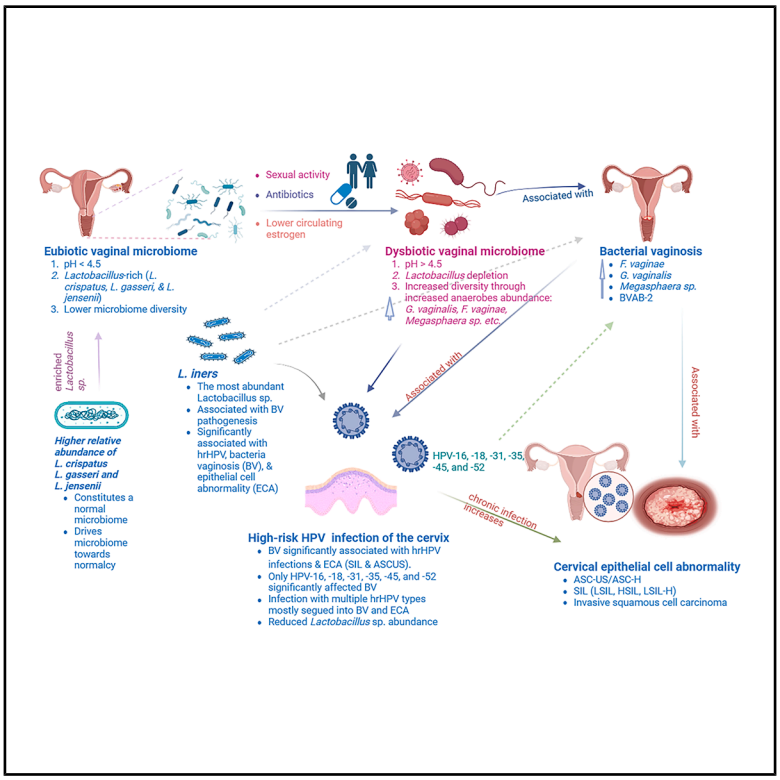
iScience

Article

Lactobacillus-rich cervicovaginal microbiome associated with lower BV, HPV, and cytology outcomes in women

Graphical abstract



Highlights

- 15 607 U.S. samples link microbiome, BV, hrHPV, and cytology outcomes
- *L. crispatus/gasseri/jensenii* dominate BV-negative, NILM specimens
- *L. iners* with anaerobes co-occurs with BV, hrHPV, and abnormal cytology
- Age, hrHPV, and L. crispatus abundance predict BV/cytology (AUROC≈0.97)

Authors

John Osei Sekyere, Jason Trama, Martin Adelson, ..., Rachel Schuster, Jing Jing Yang, Eli Mordechai

Correspondence

joseisekyere@mdlab.com

In brief

Microbiome; Female reproductive endocrinology; Machine learning

Osei Sekyere et al., 2025, iScience *28*, 113473
October 17, 2025
https://doi.org/10.1016/j.isci.2025.113473

CellPress

iScience

CellPress
OPEN ACCESS

Article

Lactobacillus-rich cervicovaginal microbiome associated with lower BV, HPV, and cytology outcomes in women

John Osei Sekyere,[1,2,3,*] Jason Trama,[1] Martin Adelson,[1] Charulata Trikannad,[1] Desiree DiBlasi,[1] Rachel Schuster,[1] Jing Jing Yang,[1] and Eli Mordechai[1]

[1]Institute of Biomarker Research, Department of Clinical Development, Medical Diagnostic Laboratories, Hamilton Township, NJ 08690, USA
[2]Department of Medical Microbiology, School of Medicine, University of Pretoria, Pretoria 0084, South Africa
[3]Lead contact
*Correspondence: joseisekyere@mdlab.com
https://doi.org/10.1016/j.isci.2025.113473

SUMMARY

The cervicovaginal microbiome modulates susceptibility to bacterial vaginosis (BV), high-risk human papillomavirus (hrHPV) infection, and epithelial cell abnormalities that precede cervical cancer. We retrospectively analyzed 15 607 qPCR-profiled cervicovaginal specimens from U.S. women (ages 14–95; 32 states) and integrated microbiome abundances, hrHPV genotyping, Pap-cytology, and demographics. BV was present in 53% and hrHPV in 11% of samples. Lactobacillus crispatus, L. gasseri, and L. jensenii were enriched in BV-negative and cytologically normal (NILM) samples, whereas L. iners and BV-associated anaerobes co-occurred with hrHPV and abnormal cytology. Machine-learning models confirmed age, hrHPV status, and L. crispatus abundance as the strongest multivariate predictors of BV and cytological outcomes (BV AUROC ≈0.97). Interaction analyses revealed synergistic associations between specific hrHPV genotypes and Gardnerella/Fannyhessea that further increased cytological risk. These findings underscore the clinical value of microbiome profiling and support probiotic strategies that promote protective Lactobacillus communities to reduce BV and hrHPV-related cervical pathology.

INTRODUCTION

The normal vaginal microbiome is comprised of abundant H_2O_2-producing *Lactobacillus* spp. and a low abundance of other anaerobic bacteria, which maintain an acidic vaginal microenvironment (pH < 4.5).[1] Under the regulation of circulating estrogen, glycogen is deposited into the vaginal lumen by epithelial cells and is metabolized by *Lactobacillus* spp. into lactic acid.[1,2] Classical *Lactobacilli* such as *L. crispatus* ferment these polysaccharides, whereas *L. iners* lacks the full glycogen-utilization arsenal and instead relies on host-derived maltose and glucose, a nuance increasingly recognized in recent metabolomic studies.[3–6] The lactic acid produced by *Lactobacillus* spp. lowers the local pH, thereby creating a hostile niche for pathogens while supporting acid-tolerant commensals, keeping the diversity of the vaginal microbiome low.[7] The inability of some *Lactobacillus* spp. such as *L. iners* to transform glycogen to lactic acid stems from its lack of the machinery to ferment glycogen. Hence, they instead rely on host enzymes' breakdown products. Thus, the traditional view of vaginal glycogen fueling *Lactobacilli* may not uniformly apply to all species.

Large-scale sequencing studies have revealed that vaginal microbial communities fall into five reproducible Community State Types (CST I–V).[8,9] CSTs I, II, and III are dominated by *Lactobacillus crispatus, L. gasseri, or L. iners,* respectively; CST IV is diverse and anaerobe-rich, whereas CST V contains *L. jensenii*. Epidemiological work shows that CSTs rich in non-lactobacilli (particularly CST IV) correlate with bacterial vaginosis (BV), heightened genital inflammation, and reduced clearance of high-risk HPV.[10–12] These observations motivated us to quantify how individual Lactobacillus spp. and BV-associated taxa co-occur with HPV genotypes and cytological outcomes in a large, racially heterogeneous U.S. cohort.

The depletion of *Lactobacillus* spp. and an increase in vaginal pH allow other microbial species to proliferate, leading to increased microbial diversity and conditions such as vulvovaginal candidiasis and bacterial vaginosis (BV). This further predisposes the vagina and cervix to sexually transmitted infections (STIs) such as gonorrhea, human papillomavirus (HPV), and HIV infections, obstetric complications, and cervical cancer[7,13]. Notably, BV is a notable risk factor for cervical HPV infections.[14] HPV infections, on the other hand, are a major cause of cervical cancer, affecting around 26.8–38.4% of women aged 15–59 years[15,16] and 31% of men.[17]

Human papillomaviruses (HPVs) comprise about 450 distinct genotypes that infect cutaneous and mucosal epithelia.[18] Of these, 17 are currently classified as oncogenic ("high-risk") because they are detected in ≥99% of cervical cancers: these include HPV-18, -16, −31, −33, −35, −45, −52, −56, −58, −59, −66, −68, −73, and −82[19,20] The non-oncogenic

(low risk) genotypes are typically associated with anogenital warts.[17,18] Persistent high-risk HPV (hrHPV) infection is the primary etiological factor in cervical cancer: hrHPV types 16 and 18 alone cause ~50% and ~16% of cervical cancer cases, respectively, but most women who acquire a high-risk HPV will not progress to cancer because productive immune responses usually clear the infection within 1–2 years.[21] Progression to malignancy therefore depends on a constellation of cofactors, including persistent high-risk HPV infection, host genetics, hormonal milieu, smoking, and, increasingly, cervico-vaginal microbiome composition, rather than viral presence alone[22,23]

Although most women with hrHPV do not develop cervical cancer, persistent infection can drive cytological abnormalities ranging from atypical squamous cells of undetermined significance (ASC-US) to low- (LSIL) and high-grade (HSIL) squamous intra-epithelial lesions.[21,22] Based on cytology findings, the cervix's squamous cell epithelium can be classified as negative for intraepithelial lesion or malignancy (NILM) for those with no cytological abnormality, while abnormal Pap-smear results are classified into three major categories: atypical squamous cells of undetermined significance (ASC-US), atypical squamous cells–cannot exclude HSIL (ASC-H), low-grade squamous intraepithelial lesions (LSIL), high-grade squamous intraepithelial lesions (HSIL), and squamous cell carcinoma (SCC).[24]

Despite the well-established role of vaginal microbiota in modulating HPV infection and cervical dysplasia risk, significant knowledge gaps remain. Most earlier studies were limited by small sample sizes and cross-sectional designs, and they often treated the vaginal microbiome in broad strokes – for example, comparing "*Lactobacillus*-dominant" vs. "dysbiotic" communities – without capturing the finer gradations of community structure.[23] While CSTs have provided a useful shorthand, many subtypes and transitional states ("transitional BV" denotes Nugent-equivalent scores 4–6, an intermediate community recognized in both molecular[10] and clinical[25] literature) are underexplored; for instance, intermediate microbiota states (sometimes called *transitional BV*) with moderate *Lactobacillus* and some anaerobes do not neatly fit the classical CST categories. Furthermore, prior microbiome-HPV studies often focused on individual microbes or simple correlations, lacking large-scale multivariate modeling to account for the complex interplay of multiple HPV types and bacterial communities.

Indeed, few studies to date have applied machine-learning approaches to vaginal microbiome data,[23] so the predictive potential of combined microbiome and HPV information remains largely untapped. Recent research hints that specific CSTs and higher microbial diversity are associated with persistent hrHPV infection and cervical lesion development,[25,26] underscoring the need for comprehensive models. Thus, the rationale for our current study is to fill these gaps by analyzing a large cohort with an integrated approach: we profile the cervicovaginal microbiome (via CSTs and key taxa), multiple HPV genotypes, and host factors, and employ advanced multivariate and machine-learning techniques to identify patterns that simpler analyses might miss. This approach allows us to test the hypothesis that certain vaginal community compositions, including those intermediate states, synergize with hrHPV to influence cervical health outcomes, which has important implications for risk prediction and intervention.

We used a validated 22-target qPCR panel rather than 16S rRNA amplicon sequencing because it (i) distinguishes the clinically relevant Lactobacillus spp. that are indistinguishable by short-read 16S, and (ii) provides absolute, not relative, quantitation—crucial for the BV algorithm applied in this study.[10] To our knowledge, no prior investigation has combined organism-specific qPCR, high-risk HPV genotyping and paired Pap cytology for >15 000 U.S. patients to disentangle microbe–virus–host interactions at this scale.

While prior studies highlight the associations of *Lactobacilli* with reduced BV, HPV, and ECA incidence, this study uniquely examines a large cohort to evaluate demographic, microbiological, and cytological predictors in the presence and absence of *Lactobacilli*.

RESULTS

Cohort overview and specimen characteristics

We obtained 19105 clinical specimen records from the MDL database between August 1, 2021, and April 5, 2023; 3,498 specimen data were removed, resulting in a final dataset of 15, 607 specimens from 15607 persons. The specimens were obtained from healthcare providers located in 32 States and the District of Columbia (DC). Most samples were obtained from Texas (n = 2327), Arizona (n = 1790), Illinois (n = 1784), Florida (n = 1660), Louisiana (n = 1517), Michigan (n = 1511), and California (n = 1052). Females were mostly (n = 15541, 99.57%) the source of these specimens, followed by males (n = 13, 0.08%) and unknown gender (n = 53, 0.34%). The ages were between 14 and 95 years: most samples were from persons between 20 and 60 years; the highest number of samples was from age 30 (n = 689) (Table S1.1–1.3; Figure 1).

Each sample was tested for nine bacterial species and 13 HPV subtypes, with both concentrations and CT scores recorded (Table S1). Among bacteria, *Megasphaera* sp. Type 1 had the highest mean concentration, while among HPV types, HPV-18 was the highest. However, *L. iners* (n = 6813, 43.65%), *F. vaginae* (n = 5913, 37.89%), *G. vaginalis* (n = 5431, 34.80%), and *L. crispatus*(n = 5183, 33.21%) were most prevalent bacteria among the samples (Dataset S1.4) while HPV-52 (n = 226, 1.45%), HPV-51 (n = 209, 1.34%), HPV-35 (n = 188, 1.20%), HPV-56 (n = 187, 1.20%) and HPV-16 (n = 185, 1.19%) were most prevalent HPV types (Dataset S1.5).

There were more BV-positive (n = 8282, 53.07%) diagnoses than BV-negative (*n* = 6919, 44.33%) and intermediate microbiota state (*n* = 406, 2.60%) (Dataset S1.6). Contrarily, 1726 (11.06%) samples/patients were positive for hrHPV, while 13881 (88.94%) were negative for hrHPV (Dataset S1.7). Further, the Pap smear cervical cytology results had 75.28% NILM (*n* = 11749) diagnosis followed by 14.29% ASCUS (*n* = 2231), 7.14% LSIL (*n* = 1115), 1.04% RCC (*n* = 163), 0.53% HSIL (*n* = 82), 0.39% ASC-H (*n* = 61), and 0.16% AGC (*n* = 25) (Dataset S1.8).

