Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Phylogenetic harmonization of cross-cohort vaginal microbiome data for machine-learning prediction of preterm birth

Sunwha Park
Ewha Womans University
Kyeongeui Yun
HuNBiome Co
Hyebeen Byun
BiowaveW
Yeonhwa Park
BiowaveW
Young-Ah You
Ewha Womans University
Gain Lee
Ewha Womans University
Young Min Hur
Ewha Womans University
Minjung Ko
Ewha Womans University
Young Ju Kim
Ewha Womans University
Suenie Park
sueniepark@biowavew.com

BiowaveW

Article

Keywords:

Posted Date: October 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11420340/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

The vaginal microbiome is associated with preterm birth (PTB), but single-center prediction models may be limited by sample size, and public 16S rRNA datasets differ in variable regions and protocols. We integrated a Korean single-center cohort (EUMC; n = 235) with the Microbiome Preterm Birth DREAM Challenge (DC; 3,578 samples from 1,268 individuals) using phylogenetic-placement-based harmonization (MaLiAmPi). In pooled-data evaluation, the AUC increased from 0.657 to 0.731 for LASSO logistic regression and from 0.57 to 0.81 for CatBoost. Because cervical length (CL) was unavailable in DC, a classifier for short CL (sCL; <25 mm) was trained on EUMC and VMET samples with measured CL (n = 983; AUC 0.63); adding the inferred sCL feature changed the AUC from 0.81 to 0.83. A nine-feature model reached an AUC of 0.74. These estimates are not independent external validation; the sCL surrogate and nine-feature panel need validation.

Introduction

Preterm birth (PTB), defined as delivery before 37 weeks of gestation, affects approximately 10% of pregnancies worldwide and remains a major contributor to neonatal morbidity and mortality [1, 2]. Despite decades of research, PTB is biologically heterogeneous, with contributions from inflammation, hormonal and immune dysregulation, genetic and environmental factors, and microbial dysbiosis [2]. Spontaneous PTB accounts for a substantial proportion of PTB, and infection-associated pathways are thought to include ascending microbial processes involving the lower genital tract [3, 4].

High-throughput sequencing has substantially advanced characterization of the vaginal microbiome and its relationship to pregnancy outcomes [2, 5]. Vaginal microbial communities are commonly described using community state types (CSTs), including *Lactobacillus*-dominant states and more diverse communities enriched in anaerobic taxa such as *Gardnerella* and *Prevotella* [2, 6]. Across diverse cohorts, *Lactobacillus*-dominant communities, particularly those enriched in *L. crispatus*, have generally been associated with a lower risk of adverse pregnancy outcomes, whereas more diverse, dysbiotic communities have been associated with increased PTB risk [2, 7–9]. These recurrent associations have motivated efforts to translate vaginal microbiome composition into clinically useful prediction models.

Several studies have developed machine-learning models that combine vaginal microbiome features with clinical covariates such as cervical length (CL) and white blood cell (WBC) count. In a multicenter Korean cohort, a model combining microbiome markers with WBC and CL achieved an AUC of 0.84 [10], and conventional biochemical markers such as fetal fibronectin have shown limited sensitivity for PTB prediction [11]. Studies of this kind nevertheless commonly rely on cohorts of several hundred participants or fewer, limiting statistical power and raising questions about transportability across populations, sampling procedures, and sequencing platforms [2, 10].

The Microbiome Preterm Birth DREAM Challenge aggregated multiple published vaginal microbiome cohorts to address limitations of single-cohort sample size [12]. However, its constituent studies used heterogeneous 16S rRNA variable regions and sequencing protocols, and clinical covariates such as CL

were incompletely available. Directly pooling heterogeneous 16S datasets can therefore introduce technical variation that is difficult to distinguish from biological variation. Conversely, restricting analysis to samples with complete clinical covariates may discard substantial information contained in large public datasets.

Phylogenetic-placement-based approaches provide one strategy for integrating 16S rRNA data generated from different variable regions by placing ASVs onto a common reference phylogeny rather than requiring direct sequence comparability. In the present study, we integrated a Korean single-center cohort from Ewha Womans University Medical Center (EUMC) with the DREAM Challenge (DC) dataset using MaLiAmPi, compared LASSO logistic regression with CatBoost across different cohort scales, and developed a microbiome-based surrogate for short cervical length (sCL; CL < 25 mm) to explore whether this phenotype could provide complementary information when direct CL measurements were unavailable. We also evaluated whether a compact nine-feature panel retained part of the predictive performance of the full model. These analyses examine the feasibility of phylogenetically harmonized integration of heterogeneous external vaginal microbiome data for PTB prediction; they do not establish that harmonization alone caused the observed performance differences or that the resulting models are ready for clinical use.

Methods

Study population and design

This retrospective study integrated a single-center cohort of pregnant women recruited at Ewha Womans University Medical Center (EUMC), Seoul, Republic of Korea, with two external publicly available resources: the Microbiome Preterm Birth DREAM Challenge (DC) dataset and the Pregnancy Vaginal Microbiome, Epigenetics, and Term birth (VMET) cohort. The datasets were used for distinct but connected analyses: EUMC and DC for PTB prediction, and EUMC and VMET for development of the microbiome-based sCL classifier.

