





RESEARCH ARTICLE

Open Access



# Development and validation of an artificial intelligence system for grading colposcopic impressions and guiding biopsies

Peng Xue<sup>1,2†</sup> , Chao Tang<sup>3†</sup>, Qing Li<sup>4†</sup>, Yuexiang Li<sup>5</sup>, Yu Shen<sup>6</sup>, Yuqian Zhao<sup>7</sup>, Jiawei Chen<sup>5</sup>, Jianrong Wu<sup>8</sup>, Longyu Li<sup>9</sup>, Wei Wang<sup>10</sup>, Yucong Li<sup>11</sup>, Xiaoli Cui<sup>12</sup>, Shaokai Zhang<sup>13</sup>, Wenhua Zhang<sup>2</sup>, Xun Zhang<sup>14</sup>, Kai Ma<sup>5</sup>, Yefeng Zheng<sup>5</sup>, Tianyi Qian<sup>8</sup>, Man Tat Alexander Ng<sup>8</sup>, Zhihua Liu<sup>15</sup>, Youlin Qiao<sup>1,2</sup> , Yu Jiang<sup>1\*</sup>  and Fanghui Zhao<sup>2\*</sup> 

## Abstract

**Background:** Colposcopy diagnosis and directed biopsy are the key components in cervical cancer screening programs. However, their performance is limited by the requirement for experienced colposcopists. This study aimed to develop and validate a Colposcopic Artificial Intelligence Auxiliary Diagnostic System (CAIADS) for grading colposcopic impressions and guiding biopsies.

**Methods:** Anonymized digital records of 19,435 patients were obtained from six hospitals across China. These records included colposcopic images, clinical information, and pathological results (gold standard). The data were randomly assigned (7:1:2) to a training and a tuning set for developing CAIADS and to a validation set for evaluating performance.

**Results:** The agreement between CAIADS-graded colposcopic impressions and pathology findings was higher than that of colposcopies interpreted by colposcopists (82.2% versus 65.9%, kappa 0.750 versus 0.516,  $p < 0.001$ ). For detecting pathological high-grade squamous intraepithelial lesion or worse (HSIL+), CAIADS showed higher sensitivity than the use of colposcopies interpreted by colposcopists at either biopsy threshold (low-grade or worse 90.5%, 95% CI 88.9–91.4% versus 83.5%, 81.5–85.3%; high-grade or worse 71.9%, 69.5–74.2% versus 60.4%, 57.9–62.9%; all  $p < 0.001$ ), whereas the specificities were similar (low-grade or worse 51.8%, 49.8–53.8% versus 52.0%, 50.0–54.1%; high-grade or worse 93.9%, 92.9–94.9% versus 94.9%, 93.9–95.7%; all  $p > 0.05$ ). The CAIADS also demonstrated a superior ability in predicting biopsy sites, with a median mean-intersection-over-union (mIoU) of 0.758.

(Continued on next page)

\* Correspondence: [jiangyu@pumc.edu.cn](mailto:jiangyu@pumc.edu.cn); [zhaofangh@cicams.ac.cn](mailto:zhaofangh@cicams.ac.cn)

<sup>†</sup>Peng Xue, Chao Tang and Qing Li contributed equally to this work.

<sup>1</sup>Department of Epidemiology and Biostatistics, School of Population Medicine and Public Health, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China

<sup>2</sup>Department of Cancer Epidemiology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

(Continued from previous page)

**Conclusions:** The CAIADS has potential in assisting beginners and for improving the diagnostic quality of colposcopy and biopsy in the detection of cervical precancer/cancer.

**Keywords:** Artificial intelligence, Cervical cancer prevention, Colposcopy diagnosis and biopsy, Global elimination of cervical cancer

## Background

Cervical cancer results in high rates of morbidity and mortality worldwide and with a disproportionate effect on low- and middle-income countries (LMICs). Annually, over 85% of new cases and 87% of deaths occur in LMICs [1]. In 2018, the World Health Organization (WHO) called for action towards the reduction of cervical cancer cases based on proven strategies [2, 3]. One of these strategies is that 70% of women between the ages of 35 and 45 years receive screening. By 2030, 90% of women in this age group must be managed to achieve the goal of fewer than four new cases per 100,000 [4]. A major concern is the diagnostic ability to appropriately identify and manage women with abnormal screening status at the time of colposcopy. Accurate visual detection of underlying colposcopic abnormalities is critical in guiding biopsies for pathological confirmation. This detection is essential for the identification of high-grade squamous intraepithelial lesion or worse (HSIL+), lesions that require immediate treatment. However, current colposcopic evaluation presents challenges, especially in LMICs. These challenges include poor agreement (under 50%) between colposcopic impressions and pathological findings, a strong dependence on the subjective experience of operators, substantial variabilities among inter- and intra-operators, a large number of women with risk factors that warrant referral to colposcopy, and a shortage of experienced colposcopists [5–7].

In light of these challenges, the American Society for Colposcopy & Cervical Pathology (ASCCP) reviewed colposcopy standards to improve diagnostic performance by establishing a comprehensive evaluation based on available test results (cytology, human papillomavirus [HPV] status, and colposcopy impressions) [8, 9]. However, despite its widespread use, over the last decade, there has been little improvement in colposcopic performance, especially in LMICs [10].

Recently, artificial intelligence (AI) methods have shown potential in subjective imaging diagnoses for malignancies such as breast cancer, colorectal cancer, and gastrointestinal cancer [11–13]. The application of similar methods to colposcopic imaging is not yet widespread [14, 15]. In this study, we developed an AI method (Colposcopic Artificial Intelligence Auxiliary Diagnostic System [CAIADS]) for grading colposcopic

impressions and guiding biopsies. We evaluated its performance on an independent validation set and compared it to colposcopy interpretations made by colposcopists.

## Methods

### The primary goals of the use of CAIADS

One of the primary goals in the application of CAIADS was to grade colposcopic impressions in accordance with the latest ASCCP colposcopy terminology: normal/benign, low-grade, high-grade, and cancer [16, 17]. The CAIADS was also expected to dichotomously grade colposcopic impressions into two hypothetical biopsy thresholds (low-grade or worse versus normal/benign, and high-grade or worse versus a less severe impression). These categories were used to find an appropriate colposcopically guided biopsy threshold and guide biopsies for detecting the clinically relevant endpoint (pathology-confirmed HSIL+).