Age and provider states were significantly associated with BV status and cytology outcomes

There was a significant distribution of age across the different States, BV diagnosis, and hrHPV outcomes. Notably, age per bacteria species and BV status had little variation except for *L. gasseri* that was mostly detected in patients aged between 35 and 50. HPV-18 and -33 mostly occurred in older populations

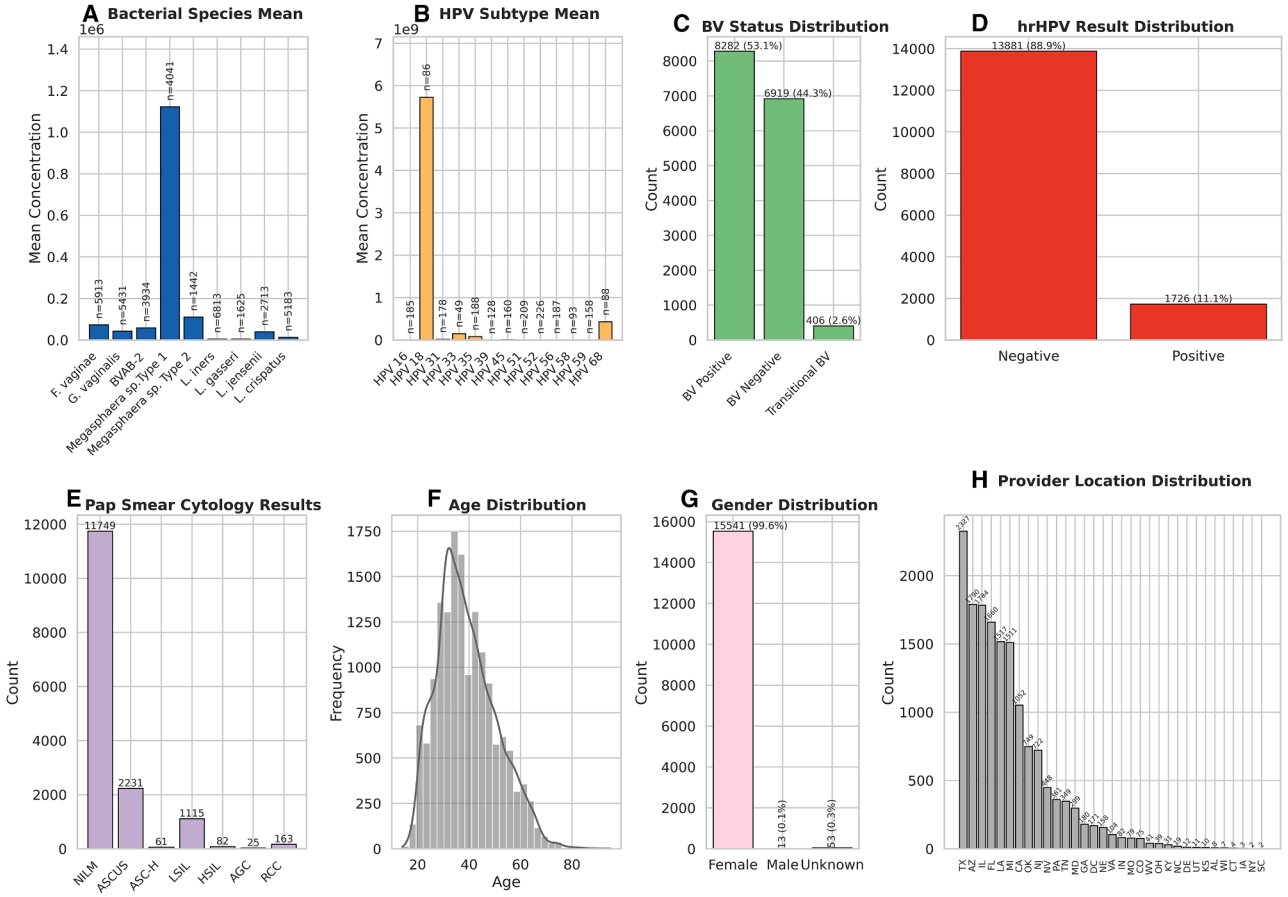


Figure 1. Epidemiological and microbiological summary

Panels A–H show the overall distributions of key demographic, cytological, microbiological and HPV variables in our 15 607-sample cohort.

(A) Bacterial Species Mean Concentration. *L. crispatus* had the highest mean concentration (∼1.1 × 10^6 copies; *n* = 5183), followed by L. iners (∼4.5 × 10^5; *n* = 6813).

(B) HPV Subtype Mean Concentration. HPV-52 was most abundant (∼6.0 × 10^8 copies; *n* = 226), while HPV-16 had *n* = 185.

(C) BV Status Distribution. 53.1% (*n* = 8282) were BV-Positive, 44.3% (*n* = 6919) BV-Negative, 2.6% (*n* = 406) Transitional.

(D) hrHPV Result Distribution. 11.1% (*n* = 1726) hrHPV-Positive vs. 88.9% (*n* = 13 881) hrHPV-Negative.

(E) Pap Smear Cytology Results. 75.3% negative for intraepithelial lesion or malignancy (NILM, *n* = 11,749), 14.3% atypical squamous cells of undetermined significance (ASCUS, *n* = 2231), 7.1% low-grade squamous intraepithelial lesion (LSIL, *n* = 1115), 1.0% reactive cellular changes (RCC, *n* = 163), ≤0.5% high-grade squamous intraepithelial lesion (HSIL), atypical squamous cells cannot exclude HSIL (ASC-H), or Atypical Glandular Cells (AGC).

(F) Age Distribution. Median age ≈38 years (range 14–95), with peak between 30 and 40 years

(G) Gender Distribution. 99.6% female (*n* = 15 541), 0.1% male (*n* = 13), 0.3% unknown (*n* = 53).

(H) Provider Location Distribution. Most samples from TX (*n* = 2327), AZ (*n* = 1790), IL (*n* = 1784), FL (*n* = 1660), LA (*n* = 1517), MI (*n* = 1511), CA (*n* = 1052).

Statistics: two-sided χ^2 (categorical) or Welch's t/ANOVA (continuous) as indicated; *q < 0.05, **q < 0.01, ***q < 0.001 after Benjamini–Hochberg FDR correction; n = biological specimens indicated in each panel. Full test details are provided in Supplementary Excel file "Statistics_Table.xlsx" and under "quantification and statistical analysis" in STAR Methods. *Data are shown as mean ± SEM where applicable; individual sample counts (n) per bar are indicated in the figure or text. Two-sided tests were used; p-values were FDR-adjusted (Benjamini–Hochberg). Significance: q < 0.05 (), q < 0.01 (), q < 0.001 ().*

than the rest while samples from Delaware, Missouri, Colorado, and DC were from older patients. Further, AGC, HSIL, and RCC were mostly in comparatively older patients. hrHPV-positive samples were in a comparatively younger population (Figure S1; Dataset S1.9–S1.15).

Gender showed significant associations only with provider location. The ages of the different genders were mostly within the same brackets. All the bacterial species (except *L. gasseri* in males) were also found in males and unknown gender patients, with 8249, 4, and 29 BV-positive specimens being from females, males, and unknown gender, respectively. Comparatively, fewer HPV types were found in males (HPV-31 and -39) and unknown gender (HPV-16, -45, −52, and −59). ECAs such as ASCUS and AGC were found in males and unknown gender, while LSIL was found in unknown gender (Figures S2–S4; Dataset S1.16–S1.23).

Except for the HPV subtypes, there was a significant association between the provider's location/state and cervical cytology, hrHPV, BV, bacterial species, and HPV subtypes. Notably, there were more BV positive (except in Colorado, California, Delaware, Michigan, Oklahoma, and Tennessee) and hrHPV-negative

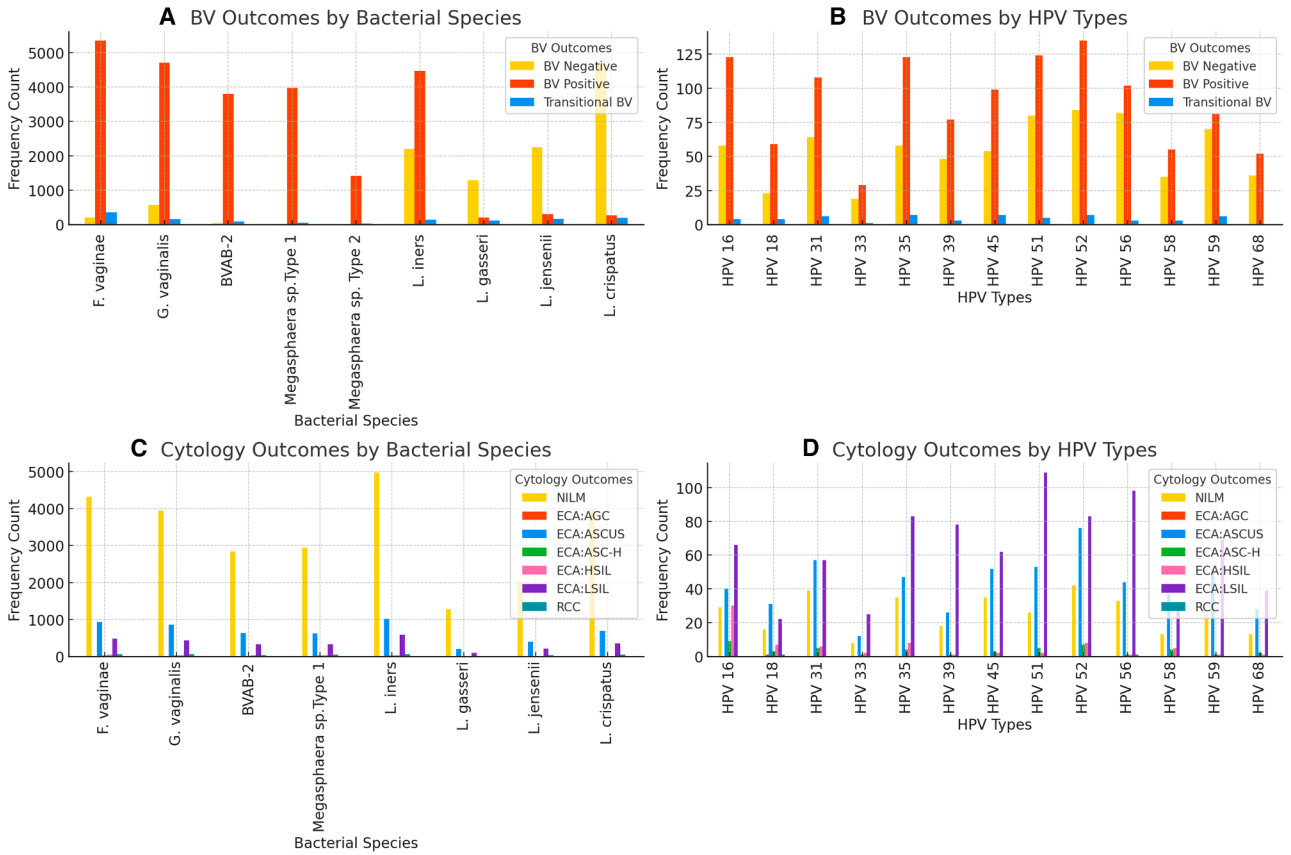


Figure 2. BV and cytology outcomes by bacterial species and HPV Genotypes

Bars show the percent of samples in each diagnostic category, stratified by species (left panels) or genotype (right panels).

(A) BV Outcomes by Bacterial Species. > 90% of *F. vaginae, G. vaginalis,* BVAB-2, *Megasphaera* spp. were BV-Positive; only 23% of L. iners and >80% of L. gasseri/jensenii/crispatus were BV-Negative.

(B) BV Outcomes by HPV Genotypes. BV-Positive ranged 55–69% across hrHPV types; HPV-52 and HPV-56 had the lowest BV co-positivity (∼59%).

(C) Cytology Outcomes by Bacterial Species. LSIL and ASCUS were most frequent cytological abnormalities among samples positive for *F. vaginae* and *G. vaginalis*, whereas NILM predominated (>70%) in *L. crispatus* -positive samples.

(D) Cytology Outcomes by HPV Genotypes. HPV-16, HPV-18, and HPV-31 showed the highest proportions of LSIL and HSIL combined (approximately 35–60%), whereas other genotypes (e.g., HPV-59, HPV-68) showed lower prevalence of cytological abnormalities (∼30%).

diagnoses in most states, while ASCUS and LSIL were relatively more prevalent in Arizona, Florida, Illinois, Louisiana, Michigan, Oklahoma, and Texas; these states also had substantial presence of almost all the HPV types. The bacteria species were commonly detected in relatively higher concentrations than the HPV types in samples from Arizona, California, DC, Florida, Georgia, Illinois, Louisiana, Maryland, New Jersey, Nevada, Oklahoma, Pennsylvania, Tennessee, and Texas (Dataset S1.24–S1.27; Figures S5–S8).

Together, these baseline observations (Figure 1; Table S1) established a predominantly female, reproductive-age cohort with high BV prevalence (53%) but comparatively low hrHPV positivity (11%), providing statistical power for the downstream stratified analyses (Specimens from males or unknown gender were retained for completeness but excluded from all inferential statistics).

Microbiota composition by bacterial vaginosis status

There was a substantially higher and statistically significant presence of *F. vaginae* (90.48%), *G. vaginalis* (86.60%), *BVAB-2* (96.54%), *and Megasphaera* sp. *Type 1* (98.42%), *Megasphaera* sp. *Type 2* (97.92%), and *L. iners* (65.54%) in BV-positive samples; specifically, 2201 BV-negative samples (32.31%) were *L. iners*-positive compared with 4465-positive *L. iners* (65.54%) in BV-positive samples. Furthermore, HPV-positive samples had higher BV-positive diagnoses (between 51.90% and 68.60%) than BV-negative diagnoses. Contrarily, *L. gasseri* (79.75%), *L. jensenii* (82.68%), and *L. crispatus*(90.91%) were mostly found in BV-negative specimens (Figures 2, 3, and S10–S15; Dataset S2.1–S2.6).

