PLOS DIGITAL HEALTH

RESEARCH ARTICLE

Assessing generalizability of an AI-based visual test for cervical cancer screening

Syed Rakin Ahmed[1,2,3,4‡*], Didem Egemen[5‡], Brian Befano[6,7], Ana Cecilia Rodriguez[5], Jose Jeronimo[5], Kanan Desai[5], Carolina Teran[8], Karla Alfaro[9], Joel Fokom-Domgue[10,11,12], Kittipat Charoenkwan[13], Chemtai Mungo[14], Rebecca Luckett[15], Rakiya Saidu[16], Taina Raiol[17,18], Ana Ribeiro[17,18], Julia C. Gage[19], Silvia de Sanjose[5,20], Jayashree Kalpathy-Cramer[1,21‡], Mark Schiffman[5‡]

1 Athinoula A. Martinos Center for Biomedical Imaging, Department of Radiology, Massachusetts General Hospital, Boston, Massachusetts, United States of America, **2** Harvard Graduate Program in Biophysics, Harvard Medical School, Harvard University, Cambridge, Massachusetts, United States of America, **3** Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **4** Geisel School of Medicine at Dartmouth, Dartmouth College, Hanover, New Hampshire, United States of America, **5** Clinical Epidemiology Unit, Clinical Genetics Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **6** Information Management Services, Calverton, Maryland, United States of America, **7** University of Washington, Seattle, Washington, United States of America, **8** Facultad de Medicina, Universidad Mayor, Real y Pontificia de San Francisco Xavier de Chuquisaca, Sucre, Bolivia, **9** Basic Health International, El Salvador, **10** Cameroon Baptist Convention Health Services, Bamenda, North West Region, Cameroon, **11** Department of Obstetrics and Gynecology, Faculty of Medicine and Biomedical Sciences, University of Yaoundé, Yaoundé, Cameroon, **12** Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, United States of America, **13** Department of Obstetrics and Gynecology, Chiang Mai University, Chiang Mai, Thailand, **14** Department of Obstetrics and Gynecology, University of North Carolina-Chapel Hill School of Medicine, Chapel Hill, North Carolina, United States of America, **15** Beth Israel Deaconess Medical Center, Boston, Massachusetts, United States of America, **16** Department of Obstetrics and Gynaecology and South African Medical Research Council Gynaecological Cancer Research Centre, University of Cape Town, Cape Town, **17** Center for Epidemiology and Health Surveillance, Oswaldo Cruz Foundation (Fiocruz), Brasília, Federal District, Brazil, **18** MARCO Clinical and Molecular Research Center, University Hospital of Brasilia/EBSERH, Federal District, Brazil, **19** Center for Global Health, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, United States of America, **20** ISGlobal, Barcelona, Spain, **21** Department of Ophthalmology, University of Colorado Anschutz, Denver, Colorado, United States of America

‡ SRA and DE share first authorship on this work. JK-C and MS are joint senior authors on this work.
* syedrakin_ahmed@fas.harvard.edu

OPEN ACCESS

Citation: Ahmed SR, Egemen D, Befano B, Rodriguez AC, Jeronimo J, Desai K, et al. (2024) Assessing generalizability of an AI-based visual test for cervical cancer screening. PLOS Digit Health 3(10): e0000364. https://doi.org/10.1371/journal.pdig.0000364

Editor: J. Mark Ansermino, University of British Columbia, CANADA

Received: September 8, 2023

Accepted: July 16, 2024

Published: October 2, 2024

Data Availability Statement: The materials used to train and generate results can be found at the following repository: https://github.com/QTIM-lab/cervix_generalizability.

Funding: This work was supported by the National Cancer Institute (NCI) of the National Institutes of Health (NIH). All NCI-affiliated staff are supported by the NCI Intramural Research Program including supplemental funding from the Cancer Cures Moonshot Initiative. Additionally, BB was supported by NCI/NIH under Grant T32CA09168.

Abstract

A number of challenges hinder artificial intelligence (AI) models from effective clinical translation. Foremost among these challenges is the lack of generalizability, which is defined as the ability of a model to perform well on datasets that have different characteristics from the training data. We recently investigated the development of an AI pipeline on digital images of the cervix, utilizing a multi-heterogeneous dataset of 9,462 women (17,013 images) and a multi-stage model selection and optimization approach, to generate a diagnostic classifier able to classify images of the cervix into "normal", "indeterminate" and "precancer/cancer" (denoted as "precancer+") categories. In this work, we investigate the performance of this multiclass classifier on external data not utilized in training and internal validation, to assess the generalizability of the classifier when moving to new settings. We assessed both the classification performance and repeatability of our classifier model across the two axes of

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

heterogeneity present in our dataset: image capture device and geography, utilizing both out-of-the-box inference and retraining with external data. Our results demonstrate that device-level heterogeneity affects our model performance more than geography-level heterogeneity. Classification performance of our model is strong on images from a new geography without retraining, while incremental retraining with inclusion of images from a new device progressively improves classification performance on that device up to a point of saturation. Repeatability of our model is relatively unaffected by data heterogeneity and remains strong throughout. Our work supports the need for optimized retraining approaches that address data heterogeneity (e.g., when moving to a new device) to facilitate effective use of AI models in new settings.

Author summary

Artificial intelligence (AI) model robustness has emerged as a pressing issue, particularly in medicine, where model deployment requires rigorous standards of approval. In the context of this work, model robustness refers to both the repeatability of model predictions across repeat images, as well as the generalizability of model performance to external data. Real world clinical data is often heterogeneous across multiple axes, with distribution shifts in one or more of these axes often being the norm. Current deep learning (DL) models for cervical cancer and in other domains exhibit poor repeatability and overfitting, and frequently fail when evaluated on external data. As recently as March 2023, the FDA issued a draft guidance on effective implementation of AI/DL models, proposing the need for adapting models to data distribution shifts. To surmount known concerns, we conducted a thorough investigation of the generalizability of a deep learning model for cervical cancer screening, utilizing the distribution shifts present in our large, multi-heterogenous dataset. We highlight optimized strategies to adapt an AI-based clinical test, which in our case was a cervical cancer screening triage test, to external data from a new setting. Given the severe clinical burden of cervical cancer, and the fact that existing screening approaches, such as visual inspection with acetic acid (VIA), are unreliable, inaccurate, and invasive, there is a critical need for an automated, AI-based pipeline that can more consistently evaluate cervical lesions in a minimally invasive fashion. Our work represents one of the first efforts at generating and externally validating a cervical cancer diagnostic classifier that is reliable, consistent, accurate, and clinically translatable, in order to triage women into appropriate risk categories.

Introduction

The development of artificial intelligence (AI) and deep learning (DL) approaches have become seemingly ubiquitous in recent years, across several clinical domains, with optimized models reporting near-clinician-level performance [1–4]. However, translation of AI models from bench to bedside remain sparse. To be clinically translatable, AI/DL models should be robust, computationally-efficient, low-cost, and blend well with existing clinical workflows, ensuring the inputs/outputs of the model and the task it performs are most relevant to the clinician for a given use case. This is often not the case with existing models, which are frequently hindered by several key methodological flaws in their design [5], thereby undermining their

validity, and hindering clinical translation. In particular, model robustness has emerged as a key challenge hindering AI model deployment from bench to clinical practice.

