Preliminary Results of an AI Classifier to Diagnose Cervical Precancer on Real-World Data

M Cremer[1,4], AT Goldstein[2], R Nissim[3], Y Misrahi[3], S Bedell[2], S Wu[5], J Sun[5], X Qu[4], J Felix[6] , D Levitz[3]

[1]Ob/Gyn and Women's Health Institute, Cleveland Clinic, Cleveland, OH, USA; [2]Center for Vulvovaginal Diseases, Washington, DC, USA; [3]MobileODT, Tel Aviv, Israel [4]Basic Health International, New York, NY, USA, and Taiyuan, China, and Xinxiang China; [5]Department of ObGyn, Shanxi Bethune Hospital, Shanxi, China; [6]Department of Pathology, Medical College of Wisconsin, Milwaukee, WI, USA

TAP TO RETURN TO KIOSK MENU

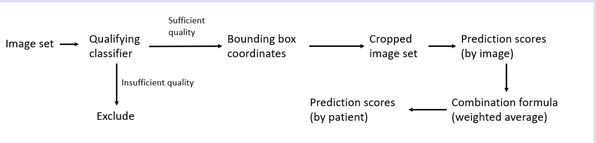
Introduction

Automated Visual Evaluation (AVE) is a new promising method utilizing machine learning of images for detecting cervical precancerous lesions. However, AVE's performance in different scenarios is not well understood

Our **objective** is to provide descriptive statistics of AVE's performance on 3 cervical cancer projects in China: 1 NIH clinical trial and 2 screening camps.

Methods

- Mobile colposcopy images were collected from 3 cervical cancer projects in China
- These 3 data sets were analyzed retrospectively using 2 classifiers
- Endpoint: CIN 2+ histopathology
- Each project received local + US IRB approval
- AVE results were compared against histopathology



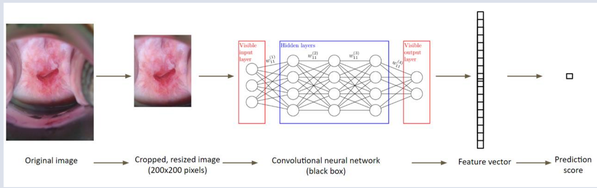
Mobile colposcopy

All mobile colposcopy images analyzed came from the Enhanced Visual Assessment (EVA) System. All images captured were stored on a secure cloud-based web portal.

Automated Visual Evaluation (AVE)

AVE involves using a machine learning classifier to predict the presence of pathology (CIN 2+) by analyzing a cervical image.

The classifier assigns scores to different features in the image and aggregates the scores into a single prediction score. This prediction is calculated in seconds, The prediction score is compared to predefined threshold values that determine if the test result is positive or negative.

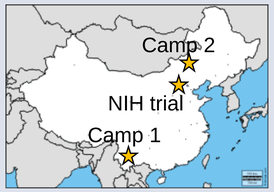


Our analysis compares 2 previously developed AVE classifiers that have been deployed clinically:

- Biopsy-trained classifier: the ground truth of each image was based on worst histopathology
- Colposcopic-impression classifier: the ground truth of each image was based on a consortium of expert opinions who reviewed the images

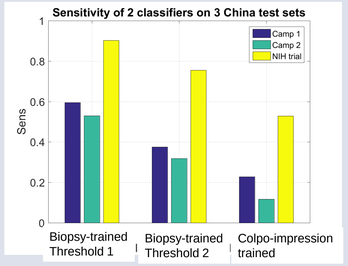
Data collection sites in China

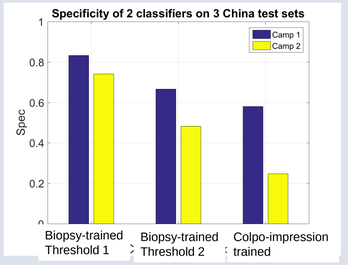
Each project took place at a different region of China, with different ethnicities, culture, and overall patient risk profile. In each site, the EVA System was used as a documentation tool. A summary of the patient characteristics at the 3 sites is below:



	NIH trial	Camp 1	Camp 2
Patient population	Urban hospital	Screening camp	Screening camp
No. patients	363	300	482
Prevalence (CIN 2+)	100%	12%	18%
Age range	30-55	All comers	All comers
Enrollment criteria	CIN 2+ biopsy	HPV+	HPV+
No. providers	4	5	5
Provider training	Expert colposcopists	Colposcopists Non-colposcopist Nurses	Colposcopists Non-colposcopist Nurses

Results





The sensitivity and specificity of 3 AVE thresholds were calculated :

- Biopsy-trained, Threshold 1: high sensitivity mode
- Biopsy-trained, Threshold 2: high specificity mode
- Human-annotated: preferred by clinicians in usability testing

Conclusion

AI has great potential as a new cervical cancer screening method. Its accuracy is at least as good as LBC and HPV genotyping from Kaiser Permanente data. Data is needed from both well designed clinical trials and real world clinical data to better ascertain the potential of AVE, both in a screening and triage setting.