To further explore how individual microbial taxa and HPV types relate to disease outcomes, we stratified BV and cytology categories by the presence of specific bacterial species and hrHPV genotypes (Figure 3). BV-positive outcomes predominated in samples with *F. vaginae*, *G. vaginalis*, BVAB2, and *Megasphaera* spp., whereas *L. gasseri*, *L. jensenii*, and *L. crispatus* were more frequent in BV-negative samples. Cytology outcomes showed that NILM was dominant in samples with lactobacilli, while

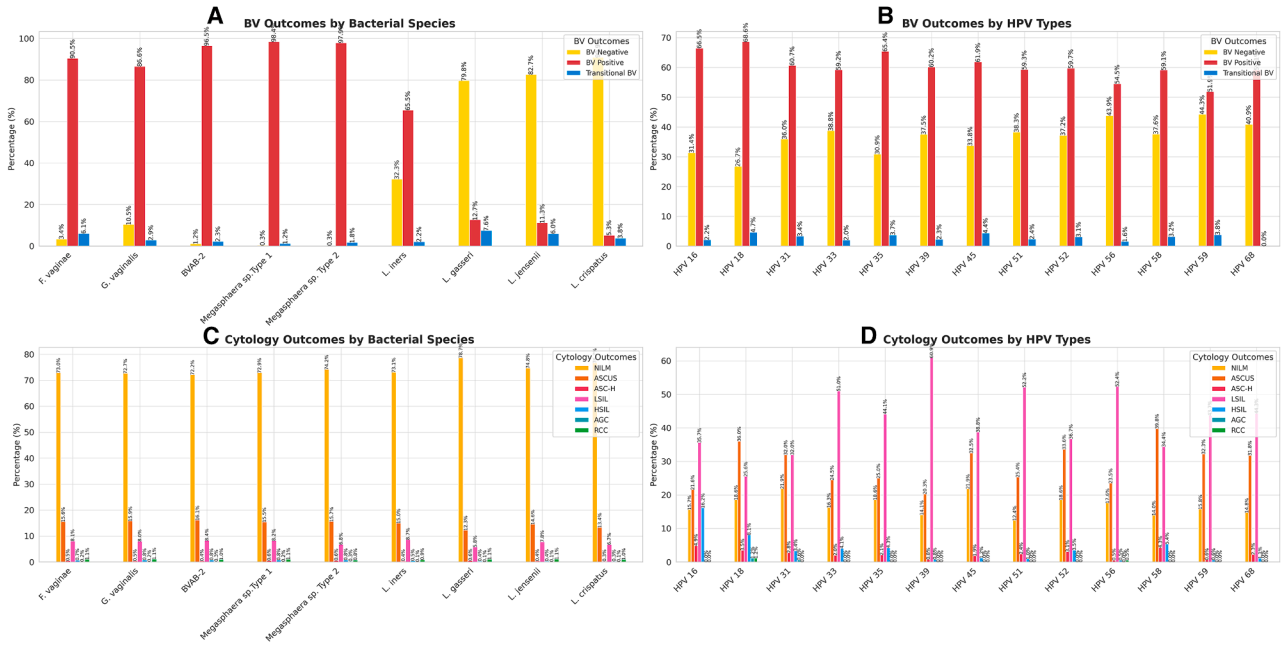


Figure 3. Associations of BV and cytology outcomes with bacterial species and HPV Types

(A) BV outcomes (BV-negative, BV-positive, intermediate/transitional BV) stratified by the detection of eight common cervicovaginal bacterial species (*Fannyhessea vaginae*, *Gardnerella vaginalis*, BVAB2, *Megasphaera* sp. type 1, *Megasphaera* sp. type 2, *Lactobacillus iners*, *Lactobacillus gasseri*, *Lactobacillus jensenii*, and *Lactobacillus crispatus*).

(B) BV outcomes stratified by the detection of 12 high-risk HPV types (HPV16, HPV18, HPV31, HPV33, HPV35, HPV39, HPV45, HPV51, HPV52, HPV56, HPV58, HPV59, and HPV68).

(C) Cytology outcomes (NILM, ASC-US, ASC-H, LSIL, HSIL, AGC, and RCC) stratified by the same bacterial species.

(D) Cytology outcomes stratified by the same 12 HPV types. Data are presented as percentages of total positive detections for each species or HPV type. Percentages above bars indicate category-specific proportions. All associations were tested using chi-square analysis; $p < 0.05$ considered statistically significant.

ASC-US, LSIL, and HSIL were more common in HPV16-, HPV18-, and HPV33-positive samples. These patterns were statistically significant (Chi-square $p < 0.05$).

The mean concentrations of *F. vaginae* (89.60%), *G. vaginalis* (93.23%), *BVAB-2* (97.80%), *Megasphaera* sp. *Type 1* (99.87%), *Megasphaera* sp. *Type 2* (99.01%), and *L. iners* (58.85%) were higher in BV-positive samples. Further, the mean concentrations of *L. gasseri* (56.33%), *L. jensenii* (83.37%), *and L. crispatus* (80.57%) were high in BV-negative specimens and very minimal (1.75%–0.3%) in BV-positive ones (Figures S10–S15; Dataset S2.1–S2.6). In contrast to prevalence data, the mean concentrations of HPV subtypes were not generally higher in BV-positive samples except HPV-33 (99.96%), HPV-35 (92.73%), HPV-39 (63.12%), and HPV-51 (59.89%); the remaining subtypes had percentage mean concentrations of 0.17%–48.48% in BV-positive samples. None of the mean concentrations of HPV subtypes per BV status was significant (Figures S10–S15; Dataset S2.1–S2.6).

Microbiota composition by cervical cytopathology

Whereas the bacterial species were all commonly detected in NILM, ASCUS, and LSIL-positive samples in descending order, their mean concentrations varied by species, with *F. vaginae, G. vaginalis, BVAB-2,* and *Megasphaera* sp. Types 1–2 being relatively higher in HSIL, ASCUS, LSIL, and AGC than in NILM in many cases. Generally, the concentrations of the above-listed species were higher than those of the *Lactobacillus* species in the HSIL, ASCUS, ASC-H, LSIL, and AGC-positive samples. However, the *Lactobacillus* species had higher mean concentrations in the NILM-positive specimens. The HPV-positive specimens, however, were mostly HSIL, ASCUS, ASC-H, LSIL, and AGC-positive than NILM-positive and (except for HPV-18, HPV-33, HPV-35) had higher concentrations in ECA-type pathologies (Figures S10–S15; Dataset S2.1–S2.6). A heatmap showing the chi-square *p*-values of these associations is shown in Figure S16.

Correlation and pairwise association analyses

The correlation matrix plot (Spearman ρ) (Figure S17; Dataset S3.1) confirmed the above findings: the bacterial pathogens were strongly and positively correlated with the presence of BV, while the three *Lactobacillus* species (*gasseri, jensenii,* and *crispatus*) were strongly negatively correlated. There was a strong negative correlation between NILM and ASCUS, SIL, HSIL, AGC, and ASC-H, and a positive correlation between the HPV subtypes and LSIL. HSIL was strongly positively correlated with HPV-16. NILM was positively correlated with age, while all BV, the ECA-types, HPV subtypes, and bacterial species were weakly negatively correlated with age. A Chi-square and ANOVA pairwise analyses of the various factors showed a significant association

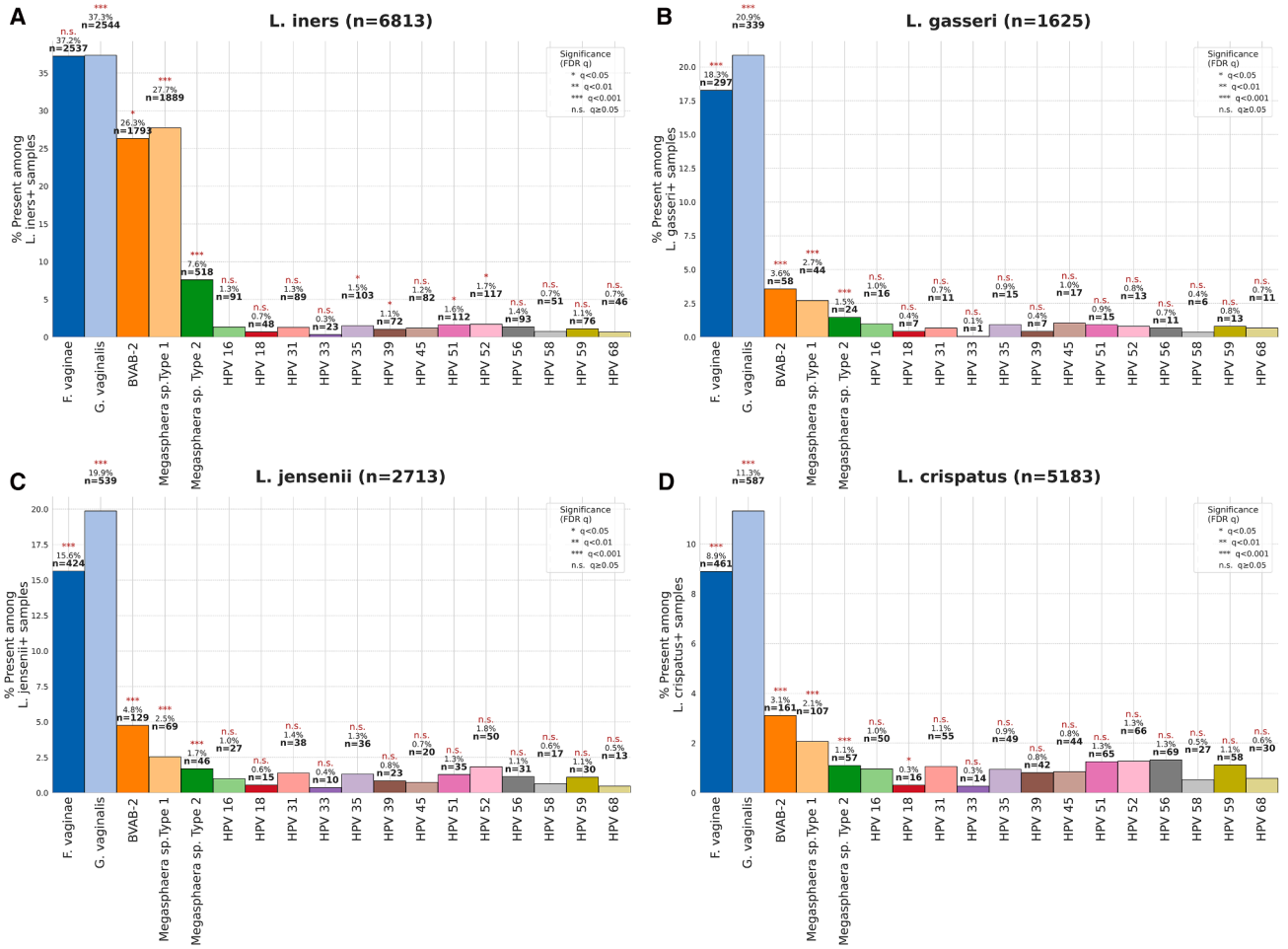


Figure 4. Frequency of other pathogens in Lactobacillus-positive samples

Panels A–D each show, for one *Lactobacillus* species, the percentage of those samples also positive for BV-associated bacteria or hrHPV subtypes, expressed as proportions of their respective *Lactobacillus*-positive denominators clearly indicated (*L. iners n* = 6813, *L. gasseri n* = 1625, *L. jensenii n* = 2713, *L. crispatus n* = 5183).

(A) *L. iners* (*n* = 6813): 65.5% co-occurrence with F. vaginae; ~0.7–1.5% with most HPV types.

(B) *L. gasseri* (*n* = 1625): Only 18.3% co-occurrence with F. vaginae; <2.1% with hrHPV.

(C) L. jensenii (*n* = 2713): 15.6% F. vaginae; <1.8% for all HPV.

(D) *L. crispatus* (*n* = 5183): 8.9% *F. vaginae*; <1.3% for hrHPV. Asterisks denote χ^2 significance (FDR q-values): *q < 0.05; **q < 0.01; ***q < 0.001; n.s. q ≥ 0.05. Statistics: two-sided χ^2 (categorical) or Welch's t/ANOVA (continuous) as indicated; *q < 0.05, **q < 0.01, ***q < 0.001 after Benjamini–Hochberg FDR correction; n = biological specimens indicated in each panel. Full test details are provided in Supplementary Excel file "Statistics_Table.xlsx" and under "quantification and statistical analysis" in STAR Methods.

between age, State, BV status, ASCUS, and LSIL, and most of the other factors tested, confirming several observations from the correlation matrix (Figures S16–S18).

Lactobacillus-pathogen associations and odds ratios

As shown in Figure 4, there was a markedly lower prevalence of bacterial pathogens and HPV types in samples containing *L. gasseri, L. jensenii, or L. crispatus* compared with those containing *L. iners.* A Spearman correlation matrix showed stronger negative correlation between *L. crispatus* and L. *gasseri* and the bacterial pathogens in terms of their concentrations (Figure S19). Notably, chi-square analysis revealed significant associations between *L. iners, L. gasseri, L. crispatus*, and various HPV subtypes and bacterial pathogens impacting BV and cervical cytology outcomes; *L. jensenii* had mostly insignificant associations with HPV. The odds ratio forest plot further showed that *Lactobacillus* species was associated with lower odds (odds ratios <1) of occurrence of certain bacteria species, HPV types, and cervical cytology outcomes (Figure 5; Dataset S3.3).

Notably, for several HPV subtypes (HPV-16) and bacterial pathogens (*G. vaginalis*), *L. crispatus* had odds ratios <1, indicating a negative association. *L. iners*, however, had odds ratios >1 for several pathogens, particularly BVAB-2 and *Megasphaera* sp. Type 1, suggesting that its presence may increase the likelihood of detecting these harmful bacteria. *L. gasseri* and *L. jensenii* showed mixed effects, with *L. gasseri* being negatively

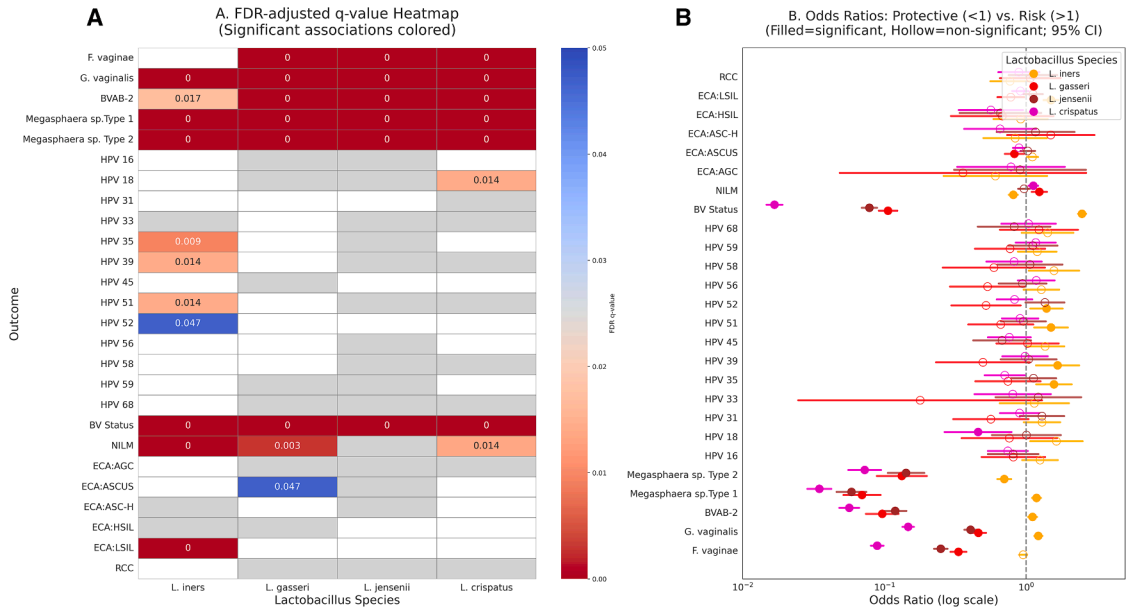


Figure 5. Statistical associations between Lactobacilli and outcomes

(A) FDR-adjusted q-value Heatmap. Most BV-associated bacteria show q < 0.001 (deep red) vs. all four *Lactobacilli*; only HPV-35/39/51/52 show q < 0.05. NILM has q = 0.003 for *L. gasseri* and q = 0.014 for *L. crispatus*.