In the context of this work, model robustness refers to two key attributes: 1. repeatability, defined as the ability of a model to generate near-identical predictions for the same patient under identical conditions, ensuring that the model produces precise, reliable outputs in the clinical setting [6]; and 2. generalizability, defined as the ability of a model to adapt well to domain expansion or, alternatively, the ability of a model to perform well on datasets that are out of distribution from the training data, i.e., having different characteristics from training data [7]. There is a paucity of work in the current DL and medical image classification literature that assess one or both of these attributes, with many models tending to overfit to the training data distribution. This is either due to 1) the absence of data heterogeneity (geography-, institution-, population- and/or device-level) in the available training data for a given use case; and/or 2) the absence of specific optimization approaches to reduce overfitting. To assess whether a model is overfit, an external dataset is required which has different characteristics from the training set. Assessing overfitting is particularly important when considering AI model deployment for use cases that are likely to involve multiple axes of data heterogeneity.

Globally, cervical cancer is the fourth most common cause of cancer morbidity and mortality, with approximately 90% of the 300,000 deaths per year occurring in low-resource settings [8–10]. Even though the causal pathway to cervical cancer is well understood, with HPV being the main cause [9,11,12], this cancer has not yet been controlled, especially in low-resource settings [13]. The primary prevention strategy is HPV vaccination, and for the secondary prevention strategy the World Health Organization (WHO) recommends screening with HPV test [14,15]. In order to triage the risk of HPV-positive individuals, visual inspection with acetic acid (VIA) is used in low-resource settings [16,17]. However, many studies have shown that expert visual evaluation has mediocre accuracy and repeatability [18,19]. Therefore, there is a need for a highly accurate, repeatable, low-cost, point-of-care visual screening test to triage the risks of HPV-positive individuals. To address this need, we previously conducted a comprehensive, multi-stage model selection and optimization approach, utilizing a large, collated multi-institution, multi-device, and multi-population dataset, in order to generate a diagnostic classifier model, termed automated visual evaluation (AVE) that is able to classify images of the cervix into "normal", "indeterminate" (interchangeably termed as "gray zone") and "precancer/cancer" (denoted as "precancer+") categories [20].

The objectives of the present study are to highlight the relative impacts of the known axes of data heterogeneity present in our dataset and to assess the generalizability of AVE on multiple external datasets; specifically, we assessed both classification performance and repeatability of AVE, utilizing various retraining and inference strategies. Our approaches are directed by the known distribution shifts present in our external dataset, in the form of device and geography. We hypothesize that these two distribution shifts present in our data will impact the classification performance of our model differently, while the repeatability of our model will be unaffected by distribution shifts. Additionally, we further hypothesize that the classification performance of our model will improve following retraining via optimized strategies with additional images from the new distribution.

Materials and methods

In this paper, we utilized a model that we developed in a prior study, following a multi-stage model selection and optimization process utilizing a multi-heterogeneous dataset, henceforth referred to as "SEED" [20]. The primary discernible axes of heterogeneity in this prior work included image capture device and geography. In the current study, we conducted a thorough

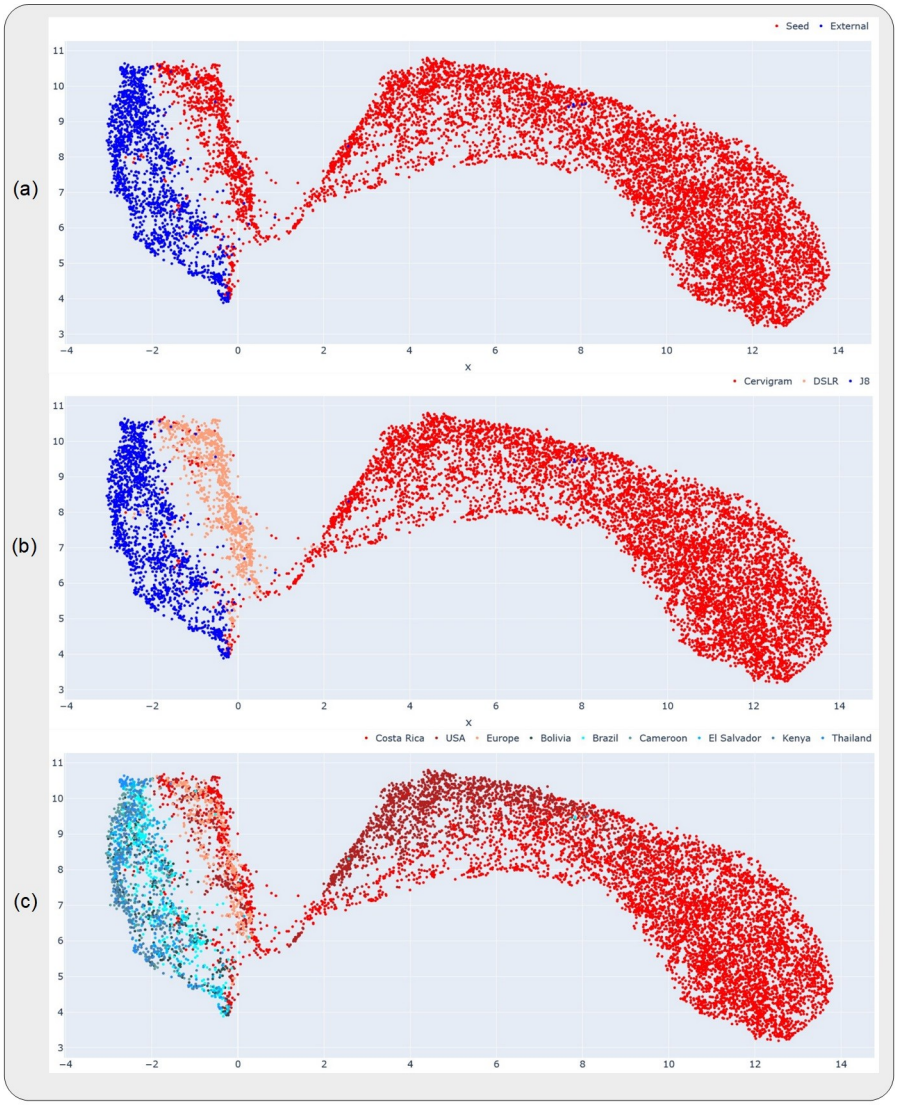

Fig 1. Uniform manifold approximation and projections (UMAP) highlighting the relative distributions of the datasets, devices and geographies investigated in this work. Each subplot highlights a different representation of the UMAP, where the color coding (highlighted in the corresponding legend at the top of each subplot) is at the (a) dataset-level, (b) device-level and (c) geography-level. The datasets and devices occupy distinct clusters in (a) and (b), while the geographies are all clustered together within the same device in (c). The x- and y-axes are in arbitrary units, representing the two UMAP components on which the higher dimensional data was projected.

https://doi.org/10.1371/journal.pdig.0000364.g001

external validation of our model by running the model on images collected from a new, external dataset, henceforth termed "EXT". The "EXT" dataset used a different image capture device, Samsung Galaxy J8, from those of the SEED (Fig 1), and also constituted six distinct geographies/countries (Table 1 and Fig 1). All of these countries are listed in the low- and middle-income countries (LMIC) classification of the World Bank and IMF [21].

