Journal of

### Cancer Research and Therapeutic Oncology

Research Article

Open Access

Application of Machine Learning to Discern Factors Contributing to ASCUS Overdiagnosis in a Predominantly Afro-Caribbean Population

### David Anthony Gaskin1, Peter Simon Chami2,3, Alain Jamar Reid4, Justin Alexander Ward1 and Pamela Suzanne Gaskin4,\*

1Integrated Pathology Services, Bridgetown, St. Michael Barbados

2The Warren Alpert Medical School of Brown University, Providence, Rhode Island, United States of America 3Faculty of Pure and Applied Sciences, University of the West Indies, Cave Hill, Bridgetown, Barbados 4Faculty of Medical Sciences, University of the West Indies, Cave Hill, Bridgetown, Barbados

**\*Corresponding Author:** Pamela Gaskin, University of the West Indies, Clinical Skills Building, Bridgetown, Barbados, Tel:

(246) 820-0105; E-mail: pamela.gaskin@cavehill.uwi.edu

**Received Date:** September 28, 2023 **Accepted Date:** October 28, 2023 **Published Date:** October 30, 2023

**Citation:** David Anthony Gaskin, Peter Simon Chami, Alain Jamar Reid, Justin Alexander Ward, Pamela Suzanne Gaskin (2023) Application of Machine Learning to Discern Factors Contributing to ASCUS Overdiagnosis in a Predominantly Afro Caribbean Population. J Cancer Res Therap Oncol 11: 1-15

**Abstract**

**Background:** Over diagnosis of ASCUS can result in unnecessary patient anxiety, inappropriate management, and financial burden.

**Objectives:** We faced the challenge of a high ASCUS/HRHPV negative rate in a community laboratory offering reflex HRH- PV testing. Methods: We utilized machine learning to predict a negative ASCUS/HRHPV outcome. We developed a model of high specificity.

**Results:** The model demonstrated age, obscuring factors and atrophy to be the most influential contributors to a negative ASCUS/HRHPV outcome. Our findings show that for women below age 30, the ASCUS HPV negative rate was low. Age group 55-59, obscuring factors and atrophy most significantly predicted an ASCUS/HRHPV negative diagnosis.

**Conclusion:** Our model had an overall rate of correct classification of 76% with a specificity of 86%. We propose that this tool can be used to risk stratify cases initially diagnosed as ASCUS especially in those laboratories that do not have resources for digital imaging analysis. It is expected that the implementation of this tool would reduce ASCUS overdiagnosis.

**Keywords:** ASCUS; HPV; Pap Smear; Natural Language Processing

©2023 The Authors. Published by the JScholar under the terms of the Crea-tive Com- mons Attribution License [http://creativecommons.org/licenses/by/3.0/,](http://creativecommons.org/licenses/by/3.0/) which per- mits unrestricted use, provided the original author and source are credited.

# Introduction

Atypical squamous cells of undetermined signifi- cance (ASCUS) is the most common abnormal diagnosis on cervical cytology. An ASCUS diagnosis is rendered when the abnormalities detected are either qualitatively or quanti- tatively insufficient to render a confident diagnosis of squa- mous intraepithelial lesion [1]. The ASCUS category signifi- cantly improves the sensitivity of cervical screening [2]. However, over diagnosis of ASCUS can result in unneces- sary patient anxiety, inappropriate management, and finan- cial burden.

Reflex HPV mRNA testing has been done on all ASCUS cases diagnosed on liquid-based cytology specimens at this institution (IPS) for a five year period and the AS- CUS HPV positivity rates are routinely calculated as part of quality control. A high-risk HPV positive ASCUS percent- age of 43.7% (40-50%) is considered acceptable for appropri- ate HPV diagnosis [3]. A rate below 40% can be considered a proxy for ASCUS over diagnosis [4,5].

There is limited data on the overall prevalence of HRHPV subtypes in Barbadian women with HPV infection. One population-based study detected high-risk HPV geno- types in an estimated 23% of screened, HPV vaccine naïve women [6]. Our reported HR HPV prevalence was high in comparison to that reported for Jamaica, Trinidad, and To- bago at the time of the study and higher than the 20.6% pre- valence of HRHPV for women in 2013-2014 reported by the CDC.We therefore expect our HRHPV rates among wo- men with ASCUS to meet or exceed 40-50%.

Multiple epidemiological and biological factors have been linked to ASCUS over diagnosis including pa- tient age, inflammation, reproductive tract infection, hormo- nal therapy, reactive atypia, atrophy, fungus, trichomonas and blood [6,7]. Most of these studies take a univariate ap- proach to the analysis of predetermined factors contribut- ing to ASCUS diagnosis. The influence of cognitive and re- porting bias on rates of ASCUS diagnosis cannot be quanti- fied effectively using univariate analytical methods.

We recognize that ASCUS is an equivocal diagnos- tic category, and the rate of diagnosis can be influenced by many highly correlated variables that are difficult to discrim-

inate using traditional qualitative statistics. Machine-learn- ing approaches to Natural Language Processing (NLP) algo- rithms have been applied in medicine and would be suitable to analyze ASCUS cytology reports in an effort to discern novel factors that contribute to the high rates of high-risk HPV (HRHPV) ASCUS discordance. This is particularly pertinent in our setting given that ASCUS HPV positivity rates are routinely calculated as part of quality control in our laboratory Integrated Pathology Services Ltd. (IPS). NLP is of particular interest in this application given that traditional qualitative means of extracting the information from pathologist notes would be very tedious. Machine- learning, combined with an NLP approach, allows for faster collation of descriptive phrases that mean the same thing and is less subject to human bias. The outcome of this type of analysis permits qualitative data to be used as variables for modelling.