(B) Odds Ratios and 95% CIs. *L. crispatus* ORs <1 for *F. vaginae* (OR≈0.10), *G. vaginalis* (≈0.12), HPV-16 (≈0.85); *L. iners* ORs >1 for BVAB-2 (≈1.25) and *Megasphaera* sp. Type 1 (≈1.30). Hollow markers denote non-significant ORs. Statistics: two-sided χ^2 (categorical) or Welch's t/ANOVA (continuous) as indicated; *q < 0.05, **q < 0.01, ***q < 0.001 after Benjamini–Hochberg FDR correction; n = biological specimens indicated in each panel. Full test details are provided in Supplementary Excel file "Statistics_Table.xlsx" and under "quantification and statistical analysis" in STAR Methods.

associated (odds ratios <1) with HPV-45 and some bacterial pathogens, while *L. jensenii* was negatively associated with HPV-35 (Figure 5).

Interaction effects and unsupervised clustering

An interaction effect analysis of all the bacterial species and HPV subtypes showed a significant deleterious effect on vaginal and cervical health when certain HPV subtypes and/or bacterial pathogens co-occurred together. Specifically, HPV-16 and HPV-31, HPV-45 and HPV-51, HPV-52 and HPV-58, HPV-16, HPV-16 and *G. vaginalis,* and HPV-68 and *F. vaginae,* among others, were associated with ECAs. However, most of the non-*L. iners Lactobacillus*-pathogen interactions with bacterial and HPV pathogens resulted in BV negative or NILM outcomes. See Supplemental Dataset S3.4–S3.9 and S5 for details (Figures S20–S25).

Principal component analysis (PCA) was undertaken to determine the variance and spread of the various bacterial species toward the two components (outcomes), P1 and P2. The loadings of PC1 are dominated by *F. vaginae, G. vaginalis*, and BVAB-2, suggesting that they contribute heavily to the variance along PC1. *L. iners* had a significant loading on PC2, meaning it influences variability in a different direction compared with other species. HPV types, such as HPV-59 and HPV-45, also had notable loadings on PC2, indicating their contribution to variance along that component (Figure S26; Dataset S4.1).

An average bacterial and HPV subtype concentrations per K-means clusters 0, 1, and 2 (Figures S27–S28) show that cluster 1 is associated with ASCUS, HSIL, BV-positive or intermediate microbiota state cases owing to the higher concentrations of *F. vaginae, G. vaginalis,* BVAB-2, *Megasphaera* sp. Type 1, *L. iners,* HPV-16, HPV-35, HPV-45, HPV-59, and HPV-68; other subtypes such as HPV-18, HPV-31, and HPV-33 also had considerable concentrations. However, cluster 2 had elevated concentrations of *L. iners, L. crispatus, L. jensenii*, and lower HPV concentrations, except HPV 31 and HPV 56, suggesting that this cluster is likely associated with a healthy vaginal microbiome, normal cytology (NILM), or HPV-negative cases (Figures S27–S28; Dataset S4.2).

Machine-learning performance and interpretability

All four supervised models, viz., Extreme Gradient Boosting (XGBoost), Random Forest, Logistic Regression, and a single Decision Tree, were trained on the full feature set of demographics plus 22-target qPCR profiles. Feature-importance plots (Figure 6) consistently identified *L. crispatus* abundance, patient age, and hrHPV result as the top three predictors across ensemble methods. Receiver-operating-characteristic analysis showed excellent discrimination for BV (median AUROC = 0.97 ± 0.01) and moderate discrimination for cytology (0.74 ± 0.02) (Figure 7).

To visualise how individual features influence BV predictions, Shapley Additive exPlanations (SHAP) were generated: the beeswarm plot (Figure 8) confirms that high *F. vaginae* and *G. vaginalis* push the model toward BV-positive, whereas high *L. crispatus, L. jensenii* and *L. gasseri* favor BV-negative outcomes. SHAP interaction values (not shown) further revealed a synergistic effect between *L. iners* and *Megasphaera* sp. *Type*

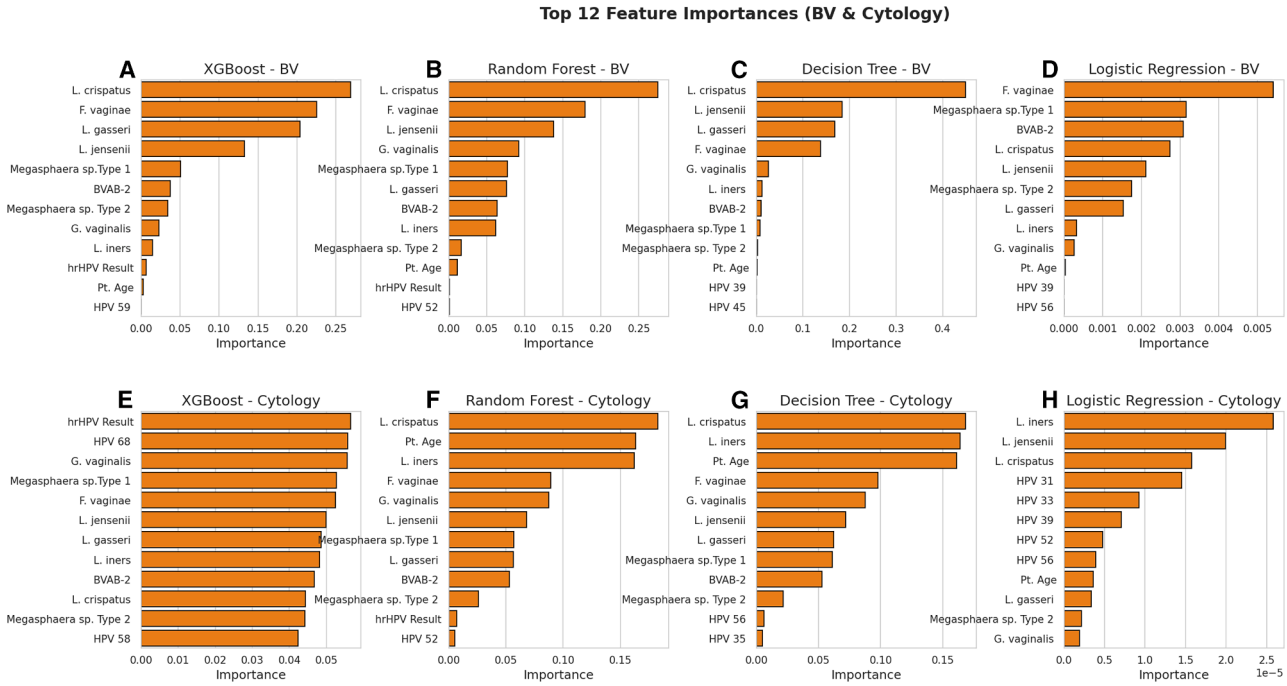


Figure 6. Top 12 feature importances for BV status and cytology

Panels show the twelve most influential predictors (bar length = relative importance) for each predictive model. BV status predictions (Panels A–D) consistently identified key bacterial species (*L. crispatus, F. vaginae*) with high importance and high model accuracy (ROC-AUC ≈0.97–1.00, see Figure 7A). In contrast, cytology predictions (Panels E–H) exhibited lower accuracy (ROC-AUC ≈0.49–0.51, see Figure 7B) and varied considerably in top predictors, highlighting the complexity of cytological outcomes.

(A) XGBoost – BV Status (top: *L. crispatus* 0.27, *F. vaginae* 0.23, *L. gasseri* 0.21).

(B) Random Forest – BV Status (top: *L. crispatus* 0.28, *F. vaginae* 0.19, *L. jensenii* 0.17).

(C) Decision Tree – BV Status (top: *L. crispatus* 0.42, *L. jensenii* 0.18, *L. gasseri* 0.18).

(D) Logistic Regression – BV Status (top: *F. vaginae* 0.0053, *Megasphaera* sp. Type 1 0.0038).

(E) XGBoost – Cytology (top: hrHPV Result 0.29, HPV-68 0.29, G. vaginalis 0.29).

(F) Random Forest – Cytology (top: *L. crispatus* 0.18, Pt Age 0.16, L. iners 0.14).

(G) Decision Tree – Cytology (top: Pt Age 0.17, *L. crispatus* 0.16, *L. iners* 0.16).

(H) Logistic Regression – Cytology (top: *L. iners* 2.6 × 10^–5, *L. jensenii* 2.0 × 10^–5).

1, echoing the logistic-interaction findings (Figure S23). Taken together, the ML results buttress the classical negatively associated pathogen-*Lactobacillus* dichotomy and quantify their relative contributions in a multivariate context.

DISCUSSION

This retrospective analysis of 15 607 U.S. cervicovaginal samples provides one of the largest qPCR-based snapshots to date of how specific microbial signatures correlate with BV, hrHPV infection, and cytology-defined abnormalities. While our cross-sectional design precludes causal inference, several consistent associations emerged that both reinforce and extend the current cervicovaginal microbiome literature. All associations reported were robust after adjustment for multiple comparisons using false discovery rate (FDR; $q < 0.05$).

Unsupervised clustering and PCA revealed three compositional states that broadly mirror the community state type framework proposed by Ravel et al. (2011). Although CSTs are typically defined by 16S profiles, our *L. crispatus –, L. gasseri–, L. iners–,* and *L. jensenii*-dominant patterns broadly correspond to CST-I/II/III/V, respectively. "Cluster 2", dominated by *Lactobacillus crispatus*, *L. gasseri*, and *L. jensenii*, aligns with CST I/II and was enriched for BV-negative and NILM cytology results. Conversely, "Cluster 1" displayed high loads of *Gardnerella vaginalis*, *Fannyhessea vaginae*, BVAB2, *Megasphaera* spp., and multiple hrHPV genotypes: an ecological profile analogous to CST IV that has been repeatedly linked to dysbiosis and cervical disease progression.[27] The transitional cluster resembled CST III, characterized by a predominance of *L. iners* with moderate anaerobe coabundance, underscoring the species' ambiguous ecological role.[28] This aligns with prior studies describing *L. iners* as a transitional or opportunistic species associated with microbiome instability.

The associations of *L. crispatus, L. gasseri,* and *L. jensenii* with hrHPV infection, BV, and cervical cytological abnormalities were determined using 15607 clinical data from 15607 anonymized patients that reported to healthcare providers in 32 states and the D. C, USA.

The molecular, pathological, and statistical analyses of this extensive cohort consistently demonstrated strong associations, especially highlighting the inverse correlation of

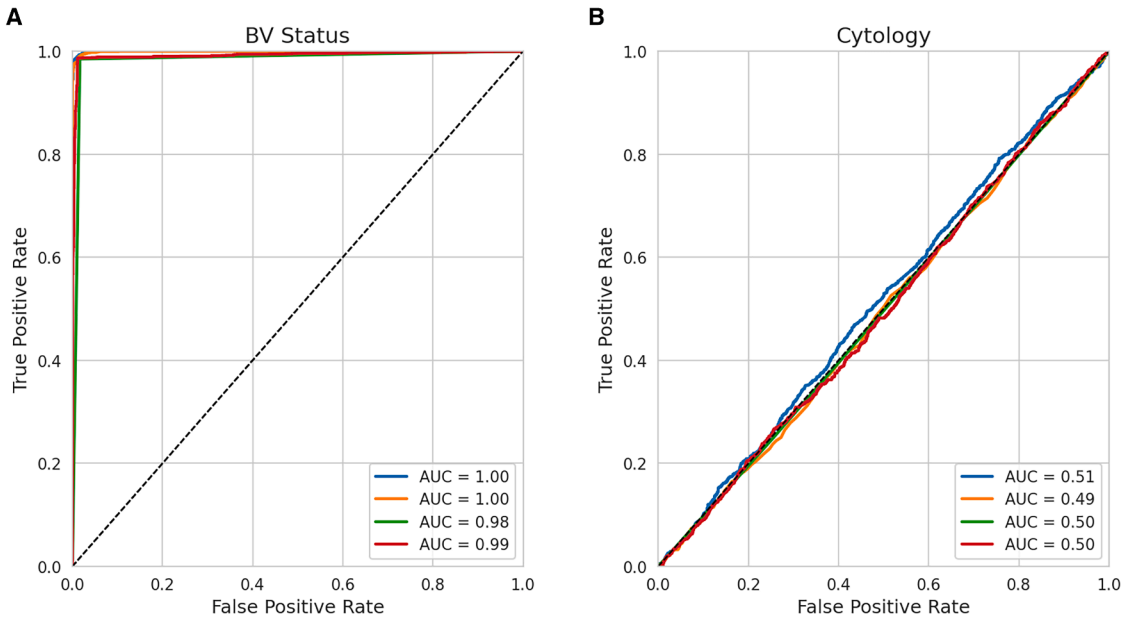


Figure 7. ROC curves for BV status and cytology predictions

(A) BV Status: Models showed excellent predictive accuracy (XGBoost and Random Forest AUC = 1.00, Decision Tree AUC = 0.98, Logistic Regression AUC = 0.99).