Dataset

Analysis population. We utilized two groups of datasets in this study: 1) a collated, multi-institutional and multi-device (cerviscope, DSLR) dataset that was previously utilized in the

Table 1. Breakdown of dataset by ground truth and geography.

DATASETS	Ground truth categories (no. (%))												GRAND TOTAL (n_i = 1669, n_w = 580) (no. (%))			
	Normal (n_i = 598, n_w = 204)				Indeterminate / Gray (n_i = 465, n_w = 157)				Precancer+ (n_i = 606, n_w = 219)							
	# images		# women		# images		# women		# images		# women		# images		# women	
Bolivia	140	(5.8%)	40	(3.1%)	118	(5.5%)	33	(2.4%)	59	(5.5%)	15	(2.4%)	317	(19.0%)	88	(15.2%)
Brazil	0	(0.0%)	0	(0.0%)	0	(0.0%)	0	(0.0%)	410	(37.9%)	154	(24.2%)	410	(24.6%)	154	(26.6%)
Cameroon	231	(9.5%)	85	(6.7%)	33	(1.5%)	13	(0.9%)	2	(0.2%)	2	(0.3%)	266	(15.9%)	100	(17.2%)
El Salvador	130	(5.3%)	49	(3.9%)	0	(0.0%)	0	(0.0%)	56	(5.2%)	21	(3.3%)	186	(11.1%)	70	(12.1%)
Kenya	0	(0.0%)	0	(0.0%)	309	(14.3%)	109	(7.8%)	36	(3.3%)	13	(2.0%)	345	(20.7%)	122	(21.0%)
Thailand	97	(4.0%)	30	(2.4%)	5	(0.2%)	2	(0.1%)	43	(4.0%)	14	(2.2%)	145	(8.7%)	46	(7.9%)
TOTAL	598	(24.6%)	204	(16.0%)	465	(21.6%)	157	(11.3%)	606	(56.0%)	219	(34.4%)	1669	(100.0%)	580	(100.0%)

Detailed breakdown of "EXT" dataset by ground truth class and geography. n_i = total # images; n_w = total # women

https://doi.org/10.1371/journal.pdig.0000364.t001

model development work, which comprised of a convenience sample combining five distinct studies–Natural History Study (NHS), ASC-US/LSIL Triage Study for Cervical Cancer (ALTS), Costa Rica Vaccine Trial (CVT), Biopsy Study in the US (Biop), and Biopsy Study in Europe (D Biop) [20]; we denote this dataset as "SEED", and 2) an external multi-geography dataset of images taken by Samsung Galaxy J8 smartphones, from six countries–Bolivia, Brazil, Cameroon, El Salvador, Kenya and Thailand; we denote this dataset as "EXT". All sites in "EXT" (except Brazil) was collected as part of the AVE Network Project, where none of the images were available/used at the initial training, validation, and testing phases of the AVE algorithm. In all six countries, cervical images were collected at the vaginal exam using a Samsung Galaxy J8 smartphone. Referral for a vaginal exam was due to human papillomavirus (HPV) positivity in Bolivia and El Salvador, with additional cervical images from El Salvador collected from a randomly selected group of HPV-negative individuals. In Cameroon, Kenya, and Thailand, images were collected from VIA positive individuals at the triage visit. In Brazil, images were collected from patients with histologically confirmed cervical intraepithelial neoplasia (CIN) 2 or worse lesions. HPV tests used in these countries were Hybrid Capture 2 (HC2) [22] for Bolivia, AmpFire [23] for Cameroon, and Care HPV [24] for El Salvador and Kenya. In Thailand, no HPV test was used for screening, however cytology was utilized in addition to VIA. Histopathologic confirmation of cervical cancer status in these countries were available in the form of CIN 2, CIN 3, adenocarcinoma in situ (AIS), and cervical cancer. In Brazil, images were collected after application of acetic acid and prior to Loop Electrosurgical Excision Procedure (LEEP).

Ground truth delineation. The ground truth values for the "EXT" dataset was assigned in a similar manner to that used for the "SEED" dataset [20]. Specifically, the three ground truth values mapped to the images, "normal", "indeterminate" and "precancer+", were based primarily on histology and HPV results. All images ≥CIN 3 were assigned to precancer/cancer. If images were CIN 2, high-risk HPV (hrHPV) positivity was used to determine classes for images from all sites except Brazil: hrHPV+ was assigned to the "precancer+" class, and hrHPV- was assigned to the "indeterminate" class. All images from Brazil were >CIN 2 and were assigned to the "precancer+" class. For images where the histopathology result is <CIN 2 or missing, the ground truth class (either "normal" or "indeterminate") was determined by a joint evaluation of a local clinician and an NCI expert colposcopist review in a site-specific manner. The final result of the ground truth distribution across each of the geographies, in terms of both the number of individuals and the number of images is depicted on Table 1.

Ethics. All study participants signed a written informed consent prior to enrollment and sample collection. All studies were reviewed and approved by the Institutional Review Boards of the National Cancer Institute (NCI) and the National Institutes of Health (NIH).

Model training and analysis

Prior to any model runs, all images were cropped with bounding boxes generated from a YOLOv5 [25] model trained for cervix detection on the "SEED" dataset images (Fig 2), resized to 256x256 pixels, and scaled to intensity values from 0 to 1. For the retraining runs, affine transformations were applied to the image for data augmentation.

We assessed the generalizability of our model by conducting two distinct sets of investigations. In the first set, we analyzed the relative impacts of device- and geography-level heterogeneities of our dataset on model performance, both visually via uniform manifold approximation and projection (UMAP), and statistically via assessing key classification performance and repeatability metrics.

First, in order to get a sense of the dataset distributions of the "SEED" and "EXT" datasets, including the distributions by device and geography, we ran out-of-the-box (OOB) inference with our initial model on the held-aside test set of the "SEED" dataset and on the full "EXT" dataset. We subsequently plotted UMAPs of the resulting features, which represent a dimension-reduced version of the features output from the model's inference run, color-coded by dataset, device, and geography (Fig 1) respectively.

We further tested the impact of device- and geography-level heterogeneity on our model performance via three distinct model runs: (i) OOB inference of AVE on a test set comprising only of "SEED" images; (ii) OOB inference of AVE on a test set comprising only of "EXT" (J8) images; and (iii) training a model using the same hyperparameters as AVE but on both "SEED" images and "EXT" images from all geographies except Bolivia and testing on Bolivia images (Fig 3).