We audit the number of ASCUS/HRHPV cases. Of concern is the fact that our HRHPV positive rates were below 40% over the over the period 2016-2020. Therefore, we successfully applied machine learning to discern factors that contributed to ASCUS over-diagnosis in a community pathology practice offering routine reflex HPV DNA test- ing.

# Materials and Methods

We conducted a retrospective study of all women receiving a Liquid based Pap Smear at Integrated Pathology Services (the sole private laboratory in Barbados where ap- proximately 98% of cytology samples are sent to Vitro Molecular Laboratories for HOLOGIC® Aptima® HPV Tests (HPV tests). HPV testing is not available locally. Data were abstracted from records for the period April 2016 and De- cember 2020.

During the period under study all women with a diagnosis of ASCUS had reflex HPV DNA testing for high-

-risk serotypes. The number of ASCUS cases and the num- ber of squamous intraepithelial lesion (SIL) cases were com- puted. The ratio of ASCUS to SIL cases was also calculated. All women who had conventional smears or total hysterec- tomies were excluded from the study. All diagnoses were re- ported using the Bethesda system.

Patients’ data were downloaded from a HIPPA- compliant secure computer at IPS that had an ASCUS diag- nosis. The data were de-identified and passed on to a field operator for cleaning and coding using Microsoft Access and Excel. These data were then used to check the figures on high-grade squamous intra-epithelial lesion (HSIL) and low-grade squamous intra-epithelial lesion (LSIL) rates routinely used as quality control.

Figure 1 shows the sequence of events outlined in the Methods. Descriptive statistics, including contingency tables were generated to conduct preliminary assessment of the factors known to affect ASCUS discordance. Keating and Wang, 2001(8) was used to categorize women’s meno- pausal status as follows: premenopausal (age ≤ 45 years); perimenopausal (ages 46-54) and postmenopausal (age ≥ 55 years). The ratio of women with HPV positive diagnosis to negative was calculated and percentages presented.



**Figure 1:** Statistical analysis Flowchart

Known intrinsic factors examined included absence or presence of inflammation, blood, infections (yeast, Trichomonas vaginalis, fungus), change in the nor- mal bacterial flora (bacterial vaginosis) and atrophy. The da- ta were then sent to the statistician for the modeling of HPV discordance with ASCUS diagnosis.

We employed NLP (a machine learning method) to categorize words and phrases used by pathologists to de- scribe specimens.The outcomes were amalgamated with all other variables for modelling.

We developed an independent ensemble predic- tive machine learning model to predict the factors that con- tribute to an ASCUS/HRHPV negative outcome.

### The Model

To model factors that predicted negative HPV with a positive ASCUS diagnosis we coded these variables as 1 for negative and 2 for positive. In modelling, we created two categorical variables; Status 0 which were ASCUS cases that had a positive HPV diagnosis (2,2) and the second vari- able Status 1, which were ASCUS cases that had a negative HPV diagnosis (2,1).

Our model was counterfactual in that it sought to determine factors that predicted a negative ASCUS/HRH- PV diagnosis.

Further, we used stacking (also known as a “stacked generalization”) which is an ensemble modeling technique that involves the combination of data from the

predictions of multiple models. These were used as features to generate a new model and make predictions. The models which are combined are known as base models. The base model predictions are used as additional features to train the final model.

Natural Language Processing (NLP) was em- ployed on the cytology reports for all the patients in the study. These reports were organized into a Document Term Matrix (DTM) which included synonymous phrases report- ed in the specimen adequacy and “other” open-ended cate- gories of the cytology reports. NLP generated text combina- tions, included “Satisfactory but limited by obscuring in- flammatory cells in areas of the smear” and all its variants. These variants formed a single semantic cluster that was de- noted “Limited by obscuring factors (Recodespeci2)” based on contextual similarities between components of the clus- ter. An example of a semantic cluster is given below:

Satisfactory but limited by obscuring inflammatory cells and processing artifact in areas of the slide.

Satisfactory but limited by obscuring inflammatory cells in areas of the smear.

Satisfactory but limited by obscuring inflammatory cells in areas of the smear.

Satisfactory cervical smear. Cells in areas of the smear obscured by inflammation.

Satisfactory cervical smear. Inflammation obscures epithelial cells.

Satisfactory cervical smear. Obscuring inflammation.

Satisfactory cervical smear. Partial obscuring by inflammatory exudate.

Satisfactory cervical smear. There is a heavy partially obscuring inflammatory exudate.

Satisfactory smear. There are obscured epithelial cells in areas due to marked inflammation.

Satisfactory smear. There are obscured

groups of epithelial cells due to marked inflammation.

We employed KeyBERT [9], which is a minimal and easy-to-use keyword extraction technique that lever- ages Biodirectional Encoder Representation from Trans- formers (BERT) embeddings to create keywords and key phrases that are most representative of the document. These phrases were then given to the Supervising pathologists, to see if they made clinical sense. In addition, it allowed us to identify if sentences were accurate and if one or more terms were semantically synonymous. This process led to the crea- tion of our new factor variables “RECODEESPECI2”. These semantic clusters were then used in the penalized regression model by a process known as stacked generalization.

We continued our ensemble model by employing a penalized regression model on our new concatenated data which we called “HPV”. For this model, we employed the Elastic Net regression.

Elastic Net is a regression method that performs variable selection and regularization both simultaneously. Regularization refers to a process of introducing additional information to solve an ill-posed problem. Alternatively, it is descried as working by biasing data towards particular val- ues (such as small values near zero). The bias is achieved by adding a tuning parameter to encourage those values, which we referred to as penalty terms.

### Penalty Terms

**L1 regularization** adds an L1 penalty equal to the absolute value of the magnitude of coefficients. In other words, it limits the size of the coefficients. L1 can yield sparse models (i.e. models with few coefficients); some coef- ficients can become zero and eliminated.