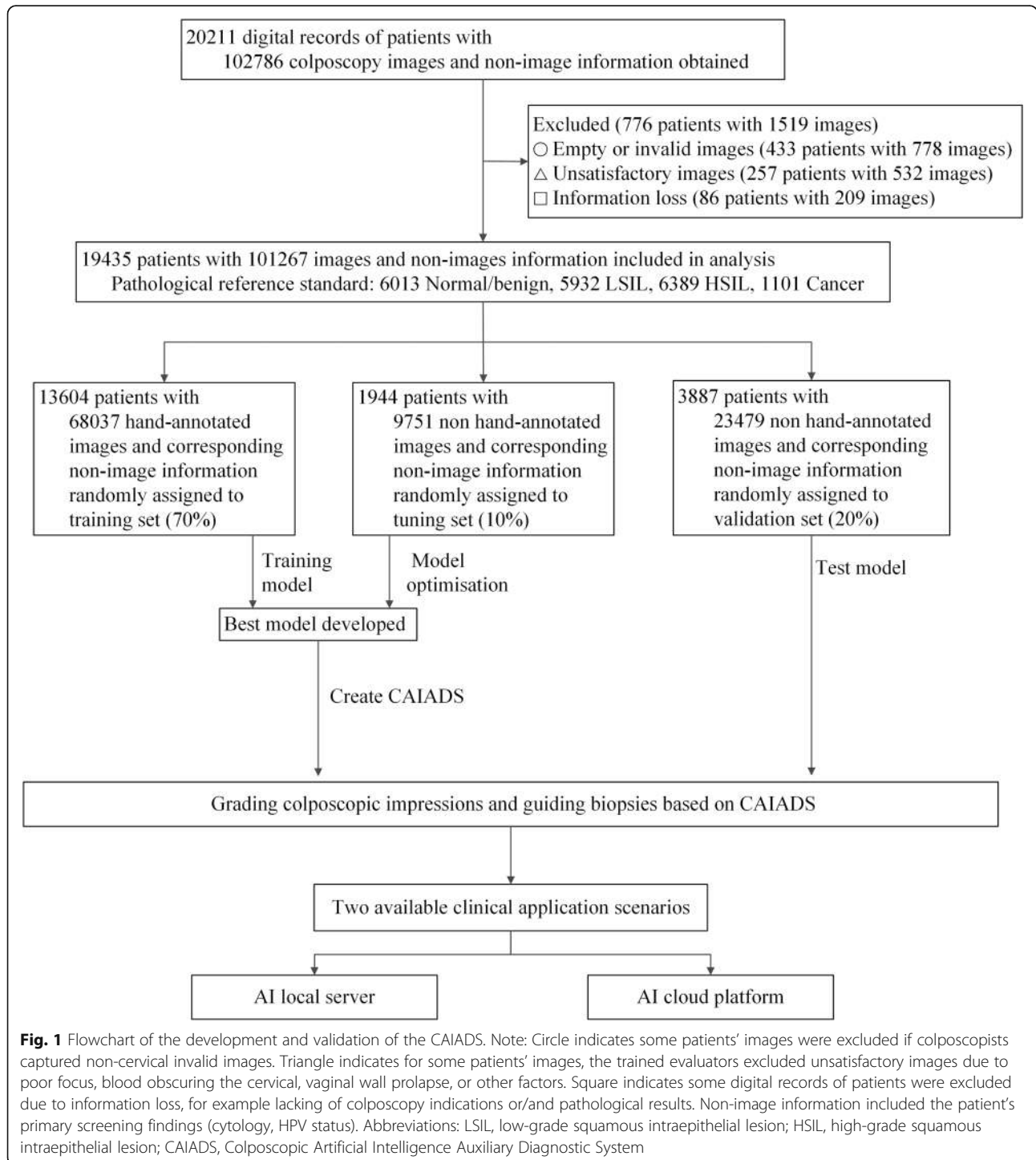
### Study patients and design

Between January 12, 2018, and December 30, 2018, anonymized digital records of patients, including colposcopic images, non-image information (cytology and HPV status), and pathological results were retrospectively obtained from archived databases of six multicenter hospitals across China (Additional file 1, Table S1), including Shenzhen Maternity and Child Healthcare Hospital (SZMCHH). The pathological results were the gold standard for developing CAIADS and validate its diagnostic performance. The study was approved by the institutional review board (IRB) of SZMCHH. The need for informed consent was waived by the IRB of SZMCHH due to the retrospective nature of archived datasets and fully anonymized personal information.

All patients aged 24–65 years with indications for the need for colposcopy underwent colposcopy imaging and biopsy, and those who were pathologically confirmed were eligible for our study. We excluded patients who lacked definitive pathological results, and we used the WHO classification system: normal/benign, low-grade squamous intraepithelial lesion (LSIL), HSIL, and cancer. All pathology slides of punch biopsies were reviewed by pathologists from SZMCHH. Any disagreement was resolved by a panel of expert pathologists.

The digital records of each patient were split into two categories: (1) those records containing at least five satisfactory colposcopic images commonly with ordinal timeslots (around 0 s, 60 s, 90 s, 120 s, and 150 s) and (2) those records containing non-image (cytology and HPV status), and quality control information conducted by trained evaluators, for which the exclusion criteria are shown in Fig. 1.

Sample images in JPEG formats are shown in Additional file 1, Figure S1. The quality control and the complete data were randomly sampled by the severity distribution of pathological results and then assigned to a training and a tuning set for developing CAIADS and to a validation set for evaluating performance in a ratio of 7:1:2. The three datasets are obtained by



random sampling according to the patient IDs, which means the patients in the validation set will not be used in the training phase.

For the training set, all selected images were automatically uploaded to an online cervix image annotation tool. These images were analyzed by a group of eight experienced colposcopists from SZMCHH. They carefully manually delineated the lesion areas and biopsy sites near the squamocolumnar junction of the cervical regions, labeling each based on the corresponding biopsy sites of the pathological results. The pathological results were the gold standard. These analyses were supervised by expert colposcopists from the National Cancer Center. The details of annotation and the annotation tool are shown in Additional file 1, Figure S2. For the tuning and validation sets, we had no manual annotations on the images. For all datasets, we made no changes to non-image information.