In the second set of the generalizability analyses, we closely assessed the overall performance of AVE on "EXT" (J8) by incrementally adding women from "EXT" to our training set

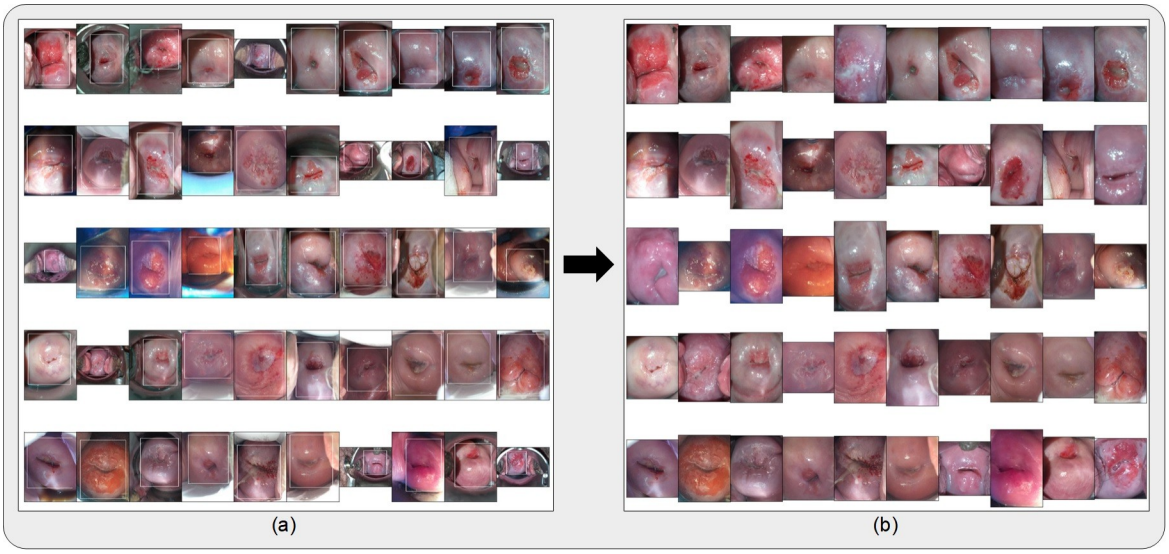

Fig 2. (a) Bounding boxes generated from running the cervix detector, highlighted in white, around 50 randomly selected images from the external ("EXT") dataset. The cervix detector utilized a YOLOv5 architecture trained on the "SEED" dataset images. (b) Bound and cropped images of the cervix which are passed onto the diagnostic classifier (AVE).

https://doi.org/10.1371/journal.pdig.0000364.g002

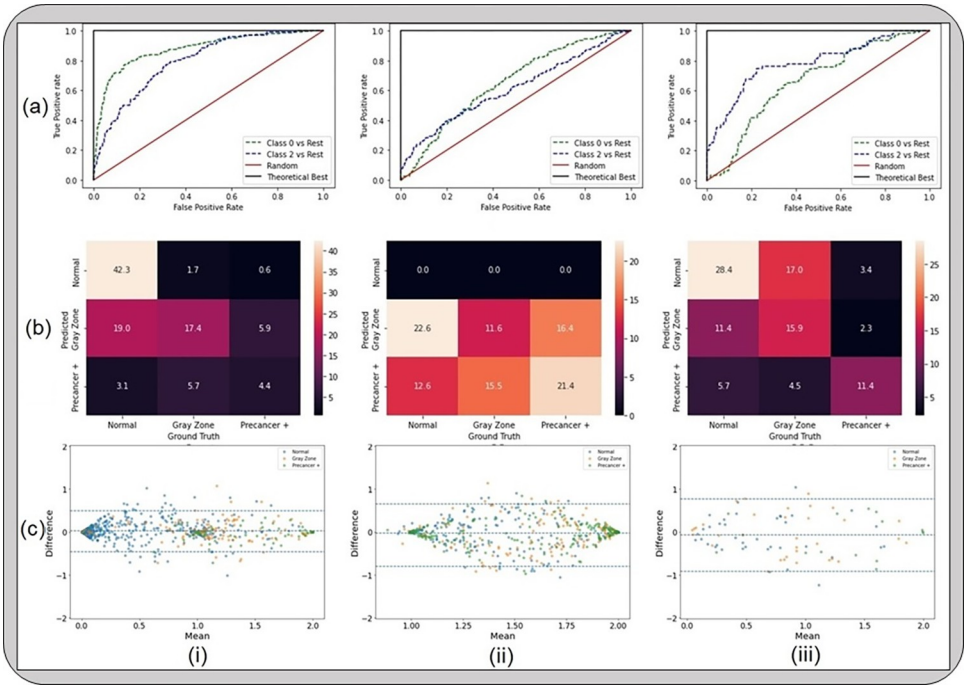

Fig 3. Results from the first set of the generalizability analyses, highlighting that device level heterogeneity impacts our model performance greater than geography level heterogeneity. The classification performance and repeatability plots depicted here include (a) receiver operating characteristics (ROC) curves; (b) confusion matrices; and (c) Bland-Altman plots, for models that were (i) trained on "SEED" and tested on a held-aside set from "SEED"; (ii) trained on "SEED" and tested on "EXT"; and (iii) trained on a dataset comprising of "SEED" + all images from "EXT" except Bolivia and tested on Bolivia images from "EXT". "Gray Zone" = "Indeterminate".

https://doi.org/10.1371/journal.pdig.0000364.g003

of "SEED" images, training on the combined set comprising of "SEED" and "EXT" images and testing on a common, held-aside set of "EXT" women (230 women, 644 images). Specifically, we added images at the woman level in two distinct ratios of ground truth– 1 normal (N): 1 indeterminate (I): 1 precancer+ (P), and 2 N: 2 I: 1 P; our intuition behind these additions were twofold: 1) we sought to minimize the number of precancer+ women needed when conducting a study utilizing a new device, and 2) we intended to mimic the ground truth balancing utilized in our model development work, which used a 2 N: 2 I: 1 P ratio of ground truths during training on "SEED", and evaluate whether matching the same balancing strategy as in the "SEED" set has any influence on the model performance. The specific increments of women added are highlighted in Fig 4 and Table 2. We assessed the classification performance of the retrained models via the area under the receiver operating characteristics curve (AUROC) (Fig 4), and the degree of extreme misclassifications (normal misclassified as precancer+ and vice versa) and total misclassifications (Table 2). We also assessed the repeatability of these models via the degree of extreme disagreement (% 2-class disagreement between image pairs across women) and the 95% limits of agreement (LoA) on a Bland-Altman plot (Table 2). S1 Text and S1 Fig further highlight the improvements in repeatability and classification performance imparted by the key innovations of our model.