The EUMC analytical cohort comprised 235 pregnant women with singleton pregnancies (149 term births [TB] and 86 preterm births [PTB]). Women with gestational diabetes mellitus, preeclampsia, or incomplete medical records were excluded according to the study protocol. Cervicovaginal fluid (CVF) was collected from the posterior vaginal fornix using sterile cotton swabs during mid-pregnancy, before vaginal examination or clinical intervention (including antibiotics, corticosteroids, or progesterone), following the collection protocol described previously [10]. Samples were profiled by 16S rRNA gene amplicon sequencing targeting the V3–V4 region on an Illumina MiSeq platform (2×300 bp). Clinical data included 48 parameters, including cervical length (CL), CST, WBC count, CRP, fetal fibronectin, and prior PTB history.

The DC resource comprised nine publicly available vaginal microbiome studies aggregated through the Microbiome Preterm Birth DREAM Challenge, with 3,578 samples from 1,268 pregnant individuals and 2,589 term versus 989 preterm outcomes [12]. The VMET resource comprised two independent sub-

cohorts with 1,028 samples and directly measured CL [15]; 961 samples were retained after MaLiAmPi quality control. Because these resources were used for different analyses, sample counts and participant counts are reported separately where available, and the analytical denominator for each model is specified in the corresponding Results section.

Table 1
Cohort characteristics.

Characteristic	EUMC	DC	VMET
n (total samples)	235	3,578	1,028 (961 retained after MaLiAmPi QC)
Term / Preterm	149 / 86	2,589 / 989	776 / 185
16S rRNA region(s)	V3-V4	Heterogeneous across 9 constituent studies (V1-V2, V1-V3, V3-V5, V4)	VMET and VMET II (V1-V3, V1-V2)
Sequencing platform	Illumina MiSeq, 2x300bp	Heterogeneous across constituent studies	Illumina MiSeq
Directly measured cervical length	Yes	No	Yes
Source	Single-center, Korea	9 pooled public studies, multinational	Pruski et al. 2021 [15]

External data acquisition and quality control

External vaginal microbiome data were obtained from two public resources: the Microbiome Preterm Birth DREAM Challenge (DC) dataset and the Pregnancy Vaginal Microbiome, Epigenetics, and Term birth (VMET and VMET2) cohorts. The DC dataset was obtained as a MaLiAmPi-processed dataset from the March of Dimes Preterm Birth Research Database (Study ID SDY2187), consistent with the data sources described in the original DREAM Challenge study [12]. For the VMET cohort, 16S rRNA raw sequencing data were downloaded from ENA [15]. Quality-control filtering was applied within the MaLiAmPi pipeline before sCL classifier training and cross-cohort integration.

Cross-cohort harmonization

To integrate vaginal microbiome data generated from different 16S variable regions across EUMC, DC, and VMET, we used MaLiAmPi v3.5.0 [13], a phylogenetic placement-based integration approach. The DC dataset consisted of MaLiAmPi-processed data provided by the DREAM Challenge study and was used without reprocessing of the original sequencing reads. For the EUMC and VMET datasets, raw amplicon sequencing reads were independently processed using MaLiAmPi v3.5.0 following the same processing framework used for the DC dataset.

Briefly, amplicon sequence variants (ASVs) were generated with DADA2; a curated 16S rRNA reference allele set (ARF/YA16S database, curated from NCBI 16S rRNA records) was recruited; a reference

phylogeny was built de novo with RAxML; and query ASVs were placed onto this phylogeny using the Evolutionary Placement Algorithm (EPA-ng). This placement-based scaffold allows samples sequenced across different variable regions to be integrated at the level of phylogenetic position rather than raw sequence identity. Taxonomic classification (phylum to species) and alpha diversity metrics (Shannon index, Simpson index) were computed in QIIME 2 (2023.09) [14].

Short cervical length (sCL) phenotype model

Because CL was identified as an important clinical predictor of PTB in the EUMC cohort (see Results) and was unavailable for the DC dataset, we developed a surrogate classifier for short cervical length (sCL), defined as CL < 25 mm. The classifier was trained using harmonized EUMC and VMET data with directly measured CL (n = 983; EUMC n = 95 [32 sCL and 63 nCL]; VMET n = 888 [202 sCL and 686 nCL]) [15], using CatBoost.

Candidate sCL-classifier features were screened for cohort-specific effects before the classifier was finalized. Features whose apparent importance was driven primarily by near-exclusive detection in one source cohort, rather than a reproducible association with sCL status across EUMC and VMET, were not retained. *Peptostreptococcus anaerobius* was excluded on this basis after ranking as the top feature in an earlier model iteration (Supplementary Table S3). The finalized classifier was applied to DC samples and to EUMC samples without directly measured CL to generate an inferred sCL/nCL label. This inferred phenotype was then evaluated as an additional feature in the PTB prediction model.

Because this label is inferred from microbiome features rather than measured by cervical ultrasonography, it is treated as a surrogate phenotype and not as an equivalent replacement for directly measured CL.

Machine learning model development

Two modeling approaches were evaluated: LASSO logistic regression implemented in SIAMCAT [18] and CatBoost gradient boosting implemented using PyCaret v3.0.2 [16, 17]. Both approaches used curated genus- and species-level relative-abundance features together with alpha-diversity features. Models were evaluated using an 80/20 train–test split with five-fold cross-validation within the training data. The split was performed at the participant level so that samples from the same participant were not assigned to both training and test sets. Performance was summarized primarily by the area under the receiver operating characteristic curve (ROC AUC); accuracy, precision, and recall were also calculated from the confusion matrix where applicable. These evaluations represent performance under the study's pooled-data split design, not independent cohort-wise external validation.