#### Development of the CAIADS algorithm

Because colposcopists analyze both images and non-image information (cytology and HPV status) during colposcopic examinations, we developed CAIADS to simulate the diagnostic judgment of colposcopists as accurate as possible. The CAIADS algorithm consists of two deep-learning-based modules for grading colposcopic impressions and guiding biopsies, respectively. A detailed description of the CAIADS algorithm is presented in Additional file 1, Supplementary Method and Figure S3 [18–20]. Briefly, the proposed CAIADS first detected the cervical area of images for the subsequent feature extraction. Then, the extracted features were fused by a graphical convolutional network. Finally, the non-image information was concatenated to the fused features of the images to yield the result of grading impressions. Additionally, the CAIADS also predicted the suspected lesion areas to limit the range for guiding biopsy sites.

The pipeline for colposcopic grading consisted of cervix detection, feature extraction, and feature fusion networks, whereas a U-Net [21] and a YOLO [22] were implemented for lesion area segmentation and biopsy site guiding, respectively. Because an accurate lesion area segmentation can effectively reduce the number of unnecessary biopsy sites that fall outside regions containing lesions, we implemented a semi-supervised framework, as shown in Additional file 1, Figure S4. The purpose of this framework was to utilize the tuning set (only with the image-level label) to further boost the segmentation performance of CAIADS. The semi-supervised framework developed on the training set was used to produce pseudo-labels for the tuning set. Then, the tuning set with pseudo-labels was mixed with the training set to fine-tune the U-Net. A subset was separated from the

training set to monitor the performance of deep-learning networks during training and to prevent overfitting. Training of the system was halted if no performance increase was observed on the separated subset.

#### Validation of the CAIADS performance

We compared the agreement of colposcopic impressions of the CAIADS and original colposcopic interpretation by using pathology as the gold standard. The original colposcopic interpretation was determined and recorded by colposcopists based on the assessment of the patient's images and non-image information. In addition, the diagnostic performance of CAIADS at different hypothetical biopsy thresholds (low-grade or worse and high-grade or worse) for the detection of pathological HSIL+ was evaluated from three aspects. Firstly, we investigated whether the diagnostic performance of CAIADS could be improved by additional non-image information, compared with grading the images alone. Secondly, we compared the performance of the CAIADS at the biopsy threshold of low-grade impression or worse versus high-grade impression or worse. Thirdly, the performance of CAIADS was compared with the original colposcopic interpretation by colposcopists. Finally, we tested the accuracy of the CAIADS in predicting biopsy sites compared with ground truth biopsy sites.

#### Statistical analyses

The ROC curve was created by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity), and we calculated AUC values. The diagnostic AUC value, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were evaluated together using 95% confidence intervals (CIs) by the Clopper-Pearson method. We defined the main metric as agreement with the pathological gold standard, measured using kappa values. The McNemar test was used to evaluate the differences in diagnostic performance including agreement, accuracy, sensitivity, and specificity. A  $p$  value less than 0.05 (two-sided) was considered to be statistically significant. Statistical analyses were done using SAS 9.4 software (SAS Institute Inc., Cary, NC, USA), Python 3.6, and scikit-learn [23].

## Results

#### Study participants

In total, 101,267 colposcopic images and non-image information from 19,435 patients were included in this study. The complete training set consisted of 68,037 images as well as non-image information from 13,604 patients with pathological results of normal/benign ( $n = 4217$ ), LSIL ( $n = 4150$ ), HSIL ( $n = 4489$ ), and cancer ( $n = 748$ ). The tuning set consisted of

9751 images as well as non-image information from 1944 patients with pathological results of normal/benign ( $n = 591$ ), LSIL ( $n = 594$ ), HSIL ( $n = 630$ ), and cancer ( $n = 129$ ). The validation set consisted of 23,479 images as well as non-image information from 3887 patients with pathological results of normal/benign ( $n = 1205$ ), LSIL ( $n = 1188$ ), HSIL ( $n = 1270$ ), and cancer ( $n = 224$ ). Detailed information of the training, tuning, and validation sets is summarized in Table 1 and Fig. 1.

### The colposcopic grading performance

Of the 3887 patients in the validation set, the CAIADS achieved an overall agreement of 82.2% for grading colposcopic impressions with the pathological gold standard (kappa 0.750). For patients pathologically confirmed as normal/benign, LSIL, HSIL, and cancer, the agreements between colposcopic impressions of the CAIADS and pathology were 95.5%, 81.6%, 66.9%, and 100%, respectively. A review of discrepant cases revealed that disagreement most often occurred when grading LSIL (81.6%) and HSIL (66.9%). Moreover, we observed that the overall agreement of the CAIADS-graded colposcopic impressions and pathology was higher than that of the original colposcopy interpretation by the colposcopists (82.2% versus 65.9%, kappa 0.750 versus 0.516,  $p <$

0.001). Confusion matrices of colposcopic grading distribution are presented in Fig. 2.

### Pathologically HSIL+ diagnostic performance

Of the 3887 patients in the validation set, there were 1494 pathologically confirmed HSIL+. The diagnostic performance of the CAIADS at different biopsy thresholds for detecting HSIL+ was evaluated from three aspects, as shown in Table 2 and Fig. 3. Firstly, compared with grading the images alone, the analyses of additional non-image information improved the diagnostic performance of CAIADS. The area under the curve (AUC) values were statistically significantly increased at different biopsy thresholds (low-grade or worse 0.681, 95% CI 0.678–0.694 versus 0.712, 0.699–0.724; high-grade or worse 0.779, 0.765–0.792 versus 0.829, 0.827–0.842; all  $p < 0.001$ ), the accuracies (low-grade or worse 63.6%; 95% CI 62.1–65.1% versus 66.7%, 65.2–68.2%; high-grade or worse 80.7%, 79.4–81.9% versus 85.5%, 84.3–86.6%; all  $p < 0.05$ ), the sensitivities (low-grade or worse 87.3%, 95% CI 85.5–88.9% versus 90.5%, 88.9–91.4%; high-grade or worse 65.8%, 63.3–68.2% versus 71.9%, 69.5–74.2%; all  $p < 0.001$ ), and the specificities (low-grade or worse 48.9%, 95% CI 46.8–50.9% versus 51.8%, 49.8–53.8%,  $p = .04$ ; high-grade or worse 90.0%, 88.7–91.2% versus 93.9%, 92.9–94.9%,  $p < 0.001$ ). Secondly, the rating of the biopsy threshold of low-grade