Finally, to aid better visualization of predictions at the individual model level, we generated the plots on Fig 4A which compare model predictions across 60 images for each of the retrained models. To generate this comparison, we first summarized each model's output as a continuous severity *score*. Specifically, we utilized the ordinality of our problem and defined the continuous severity *score* as a weighted average using softmax probability of each class as

described in Equation 3, where k is the number of classes and p_i the softmax probability of class i.

$$score = \sum_{i=0}^{k} p_i \times i$$

Put another way, the *score* is equivalent to the expected value of a random variable that takes values equal to the class labels, and the probabilities are the model's softmax probability at index i corresponding to class label i. For a three-class model, the values lie in the range 0 to 2. We next computed the average of the *score* for each image across all 16 models compared and arranged the images in order of increasing *score* within each class. From this *score*-ordered list, we randomly selected 20 images per class, maintaining the distribution of mean scores within each class, and arranged the images in order of increasing average *score* within each class in the top row of Fig 4A (i and ii), color coded by ground truth. We subsequently compared the predicted class across the models for each of these 60 images (bottom 16 rows of Fig 4A), maintaining the images in the same order as the ground truth row and color-coded by model predicted class, enabling us to gain a deeper insight into model performance.

Results

Our results highlight two critical findings in terms of model generalizability, which, we believe, hold relevance even outside of cervical imaging, as noted below:

a. Device-level heterogeneity impacts model performance greater than geography level heterogeneity. Our model performs well out of the box (no retraining) on external datasets where the axis of heterogeneity is geography only vs. device, i.e., on images from a different geography but sharing a device that is represented in the training dataset. However, the repeatability of our model is unaffected by data heterogeneity and is strong throughout.

b. Incremental retraining with inclusion of new device images to the training dataset progressively improves classification performance and class discrimination on images from a new device previously not incorporated in the training dataset, up to a point of saturation.

First, the UMAPs on Fig 1A and 1B highlight that the "EXT" dataset and its corresponding J8 device (blue) occupy a different cluster from the "SEED" dataset and its corresponding devices (red and salmon), while Fig 1C highlights the geography level distribution. Taken together, Fig 1A and 1B and Fig 1C suggests the relatively greater impact of device-level heterogeneity on model performance than geography-level heterogeneity, given that within the same device, different geographies do not occupy distinct clusters on Fig 1C, unlike the corresponding device level clusters on Fig 1B, which are distinct. S2 Fig highlights the device and geography level characteristics of our "SEED" and "EXT" datasets.

This is further reinforced by Fig 3, which highlights the results from the model runs designed to investigate the relative impacts of device- and geography-level heterogeneity. Fig 3 illustrates that, for our model, device level heterogeneity impacts model performance greater than geography level heterogeneity. Specifically, column (i) of Fig 3 highlights that our model performs well when running out-of-the-box inference on images that are acquired using devices that are represented in the SEED data utilized in training our model (AUROC Normal vs. Rest = 0.88, AUROC Precancer+ vs. Rest = 0.82). However, when tested on a new device, J8, out-of-the-box (OOB) inference using the same model trained on seed data fails, as indicated by the poor classification performance of our model (AUROC Normal vs. Rest = 0.65; AUROC Precancer+ vs. Rest = 0.60; no normal predictions as highlighted by the confusion

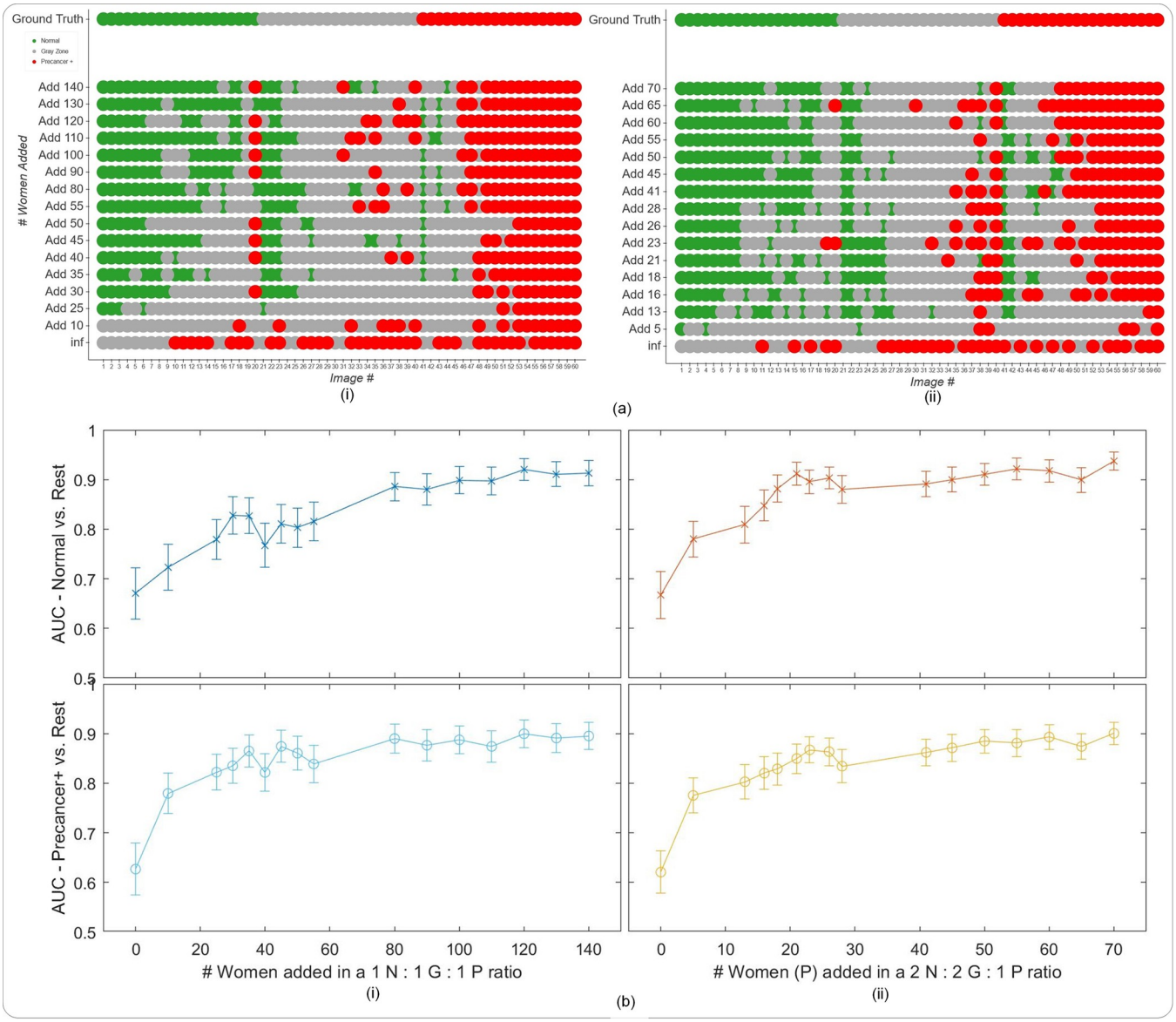

Fig 4. Results from the second set of generalizability analysis, highlighting that retraining can improve performance on a new device previously not present in the "SEED". (a) Model level comparison across models representing incremental additions of "EXT" (J8) images at the woman level to the training set of "SEED" images, with the "EXT" images added in (i) a 1n normal (N): 1n indeterminate (I): 1n precancer+ (P) ratio; and (ii) a 2n N: 2n I: 1n P ratio of ground truth classes at the woman level, where n = # of precancer+ women added (y-axes) (b) Plots of area under receiver operating characteristics curve (AUC) vs. # women added to the training set per ground truth class, in the same ratios as in (a). For example, in (ii), the x-axis represents the # precancer+ (P) women added (n) in the ratio 2n N: 2n I: 1n P to the training set. The top row plots the Normal (class 0) vs. Rest AUC, while the bottom row plots the Precancer+ (class 2) vs. rest AUC, respectively, on the y-axis. In panel (a) "normal" = green, "indeterminate" / "gray zone" = gray and "precancer+" = red.