For CatBoost, ROC AUC values reported in the text and tables refer to the class-specific ROC AUC for the positive class (PTB in the PTB prediction models and sCL in the sCL classifier). The values are reported as point estimates from the model evaluation described above. Micro-average and macro-average curves may also appear in the software-generated ROC plots, but they are not the AUC values reported in

the manuscript. For SIAMCAT, AUC refers to the area under the ROC curve generated by the package. AUC values are presented descriptively as point estimates without confidence intervals.

To evaluate the effect of cohort expansion independently of algorithm choice, PTB models were compared using EUMC alone and harmonized EUMC + DC data. The CatBoost model was subsequently used for the sCL integration and simplified-panel analyses.

A simplified model was constructed using features with importance scores > 1.0 across the EUMC-only, EUMC + DC, and EUMC + DC-with-inferred-sCL model configurations. The resulting nine-feature set comprised *Gardnerella*, *Prevotella*, *Lactobacillus*, *L. coleohominis*, *L. crispatus*, *L. iners*, *L. jensenii*, *L. reuteri*, and the Shannon diversity index. The performance of this reduced feature set is reported as an evaluation within the present pooled-data modeling framework and should not be interpreted as independent validation of a finalized assay.

Statistical analysis

Continuous clinical variables were compared between TB and PTB groups using the Wilcoxon rank-sum test. Differences in microbial distribution between groups were assessed using the Wilcoxon rank-sum test. Conditional entropy analysis was used to rank clinical variables according to their information content for the TB/PTB outcome label.

Ethics and data availability

This study was approved by the Ethical Research Committee of Ewha Womans University Mokdong Hospital (IRB No. 2018-07-007). All participants provided written informed consent. Raw 16S rRNA sequencing data for the EUMC cohort will be deposited in the NCBI Sequence Read Archive (SRA). The DC and VMET datasets were obtained from the March of Dimes Preterm Birth Research Database (Study ID SDY2187) and ENA, respectively [12, 15].

Results

PTB prediction performance varies with algorithm choice and cohort integration

We first evaluated the effects of algorithm choice and external cohort integration on PTB prediction performance. EUMC and DC vaginal microbiome data were harmonized using the MaLiAmPi phylogenetic-placement framework before the corresponding model comparisons. This common preprocessing framework was used to reduce methodological differences attributable to taxonomic and phylogenetic processing across cohorts.

Two modeling approaches were compared: LASSO logistic regression implemented in SIAMCAT and gradient-boosted decision trees implemented using CatBoost through PyCaret v3.0.2. Each approach was evaluated using the EUMC cohort alone and the harmonized EUMC + DC dataset. With SIAMCAT, the

observed AUC was 0.657 in the EUMC-only analysis and 0.731 in the EUMC + DC analysis. With CatBoost, the observed AUC was 0.57 for EUMC alone (n = 235; taxonomic relative-abundance and alpha-diversity features only) and 0.81 for the harmonized EUMC + DC analysis (n = 3,813). These are point estimates under the specified pooled-data evaluation; the changes should not be interpreted as proof that cohort integration alone caused the increase in discrimination.

On the EUMC-only dataset, the CatBoost point-estimate AUC was lower than that of LASSO logistic regression, whereas the ordering was reversed in the EUMC + DC analysis. The observed change in AUC differed between algorithms; however, these results do not establish a general advantage of CatBoost or a formal algorithm-by-cohort-size interaction. Cohort composition, sample size, and study-specific technical characteristics change together in the pooled analysis, so their individual contributions cannot be isolated from these comparisons alone.

The highest-importance features in the EUMC + DC CatBoost model included *Lactobacillus jensenii*, *Peptoniphilus harei*, *Prevotella*, and *Lactobacillus crispatus*. No raw clinical variables were included as inputs in this microbiome-only model. On the basis of this comparison, CatBoost was selected for the subsequent analyses, including integration of the microbiome-derived sCL phenotype and evaluation of the simplified feature panel.

Microbiome-based inference of short cervical length as an additional model feature

CL was unavailable for the DC cohort. To extend CL-related information to this cohort without restricting PTB modeling to participants with direct measurements, we developed a surrogate classifier for short cervical length (sCL; CL < 25 mm). The CatBoost sCL classifier was trained on harmonized EUMC and VMET data with directly measured CL (n = 983) and achieved an AUC of 0.63, indicating modest discrimination. *Gardnerella* and *L. crispatus* were among the highest-importance features, with importance scores > 10 in the reported model.

The classifier was applied to DC samples (n = 3,578) and to EUMC samples without directly measured CL (n = 140) to generate an inferred sCL/nCL label. Adding this microbiome-derived surrogate to the harmonized EUMC + DC CatBoost model was associated with an increase in the point-estimate AUC from 0.81 to 0.83. Because the added feature is a model-derived surrogate rather than a direct CL measurement, and because the downstream result depends on the performance and applicability of the sCL classifier, this difference should be interpreted as exploratory and requires independent evaluation in a dataset with directly measured CL.

A nine-feature simplified model retains part of the predictive performance of the full taxonomic model

The full harmonized EUMC + DC CatBoost model achieved a point-estimate AUC of 0.81 without the inferred sCL feature and 0.83 after sCL integration. Because a model containing several hundred taxonomic features is difficult to interpret and is not directly aligned with a targeted molecular assay, we evaluated a simplified feature set.