(B) Cytology: All models performed near chance levels (AUC = 0.49–0.51), reflecting challenges inherent in predicting complex cytological abnormalities when aggregated into binary outcomes. Curves show mean ROC with 95% CI (shaded). For bar/beeswarm SHAP plots, points are individual samples; bars indicate mean(|SHAP|).

L. crispatus with pathogenic bacterial and viral markers and, to some extent, *L. gasseri* and *L. jensenii.* Specifically, *L. crispatus* consistently emerged as the *Lactobacillus* species most strongly associated with reduced presence of pathogenic bacterial and viral markers. It reduces the likelihood of bacterial pathogens such as *G. vaginalis* and *F. vaginae,* as well as hrHPV subtypes such as HPV 16 and HPV 18. The odds ratios (OR < 1), significant *p*-values, and lower presence counts all support this significant association.

L. iners was specifically associated with higher frequencies of bacterial pathogens such as *F. vaginae, G. vaginalis,* BVAB-2, and *Megasphaera* sp. Type 1, which is often associated with bacterial vaginosis and dysbiosis. The odds ratios >1 and higher presence counts suggest that *L. iners* does not contribute to a healthy microbiome. *L. gasseri* and *L. jensenii* showed protective effects, particularly against some HPV subtypes such as HPV 45 and HPV 35, but their impact was less pronounced compared *to L. crispatus*. Indeed, these findings are not new as other studies have already shown these protective effects.[10,17,29] However, this study uses a larger number of cervicovaginal samples from different racial backgrounds and States in the United States to establish these earlier findings.

The higher relative abundance of *L. iners* in the vaginal microbiome is already established, as well as the association of *F. vaginae, G. vaginalis,* BVAB-2 and *Megasphaera* sp. Type 1 and 2 with BV.[10,27,28,30] Furthermore, these pathogens also have higher concentrations and prevalence in ECAs and could be risk factors for getting cervical cancer (as shown by the four machine learning models) as they indicate vaginal dysbiosis.[30,31]

Additionally, while the effect of hrHPV on ECAs (also found in this study) is known, this study further shows a higher prevalence of HPV-subtypes in BV-positive specimens, with most hrHPV types having higher concentrations in BV-positive specimens than BV-negative ones. Indeed, the higher prevalence and concentrations of *Lactobacillus* spp. in NILM-positive samples support their association with lower risk of ECAs[29,32–34]

Specifically, ensemble machine-learning models such as Random Forest and XGBoost demonstrated high predictive accuracy for BV status (ROC-AUC ≈0.97), but moderate accuracy for cervical cytological abnormalities (ROC-AUC ≈0.74). These differential performances highlight the multifactorial complexity underlying cervical cytological abnormalities compared to the more microbiome-specific outcome of BV.

Analysis of interaction effects revealed significant increases in the risk of BV and cervical cytological abnormalities among patients co-infected with multiple hrHPV genotypes and/or pathogenic bacterial species (Data S1; Figures S20–S25).[33]

The observed heightened risk of BV and ECA outcomes in cases of multiple hrHPV and BV-pathogen co-occurrence aligns with existing literature suggesting that concurrent infections may exacerbate immune dysregulation and tissue damage[35.] Co-occurrence may impair the local immune environment by reducing innate immunity, disrupting cytokine signaling, and inducing epithelial microtrauma.[35,36] This could create a microenvironment conducive to HPV persistence and progression to epithelial cell abnormalities. The synergistic effects between hrHPV and BV-associated pathogens, such as *G. vaginalis* and *F. vaginae*, further highlight the role of microbial community

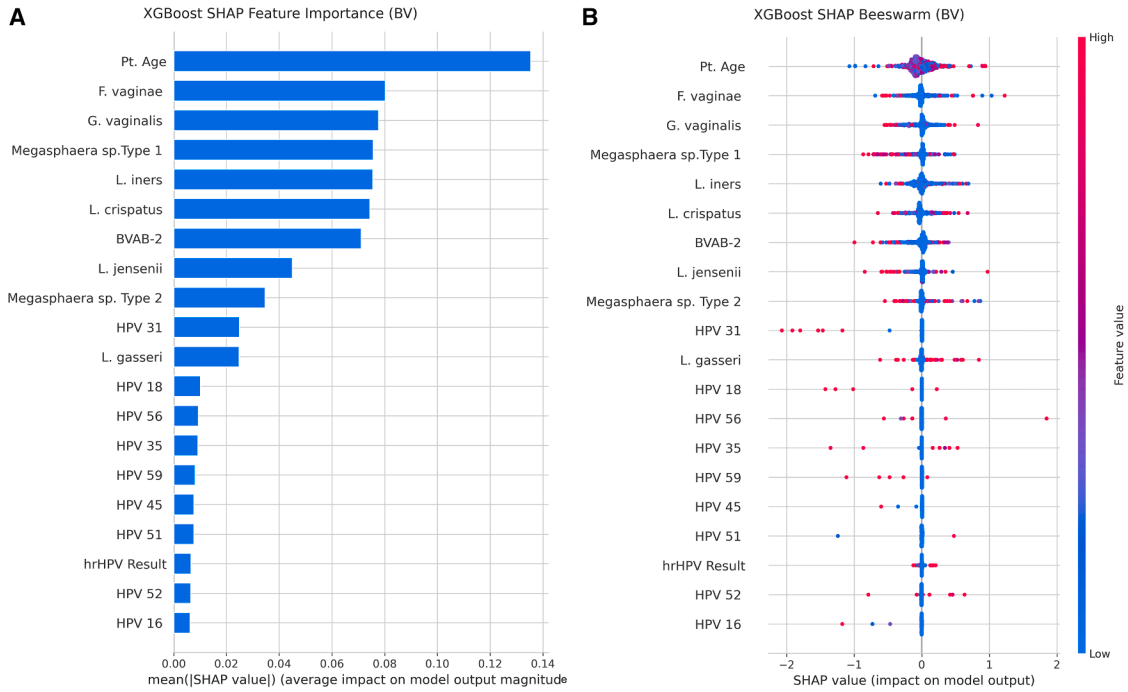


Figure 8. XGBoost SHAP interpretability for BV Status

(A) Mean(|SHAP|) Bar Chart. Age has the largest average impact (0.135), followed by *F. vaginae* (0.080), *L. crispatus* (0.078), *L. iners* (0.077), *G. vaginalis* (0.075). (B) SHAP Beeswarm Plot. High patient age and high F. vaginae values push strongly toward BV-Positive (red points to the right), whereas high *L. crispatus* values push toward BV-Negative (blue points to the left).

disruptions in cervical pathology.[37,38] Addressing these co-occurrences through integrated microbiome-focused therapies and HPV-targeted interventions could potentially mitigate risks associated with these interactions.

The associations between *Lactobacillus* spp. and healthier microbiome states provide support for the ongoing exploration of probiotics or prebiotics as adjunctive therapies to standard clinical care for BV and HPV-associated cervical abnormalities.[31,32,34,39]

Clinical trials have demonstrated that *L. crispatus* significantly reduced BV and vulvovaginal candidiasis compared to placebo.[29] The findings regarding *L. iners* are consistent with previous studies indicating its ambivalent role in the vaginal microbiome. Unlike L. *crispatus*, *L. iners* has been characterized as a transitional species, often found in microbial communities undergoing dysbiosis.[40,41] Its metabolic activities, including the production of lactic acid in lower quantities than other *Lactobacillus* species, may not sufficiently inhibit pathogenic growth.[41–43] Furthermore, *L. iners* can produce pro-inflammatory lipoteichoic acid, potentially exacerbating inflammatory conditions and facilitating pathogenic colonization.[44,45] These properties might explain its association with higher frequencies of BV pathogens and hrHPV subtypes. Future research should explore its dualistic nature in greater detail, as it may serve as a biomarker for an unstable vaginal microbiome.

Age and state were also significantly associated with the prevalence of BV and hrHPV subtypes: AGC, HSIL, and RCC were found in populations older than the average.[10,46–48] The higher presence of BV, hrHPV, and cervical cytology outcomes in some of the states should be investigated further. Most of the BV- and hrHPV-positive cases falling within ages 20–50 agree with other studies, as this age group is the most sexually active.[49–52] Although male samples (n = 13) showed the presence of BV-associated bacteria and certain hrHPV genotypes (HPV-31, HPV-39), these were excluded from inferential analyses due to limited sample size and scope. Nevertheless, this finding underscores potential sexual transmissibility, warranting dedicated studies with larger male cohorts ([12,53,54]. Given our limited male cohort (n = 13), results were descriptive only and should be confirmed through dedicated studies with larger male populations. hrHPV in males has been reported to be increasing globally, with a prevalence of 21%, which is more than the prevalence found in this study.[17] An expanded study using more clinical samples from males will provide better prevalence data on hrHPV among males in the USA.

The overall prevalence of ECAs was less than 15%, while BV was diagnosed in 53.07% of cases. This elevated BV rate may reflect sampling bias toward patients suspected of BV. Notably, despite HPV-16 and HPV-18 being traditionally recognized as the most oncogenic hrHPV types,[17,55,56] HPV-52, HPV-51, HPV-35, and HPV-56 were more prevalent in this cohort, with HPV-16 ranking fifth in prevalence.

The finding that HPV-16 and HPV-18 were not the most prevalent subtypes in this cohort may reflect demographic, geographic, or behavioral factors specific to the study population. Studies suggest that HPV-52, HPV-51, and other non-16/18 hrHPV types are increasingly common in specific regions, such as East Asia and parts of North America, possibly due to variations in sexual behaviors or reduced cross-protection

from current vaccines.[57,58] The lower prevalence of HPV-16 and HPV-18 could also be influenced by HPV vaccination programs that specifically target these high-risk types. However, the clinical significance of other hrHPV types, such as HPV-52 and HPV-51, underscores the importance of including a broader range of subtypes in screening and vaccination strategies.[59]

Lebeau et al. (2022) recently showed that HPV infections alter the cervicovaginal microbiome through the down-regulation of the host's innate peptides used by *Lactobacilli* to biosynthesize amino acids.[60] Thus, BV not only predisposes women to hrHPV infections, but the reverse can also hold true. In either case, the higher association of HPV and BV pathogens with BV and ECAs show their inter-relatedness in BV and ECA pathogenesis.[15,31,56] Notwithstanding, not all hrHPV-positive specimens were classified as ECA, showing that not all hrHPV infections affect the cervical cytology negatively; particularly when the immune system can resolve many HPV-infections efficiently.[15,16]

The protective role of *L. crispatus*, *L. gasseri*, and *L. jensenii* suggests that targeted probiotic therapies could restore and maintain a healthy cervicovaginal microbiome. Probiotic formulations containing *Lactobacilli* might help reduce the risk of BV and high-risk HPV infections, particularly in high-risk populations such as sexually active women or those undergoing treatments that disrupt the microbiome (e.g., antibiotics). Additionally, the association of *L. iners* with dysbiosis and hrHPV subtypes highlights its potential as a biomarker for identifying individuals at higher risk for cervical abnormalities. Because the ecological role of *L. iners* is strain-dependent (Holm et al., 2023; Chandra Nori et al., 2025),[61,62] we caution that its utility as a standalone screening biomarker remains preliminary.

Future directions. Prospective longitudinal cohorts are needed to test whether restoring *L. crispatus* dominated communities via targeted probiotics, prebiotics, or live biotherapeutics can reduce hrHPV persistence and progression to CIN. Randomized trials should incorporate multiomics profiling to elucidate functional pathways (e.g., lactic acid isomer ratios, tryptophan metabolism, mucosal immunity) that mediate microbiota–HPV interactions. Finally, integrating microbiome metrics into risk-based screening algorithms may refine colposcopy referrals and reduce overtreatment.

Conclusions

Within the constraints of this observational study, our findings support existing evidence that *L. crispatus* -dominant microbiota are strongly associated with cervicovaginal health, whereas microbiomes enriched with *L. iners* or BV-associated anaerobes correlate with increased dysbiosis and HPV-associated cervical abnormalities. Future longitudinal studies and interventional trials are needed to clarify causal mechanisms and therapeutic potential.

Limitations of the study

This retrospective analysis used anonymised laboratory submissions, limiting access to behavioral, contraceptive, and racial covariates that may confound microbiome–HPV relationships. The observations are cross-sectional; causal directionality cannot be inferred. Although the qPCR panel accurately quantifies key BV- and Lactobacillus-associated taxa, it does not capture the full breadth of the vaginal metagenome, and results were not independently validated by 16S or shotgun metagenomics. Additionally, because BV diagnosis in our study relies on qPCR abundances, some degree of circularity is inherent in associations between microbial profiles and BV outcomes; however, statistical and machine-learning analyses partially mitigate this concern. Finally, the cohort derives from a single commercial laboratory, and geographic clustering of providers could bias prevalence estimates. Additionally, information on patient race and ethnicity was not available, limiting analysis of possible disparities or confounding.

RESOURCE AVAILABILITY

Lead contact

Information and requests for resources, data, or reagents should be directed to the Lead Contact, *John Osei Sekyere* (joseisekyere@mdlab.com).

Materials availability

This study did not generate new unique reagents. All clinical specimens were obtained and analyzed as part of routine diagnostic testing; no materials are available for distribution beyond the described data.

Data and code availability

- The full de-identified dataset supporting this study, viz., raw qPCR CT values, calculated DNA concentrations, BV index calls, HPV genotypes, cytology results, and demographics, has been deposited in Mendeley Data (OSEI SEKYERE, 2025) under the https://doi.org/10.17632/pn7h6yjt6k.1. All Supplemental Tables (Datasets S1, S2, S3, and S4) are also included in that archive.
- Reproducible Python notebooks and utility scripts that generate every main-text and supplemental figure (including ROC curves, SHAP bar/beeswarm plots, and FDR-corrected heat-maps) are publicly available at https://github.com/joseiky/Data-analytics/tree/main/HPV-BV-Interactions_Project.
- Any other information required to re-analyse the data reported in this article is available from the lead contact upon reasonable request.

ACKNOWLEDGMENTS

The authors are grateful to the technicians at Medical Diagnostic Laboratories for their direct and indirect assistance. The material and financial resources provided by Medical Diagnostic Laboratories toward this project are warmly acknowledged and deeply appreciated. We are also grateful to Annette Daughtry for assisting with the review of the initial draft. This study was funded by Medical Diagnostic Laboratories (within the Genesis Global Group), Hamilton Township, New Jersey.