https://doi.org/10.1371/journal.pdig.0000364.g004

matrix) on column (ii) of Fig 3. Column (iii) of Fig 3 highlights that our model performs well when trained on images from a training set that includes the seed data and J8 images from all geographies except Bolivia and tested on J8 images from Bolivia (AUROC Normal vs. Rest = 0.70, AUROC Precancer+ vs. Rest = 0.79). This trend in classification performance is also reflected in the confusion matrices on row (b) of Fig 3, where column (i) and column (iii) have extreme misclassification rates of 3.7% and 9.1% respectively, while column (ii) shows the model making only "indeterminate" and "precancer+" predictions, and no "normal"

predictions. Finally, row (c) illustrates the strong repeatability performance of our model in all of the cases (i)–(iii), highlighted by the small width of the 95% limits of agreement (95% LoA) on each corresponding Bland-Altman plot (95% LoA = 0.24, 0.36, 0.42 respectively). Each point on a Bland-Altman plot refers to a single woman, with the y-axis representing the maximum difference in the continuous classification *score* (see METHODS) across repeat images per woman, and the x-axis plotting the mean of the corresponding *score* across all repeat images per woman. Repeatability is evaluated using the 95% LoA, highlighted by the blue dotted lines on either side of the mean (central blue dotted line); for each of the Bland-Altman plots in row (c), the 95% LoA is quite narrow, with most points clustered around 0 on the y-axis suggesting that *score* values of the model on repeat images taken on the same visit for each woman are quite similar. Taken together, these results suggest that the classification performance of our model is affected more by device differences than differences in geography, while repeatability is relatively unaffected and quite strong throughout.

Fig 4 illustrates that, given the impact of device level heterogeneity on the performance of our model, retraining can improve performance on a new device previously not present in the "SEED". Specifically, incremental retraining with inclusion of J8 images to the seed data, where training set = "SEED" images + J8 images, progressively improves classification performance and class discrimination on a held-aside test set consisting only of J8 images, up until a point of saturation. Panel (a) of Fig 4 highlights this finding by providing a detailed comparison of model performance at the individual image level. Here (i) represents the case where the J8 images were added in a 1n N: 1n G: 1n P ratio of ground truth class at the woman level, while (ii) represents J8 additions in a 2n N: 2n G: 1n P ratio of ground truth classes at the woman level (the y-axis represents n, or the number of precancer+ women added). In both cases, incremental addition of new device images to the training set improves class discrimination; this improvement is achieved with fewer precancer+ cases added to the training set in (ii), with the 2:2:1 ratio. Panel (b) plots the AUROCs (both normal vs. rest and precancer+ vs. rest) against number of women added in the training set for each of the two corresponding ratios together with bootstrapped confidence intervals for each AUROC value, further reinforcing the finding that our model performance on J8 images improves with increased representation of J8 images in the training set in a saturating fashion. As we add more J8 images at the woman level, both the normal vs. rest and precancer+ vs. rest AUROCs increase up to a plateau of around 0.9.

Table 2 highlights key classification (% extreme misclassifications and % total misclassifications) and repeatability (% extreme disagreement and 95% LoA) metrics for the case where J8 images are added to the training set in a 2 N: 2 G: 1 P ratio at the woman level. Specifically, the decrease in % total misclassifications with progressive addition of J8 images in the training set further illustrates the improvement in classification performance. On the other hand, the repeatability of our model is quite strong and relatively consistent throughout, as highlighted by the consistently low % extreme disagreement and 95% LoA values in Table 2. Additionally, model performance on the original "SEED" set images remains consistently strong regardless of the number of women added from the "EXT" dataset, as highlighted by the high AUROC values, the low % extreme misclassification and % total misclassifications, and the low % extreme disagreement and 95% LoA across all increments of "EXT" on S1 Table. This suggests that our model does not exhibit any catastrophic forgetting.

Discussion

The use of AI models as possible biomarkers continue to be hindered by key factors that affect their clinical translation. To be effective, any biomarker needs to: 1. generate reproducible test

Table 2. Classification and Repeatability Metrics.

# added	Classification		Repeatability	
	% ext. mis.	% tot. mis.	% ext. dis.	95% LoA
Add 00 (inf)	N/A	N/A	N/A	N/A
Add 05	N/A	N/A	N/A	N/A
Add 13	7.8%	65.7%	0.0%	0.3
Add 16	9.6%	53.0%	1.9%	0.4
Add 18	7.9%	55.2%	1.0%	0.4
Add 21	4.4%	51.3%	0.5%	0.4
Add 23	7.4%	39.1%	1.0%	0.4
Add 26	4.8%	46.1%	0.0%	0.4
Add 28	6.0%	55.2%	0.5%	0.4
Add 41	8.7%	37.8%	1.9%	0.4
Add 45	7.0%	44.8%	1.5%	0.4
Add 50	11.3%	39.1%	1.9%	0.4
Add 55	9.6%	38.3%	2.4%	0.4
Add 60	6.5%	33.9%	1.9%	0.4
Add 65	6.0%	39.1%	1.0%	0.4
Add 70	6.0%	27.8%	1.9%	0.4

Relevant classification performance metrics, including % extreme misclassifications (% ext. mis.) and % total misclassifications (% tot. mis.), and repeatability metrics, including % extreme disagreement (% ext. dis.) and 95% limits of agreement (LoA) on a Bland Altman plot, for each of the model runs involving incremental additions of images from the "EXT" (J8) dataset at the woman level. Here the metrics are presented for the incremental additions in a 2n normal (N): 2n indeterminate (I): n precancer+ (P) ratio of ground truth class, where n = # of precancer + women added, as shown on the leftmost column. All values are rounded to 1 decimal place.

https://doi.org/10.1371/journal.pdig.0000364.t002

results; 2. acknowledge uncertainty, particularly when the underlying predictive task has pre-existing uncertainty (e.g., ASCUS in the Bethesda system); and 3. acknowledge the need for, or the lack of, generalizability to data heterogeneities. In this work, we address each of these properties in turn via first investigating the key axes of heterogeneities present in the underlying data, and subsequently demonstrating that the key design innovations of our multiclass AVE model are optimized for improved repeatability and classification performance and can translate well into new settings in order to facilitate clinical decision-making.