The nine features were retained because they had importance scores > 1.0 across the EUMC-only, EUMC + DC, and EUMC + DC-with-inferred-sCL model configurations. The panel comprised *Gardnerella*, *Prevotella*, *Lactobacillus*, *L. coleohominis*, *L. crispatus*, *L. iners*, *L. jensenii*, *L. reuteri*, and the Shannon diversity index. All nine features differed between TB and PTB in the pooled EUMC + DC dataset (Wilcoxon rank-sum test, $p < 0.05$; examples: *L. jensenii*, $p = 5 \times 10^{-21}$; *L. crispatus*, $p = 3 \times 10^{-10}$; *Gardnerella*, $p = 3 \times 10^{-8}$; Supplementary Table S1). These group differences are descriptive and do not, by themselves, establish the predictive or causal importance of each feature.

A CatBoost model restricted to these nine features and trained on the harmonized EUMC + DC dataset (n = 3,813; without inferred sCL) achieved an AUC of 0.74. This point estimate was lower than that of the full model (0.81), while the number of input features was substantially reduced. The nine-feature set should therefore be regarded as a candidate panel for targeted assay development and independent analytical and clinical validation, rather than as a clinically validated diagnostic panel.

Table 2
Model performance summary across the analysis pipeline.

Stage	Dataset	n	Algorithm	Feature set	AUC
Prior single-cohort model [10]	Multicenter Korean	150	Random forest	Microbiome + WBC + CL	0.84
Algorithm comparison — LASSO	EUMC alone → EUMC + DC	235 → 3,813	SIAMCAT	Taxonomic	0.657 → 0.731
Algorithm comparison — CatBoost	EUMC alone → EUMC + DC	235 → 3,813	CatBoost	Taxonomic + diversity	0.57 → 0.81
sCL classifier	EUMC+VMET	983	CatBoost	Taxonomic + diversity	0.63
+sCL integration	EUMC + DC	3,813	CatBoost	Taxonomic + diversity + predicted sCL	0.83
Nine-feature simplified model	EUMC + DC	3,813	CatBoost	9 taxa	0.74

Taxonomic and clinical differences between term and preterm birth

The nine-feature panel reflected a broader compositional pattern: the PTB group had lower mean relative abundance of *Lactobacillus* taxa (the genus and *L. crispatus*, *L. jensenii*, *L. iners*, *L. reuteri*, and *L. coleohominis*) and higher mean relative abundance of *Prevotella*, together with a higher mean Shannon index. *Gardnerella* relative abundance differed statistically between groups but was numerically lower in PTB than TB (6.42% vs 7.13%; Supplementary Table S1). These observations are broadly consistent with previously reported associations between vaginal microbial community composition and PTB [2, 7–9], but they do not establish a causal role for any individual taxon.

Conditional-entropy analysis of the EUMC clinical variables (Supplementary Figure S1A) showed the lowest remaining uncertainty about the TB/PTB label for delivery week and weight, followed by sampling week, CL, and WBC. Delivery week and weight are outcome-proximal variables and are presented descriptively; they are not proposed as predictors for an antenatal model. CL and WBC differed between TB and PTB groups (Supplementary Figure S1B,C and Supplementary Table S2), consistent with the previously reported relevance of CL [10] and with its use as the target phenotype for the sCL classifier. This conditional-entropy analysis was descriptive and separate from model feature selection.

Among the clinical parameters, CL, CRP, and WBC differed between the TB and PTB groups, whereas age and BMI did not (Supplementary Table S2).

Discussion

The principal finding was that integration of the external DREAM Challenge microbiome dataset with the EUMC cohort was associated with higher point-estimate AUCs for PTB prediction in the two evaluated modeling approaches. The observed AUC changed from 0.657 to 0.731 for LASSO logistic regression and from 0.57 to 0.81 for CatBoost. The magnitude and direction of the change differed by algorithm. Because these values were obtained under the specified pooled-data evaluation design, they should be interpreted as internal performance estimates rather than evidence of generalizable improvement in an independent population.

The sCL analysis provided a second methodological observation. A microbiome-based classifier trained on EUMC and VMET data with directly measured CL achieved an AUC of 0.63 and generated an inferred sCL feature for samples without direct CL measurement. Addition of this surrogate was associated with an increase in the point-estimate AUC of the EUMC + DC CatBoost model from 0.81 to 0.83. Because the added feature was model-derived, this result should be regarded as exploratory proof of concept, not evidence that predicted sCL can substitute for directly measured CL or provides established clinical benefit.

Finally, a nine-feature panel substantially reduced the feature space while reaching an AUC of 0.74 in the pooled analysis, lower than that of the full model (0.81). This provides a rationale for further assay-oriented evaluation, but not for immediate clinical use.

Previous single-cohort machine-learning studies have combined vaginal microbiome features with clinical covariates and reported AUCs in the approximate range of 0.72–0.84 using algorithms such as random forest, XGBoost, and GUIDE [10, 19]. These studies suggest that microbiome information may contribute to PTB prediction, particularly when combined with clinically informative variables such as CL and other maternal or obstetric characteristics. Numerical AUC comparisons across studies are contextual only because the cohorts, predictors, endpoints, and validation designs differ; the present estimates should not be interpreted as directly comparable to those of the previous Korean model [10].

Other studies have extended this strategy by incorporating additional biological layers. For example, cervicovaginal cytokine measurements have been combined with microbiome profiles to predict preterm delivery in other populations, including an Asian cohort in which microbiota and cytokine features jointly contributed to prediction [20]. Non-invasive approaches based on bacterial DNA signatures have also been evaluated in large prospective cohorts, such as the Predict1000 study, while microbiome-based approaches have been explored for related obstetric outcomes such as chorioamnionitis [21, 22]. Collectively, these studies illustrate a broader movement toward combining microbiome-derived information with clinical or host-derived features to improve prediction of adverse pregnancy outcomes.