AUTHOR CONTRIBUTIONS

J.O.S.: conceptualization, methodology, formal analysis, visualization, writing – original draft, writing – review and editing, and data curation. J.T.: conceptualization, methodology, supervision, and writing – review and editing. M.A.: supervision, conceptualization, methodology, and writing – review and editing. C.T.: investigation, data curation, and validation. D.D.: investigation and data curation. R.S.: investigation, data curation. J.J.Y.: clinical data interpretation, investigation, data curation, and writing – review and editing. E.M.: supervision, resources, funding acquisition, and writing – review and editing. All authors approved the final article.

DECLARATION OF INTERESTS

The authors are employees of Medical Diagnostic Laboratories, LLC (MDL), a division of Genesis Global Group, which funded this study. The funder had

no role in study design, data collection and analysis, decision to publish, or preparation of the article. The authors declare no additional competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - Human subjects and specimens
  - Ethics statement — human subjects and specimens
  - Clinic and sample collection details
- METHOD DETAILS
  - Sample processing and data retrieval
  - qPCR analysis and bacterial vaginosis classification
  - Quality control
  - Bacterial vaginosis (BV) diagnosis
  - Pap smear cytology
  - Exploratory, association and interaction analyses
  - Prevalence comparisons
  - Correlation analyses
  - Group mean comparisons
  - Pairwise multi-factor associations
  - Odds ratios
  - Interaction analysis
  - Principal component analysis (PCA) and clustering
  - Machine learning models
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Statistical significance criteria

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2025.113473.

Received: March 19, 2025
Revised: June 26, 2025
Accepted: August 28, 2025
Published: September 2, 2025

REFERENCES

1. Ahrodia, T., Das, S., Bakshi, S., and Das, B. (2022). Structure, functions, and diversity of the healthy human microbiome. Prog. Mol. Biol. Transl. Sci. *191*, 53–82.
2. Mirmonsef, P., Hotton, A.L., Gilbert, D., Gioia, C.J., Maric, D., Hope, T.J., Landay, A.L., and Spear, G.T. (2016). Glycogen Levels in Undiluted Genital Fluid and Their Relationship to Vaginal pH, Estrogen, and Progesterone. PLoS One *11*, e0153553.
3. Stewart-Tull, D.E.S. (1964). Evidence that vaginal lactobacilli do not ferment glycogen. Am. J. Obstet. Gynecol. *88*, 676–679.
4. Jenkins, D.J., Woolston, B.M., Hood-Pishchany, M.I., Pelayo, P., Konopaski, A.N., Quinn Peters, M., France, M.T., Ravel, J., Mitchell, C.M., Rakoff-Nahoum, S., et al. (2023). Bacterial amylases enable glycogen degradation by the vaginal microbiome. Nat. Microbiol. *8*, 1641–1652. https://www.nature.com/articles/s41564-023-01447-2.
5. Navarro, S., Abla, H., Delgado, B., Colmer-Hamood, J.A., Ventolini, G., and Hamood, A.N. (2023). Glycogen availability and pH variation in a medium simulating vaginal fluid influence the growth of vaginal Lactobacillus species and Gardnerella vaginalis. BMC Microbiol. *23*, 186. https://bmcmicrobiol.biomedcentral.com/articles/10.1186/s12866-023-02916-8.
6. Lebeau, A., Bruyere, D., Roncarati, P., Peixoto, P., Hervouet, E., Cobraiville, G., Taminiau, B., Masson, M., Gallego, C., Mazzucchelli, G., et al. (2022). HPV infection alters vaginal microbiome through down-regulating host mucosal innate peptides used by Lactobacilli as amino acid sources. Nat. Commun. *13*, 1076. https://www.nature.com/articles/s41467-022-28724-8.
7. Navarro, S., Abla, H., Delgado, B., Colmer-Hamood, J.A., Ventolini, G., and Hamood, A.N. (2023). Glycogen availability and pH variation in a medium simulating vaginal fluid influence the growth of vaginal Lactobacillus species and Gardnerella vaginalis. BMC Microbiol. *23*, 186.
8. Ravel, J., Gajer, P., Abdo, Z., Schneider, G.M., Koenig, S.S.K., McCulle, S.L., Karlebach, S., Gorle, R., Russell, J., Tacket, C.O., et al. (2011). Vaginal microbiome of reproductive-age women. Proc. Natl. Acad. Sci. USA *108*, 4680–4687. https://www.pnas.org/doi/abs/10.1073/pnas.1002611107.
9. Vodstrcil, L.A., Muzny, C.A., Plummer, E.L., Sobel, J.D., and Bradshaw, C.S. (2021). Bacterial vaginosis: drivers of recurrence and challenges and opportunities in partner treatment. BMC Med. *19*, 194. https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-021-02077-3.
10. Oyenihi, A.B., Haines, R., Trama, J., Faro, S., Mordechai, E., Adelson, M.E., and Osei Sekyere, J. (2024). Molecular characterization of vaginal microbiota using a new 22-species qRT-PCR test to achieve a relative-abundance and species-based diagnosis of bacterial vaginosis. Front. Cell. Infect. Microbiol. *14*, 1409774.
11. Tjagur, S., Mändar, R., and Punab, M. (2020). Profile of sexually transmitted infections causing urethritis and a related inflammatory reaction in urine among heterosexual males: A flow-cytometry study. PLoS One *15*, e0242227.
12. Bruni, L., Albero, G., Rowley, J., Alemany, L., Arbyn, M., Giuliano, A.R., Markowitz, L.E., Broutet, N., and Taylor, M. (2023). Global and regional estimates of genital human papillomavirus prevalence among men: a systematic review and meta-analysis. Lancet. Glob. Health *11*, e1345–e1362. https://pmc.ncbi.nlm.nih.gov/articles/PMC10447222/.
13. Han, Y., Liu, Z., and Chen, T. (2021). Role of Vaginal Microbiota Dysbiosis in Gynecological Diseases and the Potential Interventions. Front. Microbiol. *12*, 643422.
14. Martins, B.C.T., Guimarães, R.A., Alves, R.R.F., and Saddi, V.A. (2022). Bacterial vaginosis and cervical human papillomavirus infection in young and adult women: a systematic review and meta-analysis. Rev. Saude Publica *56*, 113.
15. Dunne, E.F., Unger, E.R., Sternberg, M., McQuillan, G., Swan, D.C., Patel, S.S., and Markowitz, L.E. (2007). Prevalence of HPV Infection Among Females in the United States. JAMA *297*, 813–819.
16. Lewis, R.M., Laprise, J.F., Gargano, J.W., Unger, E.R., Querec, T.D., Chesson, H.W., Brisson, M., and Markowitz, L.E. (2021). Estimated prevalence and incidence of disease-associated HPV types among 15–59-year-olds in the United States. Sex. Transm. Dis. *48*, 273–277.
17. Bruni, L., Albero, G., Rowley, J., Alemany, L., Arbyn, M., Giuliano, A.R., Markowitz, L.E., Broutet, N., and Taylor, M. (2023). Global and regional estimates of genital human papillomavirus prevalence among men: a systematic review and meta-analysis. Lancet. Glob. Health *11*, e1345–e1362.
18. McBride, A.A. (2022). Human papillomaviruses: diversity, infection and host interactions. Nat. Rev. Microbiol. *20*, 95–108. https://pubmed.ncbi.nlm.nih.gov/34522050/.
19. Wei, F., Georges, D., Man, I., Baussano, I., and Clifford, G.M. (2024). Causal attribution of human papillomavirus genotypes to invasive cervical cancer worldwide: a systematic analysis of the global literature. Lancet *404*, 435–444. https://pubmed.ncbi.nlm.nih.gov/39097395/.
20. Human Papillomavirus (HPV) | HPV | CDC [Internet]. [cited 2025 Jun 24]. Available from: https://www.cdc.gov/hpv/index.html.
21. Centers for Disease Control and Prevention (CDC) (2021). In Human Papillomavirus, 14th Editi, E. Meites, J. Gee, E.R. Unger, and L.E. Markowitz, eds. (Centers for Disease Control and Prevention (CDC)), p. 14.
22. Lofgren, S.M., Tadros, T., Herring-Bailey, G., Birdsong, G., Mosunjac, M., Flowers, L., and Nguyen, M.L. (2015). Progression and Regression of Cervical Pap Test Lesions in an Urban AIDS Clinic in the Combined

Antiretroviral Therapy Era: A Longitudinal, Retrospective Study. AIDS Res. Hum. Retroviruses *31*, 508–513.

23. Zhang, Y., Chen, S., Chen, X., Zhang, H., Huang, X., Xue, X., Guo, Y., Ruan, X., Liu, X., Deng, G., et al. (2021). Association Between Vaginal Gardnerella and Tubal Pregnancy in Women With Symptomatic Early Pregnancies in China: A Nested Case-Control Study. Front. Cell. Infect. Microbiol. *11*, 761153. https://pmc.ncbi.nlm.nih.gov/articles/PMC8801712/.
24. Lofgren, S.M., Tadros, T., Herring-Bailey, G., Birdsong, G., Mosunjac, M., Flowers, L., and Nguyen, M.L. (2015). Progression and Regression of Cervical Pap Test Lesions in an Urban AIDS Clinic in the Combined Antiretroviral Therapy Era: A Longitudinal, Retrospective Study. AIDS Res. Hum. Retroviruses *31*, 508–513.
25. Li, Y., and Wu, X. (2024). Vaginal microbiome distinction in women with HPV+, cervical intraepithelial neoplasia, and cervical cancer, a retrospective study. Front. Cell. Infect. Microbiol. *14*, 1483544. https://doi.org/10.3389/fcimb.2024.1483544.
26. Kazlauskaitė, J., Žukienė, G., Rudaitis, V., and Bartkevičienė, D. (2025). The Vaginal Microbiota, Human Papillomavirus, and Cervical Dysplasia—A Review. Medicina *61*, 847. https://www.mdpi.com/1648-9144/61/5/847/htm.
27. Molina, M.A., Steenbergen, R.D.M., Pumpe, A., Kenyon, A.N., and Melchers, W.J.G. (2024). HPV integration and cervical cancer: a failed evolutionary viral trait. Trends Mol. Med. *30*, 890–902. https://pubmed.ncbi.nlm.nih.gov/38853085/.
28. Pimenoff, V.N., Gray, P., Louvanto, K., Eriksson, T., Lagheden, C., Söderlund-Strand, A., Dillner, J., and Lehtinen, M. (2023). Ecological diversity profiles of non-vaccine-targeted HPVs after gender-based community vaccination efforts. Cell Host Microbe *31*, 1921–1929.e3. https://pubmed.ncbi.nlm.nih.gov/37944494/.
29. Mändar, R., Sõerunurk, G., Štšepetova, J., Smidt, I., Rööp, T., Kõljalg, S., Saare, M., Ausmees, K., Le, D.D., Jaagura, M., and Piiskop, S. (2023). Impact of Lactobacillus crispatus-containing oral and vaginal probiotics on vaginal health: a randomised double-blind placebo controlled clinical trial. Benef. Microbes *14*, 143–152.
30. Lebeau, A., Bruyere, D., Roncarati, P., Peixoto, P., Hervouet, E., Cobraiville, G., Taminiau, B., Masson, M., Gallego, C., Mazzucchelli, G., et al. (2022). HPV infection alters vaginal microbiome through down-regulating host mucosal innate peptides used by Lactobacilli as amino acid sources. Nat. Commun. *13*, 1076.
31. Sharifian, K., Shoja, Z., and Jalilvand, S. (2023). The interplay between human papillomavirus and vaginal microbiota in cervical cancer development. Virol. J. *20*, 73.
32. Cascardi, E., Cazzato, G., Daniele, A., Silvestris, E., Cormio, G., Di Vagno, G., Malvasi, A., Loizzi, V., Scacco, S., Pinto, V., et al. (2022). Association between Cervical Microbiota and HPV: Could This Be the Key to Complete Cervical Cancer Eradication? Biology *11*, 1114.
33. Zeng, M., Li, X., Jiao, X., Cai, X., Yao, F., Xu, S., Huang, X., Zhang, Q., and Chen, J. (2022). Roles of vaginal flora in human papillomavirus infection, virus persistence and clearance. Front. Cell. Infect. Microbiol. *12*, 1036869.
34. Chee, W.J.Y., Chew, S.Y., and Than, L.T.L. (2020). Vaginal microbiota and the potential of Lactobacillus derivatives in maintaining vaginal health. Microb. Cell Fact. *19*, 203.
35. Sharifian, K., Shoja, Z., and Jalilvand, S. (2023). The interplay between human papillomavirus and vaginal microbiota in cervical cancer development. Virol. J. *20*, 73.
36. Gillet, E., Meys, J.F., Verstraelen, H., Bosire, C., De Sutter, P., Temmerman, M., and Broeck, D.V. (2011). Bacterial vaginosis is associated with uterine cervical human papillomavirus infection: A meta-analysis. BMC Infect. Dis. *11*, 10.
37. Chen, Y., Hong, Z., Wang, W., Gu, L., Gao, H., Qiu, L., and Di, W. (2019). Association between the vaginal microbiome and high-risk human papillomavirus infection in pregnant Chinese women. BMC Infect. Dis. *19*, 677.
38. Oyenihi, A.B., Haines, R., Trama, J., Faro, S., Mordechai, E., Adelson, M.E., and Osei Sekyere, J. (2024). Molecular Characterization of Vaginal Microbiota Using a New 22-Species qRT-PCR Test to Achieve a Relative-abundance and Species-based Diagnosis of Bacterial Vaginosis. Preprint at bioRxiv *14*, 1409774.
39. Molina, M.A., Melchers, W.J.G., Núñez-Samudio, V., and Landires, I. (2024). The emerging role of Lactobacillus acidophilus in the cervicovaginal microenvironment. Lancet Microbe *5*, e6–e7.
40. Macklaim, J.M., Clemente, J.C., Knight, R., Gloor, G.B., and Reid, G. (2015). Changes in vaginal microbiota following antimicrobial and probiotic therapy. Microb. Ecol. Health Dis. *26*, 27799. https://doi.org/10.3402/mehd.v26.27799.
41. Vaneechoutte, M. (2017). Lactobacillus iners, the unusual suspect. Res. Microbiol. *168*, 826–836.
42. Edwards, V.L., Smith, S.B., McComb, E.J., Tamarelle, J., Ma, B., Humphrys, M.S., Gajer, P., Gwilliam, K., Schaefer, A.M., Lai, S.K., et al. (2019). The Cervicovaginal Microbiota-Host Interaction Modulates Chlamydia trachomatis Infection. mBio *10*, e01548-19.
43. Zheng, N., Guo, R., Wang, J., Zhou, W., and Ling, Z. (2021). Contribution of Lactobacillus iners to Vaginal Health and Diseases: A Systematic Review. Front. Cell. Infect. Microbiol. *11*, 792787.
44. Borges, S., Silva, J., and Teixeira, P. (2014). The role of lactobacilli and probiotics in maintaining vaginal health. Arch. Gynecol. Obstet. *289*, 479–489.
45. Tamarelle, J., Ma, B., Gajer, P., Humphrys, M.S., Terplan, M., Mark, K.S., Thiébaut, A.C.M., Forney, L.J., Brotman, R.M., Delarocque-Astagneau, E., et al. (2020). Nonoptimal Vaginal Microbiota After Azithromycin Treatment for Chlamydia trachomatis Infection. J. Infect. Dis. *221*, 627–635.
46. Wilson, J.D., Lee, R.A., Balen, A.H., and Rutherford, A.J. (2007). Bacterial vaginal flora in relation to changing oestrogen levels. Int. J. STD AIDS *18*, 308–311.
47. Vodstrcil, L.A., Muzny, C.A., Plummer, E.L., Sobel, J.D., and Bradshaw, C.S. (2021). Bacterial vaginosis: drivers of recurrence and challenges and opportunities in partner treatment. BMC Med. *19*, 194.
48. Shen, L., Zhang, W., Yuan, Y., Zhu, W., and Shang, A. (2022). Vaginal microecological characteristics of women in different physiological and pathological period. Front. Cell. Infect. Microbiol. *12*, 959793.
49. Bradshaw, C.S., Vodstrcil, L.A., Hocking, J.S., Law, M., Pirotta, M., Garland, S.M., De Guingand, D., Morton, A.N., and Fairley, C.K. (2013). Recurrence of Bacterial Vaginosis Is Significantly Associated With Posttreatment Sexual Activities and Hormonal Contraceptive Use. Clin. Infect. Dis. *56*, 777–786.
50. Bakus, C., Budge, K.L., Feigenblum, N., Figueroa, M., and Francis, A.P. (2023). The impact of contraceptives on the vaginal microbiome in the non-pregnant state. Front. Microbiomes *1*, 1055472.
51. Abou Chacra, L., Fenollar, F., and Diop, K. (2021). Bacterial Vaginosis: What Do We Currently Know? Front. Cell. Infect. Microbiol. *11*, 672429.
52. Abou Chacra, L., and Fenollar, F. (2021). Exploring the global vaginal microbiome and its impact on human health. Microb. Pathog. *160*, 105172.
53. Swidsinski, S., Moll, W.M., and Swidsinski, A. (2023). Bacterial Vaginosis—Vaginal Polymicrobial Biofilms and Dysbiosis. Dtsch. Arztebl. Int. *120*, 347–354.
54. Schwartz, L.M., Castle, P.E., Follansbee, S., Borgonovo, S., Fetterman, B., Tokugawa, D., Lorey, T.S., Sahasrabuddhe, V.V., Luhn, P., Gage, J.C., et al. (2013). Risk factors for anal HPV infection and anal precancer in HIV-infected men who have sex with men. J. Infect. Dis. *208*, 1768–1775.
55. Nang, D.W., Tukirinawe, H., Okello, M., Tayebwa, B., Theophilus, P., Sikakulya, F.K., Fajardo, Y., Afodun, A.M., and Kajabwangu, R. (2023). Prevalence of high-risk human papillomavirus infection and associated factors among women of reproductive age attending a rural teaching hospital in western Uganda. BMC Womens Health *23*, 209.
56. Naidoo, K., Abbai, N., Tinarwo, P., and Sebitloane, M. (2022). BV associated bacteria specifically BVAB 1 and BVAB 3 as biomarkers for HPV risk