Our work demonstrates proof of principle on adapting an AI-based clinical test, which in our case was a cervical cancer screening triage test, to a new setting. Both "internal" and "external" validation of AI models, particularly for models that are intended for clinical translation and deployment across heterogeneous data, are essential for fair evaluation of model performance [26]. In the context of our work, "internal" validation refers to assessing model performance on data that shares similar distributional characteristics to the training data (e.g., same device, same geography, same population), while "external" validation uses datasets that are out-of-distribution [27]. In the large majority of cases, the training data that is available for an AI model is homogeneous and does not often match the intended use case. Additionally, data drift or covariate shift, a phenomenon where the distribution of input data to an AI model changes over time, can significantly impact model performance following deployment [7,28,29]. This is particularly consequential in a clinical setting, where an inaccurate model prediction can lead to a cascade of potentially harmful downstream clinical decisions which might impact the health and safety of a patient. In this work, we posit that assessing AI model

performance requires thorough consideration of both repeatability of predictions, and the discrimination ability of the model, when evaluated on "external" data from a new setting.

In this work, we demonstrate that our model is able to discriminate between classes ("normal", "indeterminate", "precancer+") well when evaluated on external data without retraining, provided that the axis of heterogeneity is geography only. If the external data is from a new device, our model performance improves as we incrementally add images collected from additional individuals from the external dataset and retrain on the collated training set. The specific retraining approach used, in particular, a ground truth ratio of women added to match the corresponding ratio in the "SEED" used for baseline model training, also determines the extent of this improvement. Additionally, as Fig 4 highlights, this performance improvement eventually reaches a saturation point. Overall, these findings have important implications for clinical deployment: in order to deploy our model to a new setting which uses a different image capture device from the family of devices utilized in model training, we would need to retrain our model, via optimized strategies, with a small portion of labelled images acquired using the new device; however, this is not needed if the new setting only differs in terms of geography. We can therefore expect our model to generalize well across diverse geographies without the need for retraining, provided that the image capture device used is represented in the training set. This is a critical and impactful result, which implies that standardizing an image capture device should minimize the need for retraining.

Our work also sheds light on the potential importance of local retraining and the need for adapting models to local needs and constraints. The concept of creating generalizable models may not necessarily fit with AI as applied to healthcare. Using global models that are expected to generalize across all known axes of data heterogeneity might, in fact, further worsen the bias and health disparities between high-resource and low-resource settings. Retraining local models that are tailored to specific regions or institutions, which may be using a specific image-capture device, may enable better coordination of the various stakeholders involved in model design and implementation, and might better enable clinical interventions that are specific to the local context. There may also be greater buy-in from local providers when models are retrained and adapted locally.

Despite the heterogeneous nature of our dataset, our work may be limited by the number of external devices utilized. Forthcoming work will evaluate our retraining approaches and assess model performance on additional external devices. Further, despite our utilization of a large, comprehensive dataset, our work may also be limited by the known number of axes of data heterogeneity investigated. Future work will investigate the impact of additional axes of data heterogeneity on model performance by acquiring new, prospective data, and additionally optimize our model for use on edge devices, thereby promoting the possibility of translation into relevant clinical settings.

Supporting information

S1 Text. Supplementary Methods And Results–Repeatability And Classification Performance Analysis.
(DOCX)

S1 Fig. Results from the repeatability and classification performance analysis. (a) Bland-Altman plots; and (b) Test-Retest score plots for each of the four models under investigation namely (i) binary; (ii) binary with MC dropout; (iii) multiclass; and (iv) multiclass with MC dropout (our model), in order to assess the relative impact of the key design choices of our model. Panel (c) (1) highlights the % extreme disagreement (proportion of women for whom the model predicts "normal" for image 1 and "precancer+" for image 2 and vice-versa) for the

each of the four models (repeatability), while Panel (c) (2)–(4) highlights relevant classification metrics, including (2) the % extreme misclassification (precancer+ misclassified as normal and vice-versa); (3) the % precancer+ misclassified as normal; and (4) the % normal misclassified as precancer+, for each of the four models. "Gray Zone" = "Indeterminate".
(TIF)

S2 Fig. Dataset Overview. The top panel highlights the five different studies (NHS, ALTS, CVT, Biop and D Biop) used to generate the final collated "SEED" dataset (top right) on which our model was trained and internally validated. The bottom panel highlights the six different countries / geographies included in the "EXT" dataset, all comprising of images acquired using a Samsung Galaxy J8 smartphone, on which our model was externally validated.
(TIF)

S1 Table. Classification and repeatability metrics on a held-aside test set of 8,734 images from the "SEED" dataset. Our model performs consistently well on "internal" data, even when retrained with added "external" images from the "EXT" dataset. Metrics are reported for each of the model runs involving incremental additions of images from the "EXT" (J8) dataset at the woman level, in a 2n normal (N): 2n indeterminate (I): 1n precancer+ (P) ratio of ground truth class, where n = # of precancer+ women added, as shown on the leftmost column. % values are rounded to 1 decimal place, while numeric values are rounded to 2 decimal places.
(DOCX)

Author Contributions

Conceptualization: Syed Rakin Ahmed, Didem Egemen, Brian Befano, Ana Cecilia Rodriguez, Jayashree Kalpathy-Cramer, Mark Schiffman.

Data curation: Syed Rakin Ahmed, Didem Egemen, Brian Befano, Jose Jeronimo, Kanan Desai, Carolina Teran, Karla Alfaro, Joel Fokom-Domgue, Kittipat Charoenkwan, Chemtai Mungo, Rebecca Luckett, Rakiya Saidu, Taina Raiol, Ana Ribeiro, Julia C. Gage, Silvia de Sanjose.

Formal analysis: Syed Rakin Ahmed, Didem Egemen.

Funding acquisition: Jayashree Kalpathy-Cramer, Mark Schiffman.

Investigation: Syed Rakin Ahmed.

Methodology: Syed Rakin Ahmed, Didem Egemen, Brian Befano, Ana Cecilia Rodriguez, Jose Jeronimo.

Project administration: Jayashree Kalpathy-Cramer, Mark Schiffman.

Resources: Syed Rakin Ahmed.

Software: Syed Rakin Ahmed.

Supervision: Jayashree Kalpathy-Cramer, Mark Schiffman.

Validation: Syed Rakin Ahmed.

Visualization: Syed Rakin Ahmed.

Writing – original draft: Syed Rakin Ahmed, Didem Egemen.

Writing – review & editing: Syed Rakin Ahmed, Didem Egemen, Brian Befano, Ana Cecilia Rodriguez, Jose Jeronimo, Kanan Desai, Carolina Teran, Karla Alfaro, Joel Fokom-

Domgue, Kittipat Charoenkwan, Chemtai Mungo, Rebecca Luckett, Rakiya Saidu, Taina Raiol, Ana Ribeiro, Julia C. Gage, Silvia de Sanjose, Jayashree Kalpathy-Cramer, Mark Schiffman.