At the multi-cohort scale, the Microbiome Preterm Birth DREAM Challenge was established to facilitate the development and comparison of machine-learning approaches using pooled, publicly available vaginal microbiome datasets [12]. The Challenge provided an important opportunity to evaluate whether models trained across heterogeneous cohorts could identify reproducible microbiome-associated patterns of PTB. At the same time, the use of multiple publicly available datasets introduces methodological challenges that are less prominent in single-cohort studies, including differences in sample collection, sequencing procedures, 16S rRNA variable regions, taxonomic assignment, and availability of clinical metadata. Thus, increasing the number of samples through cohort integration does not by itself ensure that the resulting feature space is directly comparable across studies.

The present study differs from this previous body of work in two methodological respects. First, rather than assuming that amplicon data generated from different 16S rRNA variable regions are directly comparable, we applied phylogenetic-placement-based harmonization using MaLiAmPi. This approach places ASVs onto a common reference phylogeny and provides a framework for integrating data across variable regions. It may reduce incompatibility associated with variable-region differences, but it does not remove all technical, cohort, or biological heterogeneity. This distinction is important when interpreting pooled-data model performance across studies with different collection and sequencing procedures.

Second, the present study addressed a different form of heterogeneity arising from incomplete clinical metadata. Previous single-cohort models that incorporated clinical variables such as CL generally required those measurements to be available for the participants included in the model [10, 19]. In large secondary-use datasets, however, clinically informative variables may not have been collected, may not be available in the public dataset, or may have been measured inconsistently across cohorts. In the present study, CL was available in the EUMC dataset but was not directly available for the DC samples. Rather than treating this difference solely as missing data and excluding the corresponding samples from analyses involving CL, we evaluated whether a microbiome-based machine-learning classifier could infer the clinically relevant sCL phenotype (CL < 25 mm). The resulting inferred sCL feature was then evaluated as an additional model-derived clinical feature in samples lacking direct CL measurements.

This strategy should be distinguished from conventional imputation of a missing numerical value. The objective was not to reconstruct an individual participant's exact cervical length in millimeters, but to

derive a microbiome-based surrogate of the clinically defined sCL phenotype that could provide related information in samples for which direct CL measurement was unavailable. In this sense, the approach represents a form of clinical-phenotype harmonization across cohorts, complementing the taxonomic and phylogenetic harmonization performed for the microbiome data. The sCL analysis therefore provides a methodological extension of previous work in which CL and other clinical variables were directly incorporated into PTB prediction [10].

The combined results suggest that these two forms of harmonization may be useful when integrating heterogeneous microbiome datasets. Microbiome harmonization through phylogenetic placement allowed data generated from different 16S rRNA variable regions to be represented within a common analytical framework, while the microbiome-derived sCL feature provided an opportunity to retain information related to a clinically relevant covariate that was not directly measured in one of the external cohorts. Importantly, the latter result should be interpreted as a proof-of-concept rather than as evidence that an inferred sCL phenotype can replace directly measured CL. Direct cervical-length measurement remains the clinical reference for defining cervical shortening, and the inferred feature should instead be regarded as a model-derived surrogate whose independent validity requires further evaluation.

The present findings should also be interpreted in the context of differences among the published studies. Reported AUCs cannot be compared directly across cohorts because performance is influenced by population characteristics, gestational age at sampling, sample-collection procedures, sequencing and taxonomic-analysis methods, endpoint definitions, clinical covariates, class distributions, and validation strategies. In particular, an apparent increase in AUC after combining cohorts does not by itself establish that the integrated model will perform better in a new population. The difference observed between LASSO logistic regression and CatBoost in the present analysis likewise indicates a model- and dataset-dependent response to cohort expansion rather than establishing a general superiority of nonlinear models.

Accordingly, the principal contribution of the present study is methodological rather than a claim of universal clinical superiority. The results illustrate how heterogeneous vaginal microbiome datasets may be integrated by addressing both technical differences in microbiome measurements and incompleteness of clinically relevant metadata. The use of a microbiome-derived sCL surrogate further suggests a possible strategy for retaining clinically informative phenotype-related information when direct measurements are unavailable. These approaches may be particularly relevant to secondary-use microbiome datasets, although their robustness, calibration, and generalizability should be evaluated in independent cohorts with prospectively collected clinical measurements and standardized sampling and sequencing procedures.

The microbiome findings in the present study are broadly consistent with the literature on vaginal microbiome composition and PTB, in which reduced representation of *Lactobacillus*-dominated communities and increased representation of taxa such as *Gardnerella* and *Prevotella* have been associated with PTB-related dysbiosis across diverse populations [2, 7–9]. In our data, the PTB-

associated microbiome pattern included lower relative abundance of selected *Lactobacillus* taxa and higher relative abundance of *Prevotella*, together with higher Shannon diversity in PTB. *Gardnerella* differed significantly between groups, but its mean relative abundance was not higher in PTB in the pooled dataset; this finding warrants further examination, for example at the species or clade level. The nine-feature panel captured part of this compositional pattern in a more compact form.

The pooled-model AUC estimates (0.81 without the inferred sCL feature and 0.83 with the inferred sCL feature) were within the range of values reported in previous single- or multicenter studies (approximately 0.72–0.84) and did not exceed that of a previous Korean model that incorporated directly measured CL and WBC (0.84) [10]. However, these numerical comparisons should be interpreted cautiously because the cohorts, sampling characteristics, microbiome features, clinical covariates, endpoint definitions, and validation strategies differed across studies. In particular, the present inferred sCL feature should not be considered equivalent to directly measured CL. The practical contribution of the present approach is therefore not to demonstrate replacement of standard clinical measurements, but to provide a potential strategy for retaining clinically relevant phenotype-related information when direct clinical covariates are incompletely available in retrospective or secondary-use microbiome datasets.