**L2 regularization** adds an L2 penalty equal to the square of the magnitude of coefficients. L2 will *not* yield sparse models and all coefficients are shrunk by the same factor (none are eliminated).

**Elastic nets** combine L1 & L2 methods, but do add a hyperparameter [10].

Hence, the term regularization is the main concept

#### behind the elastic net. Regularization comes into the picture when the model is overfitted. Overfitting is a problem that occurs when the model is performing well with the training dataset, but with the test, dataset model is giving errors; in this situation the regularization is a technique to reduce the errors by fitting a function appropriately in the training da- taset. These functions, as stated prior, are referred to as “penalties”, of which there are two types: L1 and L2. Penal- ized regression methods keep all the predictor variables in the model but constrain (regularize) the regression coeffi-

cients by shrinking them toward zero. If the amount of shrinkage is large enough, these methods can also perform variable selection by shrinking some coefficients to zero.

We developed an optimal model to predict factors that were associated with negative ASCUS/HRHPV diagnos- es and HPV results.

### The Elastic Net Model

Elastic Net aims at minimizing the following loss

function:

where α is the mixing parameter between ridge (α = 0) and lasso (α = 1).

Now, there are two parameters to tune: λ and α.

### Tuning

The ultimate goal of machine learning is to make a machine system that can automatically build models from data without requiring tedious and time-consuming human involvement. As we recognize, one of the difficulties is that learning algorithms (eg. decision trees, random forests, clus- tering techniques, etc.) requires one to set parameters be- fore we use the models (or at least to set constraints on those parameters).

How those parameters are set can depend on a whole host of factors. That said, the goal, is usually to set those parameters to optimal values that enable one to com- plete a learning task in the best way possible. Thus, tuning an algorithm or machine learning technique, can be simply thought of as process which one goes through in which they optimize the parameters that impact the model in order to enable the algorithm to perform the best.

Cross-validation is a statistical method used to esti- mate the performance (or accuracy) of machine learning models. It is used to protect against overfitting in a predic-

tive model, particularly in a case where the amount of data may be limited. In cross-validation, we make a fixed num- ber of folds (or partitions) of the data, run the analysis on each fold, and then average the overall error estimate. For our model we chose to carry out a 10-fold cross-validation with 10 repeats. We did have to invoke parallel processing using the doParallel [11] package in R to get the model, in- cluding the Cross validation to run down from over 20 min- utes to under 5 minutes.

### Optimal Model

The HPV dataset that was created with the addi- tion of the RECODEESPECI2 variable provides a problem with many potential candidate predictor variables occur- ring. These problems require the performance of a statisti- cal model selection to find an optimal model, one that is as simple as possible while still providing good predictive per- formance. Penalized regression will allow the creation of a regression model that is penalized, for having too many vari- ables in the model, by adding a constraint in the equation (- James et al. 2014, P. Bruce and Bruce (2017)). This is re- ferred to as shrinkage regression. The consequence of im- posing this penalty, is to reduce (i.e., shrink) the coefficient values towards zero. This allows the less contributive vari- ables to have a coefficient close to zero or equal zero. We should note that shrinkage requires the selection of a tuning

#### parameter (lambda) that determines the amount of shrink- age.

The NLP analysis was carried out using Python programing language Version 3.6. All other analysis was car- ried out using the R-Programming language, version 4.0.3 (2020-10-10).

**Results**

#### Eighteen thousand two-hundred and twenty-sev- en (18,227) cervical smears were reviewed for the study peri- od of which 958 were classified as ASCUS. Of these cases, 273 or 28.5% tested positive for high-risk HPV mRNA. Five hundred and seventy-seven (577) women had a diagnosis of squamous intra-epithelial lesion (SIL) (519 LSIL and 58 HSIL) giving an ASCUS: SIL ratio of 1.66.

Of the ASCUS cases, age was unavailable for 16 participants, which reduced our sample to 942 (98%) of the ASCUS population, 622 (66%), 172 (18.3%) and 148 (15.7%) of these were classified as pre, peri and postmeno- pausal respectively.Age ranged from 16-91 years and the mean age of patients with ASCUS was 40.7 years (95% CI 39.85, 41.47); premenopausal 33.28 years (95% CI 32.69,

33.86); perimenopausal 49.5 years, (95% CI 49.2 – 49.9),

and postmenopausal 61.37 years (95% CI 60.34, 62.41). The prevalence of high risk HPV (HRHPV) was premenopausal

228 (24.2%); perimenopausal 27 (2.9%) and postmeno- pausal 18 (1.9%).

When the cases were segregated based on addition- al cytological findings: Reactive atypia, candida and shift in bacterial flora were significantly different across the AS- CUS/HRHPV negative vs ASCUS/HRHPV positive groups (Table 1).

**Table 1:** Comparison of additional cytological findings of ASCUS/HRHPV negative vs. ASCUS/HRHPV positive groups (N=942)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Intrinsic factors | ASCUS/HRHPV(-ve) | ASCUS/HRHPV(+ve) | Chi-squared | P -value |
| Inflammation | 669 (71%) | 273 (29%) |  | NS |
| Absent | 117 (12.4%) | 51 (5.4%) |  |  |
| Present | 552 (58%) | 273 (23.6%) | 0.188 |  |
| Reactive atypia | 669 (71%) | 273 (29%) |  |  |
| Absent | 591 (62.7%) | 227 (24.1%) |  |  |
| Present | 78 (8.3%) | 46 (4.9%) | 4.57 | 0.033\* |
| Atrophy | 669 (71%) | 273 (29%) |  |  |
| Absent | 645 (68.5%) | 267 (28.3%) |  |  |
| Present | 24 (2.5%) | 6 (0.6%) | 1.214 | 0.27 |
| Fungal infection | 669 (71%) | 273 (29%) |  |  |
| Absent | 634 (67.3%) | 248 (26.3%) |  |  |
| Present | 35 (3.7%) | 25 (2.7%) | 5.01 | 0.025\* |
| Trichomonas vaginalis | 669 (71%) | 273 (29%) |  |  |
| Absent | 667 (70.8) | 270 (28.7%) |  |  |
| Present | 2 (0.2%) | 3 (0.3%) | 2.35 | NS |
| Blood | 669 (71%) | 273 (29%) |  |  |
| Absent | 499 (53%) | 194 (20.6%) |  |  |
| Present | 170 (18%) | 79 (8.4%) | 1.24 | NS |
| Shift in flora | 669 (71%) | 273 (29%) |  |  |
| Absent | 578 (61.4%) | 203 (21.5%) |  |  |
| Present | 91 (9.7%) | 70 (7.4%) | 19.83 | <0.001\* |