**Table 1** Basic characteristics

Characteristic	No. (%)		
	Training set	Tuning set	Validation set
Images, total no.	68,037	9751	23,479
Patients, total no.	13,604	1944	3887
Age (years)			
24–29	2694 (19.8)	399 (20.5)	988 (25.4)
30–49	8815 (64.8)	1211 (62.3)	2058 (53.0)
50–65	2095 (15.4)	334 (17.2)	841 (21.6)
Referral colposcopy indications <sup>a</sup>			
Primary screening results <sup>b</sup>	9989 (73.4)	1254 (64.5)	2337 (60.1)
Suspicious clinical symptoms <sup>c</sup>	3687 (27.1)	735 (37.8)	1650 (42.4)
Distribution of pathological results <sup>d</sup>			
Normal/benign	4217 (31.0)	591 (30.4)	1205 (31.0)
LSIL	4150 (30.5)	594 (30.6)	1188 (30.5)
HSIL	4489 (33.0)	630 (32.4)	1270 (32.7)
Cancer	748 (5.5)	129 (6.6)	224 (5.8)

**Abbreviations:** HR-HPV high-risk human papillomavirus, ASC-US atypical squamous cells of undetermined significance, LSIL low-grade squamous intraepithelial lesion, HSIL high-grade squamous intraepithelial lesion

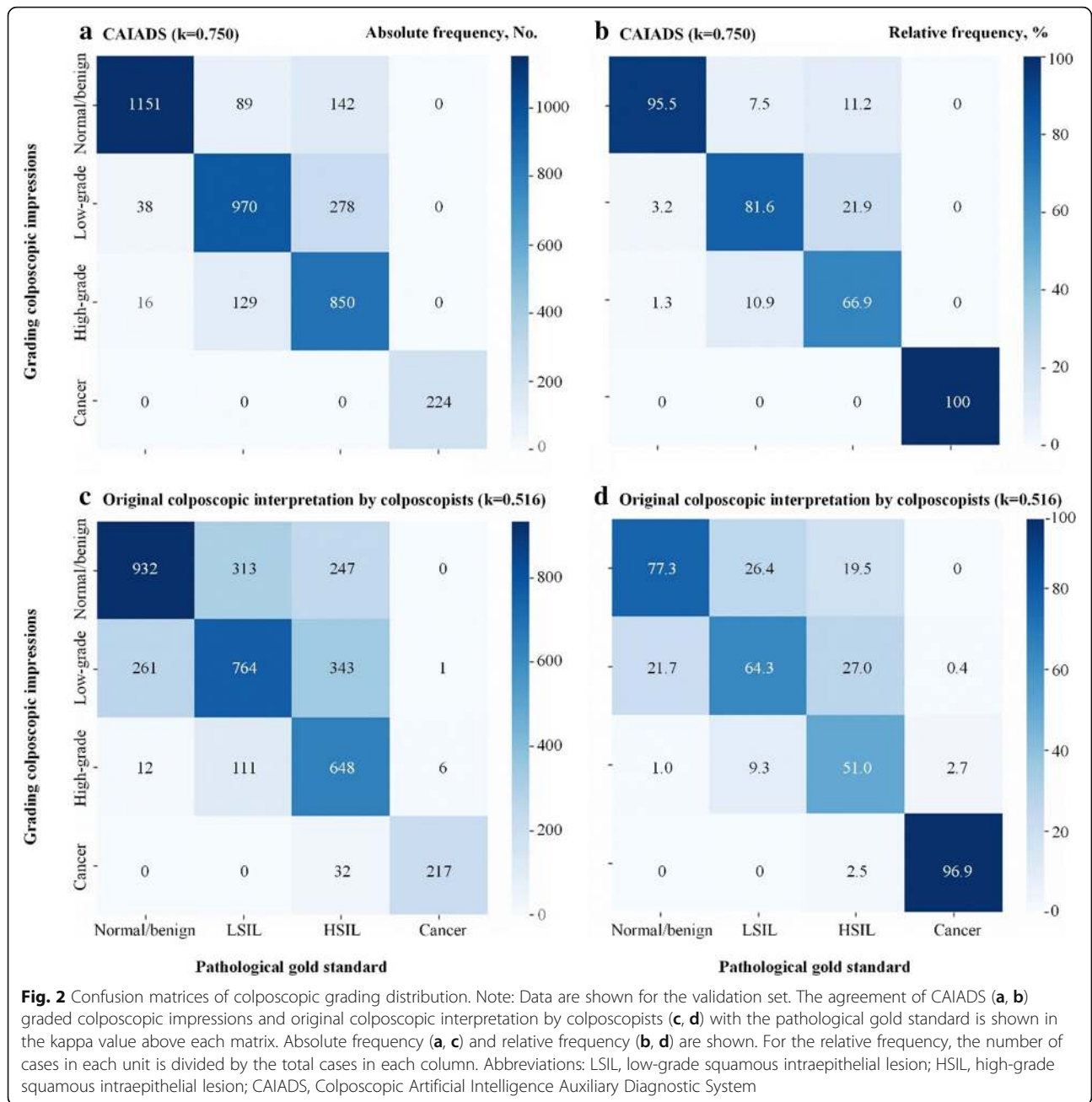
<sup>a</sup>The percentages sum to over 100% due to the overlap of different colposcopy indications

<sup>b</sup>Primary screening results mainly included cytology (HSIL, LSIL, ASC-US, negative) and/or HPV status (HR-HPV positive without genotyping, HPV16/18, other HR-HPV positive, negative)

<sup>c</sup>Suspicious clinical symptoms mainly included abnormal genital tract bleeding, suspicious cervical abnormality, unexplained cervicovaginal discharge, and other factors

<sup>d</sup>When multiple lesions were present in a patient, the highest grade was used as the final pathological diagnosis





or worse was statistically more sensitive (90.5%, 95% CI 88.9–91.4% versus 71.9%, 69.5–74.2%,  $p < 0.001$ ), albeit less specific than those rated high-grade or worse (51.8%, 95% CI 49.8–53.8% versus 93.9%, 92.9–94.9%,  $p < 0.001$ ). Thirdly, we observed that the AUC values of the CAIADS were higher than the original colposcopy interpreted by colposcopists using either a biopsy threshold (low-grade or worse 0.712, 0.699–0.724 versus 0.678, 0.663–0.691; high-grade or worse 0.829, 0.827–0.842 versus 0.777, 0.763–0.790; all  $p < 0.001$ ), the accuracies (low-grade or worse 66.7%, 95% CI 65.2–