An important methodological observation was the identification and exclusion of *Peptostreptococcus anaerobius* as a candidate sCL feature whose apparent importance was driven largely by cohort-specific detection in VMET (Supplementary Table S3). This finding illustrates that phylogenetic-placement-based harmonization does not necessarily eliminate all cohort-related technical or sampling effects. Candidate biomarkers emerging from pooled microbiome datasets therefore require cohort-stratified assessment before being interpreted as biologically meaningful or advanced toward assay development. This observation also supports the broader need for standardized sampling, sequencing, preprocessing, and cross-study validation in microbiome-based prediction research [2, 23]. Harmonization should be viewed as a strategy for reducing technical incompatibility, not as a guarantee that all residual cohort effects have been removed.

Strengths of this study include the explicit comparison of a regularized linear model and a gradient-boosted tree model across cohort scales, the application of a common phylogenetic-placement framework to all three data sources, the cohort-stratified screening of candidate sCL features that led to the exclusion of *P. anaerobius*, and the explicit labeling of the sCL feature as a model-derived surrogate rather than a measured clinical variable.

Several limitations should be considered when interpreting the present findings. First, the sCL feature incorporated into the PTB prediction model was model-derived rather than directly measured for most DC samples and for a subset of EUMC samples. The estimated sCL values therefore depend on the performance and transportability of the sCL classifier and should not be interpreted as a replacement for ultrasound measurement. In addition, the dataset audit identified overlap between VMET-derived BioProjects (PRJEB11895 and PRJEB12577) represented in the DC dataset and the VMET data used to

train the sCL classifier. Samples from these overlapping BioProjects were not excluded from the analysis. Consequently, the sCL classifier may have been applied to samples represented in its training data, potentially introducing information leakage and leading to optimistic estimates of downstream PTB prediction performance. The increase in point-estimate AUC from 0.81 to 0.83 should therefore be interpreted cautiously and does not, by itself, establish a meaningful improvement in predictive performance. Independent evaluation using samples excluded from the sCL classifier training data, together with directly measured cervical length (CL), is needed before this surrogate approach is considered for downstream PTB prediction.

Second, the primary model evaluation used an 80/20 participant-level train–test split of the pooled EUMC + DC dataset with five-fold cross-validation in the training data, rather than a cohort-wise external holdout or leave-one-study-out validation strategy. The reported AUCs therefore describe performance under the specified pooled-data evaluation design. They do not establish transportability to an entirely independent population, laboratory, sequencing protocol, or clinical setting. Cohort-wise external validation, preferably using prospectively collected samples and standardized clinical and laboratory procedures, will be necessary to establish generalizability.

Third, residual technical and biological heterogeneity across the constituent DC studies cannot be excluded despite the use of phylogenetic-placement-based harmonization. MaLiAmPi-based placement addresses an important component of heterogeneity arising from differences in 16S rRNA variable regions, but it does not eliminate other sources of inter-study variation, including specimen collection and storage, DNA extraction, amplification and sequencing procedures, sequencing depth, taxonomic reference databases, gestational timing, clinical characteristics, and population composition. In addition, cohort of origin was not explicitly modeled as a covariate in the primary analysis. Consequently, some of the observed predictive signal may still reflect cohort-specific biological or technical characteristics rather than features that are fully transportable across populations. The exclusion of *P. anaerobius*, whose apparent importance in the sCL classifier reflected near-exclusive detection in VMET, illustrates this residual cohort variation.

Finally, the EUMC cohort was collected at a single Korean center and was integrated with predominantly non-Korean public datasets. Differences in geographic background, demographic characteristics, clinical management, gestational age at sampling, microbiome composition, and PTB case mix may affect the transportability of the resulting models. Moreover, the current study is based on retrospective or secondary-use datasets and does not establish prospective clinical utility, calibration, or clinically actionable decision thresholds. The nine-feature panel should therefore be regarded as a candidate feature set for subsequent analytical and clinical validation rather than as a clinically validated diagnostic assay.

Future work should prioritize independent cohort-wise external validation, leakage-resistant model evaluation, and assessment of calibration and clinical utility, including decision-curve analysis where appropriate. The microbiome-derived sCL approach should be evaluated independently in cohorts with

directly measured CL, with non-overlapping training and evaluation participants, before the inferred phenotype is incorporated into downstream PTB models. The nine-feature panel should then undergo analytical validation using targeted quantitative molecular assays, including assessment of sensitivity, specificity, precision, reproducibility, and cross-reactivity. Decision thresholds should be prespecified and evaluated in independent clinical samples rather than selected from the model-development dataset.

Third, the nine-feature panel provides a defined candidate set for assay-oriented development. Its analytical performance, including sensitivity, specificity, precision, reproducibility, and cross-reactivity, should be evaluated using targeted quantitative molecular assays, and prespecified decision thresholds should be established and evaluated in independent clinical samples rather than derived from the datasets used for model development. Fourth, future biological studies should move beyond taxonomic association toward strain-level and functional characterization of the candidate organisms and communities, with prospective evaluation of any incremental value of host inflammatory markers. Finally, when cohorts contain partially overlapping clinical and molecular measurements, a well-characterized reference cohort could potentially be used to derive surrogate phenotypes from another data modality, for example using metabolites, cytokine profiles, or extracellular-vesicle-associated features; such extensions would require independent validation to establish that the inferred features capture reproducible biological relationships rather than cohort-specific correlations.