Preliminary Results of an AI Classifier to Diagnose Cervical Precancer on Real-World Data

M Cremer[1,4], AT Goldstein[2], R Nissim[3], Y Misrahi[3], S Bedell[2], S Wu[5], J Sun[5], X Qu[4], J Felix[6] , D Levitz[3]

[1]Ob/Gyn and Women's Health Institute, Cleveland Clinic, Cleveland, OH, USA; [2]Center for Vulvovaginal Diseases, Washington, DC, USA; [3]MobileODT, Tel Aviv, Israel [4]Basic Health International, New York, NY, USA, and Taiyuan, China, and Xinxiang China; [5]Department of ObGyn, Shanxi Bethune Hospital, Shanxi, China; [6]Department of Pathology, Medical College of Wisconsin, Milwaukee, WI, USA

Biopsy-trained classifier

- Trained on images biopsy correlated images from 832 patients + 32,156 unlabeled images
- Data came from 14 countries
  - No standardization of histopathology
- Many positives came from China
- Based on Semi-supervised learning using a ResNet architecture running on TensorFlow
- *Global sources of data for 2 classifiers:*

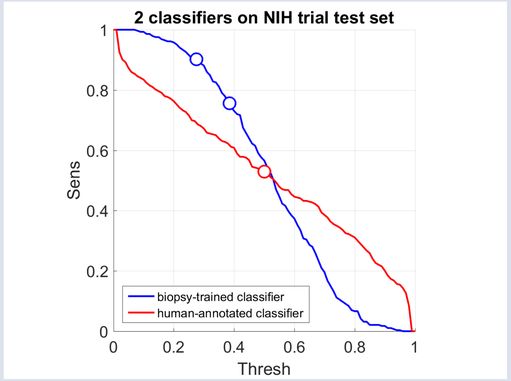
Biopsy-trained classifier

Colpo-impression classifier

Colposcopic-impression classifier

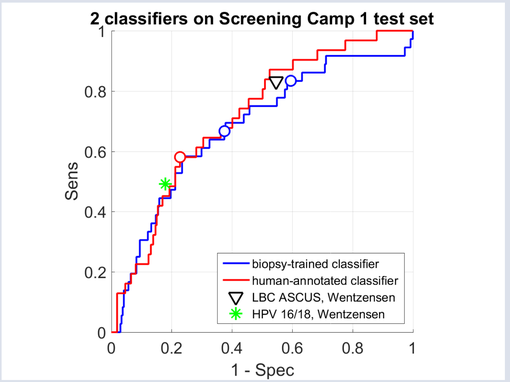
- Presented at IPVC 2018 by Demarco et al
- Trained on manually annotated images from 1473 patients
  - 3 reviewers
  - Standardized reviews with adjudication process
- Data came from 17 countries, with heavy representation from Kenya and India.
- No data came from China
- Based on Faster RCNN architecture running on Caffe

Performance of 2 classifiers



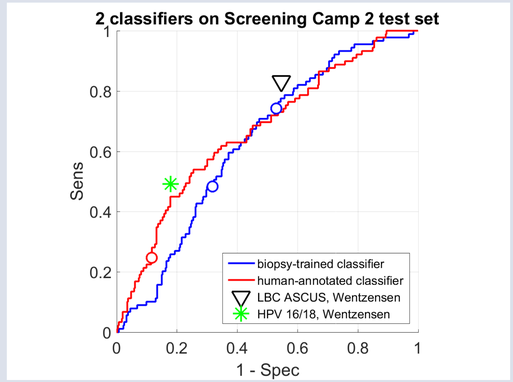
(In the NIH trial, only sensitivity is calculated)

A large difference is observed between 2 classifiers. In the biopsy-trained model, most of the patients are classified as positive in both high sensitivity and high specificity modes. In the human annotation model, sensitivity is low (0.53) at the operating threshold of 0.5.



In Camp 1, the 2 classifiers performed similarly. The biopsy-trained model was a bit more specific (left side), while the human annotation model was a bit more sensitive (top of curve).

The patient population in the camp included many postmenopausal patients, and many with severe vaginal atrophy.



In Camp 2, the human annotated classifier had higher specificity than the biopsy-trained classifier.

The patient population in the screening camp included many postmenopausal women.

The performance of AVE matched those of comparable triage tests for HPV+ patients (cytology, HPV 16/18 genotyping) from high end US clinics.

The largest variation in performance between the 2 classifiers was in the NIH trial data. Interestingly, the biopsy-trained model was more specific in Camp 2 but more sensitive in Camp 1, while the human annotation model was more sensitive in Camp 2 and more specific in Camp 1.

References

Wentzensen et al, J AMA Int Med 2019
Belinson et al, presentation at ASCCP 2019
Goldstein et al, poster at ASCCP 2020
Demarco et al, presentation at IPVC 2018
Goldstein et al, presentation at ASCCP 2019

Funding

This retrospective analysis was funded by MobileODT. The NIH trial was funded by NIH grant UH3CA189883, with supplemental funding by the Rising Tide Foundation. The 2 screening camps were funded by the Gynecologic Cancers Research Foundation.