References

1. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nat 2017 5427639. 2017; 542: 115–118. https://doi.org/10.1038/nature21056 PMID: 28117445
2. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med 2019 251. 2019; 25: 65–69. https://doi.org/10.1038/s41591-018-0268-3 PMID: 30617320
3. Piccialli F, Somma V Di, Giampaolo F, Cuomo S, Fortino G. A survey on deep learning in medicine: Why, how and when? Inf Fusion. 2021; 66: 111–137. https://doi.org/10.1016/J.INFFUS.2020.09.006
4. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019 251. 2019; 25: 44–56. https://doi.org/10.1038/s41591-018-0300-7 PMID: 30617339
5. Gidwani M, Chang K, Patel JB, Hoebel KV, Ahmed SR, Singh P, et al. Inconsistent Partitioning and Unproductive Feature Associations Yield Idealized Radiomic Models. 2022 [cited 3 Jan 2023]. https://doi.org/10.1148/radiol.220715 PMID: 36537895
6. Lemay A, Hoebel K, Bridge CP, Befano B, De Sanjosé S, Egemen D, et al. Improving the repeatability of deep learning models with Monte Carlo dropout. 2022 [cited 13 Nov 2022]. https://doi.org/10.1038/s41746-022-00709-3 PMID: 36400939
7. Lu C, Ahmed SR, Singh P, Kalpathy-Cramer J. Estimating Test Performance for AI Medical Devices under Distribution Shift with Conformal Prediction. 2022 [cited 13 Nov 2022]. https://doi.org/10.48550/arxiv.2207.05796
8. Wentzensen N, Lahrmann B, Clarke MA, Kinney W, Tokugawa D, Poitras N, et al. Accuracy and Efficiency of Deep-Learning–Based Automation of Dual Stain Cytology in Cervical Cancer Screening. JNCI J Natl Cancer Inst. 2021; 113: 72–79. https://doi.org/10.1093/jnci/djaa066 PMID: 32584382
9. 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: 664–670. https://doi.org/10.1002/ijc.30716 PMID: 28369882
10. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021; 71: 209–249. https://doi.org/10.3322/caac.21660 PMID: 33538338
11. Schiffman M, Doorbar J, Wentzensen N, De Sanjosé S, Fakhry C, Monk BJ, et al. Carcinogenic human papillomavirus infection. Nat Rev Dis Prim 2016 21. 2016; 2: 1–20. https://doi.org/10.1038/nrdp.2016.86 PMID: 27905473
12. Schiffman MH, Bauer HM, Hoover RN, Glass AG, Cadell DM, Rush BB, et al. Epidemiologic Evidence Showing That Human Papillomavirus Infection Causes Most Cervical Intraepithelial Neoplasia. JNCI J Natl Cancer Inst. 1993; 85: 958–964. https://doi.org/10.1093/jnci/85.12.958 PMID: 8388478
13. Schiffman M, Castle PE, Jeronimo J, Rodriguez AC, Wacholder S. Human papillomavirus and cervical cancer. Lancet. 2007; 370: 890–907. https://doi.org/10.1016/S0140-6736(07)61416-0 PMID: 17826171
14. WHO Guidelines Approved by the Guidelines Review Committee. WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention. Geneva World Heal Organ. 2013.
15. World Health Organization. Global strategy to accelerate the elimination of cervical cancer as a public health problem and its associated goals and targets for the period 2020–2030. United Nations Gen Assem. 2020; 2: 1–56. Available: https://www.who.int/publications/i/item/9789240014107
16. Belinson J. Cervical cancer screening by simple visual inspection after acetic acid. Obstet Gynecol. 2001; 98: 441–444. https://doi.org/10.1016/s0029-7844(01)01454-5 PMID: 11530126
17. Ajenifuja KO, Gage JC, Adepiti AC, Wentzensen N, Eklund C, Reilly M, et al. A Population-Based Study of Visual Inspection With Acetic Acid (VIA) for Cervical Screening in Rural Nigeria. Int J Gynecol Cancer. 2013; 23: 507–512. https://doi.org/10.1097/IGC.0b013e318280f395 PMID: 23354369
18. Massad LS, Jeronimo J, Schiffman M. Interobserver agreement in the assessment of components of colposcopic grading. Obstet Gynecol. 2008; 111: 1279–1284. https://doi.org/10.1097/AOG.0b013e31816baed1 PMID: 18515509

19. Silkensen SL, Schiffman M, Sahasrabuddhe V, Flanigan JS. Is It Time to Move Beyond Visual Inspection With Acetic Acid for Cervical Cancer Screening? Glob Heal Sci Pract. 2018; 6: 242–246. https://doi.org/10.9745/GHSP-D-18-00206 PMID: 29959268

20. Ahmed SR, Befano B, Lemay A, Egemen D, Rodriguez AC, Angara S, et al. Reproducible and clinically translatable deep neural networks for cervical screening. Sci Reports 2023 131. 2023; 13: 1–18. https://doi.org/10.1038/s41598-023-48721-1 PMID: 38066031

21. Low & middle income | Data. [cited 20 Aug 2023]. Available: https://data.worldbank.org/country/XO

22. Arbyn M, Sasieni P, Meijer CJLM, Clavel C, Koliopoulos G, Dillner J. Chapter 9: Clinical applications of HPV testing: A summary of meta-analyses. Vaccine. 2006; 24: S78–S89. https://doi.org/10.1016/J.VACCINE.2006.05.117 PMID: 16950021

23. Christine N, Juliana A, Juma MH, Heijmans R, Ouburg S, Ali M, et al. Detection of high-risk human papillomavirus (HPV) by the novel AmpFire isothermal HPV assay among pregnant women in Pemba Island, Tanzania. Pan Afr Med J. 2020; 37: 37. https://doi.org/10.11604/pamj.2020.37.183.23367 PMID: 33447338

24. Jeronimo J, Holme F, Slavkovsky R, Camel C. Implementation of HPV testing in Latin America. J Clin Virol. 2016; 76: 69–73. https://doi.org/10.1016/j.jcv.2015.11.035 PMID: 26699418

25. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: Unified, real-time object detection. Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit. 2016;2016-December: 779–788. https://doi.org/10.1109/CVPR.2016.91

26. Egemen D, Perkins RB, Cheung LC, Befano B, Rodriguez AC, Desai K, et al. Artificial intelligence–based image analysis in clinical testing: lessons from cervical cancer screening. JNCI J Natl Cancer Inst. 2024; 116: 26–33. https://doi.org/10.1093/jnci/djad202 PMID: 37758250

27. Bengio Y, Bastien F, Bergeron A, Boulanger–Lewandowski N, Breuel T, Chherawala Y, et al. Deep Learners Benefit More from Out-of-Distribution Examples. JMLR Workshop and Conference Proceedings; 2011. pp. 164–172. Available: https://proceedings.mlr.press/v15/bengio11b.html

28. Žliobaitė I, Pechenizkiy M, Gama J. An Overview of Concept Drift Applications. Stud Big Data. 2016; 16: 91–114. https://doi.org/10.1007/978-3-319-26989-4_4/COVER

29. Hoens TR, Polikar R, Chawla N V. Learning from streaming data with concept drift and imbalance: An overview. Prog Artif Intell. 2012; 1: 89–101. https://doi.org/10.1007/S13748-011-0008-0/METRICS