and progression of cervical neoplasia. Infect. Dis. Obstet. Gynecol. *2022*, 9562937.

57. Bruni, L., Albero, G., Rowley, J., Alemany, L., Arbyn, M., Giuliano, A.R., Markowitz, L.E., Broutet, N., and Taylor, M. (2023). Global and regional estimates of genital human papillomavirus prevalence among men: a systematic review and meta-analysis. Lancet. Glob. Health *11*, e1345–e1362.

58. Clifford, G.M., Smith, J.S., Plummer, M., Muñoz, N., and Franceschi, S. (2003). Human papillomavirus types in invasive cervical cancer worldwide: a meta-analysis. Br. J. Cancer *88*, 63–73.

59. Zhao, F.H., Lin, M.J., Chen, F., Hu, S.Y., Zhang, R., Belinson, J.L., Sellors, J.W., Franceschi, S., Qiao, Y.L., and Castle, P.E.; Cervical Cancer Screening Group in China (2010). Performance of high-risk human papillomavirus DNA testing as a primary screen for cervical cancer: a pooled analysis of individual patient data from 17 population-based studies from China. Lancet Oncol. *11*, 1160–1171.

60. Lebeau, A., Bruyere, D., Roncarati, P., Peixoto, P., Hervouet, E., Cobraiville, G., Taminiau, B., Masson, M., Gallego, C., Mazzucchelli, G., et al. (2022). HPV infection alters vaginal microbiome through down-regulating host mucosal innate peptides used by Lactobacilli as amino acid sources. Nat. Commun. *13*, 1076.

61. Nori, S.R.C., Walsh, C.J., McAuliffe, F.M., Moore, R.L., Van Sinderen, D., Feehily, C., and Cotter, P.D. (2025). Strain-level variation among vaginal Lactobacillus crispatus and Lactobacillus iners as identified by comparative metagenomics. NPJ Biofilms Microbiomes *11*, 49. https://www.nature.com/articles/s41522-025-00682-1.

62. Holm, J.B., France, M.T., Gajer, P., Ma, B., Brotman, R.M., Shardell, M., Forney, L., and Ravel, J. (2023). Integrating compositional and functional content to describe vaginal microbiomes in health and disease. Microbiome *11*, 259. https://doi.org/10.1186/s40168-023-01692-x.

63. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.

64. Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), pp. 785–794.

65. Lundberg, S.M., and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems, *30*, I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds. (Curran Associates, Inc.), pp. 4765–4774.

66. Lemaître, G., Nogueir, F., and Aridas, C.K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. *18*, 1–5. http://jmlr.org/papers/v18/16-365.

67. Hunter, J.D. (2007). Matplotlib: A 2D graphics environment. Comput. Sci. Eng. *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55.

68. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
De-identified HPV–BV dataset (15 607 records)	Mendeley Data	https://doi.org/10.17632/pn7h6yjt6k.1
Software and algorithms		
Full analysis & plotting code	This paper (GitHub)	https://github.com/joseiky/Data-analytics/tree/main/HPV-BV-Interactions_Project
Python 3.10 (base language)	Python Software Foundation	https://www.python.org/
scikit-learn v 1.4.2	Pedregosa et al.[63]	PyPI: scikit-learn = = 1.4.2
XGBoost v 1.7.6	Chen and Guestrin[64]	PyPI: xgboost = = 1.7.6
shap v 0.43.0	Lundberg and Lee[65]	PyPI: shap = = 0.43.0
imbalanced-learn v 0.11.0	Lemaître et al.[66]	PyPI: imbalanced-learn = = 0.11.0
Matplotlib v 3.7.3	Hunter[67]	PyPI: matplotlib = = 3.7.3
PyPI: pandas = = 1.5.3	pandas v 1.5.3	Wes McKinney 2010
PyPI: statsmodels = = 0.14.0	statsmodels v 0.14.0	Seabold & Perktold 2010
SciPy v1.9	Virtanen et al.[68]	PyPI: scipy = = 1.9.0

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human subjects and specimens

This study is a retrospective analysis of 15,607 cervicovaginal specimens obtained from routine clinical diagnostic submissions. The specimens came from 15,541 females (99.57% of samples), 13 males (0.08%), and 53 individuals of unspecified gender (0.34%), all of whom were patients of physicians across 32 U.S. states and Washington, D.C. during the period August 1, 2021 to April 5, 2023. The age of subjects ranged from 14 to 95 years, with a median in the late 20s (most samples were from individuals 20–60 years old). Because this was a de-identified dataset obtained from an internal laboratory database without patient identifiers, specific demographic or clinical information beyond age, sex, and provider location was not available. All samples were leftover de-identified clinical specimens; ethical approval: use of such data is exempt from IRB review under U.S. federal regulations (no human subjects contact, and all data were pre-existing and de-identified). (The study was performed in accordance with institutional guidelines for research on de-identified clinical datasets).

Ethics statement — human subjects and specimens

This study analysed only pre-existing, fully de-identified clinical laboratory records; no direct patient intervention or access to individually identifiable information occurred. Accordingly, the activity does not constitute ‘‘human-subjects research’’ under the U.S. Common Rule [45 CFR 46.102(e)], and is specifically exempt as ‘‘secondary research using information recorded in such a manner that subjects cannot be readily identified’’ [45 CFR 46.104(d)(4)]. The dataset also satisfies the HIPAA Privacy Rule safe-harbor de-identification standard [45 CFR 164.514(b)(2)]. Institutional policy therefore does not require Institutional Review Board (IRB) review or written informed consent.

Aggregate demographic variables—sex, age, and self-reported ethnicity—were retained solely for statistical analysis and are reported in Table S1 and Figure 1. The influence of sex and age on study outcomes is presented in the Results. All procedures complied with federal regulations and institutional guidelines.

Clinic and sample collection details

Cervical and vaginal swab specimens were collected by health care providers as part of routine screening or diagnostic work-ups. In a few cases, for male patients, urine, anal swabs, or semen samples were submitted for HPV/BV testing (the male sample subset is extremely small). After collection, swabs were placed in appropriate transport media and shipped to Medical Diagnostic Laboratories, LLC (Hamilton, NJ). Upon arrival, samples were processed for molecular testing and cytology as described below. No interventions or treatments were administered as part of this study.

METHOD DETAILS

Sample processing and data retrieval

We retrieved electronic laboratory records for all specimens received in the specified time frame that had complete results for the tests of interest. The data fields obtained for each sample included[1]: patient demographics – age, sex, and provider state[2]; qPCR CT values and calculated DNA concentrations for each of 22 microbial targets (see "qPCR Analysis" below)[3]; the diagnosed BV status for the sample (positive, negative, or transitional)[4]; the hrHPV result (positive for any high-risk HPV genotype, or negative); and[5] the Pap smear cytology result, as reported by certified cytopathologists. These data were exported from the laboratory information system into a spreadsheet, then imported into Python for analysis.

qPCR analysis and bacterial vaginosis classification

Each specimen was tested with a quantitative PCR (qPCR) panel targeting 22 vaginal microbiome-related species (Oyenihi et al., 2024).[10] This panel included nine bacterial taxa commonly associated with BV or normal flora: *Fannyhessea vaginae* (formerly *Atopobium vaginae*), *Gardnerella vaginalis*, BVAB-2, *Megasphaera* sp. *type 1*, *Megasphaera* sp. *type 2*, *Lactobacillus iners*, *Lactobacillus gasseri*, *Lactobacillus jensenii*, and *Lactobacillus crispatus.* The HPV PCR was undertaken using a proprietary in-house PCR (HPV Type-Detect® 2.0) and the Bio-Plex® instrument (Bio-Rad, USA) and reagents to determine the 13 high-risk HPV types – HPV-16, -18, −31, −33, −35, −39, −45, −51, −52, −56, −58, −59, and −68. Briefly, the L1 major capsid region of the HPV genome is first amplified by PCR and subsequently differentiated into the actual HPV types using the Bio-Plex reagents and analysis system, which uses microspheres that are internally dyed with different fluorophores, allowing for the simultaneous analysis of multiple targets. Specific oligonucleotide probes attached to the microsphere beads hybridize to the amplified HPV DNA, enabling detection and identification of the different genotypes (www.mdlab.com/laboratorians/vol1_3.pdf).

All qPCRs were performed using the Bio-Rad's CFX384 real-time PCR machine, following the manufacturer's recommended protocols for TaqMan chemistry.

DNA was extracted from swabs using standard automated methods. qPCR was performed using TaqMan assays specific to each target; CT values were converted to genomic DNA concentrations (copies per mL) via calibration curves. Raw CT values and DNA concentrations for all 22 targets are supplied in Table S1; the exact BV-index decision rules (numerator/denominator ratios and cut-points) are implemented in the shared python scripts (in Github: https://github.com/joseiky/Data-analytics/tree/main/HPV-BV-Interactions_Project) for transparency. PCR positivity for each pathogen was defined using a CT (cycle threshold) cut-off of $\leq$33. Samples yielding a CT score of 34 or higher, or a CT score of zero (indicating no amplification), were deemed negative for the respective pathogen.

Quality control

Each run included positive controls for each target and a human housekeeping gene control to ensure adequate DNA yield. All qPCR assays had analytical sensitivity >95% at ~10–100 copies/mL and were validated according to CLIA guidelines. Each pathogen-specific qPCR used plasmid standards to extrapolate pathogen DNA concentrations. The reported concentrations reflect true pathogen abundance rather than differences in PCR efficiency.

Bacterial vaginosis (BV) diagnosis

We applied the proprietary MDL-BV diagnostic index (developed in Oyenihi et al., 2024) to classify samples as BV-positive, BV-negative, or "intermediate microbiota state." This algorithm considers the relative abundances of the BV-associated bacteria vs. Lactobacilli. In brief, a sample was called BV-positive if it showed high concentrations of BV-associated anaerobes (e.g., *Gardnerella*, *Fannyhessea*, BVAB) and low Lactobacillus, BV-negative if dominated by Lactobacillus (especially *L. crispatus*/*L. jensenii*) with low pathogen load, and intermediate microbiota state for intermediate cases. The "intermediate microbiota state" category corresponds to an intermediate microbiota state (analogous to Nugent score 4–6) and is clinically relevant as shown in recent studies (Oyenihi et al., 2024). We retained this category in analyses to capture microbiome states that are neither clearly healthy nor BV, as recommended in the literature.

