Supplementary Material

Supplementary Methods:

Concentration Curve

The concentration curve is used in epidemiology and health economics to demonstrate the distribution of health variables, like disease risk, health outcomes, or healthcare utilization, in a population. Gail[1] explains when the resources are limited the concentration curve can be a useful tool to determine how to allocate resources to maximize the health benefits in the population. Cheung et al.[2] and Katki et al.[3] use concentration curve to compare different screening strategies for oral cancer screening and lung cancer screening, respectively. To plot a concentration curve, a dataset representative of the full population is required. The x-axis represents the cumulative percentage of the full population (0% to 100%) and the dataset is sorted from highest risk to lowest risk (or any other criteria on which the screening strategy is based, for example in Cheung et al.[2] risk-based screening vs age-based screening are compared so in one of the concentration curves the population is ranked from oldest age to youngest). The y-axis shows what percentage of expected disease can be prevented if x% of the highest-risk group in the population is screened or treated.

In our study, we are using concentration curve to help local decision-makers to create their local risk-based guidelines in each setting. Among HPV-positive individuals, we have 12 risk strata (4 HPV genotype groups, HPV 16 positive, else HPV 18/45 positive, else HPV 31/33/35/52/58 positive, else HPV 39/51/56/59/68 positive crossed with 3 AVE test results, precancer/cancer-indeterminate-normal) that we can rank from highest to lowest risk of having cervical precancer/cancer within that specific population. The x-axis in Figure 3 presents the cumulative percentage of the population while the y-axis shows what percentage of expected precancers/cancers can be treated if x% of the highest-risk patients are referred for treatment. The curve starts with the most severe risk stratum, which is HPV 16 positive and AVE=precancer/cancer, in this hypothetical population that we are illustrating in Figure 3. Patients falling into this stratum represent only 1.4% of the total screening population but 37% of the expected precancer/cancer cases. Adding the second most severe risk stratum, HPV High-Risk Medium positive (in this example this is the combination of HPV 18/45 positive and HPV 31/33/35/52/58 positive type groups) and AVE=precancer/cancer, increases the referral individuals to 3.2% of the population and increases precancer/cancer treatment to 71% of total CIN2+ cases. Adding the 3rd highest risk stratum (HPV 16, AVE indeterminate) refers to only 3.7% of the screening population but 80% of all precancers and cancers are cumulatively referred for treatment. A potential cut-point in this example would be referring 5.9% of the screening population for treatment and in return eliminating 94% of the expected CIN2+ cases by treating them (this level is shown with a blue arrow in Figure 3). This could be a reasonable option in settings with very few planned screens per lifetime, for which treatment is favored over expectant management for most patients. However, as mentioned in the main manuscript, we expect local decision-makers to take into account available resources, treatment options, and risk tolerance levels in that specific region and determine their management thresholds. The remaining 41% of HPV-positive patients (4.1% of the total screening population), have low-risk combinations of HPV genotype and AVE results and are therefore less likely to benefit from treatment. Approximately 90% of the population has HPV-negative result and would therefore not undergo triage testing. We should note that, in this example, high-risk HPV genotypes were grouped under 3 categories (i.e., HPV 16 positive, else HPV HR medium positive; which are types 18/45/31/33/35/52/58; else HPV HR low; which are types 39/51/56/59/68); therefore we have 9 risk strata among HPV-positive individuals and 12 risk strata overall when including HPV-negative individuals (therefore 12 dots in the figure in total).

References:

1. Gail MH. *Applying the Lorenz Curve to Disease Risk to Optimize Health Benefits under Cost Constraints*. http://www.cancer.gov/bcrisktool/
2. Cheung LC, Ramadas K, Muwonge ; Richard, et al. Risk-Based Selection of Individuals for Oral Cancer Screening. *J Clin Oncol*. 2021;39:663-674. doi:10.1200/JCO.20
3. Katki HA, Kovalchik SA, Berg CD, Cheung LC, Chaturvedi AK. Development and validation of risk models to select ever-smokers for ct lung cancer screening. *JAMA - Journal of the American Medical Association*. 2016;315(21):2300-2311. doi:10.1001/jama.2016.6255

Supplementary Table 1: What makes a good medical test?