\*p < 0.05

**The Elastic Net Model**

#### Our model used both coded and NLP elicited vari- ables. Accuracy was used to select the optimal model using the largest value; and here the best accuracy obtained by the optimal model was 72%. The final values used for the model were alpha = 1 and lambda = 0.02 and a Kappa statistic of 0.3.

Age group in years contributed significantly to the model with influence (regression coefficients) as follows: Age groups 20-24 (-1.53), 25-29 (-1.13) and 30-34 (-0.44)

#### were significant negative contributors. On the other hand, older age contributed positively: Age groups 45-49 (0.17),

55-59 (0.4), 60-64 (0.09) and 65-69 (0.02) such that the like- lihood of an ASCUS/HRHPV negative outcome was highest at age group 55-59 years.

Table 2 shows significant Pathological Descrip- tions from the Elastic- Net model it demonstrates that among pathological descriptions, the term “Limited by obs- curing factors Recodespeci2).had an influence of 0.39.

We presented an estimate of the model’s accuracy. We confirmed this by use of the kappa statistic which com- pares the observed accuracy with the expected accuracy and is accepted as a superior measure of the performance of the model. Accuracy was used to select the optimal model using the largest value.

**Table 2:** Significant contributing cytological factors of the optimal predictive model for an ASCUS/HRHPV negative outcome

|  |  |  |
| --- | --- | --- |
| Contributing factors | Combination codea | Influenceb |
| Limited by obscuring factors (Recodespeci2) |  | 0.3914509 |
| Atrophic epithelial changes. Moderate inflammation. Blood++. | 1 |  |
| (Old) blood++ | 1 | 0.2688663 |
| Atrophic epithelial changes are evident. Blood. | 2 |  |
| Atrophy. Blood. Moderate inflammation. | 2 | 0.2444535 |
| Shift in vaginal flora. Mild inflammation. | 3 |  |
| Atrophic smear.Moderate to severe inflammation.Blood+. | 3 | 0.0676442 |
| Abundant blood. Shift in vaginal bacterial flora. | 4 |  |
| Blood + | 4 | 0.0676427 |
| Abundant blood. Shift in vaginal flora. | 5 |  |
| Background blood. | 5 | 0.0676427 |
| Blood++. Moderate inflammation. |  | 0.0676426 |
| blood2 |  | 0.051097 |
| Atrophic smear. Mild inflammation. | 6 |  |
| (Old) blood++ | 6 | 0.0169809 |
| Atrophic smear.Moderate to severe inflammation.Blood+. |  | 0.0169792 |

aSimilar combination codes refer to combinations of cytological descriptions that together significantly contributed to the model.bAll coeffi- cients are significant at p<0.002.

## Accuracy

#### Accuracy was used to select the optimal model us- ing the largest value; and here the best accuracy obtained by the optimal model was 72% with a Kappa statistic of 0.3.

The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance). The kappa statistic is used not only to evaluate a single classifier, but also to evaluate classifiers amongst themselves. In addition, it considers random chance (agree-

#### ment with a random classifier), which generally means it is less misleading than simply using accuracy as a metric. The expected accuracy for an ASCUS/HRHPV positive outcome is 50%. The ideal frequency of an ASCUS/HRHPV positive outcome is 50%. Our internal lab assessment obtained a fre- quency that was much lower (28%). We therefore needed a model to determine those factors that were causing us to have a low frequency of ASCUS/HRHPV positive results. This model would have to identify the factors that would lead to an ASCUS/HRHPV negative outcome. It should be borne in mind that the model predicted a negative outcome which is the absence of HPV. For example, an Observed Ac-

curacy of 80% is a lot less impressive with an Expected Accu- racy of 75% versus an Expected Accuracy of 50%. Computa- tion of Observed Accuracy and Expected Accuracy is inte- gral to comprehension of the kappa statistic and is most easi- ly illustrated through use of a confusion matrix (Table 3).

The most complete model attained an overall rate of correct classification estimated at 76%, as shown by our Confusion Matrix. Further, we obtained a 10 % sensitivity and 86% specificity. We reiterate that in the current study we modelled a negative HRHPV outcome. Sensitivity gives the “True Positive Rate”, therefore, we expected our model to have poor sensitivity.

**Table 3:** The Confusion Matrix

|  |  |  |
| --- | --- | --- |
| **p** | **0** | **1** |
| **0** | 78 | 33 |
| **1** | 191 | 634 |

# Discussion

In pathology many studies involve the use of quali- tative analyses of reports followed by collation of the data. An important part of this study is that we used a machine learning algorithm to analyze the text and enhance the likeli- hood of drawing objective conclusions. Natural language processing can extract data from unstructured text such as electronic health records (EHRs), physician's notes, and medical literature. This information is fed into Machine learning algorithms, particularly named entity recognition (NER) models, to identify and categorize specific clinical en- tities, such as diseases, medications, symptoms, and patient demographics.

