or lugol iodine to detect cervical precancer: a meta-analysis. *BJOG*. 2017;125(5):545–553.
11. Jeronimo J, Schiffman M. Colposcopy at a crossroads. *Am J Obstet Gynecol*. 2006;195(2):349–353.
12. Perkins MJ, RB, Smith KM, Gage JC.et al. ASCCP Colposcopy Standards: Risk-Based Colposcopy Practice. J Low Genit Tract Dis. 2017 Oct;21(4):230–234. doi: 10.1097/LGT.0000000000000334. PubMed PMID: 28953111.
13. Ting DSW, Cheung CY, Lim G, et al. Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*. 2017;318(22):2211–2223.
14. Xu T, Zhang H, Xin C, et al. Multi-feature based benchmark for cervical dysplasia classification evaluation. *Pattern Recogn*. 2017;63:468–475.

15. Song D, Kim E, Huang X, et al. Multimodal entity coreference for cervical dysplasia diagnosis. *IEEE Trans Med Imaging*. 2015;34(1):229–245.
16. Ren SQ, He KM, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell*. 2017;39(6):1137–1149.
17. Bratti MC, Rodriguez AC, Schiffman M, et al. Description of a seven-year prospective study of human papillomavirus infection and cervical neoplasia among 10000 women in Guanacaste, Costa Rica. *Rev Panam Salud Publica*. 2004;15(2):75–89.
18. Herrero R, Schiffman MH, Bratti C, et al. Design and methods of a population-based natural history study of cervical neoplasia in a rural province of Costa Rica: the Guanacaste Project. *Rev Panam Salud Publica*. 1997;1(5):362–375.
19. Rodriguez AC, Avila C, Herrero R, et al. Cervical cancer incidence after screening with HPV, cytology, and visual methods: 18-year follow-up of the Guanacaste cohort. *Int J Cancer*. 2017;140(8):1926–1934.
20. Schneider DL, Herrero R, Bratti C, et al. Cervicography screening for cervical cancer among 8460 women in a high-risk population. *Am J Obstet Gynecol*. 1999;180(2):290–298.
21. Schneider DL, Burke L, Wright TC, et al. Can cervicography be improved? An evaluation with arbitrated cervicography interpretations. *Am J Obstet Gynecol*. 2002;187(1):15–23.
22. Jeronimo J, Long LR, Neve L, Michael B, Antani S, Schiffman M. Digital tools for collecting data from cervigrams for research and training in colposcopy. *J Low Genit Tract Dis*. 2006;10(1):16–25.
23. Hutchinson ML, Zahniser DJ, Sherman ME, et al. Utility of liquid-based cytology for cervical carcinoma screening: results of a population-based study conducted in a region of Costa Rica with a high incidence of cervical carcinoma. *Cancer*. 1999;87(2):48–55.
24. Sherman ME, Schiffman M, Herrero R, et al. Performance of a semiautomated Papanicolaou smear screening system: results of a population-based study conducted in Guanacaste, Costa Rica. *Cancer*. 1998;84(5):273–280.
25. Castle PE, Schiffman M, Gravitt PE, et al. Comparisons of HPV DNA detection by MY09/11 PCR methods. *J Med Virol*. 2002;68(3):417–423.
26. Bouvard V, Baan R, Straif K, et al. A review of human carcinogens—Part B: biological agents. *Lancet Oncol*. 2009;10(4):321–322.
27. He H, Garcia EA. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*. 2009;21(9):1263–1284.
28. Yosinski J, Clune J, Bengio Y, Lipson H, How transferable are features in deep neural networks? In: Proceedings of the 27th International Conference on Neural Information Processing Systems, Volume 2; December 8–13, 2014; Montreal, Canada.
29. Russakovsky O, Deng J, Su H, et al. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015;115(3):211–252.
30. Wong SC, Gatt A, Stamatescu V, McDonnell MD. Understanding data augmentation for classification: when to warp? In: 2016 International Conference on Digital Image Computing: Techniques and Applications (Dicta); November 30–December 2, 2016:59–64. Gold Coast, Australia.
31. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44(3):837–845.
32. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3(1):32–35.
33. Verdoodt F, Jentschke M, Hillemanns P, Racey CS, Snijders PJ, Arbyn M. Reaching women who do not participate in the regular cervical cancer screening programme by offering self-sampling kits: a systematic review and meta-analysis of randomised trials. *Eur J Cancer (Oxf, Engl: 1990)*. 2015;51(16):2375–2385.
34. Schiffman M, Glass AG, Wentzensen N, et al. A long-term prospective study of type-specific human papillomavirus infection and risk of cervical neoplasia among 20,000 women in the Portland Kaiser Cohort Study. *Cancer Epidemiol Biomarkers Prev*. 2011;20(7):1398–1409.
35. Chen HC, Schiffman M, Lin CY, et al. Persistence of type-specific human papillomavirus infection and increased long-term risk of cervical cancer. *J Natl Cancer Inst*. 2011;103(18):1387–1396.