68.2% versus 64.1%, 62.6–65.6%; high-grade or worse 85.5%, 84.3–86.6% versus 81.6%, 80.4–82.8%; all  $p < 0.05$ ), and the sensitivities (low-grade or worse 90.5%, 95% CI 88.9–91.4% versus 83.5%, 81.5–85.3%; high-grade or worse 71.9%, 69.5–74.2% versus 60.4%, 57.9–62.9%; all  $p < 0.001$ ), whereas the specificities were similar (low-grade or worse 51.8%, 95% CI 49.8–53.8% versus 52.0%, 50.0–54.1%,  $p = 0.91$ ; high-grade or worse 93.9%, 92.9–94.9% versus 94.9%, 93.9–95.7%,  $p = 0.17$ ). Overall, the CAIADS achieved higher diagnostic sensitivity and similar specificity compared with the original

**Table 2** The diagnostic performance for detecting pathological HSIL+ at different hypothetical biopsy thresholds

	Accuracy, % (95% CI)	Sensitivity, % (95% CI)	Specificity, % (95% CI)	Positive predictive, value % (95% CI)	Negative predictive, value % (95% CI)
<b>Analysis of images alone by CAIADS</b>					
Normal/benign versus low-grade or worse	63.6 (62.1–65.1)	87.3 (85.5–88.9)	48.9 (46.8–50.9)	51.6 (49.6–53.6)	86.0 (84.1–87.8)
Less severe impressions <sup>a</sup> versus high-grade or worse	80.7 (79.4–81.9)	65.8 (63.3–68.2)	90.0 (88.7–91.2)	80.4 (78.0–82.6)	80.8 (79.3–82.3)
<b>Analysis of both images and non-image information<sup>b</sup> by CAIADS</b>					
Normal/benign versus low-grade or worse	66.7 (65.2–68.2)	90.5 (88.9–91.4)	51.8 (49.8–53.8)	54.0 (52.0–55.9)	89.7 (88.0–91.3)
Less severe impressions <sup>a</sup> versus high-grade or worse	85.5 (84.3–86.6)	71.9 (69.5–74.2)	93.9 (92.9–94.9)	88.1 (86.2–89.9)	84.3 (82.8–85.6)
<b>Analysis of both images and non-image information<sup>b</sup> by colposcopists</b>					
Normal/benign versus low-grade or worse	64.1 (62.6–65.6)	83.5 (81.5–85.3)	52.0 (50.0–54.1)	52.1 (50.0–54.1)	83.5 (81.5–85.3)
Less severe impressions <sup>a</sup> versus high-grade or worse	81.6 (80.4–82.8)	60.4 (57.9–62.9)	94.9 (93.9–95.7)	88.0 (85.9–89.9)	79.3 (77.8–80.8)

*Abbreviations:* CAIADS Colposcopic Artificial Intelligence Auxiliary Diagnostic System  
<sup>a</sup>Less severe impressions included normal/benign and low-grade  
<sup>b</sup>Non-image information included patient’s primary screening findings (cytology, HPV status)

colposcopy interpreted by colposcopists for detecting HSIL+.

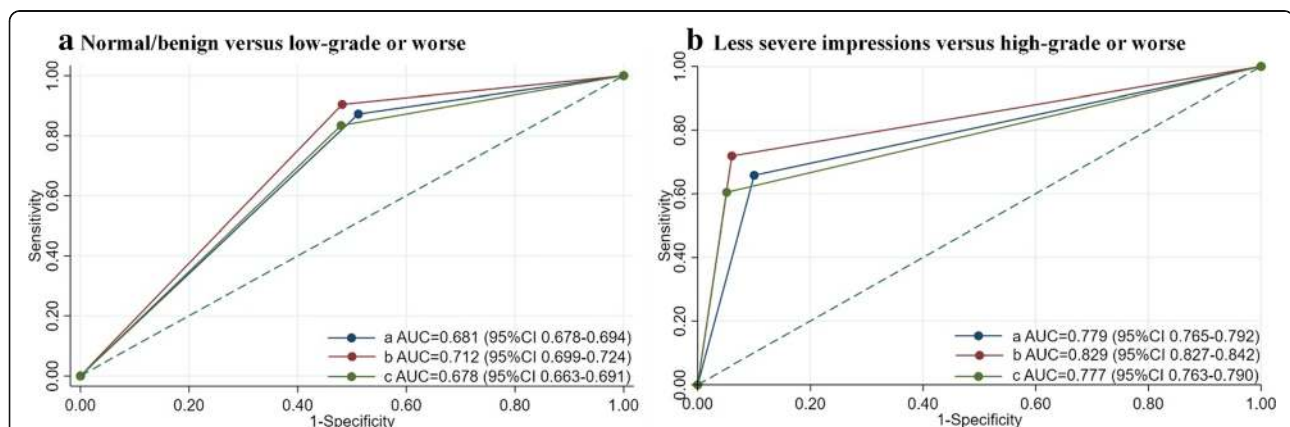
represent the predicted and ground truth biopsy sites, respectively.

**Biopsy sites predicting performance**

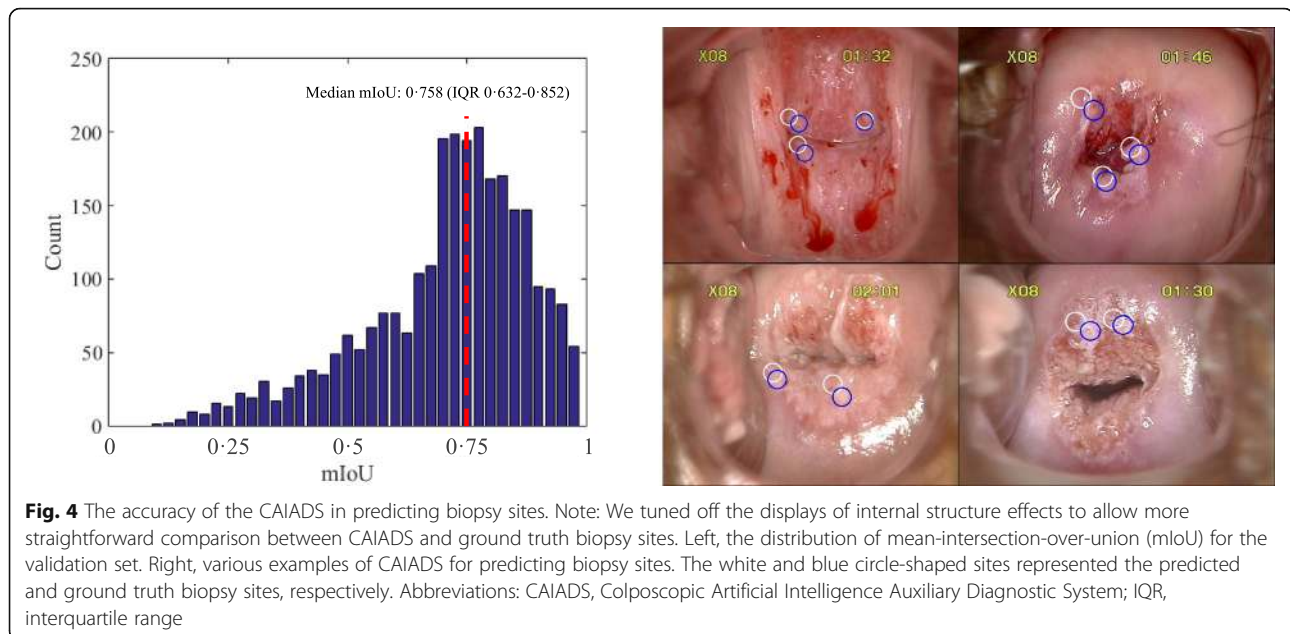
The accuracy of the CAIADS in predicting biopsy sites was evaluated using ground truth biopsy sites. Figure 4 shows the distribution of mean-intersection-over-union (mIoU) for the validation set, and various examples of biopsy sites predicted by CAIADS. A median mIoU of 0.758 (interquartile range 0.632–0.852) was achieved by CAIADS on the validation set. The higher mIoU value represented the more accurate performance for biopsy site prediction. The white and blue circle-shaped sites