NLP and modelling techniques such as elastic net have been shown to reduce data omission which is a source of ascertainment bias in analysis of electronic health databases [12]. The application of NLP is evolving to allow accurate diagnostic classification of cases using pathology re- ports. It is potentially a cheaper alternative to digital image analysis for diagnostic classification. Non-peer reviewed evi- dence suggests that NLP for text-only analysis may achieve slightly better accuracy than image-only analysis (95.7% vs 93.77%) in diagnostic classification of malignancies based on histopathology reports [13]. The combination of text

and digital imaging analysis is likely to have the best out- come in digital pathology [14,15]. NLP has been applied to the identification of reported HPV positive test results in cervical and anal smears with 97% sensitivity and 95% pre- diction [14]. Our model allowed predictions to be made from a dataset with multiple highly correlated variables and permitted us to assess the strong influence of the unexpect- ed variable of “obscuring factors”. To our knowledge, this is the first application of this process to predict HPV negative test results in ASCUS cases.

Approximately 29% of our ASCUS cases had HRH- PV genotypes and the ratio of ASCUS to SIL cases was 1.66.The percentage of ASCUS/HRHPV positive cases was below the benchmark of 40-50% but the ASCUS:SIL ratio did not exceed the 3:1 threshold recommended by the Beth- esda system. The ASCUS:SIL ratio is a subjective standard of quality because it is based entirely on interpretative diag- noses. By contrast, the percentage of ASCUS cases that are positive for high-risk HPV genotypes is an objective assess- ment of the risk of dysplasia as it is measured against an ex- ternal objective standard [16]. Our documented percentage of ASCUS HPV positive cases of 28.5% likely represents overdiagnosis of ASCUS despite the relatively low AS- CUS:SIL ratio. We therefore inferred that any statistically significant differences observed between the ASCUS/HRH-

PV negative and ASCUS/HRHPV positive groups in our co- hort could be at least partially attributed to factors associat- ed with overcalling ASCUS.

Notably, the rate of ASCUS/HRHPV positive diag- nosis for the study period was 36.7 % among premeno- pausal women which was close to the recommended North American benchmark of (40-50%) and reflected the preva- lence of 39.4% for SIL lesions following a diagnosis of AS- CUS demonstrated in a Barbadian study [17]. In keeping with expectations, premenopausal women were significant negative contributors to a discordant diagnosis of AS- CUS/HRHPV negative, indicating that the probability of AS- CUS overdiagnosis in this subpopulation is low. Converse- ly, women aged 55-59 contributed the most to a discordant ASCUS/HRHPV negative diagnosis indicating that overdi- agnosis of ASCUS is more likely among postmenopausal wo- men. These findings are corroborated by large North Ameri- can, Chinese and Norwegian population-based studies that show a decreasing trend of HR HPV positivity among AS- CUS cases with age [16,18,19]. In contrast to our machine learning application, the age groups of these epidemiologi- cal studies were predetermined and spanned about one de- cade. The lowest HPV positive rate was reported in women ages 50-60 years. Only the North American cohort included the ASCUS/HRHPV negative rate which peaked in women 70+ years [16]. These studies described the general trend of HPV outcomes in ASCUS with age for predetermined age subgroups, but none was able to discern specific age ranges that independently predicted an HPV negative outcome. An advantage of the application of machine learning in the cur- rent study is that it allowed us to measure the extent of the influence of specific age ranges of women on prediction of a HRHPV negative outcome. Specifically, the risk of an AS- CUS/HRHPV negative diagnosis in our model peaked for menopausal women aged 55-59 years with a nadir in the cor- relation of an ASCUS/HRHPV negative diagnosis observed in women aged 65-69 years.

The decreasing influence of age on an HPV nega- tive outcome for patients over 65 years was an unexpected finding considering the decreasing trend of HPV positive status with increasing age reported in other epidemiological studies [18,19]. A secondary peak in prevalence of HRHPV subtypes was observed in persons >65 years in North Ameri-

can and South American cohorts [20,21]. This might ex- plain the decreasing contribution of older age to prediction of an ASCUS/HRHPV negative diagnosis by the model for women >65 years.

Peri and postmenopausal status has been demons- trated to contribute to ASCUS overdiagnosis across many retrospective cohort studies [8,16,22-27]. This has been at- tributed to multiple variables including age [8,22], atrophic cytomorphological changes [16,23-27] and the effects of hor- monal therapy [24]. Some of these studies used HPV geno- type results as markers of ASCUS overdiagnosis [23,25]. However, the predetermined age groups in these studies are wide and overlapping and none of them have been able to define a specific age range within the menopausal popula- tion that is independently correlated with ASCUS over diag- noses. Age groups reported range from 46-54 years in peri- menopausal women to over 55 years in post-menopausal women.

Our machine learning algorithm was able to inde- pendently discern a specific age group (55-59 years) with the greatest quantitative influence on a discordant AS- CUS/HRHPV negative outcome and further elucidated a notably decreased influence of age on a discordant AS- CUS/HRHPV negative diagnosis for women over 60 years. Atrophic epithelial changes which are highly correlated with menopausal age also significantly influenced a discor- dant ASCUS/HRHPV negative outcome. Importantly, peri- menopausal women do not contribute to an ASCUS/HRH- PV negative diagnosis in our study compared to others [8,25]. This suggests that age related cytological features in- cluding high frequency of cells with bland nuclear enlarge- ment coined “perimenopausal cells” [25] do not contribute to ASCUS over diagnosis in women under 55 in our setting.