Technical limitations inherent to 16S rRNA amplicon sequencing should also be addressed through standardized laboratory and analytical procedures. Phylogenetic-placement-based harmonization can reduce incompatibility arising from different variable regions, but it cannot correct all biases associated with sample collection and storage, DNA extraction, primer choice, sequencing depth, rarefaction, 16S rRNA gene copy number, or reference-database choice and version [23, 24]. A 23-laboratory study of two DNA reference reagents composed of 20 common gut bacteria identified several of these factors as sources of profiling bias and proposed minimum quality criteria for evaluating performance [24]. Those reagents were developed for gut microbiome profiling; vaginal microbiome-specific reference reagents have separately been proposed to improve reproducibility and comparability across laboratories [23]. Such reagents were not evaluated in the present study, but future studies could use them alongside clinical samples and predefined quality criteria to help distinguish technical bias from biological and cohort effects.

In conclusion, phylogenetic-placement-based harmonization enabled pooled analysis of heterogeneous vaginal microbiome datasets, and the observed point-estimate AUCs differed between single-cohort and pooled-data evaluations. The microbiome-derived sCL surrogate and nine-feature panel are methodological candidates for further study, not clinically validated tools. Their value will depend on independent generalization, analytical reproducibility, biological plausibility, and clinical utility demonstrated in prospectively designed validation studies.

Declarations

Funding

This study was primarily funded by BiowaveW using its own internal resources and was partly supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant numbers, RS-2026-25538257 and RS-2023-00262969).

Author Contribution

Conceptualization and study design: Young Ju Kim, Sunwha Park, Suenie Park. Clinical cohort recruitment, sample collection and clinical data curation: Sunwha Park, Young-Ah You, Gain Lee, Young Min Hur, Minjung Ko, Young Ju Kim. 16S rRNA sequencing and MaLiAmPi processing: Kyeongeui Yun. Data integration, machine-learning analysis and statistical analysis: Young-Ah You, Gain Lee, Suenie Park. Interpretation of results: Sunwha Park, Young-Ah You, Young Ju Kim, Suenie Park. Writing of the original draft: Sunwha Park. Review and editing: all authors. Supervision and funding acquisition: Young Ju Kim, Suenie Park. All authors read and approved the final manuscript.

Data Availability

Raw 16S rRNA sequencing data for the EUMC cohort will be deposited in the NCBI Sequence Read Archive (SRA). The DC and VMET datasets were obtained from the March of Dimes Preterm Birth Research Database (Study ID SDY2187) and ENA, respectively.

References

1. Goldenberg RL, Culhane JF, Iams JD, Romero R. Epidemiology and causes of preterm birth. Lancet 371, 75–84 (2008).
2. Park S, You YA, Lee G, Hur YM, Go YY, Park S, Kim YJ. Vaginal microbiome and its implications in preterm birth. Obstet Gynecol Sci 68, 449–457 (2025). doi: 10.5468/ogs.25085
3. Romero R, Dey SK, Fisher SJ. Preterm labor: one syndrome, many causes. Science 345, 760–765 (2014).
4. Chan RL. Biochemical markers of spontaneous preterm birth in asymptomatic women. Biomed Res Int 2014, 164081 (2014).
5. Yoo JY, et al. 16S rRNA gene-based metagenomic analysis reveals differences in bacteria-derived extracellular vesicles in the urine of pregnant and non-pregnant women. Exp Mol Med 48, e208 (2016).
6. DiGiulio DB, et al. Temporal and spatial variation of the human microbiota during pregnancy. Proc Natl Acad Sci U S A 112, 11060–11065 (2015).

7. Kan H, et al. Differential effect of vaginal microbiota on spontaneous preterm birth among Chinese pregnant women. Biomed Res Int 2022, 3536108 (2022).
8. Fettweis JM, et al. The vaginal microbiome and preterm birth. Nat Med 25, 1012–1021 (2019).
9. Dunlop AL, et al. Vaginal microbiome composition in early pregnancy and risk of spontaneous preterm and early term birth among African American women. Front Cell Infect Microbiol 11, 641005 (2021).
10. Park S, et al. Predicting preterm birth through vaginal microbiota, cervical length, and WBC using a machine learning model. Front Microbiol 13, 912853 (2022).
11. Conde-Agudelo A, Papageorghiou AT, Kennedy SH, Villar J. Novel biomarkers for the prediction of the spontaneous preterm birth phenotype: a systematic review and meta-analysis. BJOG 118, 1042–1054 (2011).
12. Golob JL, et al. Microbiome preterm birth DREAM challenge: crowdsourcing machine learning approaches to advance preterm birth research. Cell Rep Med 5, 101350 (2024).
13. Minot SS, et al. MaLiAmPi enables generalizable and taxonomy-independent microbiome features from technically diverse 16S-based microbiome studies. Cell Rep Methods 3, 100650 (2023).
14. Bolyen E, et al. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37, 852–857 (2019).
15. Pruski P, et al. Direct on-swab metabolic profiling of vaginal microbiome host interactions during pregnancy and preterm birth. Nat Commun 12, 5967 (2021).
16. Ali M, contributors. PyCaret: An open-source, low-code machine learning library in Python. 2024. https://github.com/pycaret/pycaret
17. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. Adv Neural Inf Process Syst 31 (2018).
18. Wirbel J, et al. Microbiome meta-analysis and cross-disease comparison enabled by the SIAMCAT machine learning toolbox. Genome Biol 22, 93 (2021).
19. Della Rosa PA, et al. A hierarchical procedure to select intrauterine and extrauterine factors for methodological validation of preterm birth risk estimation. BMC Pregnancy Childbirth 21, 306 (2021).
20. Kumar M, et al. Vaginal microbiota and cytokine levels predict preterm delivery in Asian women. Front Cell Infect Microbiol 11, 639665 (2021).
21. Payne MS, et al. A specific bacterial DNA signature in the vagina of Australian women in midpregnancy predicts high risk of spontaneous preterm birth (the Predict1000 study). Am J Obstet Gynecol 224, 206.e1-206.e23 (2021).
22. Urushiyama D, et al. Vaginal microbiome as a tool for prediction of chorioamnionitis in preterm labor: a pilot study. Sci Rep 11, 18971 (2021).
23. Da Silva AS, et al. The untapped potential of vaginal microbiome diagnostics for improving women's health. Front Cell Infect Microbiol 15, 1595182 (2025).