**Discussion**

The performance of colposcopy and directed biopsies is the major challenge in the cervical cancer screening process. Previous studies [24, 25] have suggested that even some experienced colposcopists are challenged at correctly grading colposcopic impressions. To this end, we developed a Colposcopic Artificial Intelligence Auxiliary Diagnosis System which we termed CAIADS. This system was trained and validated in 101,267



**Fig. 3** The ROC curve of diagnostic performance for detecting pathological HSIL+ at different hypothetical biopsy thresholds. Note: The hypothetical biopsy thresholds are shown: **A)** normal/benign versus low-grade or worse and **B)** less severe impressions versus high-grade or worse. **a** Analysis of images alone by CAIADS. **b** Analysis of both images and non-image information by CAIADS. **c** Analysis of both images and non-image information by colposcopists. Abbreviations: CAIADS, Colposcopic Artificial Intelligence Auxiliary Diagnostic System; AUC, area under the curve. Less severe impressions included normal/benign and low-grade. Non-image information included the patient’s primary screening findings (cytology, HPV status)



retrospective colposcopic images as well as in non-image information from 19,435 patients.

To the best of our knowledge, this is the first study using a large-scale dataset in the field of artificial intelligence-graded colposcopic impressions and guided biopsies worldwide. The CAIADS method achieved a high level of agreement (82.2%) with pathological results as the gold standard regarding grading all colposcopic impressions and was higher than the original colposcopic interpretation by colposcopists (65.9%). This level of agreement was significantly higher than that of Benedet et al. (52.0%) [26] which used a large dataset of colposcopic diagnostic studies from 84,244 British patients. In addition, we observed that CAIADS had an excellent agreement with pathology results when distinguishing between normal/benign (95.5%) and cancer (100%), thereby providing more reassurance in regard to the veracity of positive results and negative results. Despite the fact that AI was not limited by diagnostic subjectivity of colposcopists, as expected, CAIADS showed promising but suboptimal performance in grading LSIL (81.6%) and HSIL (66.9%) owing to the vague and subtle distinctions between LSIL and HSIL features (the thickness of acetowhite epithelium, vascular patterns, margins/border, etc.). For example, some LSIL lesions are condylomas caused by HPV infection, but only in a very small range contains HSIL, which it is difficult to judge. The ability to distinguish between LSIL and HSIL remains one of the most important challenges in colposcopy practice as well as for colposcopists. However, in terms of AI characteristics in iterative enhancement capability, continual receiving and learning would increase the discrimination power of the CAIADS between LSIL and

HSIL. Therefore, it should be persistently and closely observed and evaluated in further research.

In real-world clinical practice, the most important task for colposcopists is to guide colposcopic biopsy for detecting underlying HSIL+ cases for subsequent treatment. Biopsy protocols remain varied for colposcopists, and the option of biopsy thresholds remained controversial [5, 27]. In our study, the diagnostic performance of the CAIADS at different hypothetical biopsy thresholds for the detection of pathological HSIL+ was evaluated from three aspects. Firstly, the statistically significant improvements for HSIL+ detection suggested that the diagnostic performance of CAIADS can be improved by additional non-image information. As such, CAIADS could have the potential to perform a comprehensive evaluation by analysis of both images as well as non-image information for detecting HSIL+. This could reduce the risk of misdiagnosis and provide tailored colposcopic examinations individually, based on the principle of precision prevention. Secondly, we found that the identification of low-grade or worse lesions was a highly sensitive indicator for detecting HSIL+, compared with the biopsy threshold at high-grade or worse. The colposcopy-guided biopsies were required to guide subsequent treatment or management. Although the specificity would be unsatisfactory as 51.8% that some of the patients with low-grade colposcopic impression may not have HSIL+, minimizing false negative values for HSIL+ should be a priority in choosing an appropriate cutoff point. In addition, cost-effectiveness should be considered, given the high cost of cervical cancer treatment. Thirdly, the CAIADS achieved higher sensitivity and similar specificity compared with the original



colposcopic interpretation by the colposcopists in detecting HSIL+. These findings suggest that CAIADS has potential applications in assisting beginners with diagnoses, because the system extracted and learned mass and robust cervical lesion features from annotated images in terms of the pathological reports. Given the important role of choosing whether and where to place cervical biopsy sites to detect underlying disease states, automatic biopsy localization is of clinical importance. In the validation set, the CAIADS achieved a median mIoU of 0.758, which demonstrated the CAIADS could be implemented as an auxiliary biopsy location tool for colposcopists.

On the basis of robustness of CAIADS in grading colposcopic impressions and guiding biopsies, we propose the integration of CAIADS into local colposcopy clinics as an accurate and auxiliary diagnosis tool for colposcopists during colposcopic procedures. We also propose establishing a cloud-based artificial intelligence platform to provide accessible telemedical assistance for most low-resource settings, such as China (accounting for 20% of the world's population), where experienced colposcopists and colposcopy services are in short supply. Therefore, CAIADS may be expected to fill a need for standardized cervical cancer screening/diagnosis procedures, narrow the gap of diagnostic ability between tertiary hospitals and primary care hospitals, improve the quality of screening programs, and promote cooperation in scaling up coverage worldwide. Currently, CAIADS is being routinely introduced into the colposcopic clinical workflow with real-time assistance at SZMCHH, and it has been recently implemented by other hospitals, providing free access to the cloud-based artificial intelligence-aided colposcopic examination.