Misconceptions of low rates of HPV infection in women over 60 years may lead to underdiagnosis of SIL and overdiagnosis of equivocal diagnostic categories like AS- CUS or Atypical Squamous Cells cannot exclude High-- grade lesion (ASC-H).The Elastic Net model used essential- ly excluded the contribution of these women to a AS- CUS/HRHPV negative outcome within the Barbadian co- hort. These findings imply that we are either less likely to di- agnose women > 65 years with ASCUS or those ASCUS cas-

es that we diagnose in this age group are more likely to be HPV positive. An appreciation of the secondary peak in HPV prevalence in this group would encourage close evalua- tion of these slides to exclude cytomorphological changes in keeping with SIL. Prospective use of the model to predict a positive HPV outcome might be useful to determine if a se- condary peak of HRHPV positive cases is present in per- sons over 65 years in our population.

Our model can assist both cytotechnologist and pathologist by selecting those cases for which an ASCUS di- agnosis should be reconsidered and followed up by more diligent review of the cytomorphological features. Artificial intelligence applied in this way potentially provides an em- pirical means to improve the rate of equivocal diagnosis among menopausal women. Ideally, repeated use of this in- strument will enable development of an age-appropriate threshold for ASCUS diagnosis in this subpopulation. Cur- rently, the local data suggests that the threshold for an AS- CUS diagnosis should be highest in women 55-59 but can be lower in women aged 30-44 and > 60 years. It is expected that this approach will increase the ASCUS/HRHPV posi- tive rate.

Marked obscuring inflammation was second to age as an independent predictor of an HPV negative diagno- sis. Inflammation is associated with reactive nuclear changes that can be challenging to interpret leading to AS- CUS overdiagnosis. The results of the model show that the reporting of inflammation severe enough to compromise specimen adequacy is more likely to result in an AS- CUS/HRHPV negative diagnosis. The presence of obscur- ing factors which compromise the quality of the smear can lead to ASCUS overdiagnosis [28,29]. In one study, some of these cases were reclassified as unsatisfactory on review [29].The findings of the ensemble model support close re- view of those smears diagnosed as ASCUS with severe obs- curing inflammation.

Interestingly, univariate analysis demonstrated a statistically significant negative association between reactive epithelial atypia (p=0.033) and Candida infection (p=0.0025) and an ASCUS/HRHPV positive diagnosis. Can- dida infection is known to cause reactive nuclear changes that mimic ASCUS [6,27,28]. Reactive changes have been

documented as a cause of ASCUS over diagnosis which can lead to unnecessary HPV testing and follow-up colposcopy or biopsy procedures [27,28]. The studies demonstrating an association of the variables reactive change and Candida in- fection with ASCUS over diagnosis had retrospective co- horts and were limited by univariate analysis with no sup- porting data on patient HPV status [6,27,28]. Significantly, our multivariate analysis using artificial intelligence support- ed by HPV test results showed that reactive changes or Can- dida infection were not independent predictors of an AS- CUS/HRHPV negative status and therefore likely do not contribute to ASCUS overdiagnosis or a negative HRHPV status in our population.

The relationship between Candida infection and high-risk HPV is complex and a recent study from Turkey demonstrated a positive association between Candida infec- tion and HRHPV [30]. This has been linked to multiple vari- ables including higher number of sexual partners, systemic immune status and local pathogenic effects of Candida that compromise the mucosal barrier. Our combined use of mul- tivariate analysis with supportive HPV testing analysis sup- ports this complex biological link as it suggests that the pres- ence of Candida infection is not predictive of a benign HRH- PV negative ASCUS diagnosis. Therefore, nuclear changes in cases with Candida infection may not just be benign reac- tive effects and should be more closely evaluated to exclude associated HPV viral cytopathic changes. Treatment of the Candida infection detected cytologically may also be protec- tive against HPV infection.

Reactive atypia was not an independent predictor of an ASCUS/HRHPV negative outcome on multivariate analysis. Reactive atypia is a nonspecific cytomorphological finding representing a nuclear change found in multiple overlapping non-HPV related cervical pathologies. It is un- likely to be an independent predictor of ASCUS/HRHPV status. Our analysis supports this and implies that the find- ing of reactive epithelial atypia in our population does not impact ASCUS/HRHPV discordance or contribute to unne- cessary reflex HPV testing. Patients with reactive epithelial atypia are likely to have a benign clinical course and can continue routine cervical Pap smear screening.

Shift in flora was not a positive predictor of an AS-

CUS/HRHPV negative diagnosis. A metanalysis of pooled studies involving 6732 women demonstrated a positive asso- ciation between HPV infection and bacterial vaginosis [31]. Shift in normal bacterial flora is a cytological finding identi- fied in women with bacterial vaginosis. Shift in flora was a poor predictor of an ASCUS/HRHPV negative outcome in our model with an influence of 0.06. An abnormal bacterial flora would not be expected to be a predictor of an HPV neg- ative outcome in ASCUS based on a metanalysis that suggests a positive association between HPV infection and bacterial vaginosis [30]. A cause effect relationship between HPV and bacterial vaginosis has not been established but bacterial vaginosis is linked to mucus degrading enzymes that can impair the mucosal barrier and facilitate infection of epithelial cells by HPV [30]. Our results underscore that a shift in flora in ASCUS cannot be dismissed as a benign finding not associated with HRHPV and further supportive testing in the clinic including vaginal fluid ph. testing and the whiff test should be considered in these patients to con- firm bacterial vaginosis.

Machine learning algorithms analyzing digitized pathology images have shown comparable sensitivity (92.9% (75-99.8%)) and superior specificity to HRHPV (49.7% (45.6-53.8%) testing for risk classification of ASCUS [32,33]. Data driven models like the Pittsburgh Cervical can- cer screening model have been developed as risk assessment tools for persons undergoing cervical screening. The Pitts- burg Cervical screening model allowed quantification of the risk of CIN2+ lesions in patient’s post cervical screening us- ing both HPV results and histopathology findings [34]. This Pittsburg model used results of screening aided by the Ho- logic Thin prep imaging system for all Bethesda diagnostic categories. However, all these models rely on digital scann- ing instruments that are prohibitively expensive in a low-in- come setting. We developed a tool of high specificity that was able to identify factors associated with a high risk of AS- CUS/HRHPV negative diagnosis and the implementation of this tool is expected to reduce ASCUS overdiagnosis. The large sample size allowed for extensive cross validation of the model. An accuracy of 76% suggests that this model is likely to be reproducible using a similar external data set. The high specificity of the model should enable it to identify at least 86% of ASCUSHRHPV negative cases that would

not require costly molecular testing. This is expected to re- duce the financial and psychological cost to the patients of a false positive result.