Criteria:	**Explanation:**
Repeatability	Repeatability of a medical test indicates that the same patient, if tested multiple times under the same conditions, will have the same test result. Medical tests lacking repeatability will lead to untrustworthy results. If the test results are continuous, weighted kappa value can be used to compare the agreement between the two test results from the same patient. To visualize this agreement, the average of two test results versus their difference can be plotted (this visual method is called Bland-Altman plot). This graphical method can show the pattern between the test result and where it lacks repeatability if there is any pattern. In our early 2-class classification experiments, the values vary a lot between positive and negative around the cut-point value (i.e., 0.5) which was a sign that the algorithm lacks repeatability for ambiguous cases (neither completely normal nor completely a case). For tests with multiple categories, the percentage of concordant test results for each class can be assessed. When there is an ordinality between categories, disagreement at the extreme end classes should be minimized (in our example, for instance, we tried minimizing non-repeatability between normal and precancer/cancer classes). For ordinal multi-category tests, besides testing repeatability for each class, it is also important to minimize the disagreement at the extreme end classes. Therefore, we assess both quadratic weighted kappa and percentage of disagreement at the extreme classes (which clinically means the first test result is normal while the second test result is precancer/cancer when these two measurements are obtained from the same patient at the same visit)
Accuracy	The accuracy of a test is usually assessed by sensitivity and specificity. In multi-class classification (3-class in our example of cervical images; normal, indeterminate, precancer/cancer), a confusion matrix (ground truth values as one dimension and the test prediction on the other dimension) can be constructed to evaluate the test performance across different classes. The most important false classifications to be minimized are extreme misclassifications, i.e., the precancer/cancer cases predicted as being normal and individuals with normal cervix predicted as having precancer/cancer. Therefore, besides analyzing sensitivity/specificity and confusion matrix, assessing the percentage of extreme misclassifications was important to obtain the best-performing algorithm in our example. Ultimately, our goal is to accurately identify patients with high-grade biopsy results and individuals with a normal cervix based on automated visual evaluation of an image taken during the vaginal examination.
Calibration	Calibration reflects the agreement between predicted and observed outcomes in a model. To assess calibration, in our example, the expected number of individuals in each class from the AVE predictions is compared with the numbers observed from the data. Our risk model combines AVE and HPV test results, so we also evaluate calibration at this step by comparing observed and expected risks. Since we are using population representative sample to predict precancer/cancer risk, the results reflect good calibration. This assessment shows how well the model predicts the risk of cervical precancer/cancer and identifies any discrepancies between predicted and observed outcomes. Comparing continuous deep learning outcomes across observed disease risks on calibration curves is another method mostly used in machine learning

	literature and Monte Carlo drop-out method has been shown to improve model calibration.
Reduced Overfitting	In statistical modeling, overfitting refers to the phenomenon where a model fits too well to the dataset it was trained on. Overfit models result in poor performance when tested on new datasets that they have not encountered before. The primary cause of overfitting is creating models that are overly complex, incorporating too many variables or parameters. Deep learning models are prone to overfitting due to their high degree of complexity. To assess the risk of overfitting, model performance should be evaluated on both internal and external datasets, using techniques such as mean squared error (MSE) in regression models or area under the ROC curve (AUC) in classification models. By carefully monitoring for signs of overfitting and taking steps to prevent it (in our example we used Monte Carlo drop-out method to reduce overfitting), we can ensure that the models created for clinical diagnostics are robust, reliable, and can easily be adapted to new settings.
External Validation & Portability	External validation is a crucial step in assessing the reliability and generalizability of statistical models. Specifically, it refers to the ability of a model to produce repeatable, accurate, and well-calibrated results when tested on new datasets (different from the one used in training). This is particularly important for deep learning models, which are known to be at risk of overfitting to the training set. Overfit models lack generalizability to new settings, and their predictions are limited to the data they were trained on. To address this issue, external validation should be assessed by evaluating the performance on the new datasets in terms of repeatability, accuracy, and calibration measures. Our experiments have shown that external validation for deep learning models can be hard to achieve. The most important factor that affected our model's generalizability, was the image capture device used in data collection process. The algorithm cannot make accurate predictions until it has been retrained with images from the new device. Therefore, to adapt the model to a new setting, we have adopted a strategy of retraining the algorithm using a small subset of the new dataset.
Risk Stratification by Including Other Factors and/or Test Results	Risk stratification helps identifying patients at high risk from those at low risk for developing a specific disease or a condition. We are evaluating the performance of a composite triage test for those HPV positive individuals that includes HPV genotyping and AVE screening tests to improve the detection of patients at high risk of cervical precancer. While HPV genotyping test alone provides high risk stratification, the AVE test must provide additional information about an individual's risk of precancer beyond what is already provided by the HPV genotype testing to be clinically useful. To test this the AUC curves of the models with HPV genotyping only and HPV genotyping combined with AVE test can be compared. Risk models provide a useful tool for integrating multiple test results and individual factors (such as age, gender) to obtain more precise risk estimates. Obtaining overall risk profile of individuals, clinicians can make more informed decisions about patient management and treatment.