There are several limitations. First, although CAIADS showed satisfactory accuracy in the validation set, the design was retrospective. In this regard, a prospective study will be conducted to further validate its performance and provide evidence of cost-effectiveness in the clinical practice of cervical cancer screening. Second, we focused on the grading of colposcopic impressions for cervical neoplasia lesions in the current study. However, miscellaneous findings such as polyps, stenosis, and condyloma should be identified and biopsied in some cases. Therefore, CAIADS cannot replace clinician evaluations in grading colposcopic impressions and guiding biopsies but could assist colposcopists in clinical practice. CAIADS could be particularly helpful for less experienced colposcopists who practice in LMICs. In future studies, we will be considering the recruitment of patients with miscellaneous findings such as polyps, stenosis, and condyloma, to extract their lesion features for further training and validation of the CAIADS algorithm.

## Conclusions

The CAIADS achieved higher sensitivity and comparable specificity to colposcopies interpreted by colposcopists and also demonstrated satisfactory accuracy in guiding biopsy sites. The CAIADS has potential in assisting beginners and for improving the diagnostic quality of colposcopy and biopsy in the detection of cervical pre-cancer/cancer.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12916-020-01860-y>.

### Additional file 1: Table S1. Participating study sites. Table S2.

Detailed information of E-GCN architecture. **Table S3.** Look-up matrix mapping non-image information to features. **Figure S1.** Examples of colposcopy images for increasingly severe pathology. **Figure S2.** Cervix image annotation tool. **Figure S3.** Pipeline of system for colposcopic grading and guiding biopsies. **Figure S4.** Pipeline of semi-supervised learning framework for lesion segmentation.

## Abbreviations

AI: Artificial intelligence; ASCCP: American Society for Colposcopy & Cervical Pathology; AUC: Area under the curve; CAIADS: Colposcopic Artificial Intelligence Auxiliary Diagnostic System; CIs: Confidence intervals; HPV: Human papillomavirus; HSIL+: High-grade squamous intraepithelial lesion or worse; IRB: Institutional review board; LMICs: Low- and middle-income countries; LSIL: Low-grade squamous intraepithelial lesion; mIoU: Mean-intersection-over-union; NPV: Negative predictive value; PPV: Positive predictive value; SZMCHH: Shenzhen Maternity & Child Healthcare Hospital; WHO: World Health Organization

## Acknowledgements

We acknowledge the following pathologists in reviewing pathological reports in our study: Hongping Tang, Yanqing Kong, Yanrui Zhang, and Jian Shi (Department of Pathology, Shenzhen Maternity and Child Healthcare, Shenzhen, China), and the following colposcopists for assisting with the collection and the manual annotation of colposcopy images: Yanzhu Wang, Qifen Lin, Menglan Ye, Lifang Yang, Huanhua Wang, Qingmei Liao, Gaoping Zheng, and Kangyan Liu (Diagnosis and Treatment for Cervical Lesions Center, Shenzhen Maternity and Child Healthcare Hospital, Shenzhen, China). We also thank Qi Zhou (Chongqing University Cancer Hospital, Chongqing, China), Jun Lei (Jiangxi Maternal and Child Health Hospital, Nanchang, China), Jing Chen (Chengdu Women's and Children's Central Hospital, Chengdu, China), and Peipei Chen (Affiliated Cancer Hospital of Zhengzhou University, Zhengzhou, China) for assisting with the data collection, and Dr. Nasra Mohamoud Ali (a native English speaker) for providing assistance in editing this manuscript.

## Authors' contributions

All authors read and approved the final manuscript. Yu Jiang, Fanghui Zhao, and Youlin Qiao had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Concept and design: Peng Xue, Chao Tang, Qing Li, Yu Jiang, Fanghui Zhao, and Youlin Qiao. Acquisition, analysis, or interpretation of the data: Peng Xue, Chao Tang, Yuexiang Li, Qing Li, Yu Shen, Yuqian Zhao, and Jiawei Chen. Drafting of the manuscript: Peng Xue and Yuexiang Li. Critical revision of the manuscript for important intellectual content: Peng Xue, Yuexiang Li, Yuqian Zhao, Yefeng Zheng, Yu Jiang, Fanghui Zhao, and Youlin Qiao. Administrative, technical, or material support: Qing Li, Yu Shen, Jiawei Chen, Jianrong Wu, Longyu Li, Wei Wang, Yucong Li, Xiaoli Cui, Shaokai Zhang, Wenhua Zhang, Xun Zhang, Kai Ma, Yefeng Zheng, Man Tat Alexander Ng, Tianyi Qian, Zhihua Liu, Youlin Qiao, Yu Jiang, and Fanghui Zhao.

### Funding

This work was supported by the Chinese Academy of Medical Science Initiative for Innovative Medicine (grant CIFS2017-I2M-B&R-03); the National Key Technology R&D Program (grant 2018YFC1315504); Ministry of Science and Technology of China, the Key Area Research and Development Program of Guangdong Province, China (grant 2018B010111001); Science and Technology Program of Shenzhen, China (grant ZDSYS201802021814180); and Sanming Project of Medicine in Shenzhen (grant SZSM201612042).

### Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

### Ethics approval and consent to participate

The study was approved by the institutional review board (IRB) of Shenzhen Maternity and Child Healthcare Hospital (SZMCHH). The need for informed consent was waived by the IRB of SZMCHH due to the retrospective nature of archived datasets and fully anonymized personal information.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no conflicts of interest.