We are proposing that our tool driven by text anal- ysis of reported data from standardized Bethesda reports can determine with comparable accuracy to imaging digital analysis which ASCUS cases are unlikely to be HPV positive and therefore at low risk of having CIN 2+ lesions in a re- source limited setting.

### Limitations

This study would be enhanced by manual review of cytology slides and reports from Afro Caribbean women age 55-59 years. These findings may elucidate specific clini- copathological features contributing to overuse of ASCUS diagnosis within this subpopulation.

In addition, this study used a surrogate or proxy factor for cervical dysplasia as a gold standard. It is impor- tant to note that up to 43.5% of women with HR-HPV AS- CUS can have normal histology results [18]. HR-HPV test- ing has high sensitivity but low specificity for cervical dys- plasia [18,33,34] [33,35] and therefore it is an imperfect re- flection of the accuracy of cytomorphological interpretation by the pathologist. Furthermore, the false negative rate of HRHPV testing for detecting cervical intraepithelial neopla- sia 2-3 (CIN2+) lesions following an ASCUS diagnosis may increase with age [35].

In this study we used reflex HPV testing which is currently the best method for determining those ASCUS cas- es that require follow up colposcopy. In future, the model can be trained using cytology cases with cervical dysplasia confirmed on biopsy to increase specificity. This will com- pensate for the low specificity of HR-HPV testing and im- prove risk stratification of ASCUS cases by the model espe- cially among older women.

It must be borne in mind that the major question of this paper was to model a negative outcome. Therefore a low sensitivity and high specificity are expected.

Based on our findings, we propose the following al- gorithmic approach to cases initially diagnosed as ASCUS (Figure 2).



**Figure 2:** Algorithmic approach to ASCUS diagnosis based on the study findings

# Conclusion Conflict of Interest

Our findings show that for women below age 30, the ASCUS HPV negative rate was low. Age group 55-59, obscuring factors and atrophy most significantly predicted

clare.

The authors have no conflicts of interest to de-

an ASCUS/HRHPV negative diagnosis. Our model had an overall rate of correct classification of 76% with a specificity of 86%.Our findings add to the scant literature on HPV in- fection among Caribbean women.

# Ethical Considerations

Permission for this study was given by the Institu- tional Review Board, University of the West Indies/Ministry of Health and Wellness, Barbados. The study was per- formed in accordance with the Declaration of Helsinki.

# Funding

The study was not funded.

# Author Contributions

D.G. conceptualized the study and drafted the ini- tial manuscript, P.S.G. coordinated the study and with A.R. and J.W., contributed to development of methodology, writ- ing, review and revision of the paper. J.W. provided acquisi- tion of data. P.C. conducted all statistical analyses and con- tributed to revision of the paper. All authors read and ap- proved the final paper.

**References**

1. Ndifon CO, Al-Eyd G (2021) Atypical squamous cells of undetermined significance. StatPearls.
2. Pitman MB, Cibas ES, Powers CN, Renshaw AA, Frable WJ (2002) Reducing or eliminating use of the category of atypical squamous cells of undetermined significance de- creases the diagnostic accuracy of the Papanicolaou smear. Cancer Cytopathology: Interdisciplinary International Jour- nal of the American Cancer Society 96: 128-34.
3. Moriarty AT, Schwartz MR, Eversole G, Means M, Clayton A, Souers R, et al. (2008) Human papillomavirus test- ing and reporting rates: practices of participants in the Col- lege of American Pathologists Interlaboratory Comparison Program in Gynecologic Cytology in 2006. Archives of pathol- ogy & laboratory medicine 132: 1290-4.
4. Cibas ES, Zou KH, Crum CP, Kuo F (2008) Using the rate of positive high-risk HPV test results for ASC-US to- gether with the ASC-US/SIL ratio in evaluating the perfor- mance of cytopathologists. American Journal of Clinical Pathology 129: 97-101.
5. Pendse AA, Bauer AE, Dodd L, Scanga L (2018) In- creased rate of ASCUS diagnosis with concomitant request for high-risk human papillomavirus reflex testing may be due to cognitive bias. American Journal of Clinical Pathology 149: 425-33.
6. Jahic M, Jahic E (2016) Diagnostic approach to pa- tients with atypical squamous cells of undetermined signifi- cance cytologic findings on cervix. Medical Archives 70: 296.
7. Gupta S, Sodhani P (2012) Reducing “atypical squa- mous cells” overdiagnosis on cervicovaginal smears by dili- gent cytology screening. Diagnostic Cytopathology 40: 764-9.
8. Keating JT, Wang HH (2001) Significance of a diagno- sis of atypical squamous cells of undetermined significance for Papanicolaou smears in perimenopausal and postmeno- pausal women. Cancer Cytopathology 93:100-5.
9. Sharma P, Li Y (2009) Self-supervised contextual key- word and keyphrase retrieval with self-labelling.
10. Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. Journal of the royal statistical soci- ety: series B (statistical methodology) 67: 301-20.
11. Weston S (2022) doParallel: Foreach Parallel Adpator for the 'parallel' Package. 1.0.17 ed: Microsoft Corporation.
12. Khurshid S, Reeder C, Harrington LX, Singh P, Sar- ma G, Friedman SF, et al. (2022) Cohort design and natural language processing to reduce bias in electronic health re- cords research. NPJ digital medicine 5: 1-14.
13. Allada AK (2021) Histopathology Image analysis and NLP for Digital PathologyOntario, Canada.
14. Oliveira CR, Niccolai P, Ortiz AM, Sheth SS, Shapiro ED, Niccolai LM, et al. (2020) Natural language processing for surveillance of cervical and anal cancer and precancer: al- gorithm development and split-validation study. JMIR medi- cal informatics 8: e20826.
15. Tizhoosh HR, Diamandis P, Campbell CJ, Safarpoor A, Kalra S, Maleki D, et al. (2021) Searching images for consensus: can AI remove observer variability in pathology? The American journal of pathology 191: 1702-8.
16. Ko V, Nanji S, Tambouret RH, Wilbur DC (2007) Testing for HPV as an objective measure for quality assur- ance in gynecologic cytology: positive rates in equivocal and abnormal specimens and comparison with the ASCUS to SIL ratio. Cancer Cytopathology: Interdisciplinary International Journal of the American Cancer Society 111: 67-73.
17. Prussia PR, Skeete D, Joakimsen P (2000) The signifi- cance of atypical squamous cells of undetermined signifi- cance in Barbados-Poster abstract. West Indian med j 49.
18. Wang Z, Gu Y, Wang H, Chen J, Zheng Y, Cui B, et al. (2020) Distribution of cervical lesions in high-risk HPV (hr-HPV) positive women with ASC-US: a retrospective sin- gle-center study in China. Virology Journal 17: 1-10.
19. Haldorsen T, Skare GB, Ursin G, Bjørge T (2015) Re- sults of delayed triage by HPV testing and cytology in the Norwegian Cervical Cancer Screening Programme. Acta On- cologica 54: 200-9.
20. Ko EM, Tambouret R, Wilbur D, Goodman A (2011)