Pap smear cytology

All samples were evaluated via routine Pap cytology. Cytotechnologists and pathologists classified each specimen's cervical cells using the Bethesda System for Reporting Cervical Cytology, which includes categories such as NILM, ASC-US, ASC-H, LSIL, HSIL, and AGC. No cases of squamous cell carcinoma (SCC) were identified in this cohort. Negative for intraepithelial lesion or malignancy (NILM) indicates normal cytology; epithelial cell abnormalities (ECA) were categorized as: ASC-US (atypical squamous cells of undetermined significance), ASC-H (atypical squamous cells, cannot exclude HSIL), LSIL (low-grade squamous intraepithelial lesion), HSIL (high-grade squamous intraepithelial lesion), SCC (squamous cell carcinoma), or AGC (atypical glandular cells). In our dataset, a small subset of NILM cases were noted to have Reactive Cellular Changes (RCC) due to inflammation; we tracked those as a subcategory of NILM (benign reactive changes). For analysis, we often grouped all ASC-US+ lesions as

"ECA-positive" vs. NILM. Any sample with an ECA (including ASC-US or worse) was considered "abnormal cytology." In the final dataset, 75.3% were NILM, with the remainder showing ECAs (14.3% ASC-US, 7.1% LSIL, 0.5% HSIL, 0.4% ASC-H, 0.16% AGC, and 0% SCC – no frank carcinoma was diagnosed).

Data Cleaning and Preprocessing: The compiled dataset underwent several cleaning steps.

- Samples missing key qPCR results (e.g., a sample where all 22 targets failed to amplify, making BV status indeterminable) were excluded (this accounted for the removal of 3,498 records out of 19,105 initial, yielding 15,607.
- Categorical fields like gender, BV status, hrHPV result, cytology outcome were encoded numerically for analysis (e.g., Female = 0/Male = 1, BV-negative = 0/positive = 1/transitional = 2, etc.). We ensured that Intermediate microbiota state remained a distinct category in analysis rather than lumping with negative.
  - Sex: specimens from males or unknown gender (0.42% of the dataset) were retained in the public archive, demographics (Figure 1) and supplementary figures but excluded from all downstream statistics and predictive modeling to avoid sex-based bias.
- Duplicate entries (more than one record for the same patient ID and date) were removed to avoid over-representation. (Each sample was considered an independent data point; patients with multiple visits were beyond the scope of this study's analysis.)
- Missing demographic data: A few samples lacked an age or provider state. Rather than impute, we opted to exclude those samples from analyses involving age or location to maintain data integrity (they were still used in analyses not requiring those fields). In practice, the vast majority of samples had complete info.
- All numerical features (DNA concentrations) were reviewed for outliers; extremely high concentrations were double-checked against raw CT data for validity. We log-transformed DNA concentrations for certain analyses (e.g., PCA) to normalize distributions.

Exploratory, association and interaction analyses

All prevalence contrasts (categorical vs. categorical) used two-sided χ^2 or Fisher's exact tests as appropriate, followed by Benjamini–Hochberg FDR correction; adjusted *p* values are reported as q-values throughout. Continuous-variable comparisons (e.g., mean HPV load across BV strata) employed Welch's t-test or one-way ANOVA with Tukey HSD. Large-scale pairwise screening (Figure S18 heatmap) iterated over every variable pair and plotted –log10(q) to visualise significance density.

To dissect two-way microbial synergies we fitted logistic models with interaction terms ($\beta 12$) and displayed the coefficients and q-values as lollipop/bubble plots (Figures S20–S25). A positive $\beta 12$ indicates supra-additive risk; negative suggests antagonism.

All statistical routines were scripted in Python 3.10 using *Pandas* 1.5, *Numpy* 1.24, *Scipy.stats* 1.9 and *statsmodels* 0.14.

Prevalence comparisons

We used Chi-square tests (or Fisher's exact test when expected counts were low) to assess associations between categorical variables. For example, we tested whether the presence of a given Lactobacillus species was associated with BV status, hrHPV positivity, or cytology outcome (and similarly for each pathogen and each HPV type). These 2×2 or $2 \times k$ contingency analyses produced *p*-values which were then corrected for multiple comparisons (see Quantification & Statistical Analysis below).

Correlation analyses

To quantify linear relationships among continuous and binary variables (e.g., age, microbial concentrations, presence/absence flags), we computed Spearman rank correlation coefficients. A comprehensive correlation matrix was generated (shown in Figure S17) for all pairwise combinations of key variables. This helped identify clusters of positively or negatively correlated factors (for instance, strong positive correlations among BV-associated bacteria, and strong negative correlations between Lactobacilli and BV status).

Group mean comparisons

For continuous outcomes like mean HPV concentration in BV-positive vs. BV-negative groups, we performed t-tests or ANOVA (analysis of variance) as appropriate. For example, we tested if the mean viral load of each HPV subtype differed significantly between BV+ and BV– women (No significant differences were found for HPV load by BV status).

Pairwise multi-factor associations

We also conducted a global Chi-square/ANOVA analysis across all factors (Figure S18) to find any significant linkage (e.g., age vs. state, state vs. BV, etc.). This essentially screened all variable pairs with the appropriate test (Chi-square for categorical–categorical, ANOVA for continuous–categorical). The results were visualized as a heatmap (with significance indicated), which guided which associations warranted closer look.

Odds ratios

To complement prevalence comparisons, we calculated odds ratios (OR) for the occurrence of outcomes given the presence of certain microbes (especially Lactobacillus spp.). For each Lactobacillus, we performed univariate logistic regressions predicting: BV status

(positive/negative), hrHPV status, and each cytology outcome, as well as presence of each pathogen. The ORs and 95% confidence intervals from these models are presented (Figure 5B). OR < 1 indicates that Lactobacillus is associated with lower odds of that outcome, whereas OR > 1 indicates higher odds (risk factor).

Interaction analysis

We probed potential interaction effects between multiple microbes on outcomes. Using logistic regression models, we included interaction terms (e.g., *HPV16* × *G. vaginalis*) to see if co-occurrence had a synergistic impact on BV or cytology. Significant interactions were identified and plotted (Figures S20–S25), where an interaction coefficient >0 suggests an outcome odds higher than expected from individual effects (and <0 suggests a less-than-additive effect). For instance, we tested interactions like "hrHPV presence AND *G. vaginalis*" predicting abnormal cytology. Notable findings included synergistic interactions between certain HPV types and BV bacteria raising the likelihood of HSIL (e.g., HPV-16 with *G. vaginalis*).

Principal component analysis (PCA) and clustering

We applied PCA to the scaled quantitative dataset to reduce dimensionality and reveal dominant variance patterns. This included all microbial abundance measures and key clinical outcomes coded as numeric (e.g., BV status 0/1, HPV status 0/1, cytology categories mapped to an ordinal scale). We extracted principal components and examined the loading scores to identify which variables contributed most to variability. For example, we found that PC1 was driven by high levels of BV-related bacteria (e.g., *F. vaginae*, *G. vaginalis*, BVAB-2) vs. Lactobacilli, whereas L. iners and certain HPV types loaded heavily on PC2. We then performed K-means clustering (with k = 3 clusters chosen based on explained variance) on the PCA-reduced data to see if samples group into distinct microbiome profiles. Indeed, Cluster 1 corresponded to a BV-positive/HPV-rich microbiome (dysbiotic, often with ASCUS/HSIL outcomes), Cluster 2 to a Lactobacillus-dominant, healthier microbiome (mostly BV-negative, NILM cytology), and Cluster 3 to intermediate cases. These are detailed in Figures S27–S28. Such unsupervised analyses supported our main findings by showing, for instance, that *L. crispatus* and *L. jensenii* cluster with healthy outcomes, whereas *L. iners* clusters with pathogens and disease.

Machine learning models

To predict clinical outcomes from the microbiome and demographic data, we built and evaluated four supervised machine learning models: XGBoost (extreme gradient boosting), Random Forest, Logistic Regression, and Decision Tree. We built separate models for two endpoints: BV status (positive vs. negative) and cervical cytology outcome (we modeled this as binary: NILM vs. ECA, i.e., whether the sample had any abnormal cytology).

- Feature set: Model inputs included patient age, sex, and state, plus the presence/absence and/or abundance of each microbiome target (the Lactobacillus and pathogen qPCR results). In practice, we used either binary presence or the log-transformed concentration for each microbe as features. Categorical variables like state were one-hot encoded. We also included an indicator for hrHPV positivity as a feature in the cytology outcome model (since hrHPV is known to be predictive of abnormal cytology).
- Training and validation: We split the data into training and test sets (e.g., 70/30 split stratified by outcome) or used 5-fold cross-validation for robust performance estimates. Hyperparameter tuning was performed via grid search with cross-validation on the training set. For example, for XGBoost we tuned parameters such as the learning rate, max tree depth, and number of estimators; for Random Forest, we tuned number of trees and max features; etc. We selected the model settings that optimized the F1-score (harmonic mean of precision and recall) on the validation folds.
- Performance metrics: We evaluated models by precision, recall, F1-score, and accuracy on the held-out test set. As noted in the Results, the models achieved higher performance for BV (e.g., ∼90% precision/recall) than for cytology outcomes, which is understandable given the clearer microbiome signal for BV. We also plotted ROC curves and calculated the Area Under the ROC Curve (AUC) for each model (Figure 7). For BV prediction, all models had high AUCs (≈0.90–0.95), whereas for cytology prediction AUCs were more modest (≈0.70–0.80), indicating the challenge of predicting cytological abnormalities from microbiome data alone.
- Feature importance: For each model, we extracted measures of feature importance. In tree-based models (XGBoost, Random Forest, Decision Tree), importance was measured by Gini importance (total reduction in impurity from splits on that feature). In the logistic model, we considered the absolute magnitude of standardized coefficients as a proxy for importance. We found that *L. crispatus* presence, patient age, and hrHPV status were consistently among the top predictors for both BV and cytology outcomes across models. *L. jensenii*, *L. gasseri*, and certain BV bacteria (e.g., *G. vaginalis*, *F. vaginae*) were also important for BV prediction, whereas cytology prediction relied more on age, hrHPV, *L. iners* presence, and a combination of multiple minor features. (See Supplemental Table S4.5 for the full ranked feature lists per model.)
- Model explainability (SHAP values): To obtain model-agnostic explanations we computed Shapley Additive exPlanations (SHAP) with the 'shap' Python package (v0.43). For the best-performing XGBoost models (BV and Cytology tasks), we generated mean(|SHAP|) bar plots (Figure 8A) and beeswarm distributions (Figure 8B) to visualise both global and per-sample feature impact. The SHAP pipeline is provided in "shap_analysis.ipynb" and runs in ∼3 min on a single CPU core.

- Consensus and variability: We compared the feature rankings across models and found a strong consistency between the two ensemble methods (XGBoost & Random Forest): they agreed on the top ~5 features. The logistic regression and single decision tree showed some differences (likely due to each model's inherent biases and possible overfitting in the case of the single tree). This comparison underscores which predictors are robust. For instance, *L. crispatus* and age were top features in nearly all models (robust signals), whereas the importance of, say, *HPV-51* or *Megasphaera* varied (model-dependent). We have noted these nuances in the Discussion.
- Final model and visualization: To illustrate model performance, we present ROC curves for each model (Figure 7) and corresponding SHAP interpretability plots (Figure 8) as well as summarize that XGBoost had the highest precision/recall overall for both tasks. We include the statement from the Results that *"the accuracy of all models was higher for BV than for cytology"* in our descriptions, and have added that note to Figure 7's legend as well, to ensure this important point is clearly communicated.

All machine learning analysis was conducted using Python's Scikit-learn (v1.x) and XGBoost (v1.x) libraries. Model training and validation code has been provided (see "Code Availability"). shap 0.43.0 was used for Shapley value calculation and imbalanced-learn 0.11.0 for optional SMOTE-ENN experiments (not used in final models)

QUANTIFICATION AND STATISTICAL ANALYSIS

All statistical analyses were performed in Python 3.10 (Ubuntu 22.04 LTS) with key libraries: NumPy for numeric computations, Pandas for data handling, SciPy/StatsModels for statistical tests, and Seaborn/Matplotlib for visualizations. GraphPad Prism 10 was additionally used for some plotting and confirmation of statistical results.

Statistical significance criteria

We set a threshold of $p < 0.05$ (two-tailed) to determine significance for individual tests. However, given the large number of comparisons in our analyses, we implemented corrections for multiple testing where appropriate.

- For targeted hypothesis tests (e.g., comparing BV + vs. BV– for a specific bacterium's prevalence), we report the raw p-value but also mention if it remains significant after correction considering the context.
- For broad screens (e.g., the heatmaps in Figures 5A and S18 that cover many variables), we applied the Benjamini-Hochberg false discovery rate (FDR) correction to the set of p-values. Adjusted p-values (q-values) < 0.05 were considered significant. In our results tables/figures, we typically mark an association as significant only if q<0.05. For example, although Figure 4 initially showed many nominal p<0.05 associations, only those that met q<0.05 after FDR are now highlighted in the revised figure.
- In cases of multiple groups (e.g., >2 cytology categories compared), if an overall ANOVA was significant, we followed up with post-hoc tests (Tukey's HSD) to identify pairwise differences, again using p<0.05 threshold per comparison.
- Effect size reporting: Alongside p-values, we report effect sizes where relevant: odds ratios for logistic associations, correlation coefficients (Spearman's ρ), or mean differences. This is to ensure that statistically significant results are also examined for practical significance. For instance, an OR of 0.5 or 2.0 is a sizeable effect, whereas an OR of 0.85, though significant in a large sample, might be of modest practical impact.
- Confidence intervals: We provide 95% confidence intervals for key estimates (ORs, AUCs, etc.). For the machine learning models, we used cross-validation to derive confidence intervals on metrics like accuracy (noting the range across folds).
- Software for statistics: All chi-square tests, t-tests, etc., were done with SciPy or StatsModels implementations. We double-checked the Chi-square results with Yates' continuity correction for 2×2 tables when counts were low, and used Fisher's exact test when any expected count <5.

In summary, the statistical approach was chosen to be comprehensive yet cautious: we cast a wide net to find patterns in this large dataset, but we account for multiple comparisons to reduce false positives, and we interpret findings in light of effect sizes and biological plausibility. All significant findings highlighted in the Results have q<0.05 unless otherwise specified. The robustness of these findings (e.g., through consistent machine learning feature importance and concordance with prior studies) gives us confidence in the conclusions drawn.