### Author details

<sup>1</sup>Department of Epidemiology and Biostatistics, School of Population Medicine and Public Health, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100730, China. <sup>2</sup>Department of Cancer Epidemiology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing 100021, China. <sup>3</sup>School of Public Health, Dalian Medical University, Dalian, China. <sup>4</sup>Diagnosis and Treatment for Cervical Lesions Center, Shenzhen Maternity & Child Healthcare Hospital, Shenzhen, China. <sup>5</sup>Tencent Jarvis Lab, Shenzhen, China. <sup>6</sup>Zonsun Healthcare, Shenzhen, China. <sup>7</sup>Center for Cancer Prevention Research, Sichuan Cancer Hospital & Institute, Sichuan Cancer Center, School of Medicine, University of Electronic Science and Technology of China, Chengdu, China. <sup>8</sup>Tencent Healthcare, Shenzhen, China. <sup>9</sup>Jiangxi Maternal and Child Health Hospital, Nanchang, China. <sup>10</sup>Chengdu Women's and Children's Central Hospital, School of Medicine, University of Electronic Science and Technology of China, Chengdu, China. <sup>11</sup>Chongqing University Cancer Hospital, Chongqing, China. <sup>12</sup>Cancer Hospital of China Medical University, Liaoning Cancer Hospital & Institute, Shenyang, China. <sup>13</sup>Affiliated Cancer Hospital of Zhengzhou University/Henan Cancer Hospital, Zhengzhou, China. <sup>14</sup>Department of Pathology, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. <sup>15</sup>Department of Gynecology, Shenzhen Maternity & Child Healthcare Hospital, Shenzhen, China.

Received: 31 August 2020 Accepted: 19 November 2020

Published online: 22 December 2020

### References

- Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394–424.
- Brisson M, Drolet M. Global elimination of cervical cancer as a public health problem. *Lancet Oncol*. 2019;20(3):319–21.
- Zhao F, Qiao Y. Cervical cancer prevention in China: a key to cancer control. *Lancet*. 2019;393(10175):969–70.
- de Sanjose S, Holme F. What is needed now for successful scale-up of screening? *Papillomavirus Res*. 2019;7:173–5.
- Schiffman M, Wentzensen N. Issues in optimising and standardising the accuracy and utility of the colposcopic examination in the HPV era. *Ecancermedicallscience*. 2015;9:530.
- Brown BH, Tidy JA. The diagnostic accuracy of colposcopy—a review of research methodology and impact on the outcomes of quality assurance. *Eur J Obstet Gynecol Reprod Biol*. 2019;240:182–6.
- Leeson SC, Alibegashvili T, Arbyn M, Bergeron C, Carriero C, Mergui JL, et al. The future role for colposcopy in Europe. *J Low Genit Tract Dis*. 2014;18(1):70–8.
- Cheung LC, Egemen D, Chen X, Katki HA, Demarco M, Wiser AL, et al. 2019 ASCCP risk-based management consensus guidelines: methods for risk estimation, recommended management, and validation. *J Low Genit Tract Dis*. 2020;24(2):90–101.
- Silver MI, Andrews J, Cooper CK, Gage JC, Gold MA, Khan MJ, et al. Risk of cervical intraepithelial neoplasia 2 or worse by cytology, human papillomavirus 16/18, and colposcopy impression: a systematic review and meta-analysis. *Obstet Gynecol*. 2018;132(3):725–35.
- Xue P, Ng MTA, Qiao Y. The challenges of colposcopy for cervical cancer screening in LMICs and solutions by artificial intelligence. *BMC Med*. 2020; 18(1):169.
- Rodriguez-Ruiz A, Lång K, Gubern-Merida A, et al. Stand-alone artificial intelligence for breast cancer detection in mammography: comparison with 101 radiologists. *J Natl Cancer Inst*. 2019;111(9):916–22.
- Derclle L, Lu L, Schwartz LH, et al. Radiomics response signature for identification of metastatic colorectal cancer sensitive to therapies targeting EGFR pathway. *J Natl Cancer Inst*. 2020;112(9):902–12.
- Luo H, Xu G, Li C, He L, Luo L, Wang Z, et al. Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. *Lancet Oncol*. 2019;20(12):1645–54.
- Hu L, Bell D, Antani S, Xue Z, Yu K, Horning MP, et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *J Natl Cancer Inst*. 2019;111(9):923–32.
- Miyagi Y, Takehara K, Nagayasu Y, Miyake T. Application of deep learning to the classification of uterine cervical squamous epithelial lesion from colposcopy images combined with HPV types. *Oncol Lett*. 2020;19(2):1602–10.
- Waxman AG, Conageski C, Silver MI, Tedeschi C, Stier EA, Apgar B, et al. ASCCP colposcopy standards: how do we perform colposcopy? Implications for establishing standards. *J Low Genit Tract Dis*. 2017;21(4):235–41.
- Wentzensen N, Massad LS, Mayeaux EJ Jr, Khan MJ, Waxman AG, Einstein MH, et al. Evidence-based consensus recommendations for colposcopy practice for cervical cancer prevention in the United States. *J Low Genit Tract Dis*. 2017;21(4):216–22.
- Li Y, Chen J, Xue P, Tang C, Chang J, Chu C, et al. Computer-aided cervical cancer diagnosis using time-lapsed colposcopic images. *IEEE Trans Med Imaging*. 2020;39(11):3403–15.
- Shibata N, Tanito M, Mitsuhashi K, et al. Development of a deep residual learning algorithm to screen for glaucoma from fundus photography. *Sci Rep*. 2018;8(1):14665.
- Zhang S, Tong H, Xu J, et al. Graph convolutional networks: a comprehensive review[J]. *Computational Social Networks*. 2019;6(1):11.
- Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. Paper presented at: International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015. <https://pdfs.semanticscholar.org/0704/5f87709d0b7b998794e9fa912c0aba912281.pdf>. Accessed 13 Nov 2019.
- J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," *IEEE Conference on Computer Vision and Pattern Recogn*, 2016, pp. 779–788. <https://arxiv.org/pdf/1506.02640.pdf>. Accessed 25 Nov 2019.
- Pedregosa F, Varoquaux G, Gramfort A, Vincent M, Bertrand T, Olivier G, et al. Scikit-learn: machine learning in Python. *J Mach Learn Res*. 2011;12(10): 2825–30.
- Fan A, Wang C, Zhang L, Yan Y, Han C, Xue F. Diagnostic value of the 2011 International Federation for Cervical Pathology and Colposcopy Terminology in predicting cervical lesions. *Oncotarget*. 2018;9(10):9166–76.
- Jeronimo J, Schiffman M. Colposcopy at a crossroads. *Am J Obstet Gynecol*. 2006;195(2):349–53.
- Benedet JL, Matisic JP, Bertrand MA. An analysis of 84244 patients from the British Columbia cytology-colposcopy program. *Gynecol Oncol*. 2004;92(1): 127–34.
- Massad LS, Jeronimo J, Katki HA, Schiffman M. The accuracy of colposcopic grading for detection of high-grade cervical intraepithelial neoplasia. *J Low Genit Tract Dis*. 2009;13(3):137–44.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.