HPV reflex testing in menopausal women. Pathology Re- search International.

1. Herrero R, Castle PE, Schiffman M, Bratti MC, Hildesheim A, Morales J, et al. (2005) Epidemiologic profile of type-specific human papillomavirus infection and cervical neoplasia in Guanacaste, Costa Rica. The Journal of infec- tious diseases 191: 1796-807.
2. Flynn K, Rimm DL (2001) Diagnosis of “ASCUS” in women over age 50 is less likely to be associated with dyspla- sia. Diagnostic cytopathology 24: 132-6.
3. Johnston EI, Logani S (2007) Cytologic diagnosis of atypical squamous cells of undetermined significance in peri- menopausal and postmenopausal women: lessons learned from human Papillomavirus DNA testing. Cancer Cyto- pathology: Interdisciplinary International Journal of the American Cancer Society 111: 160-5.
4. Menezes GA, Wakely Jr PE, Stripe DM, Nuovo GJ (2001) Increased incidence of atypical Papanicolaou tests from ThinPreps of postmenopausal women receiving hor- mone replacement therapy. Cancer Cytopathology: Interdisci- plinary International Journal of the American Cancer Society 93: 357-63.
5. Cibas ES, Browne TJ, Mantel Bassichis MH, Lee KR (2005) Enlarged squamous cell nuclei in cervical cytologic specimens from perimenopausal women (“PM Cells”) a cause of ASC overdiagnosis. American journal of clinical pathology 124: 58-61.
6. Saad RS, Dabbs DJ, Kordunsky L, Kanbour-Shakir A, Silverman JF, Liu Y, et al. (2006) Clinical significance of cyto- logic diagnosis of atypical squamous cells, cannot exclude high grade, in perimenopausal and postmenopausal women. American journal of clinical pathology 126: 381-8.
7. Howell LP, Davis RL (1996) Follow‐up of Papanico- laou smears diagnosed as atypical squamous cells of undeter-

mined significance. Diagnostic cytopathology 14: 20-4.

1. Kumar N, Sayed S, Moloo Z (2011) Educational needs and causes of false diagnosis of atypical squamous cells of unknown significance at a university hospital. African Jour- nal of Reproductive Health 15.
2. Sebastião APM, Noronha Ld, Pinheiro DL, Collaço LM, de Carvalho NS, Bleggi‐Torres LF (2004) Influence of specimen adequacy on the diagnosis of ASCUS. Diagnostic Cytopathology 31: 155-8.
3. Oktay Gultekin E, Can B (2023) Prevalence of Candi- da albicans in High-Risk Human Papillomavirus-Positive Wo- men: A Study in Diyarbakır Province, Turkey. Canadian Jour- nal of Infectious Diseases and Medical Microbiology.
4. Gillet E, Meys JF, Verstraelen H, Bosire C, De Sutter P, Temmerman M, et al. (2011) Bacterial vaginosis is associat- ed with uterine cervical human papillomavirus infection: a meta-analysis. BMC infectious diseases 11: 1-9.
5. Tao X, Chu X, Guo B, Pan Q, Ji S, Lou W, et al. (2022) Scrutinizing high‐risk patients from ASC‐US cytology via a deep learning model. Cancer Cytopathology.
6. Pantanowitz L (2022) Improving the Pap test with ar- tificial intelligence. Wiley Online Library.
7. Klein S, Quaas A, Quantius J, Löser H, Meinel J, Peifer M, et al. (2021) Deep Learning Predicts HPV Associa- tion in Oropharyngeal Squamous Cell Carcinomas and Identi- fies Patients with a Favorable Prognosis Using Regular H&E StainsDeep Learning Predicts HPV Association in OPSCC. Clinical Cancer Research 27: 1131-8.
8. Won KH, Lee JY, Cho HY, Suh DH, No JH, Kim YB (2015) Impact of age on the false negative rate of human pa- pillomavirus DNA test in patients with atypical squamous cells of undetermined significance. Obstetrics & Gynecology Science 58: 117-23.