24. Anwar S, et al. DNA reference reagents isolate biases in microbiome profiling: a global multi-lab study. mSystems e00466-25 (2025). doi: 10.1128/msystems.00466-25.

Figures

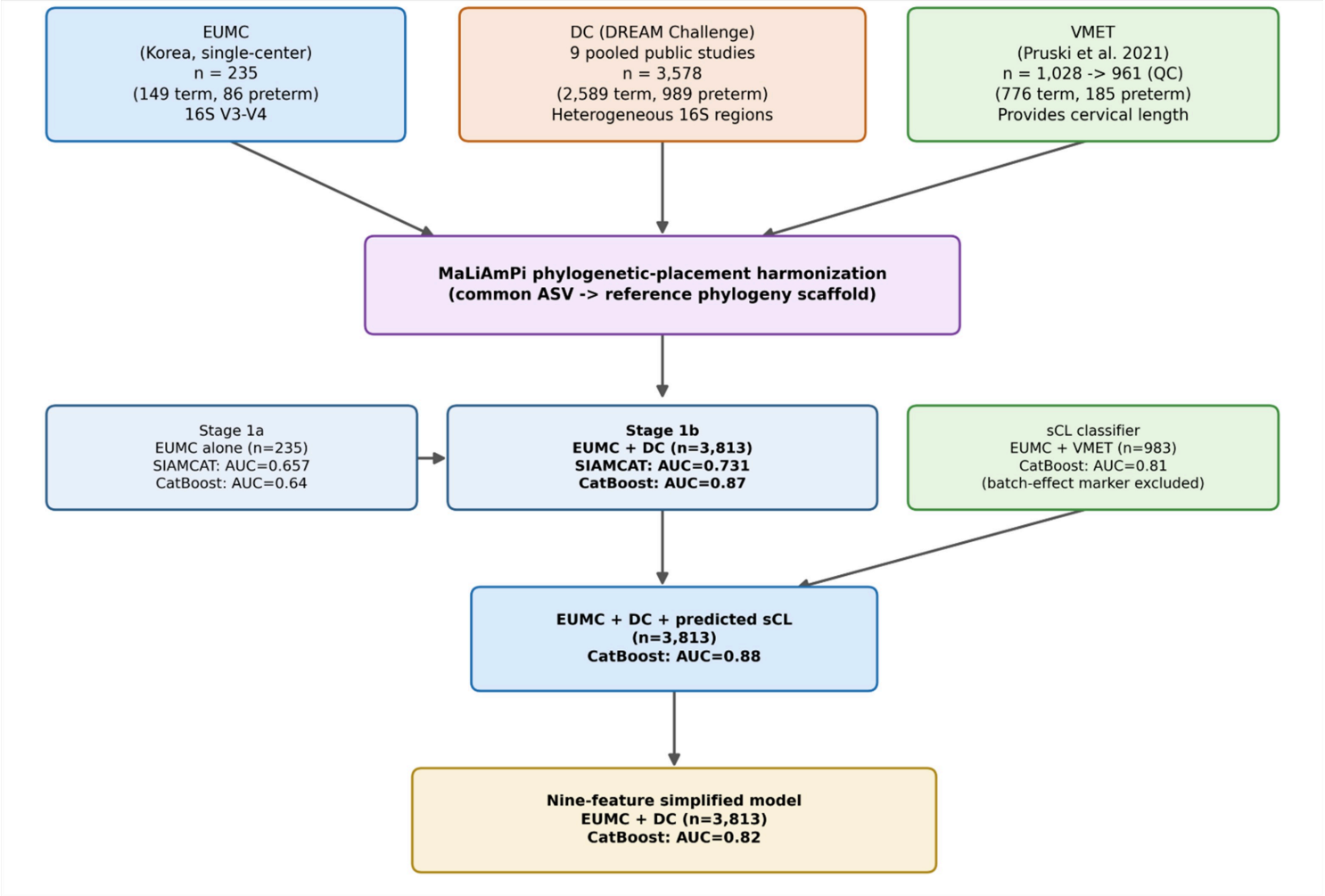


Figure 1

Study design and cohort flow. Schematic of the three source cohorts (EUMC, DC, VMET), the MaLiAmPi harmonization step, and the sequential model pipeline (EUMC alone → EUMC+DC → EUMC+DC+sCL → nine-feature model), with the AUC achieved at each stage

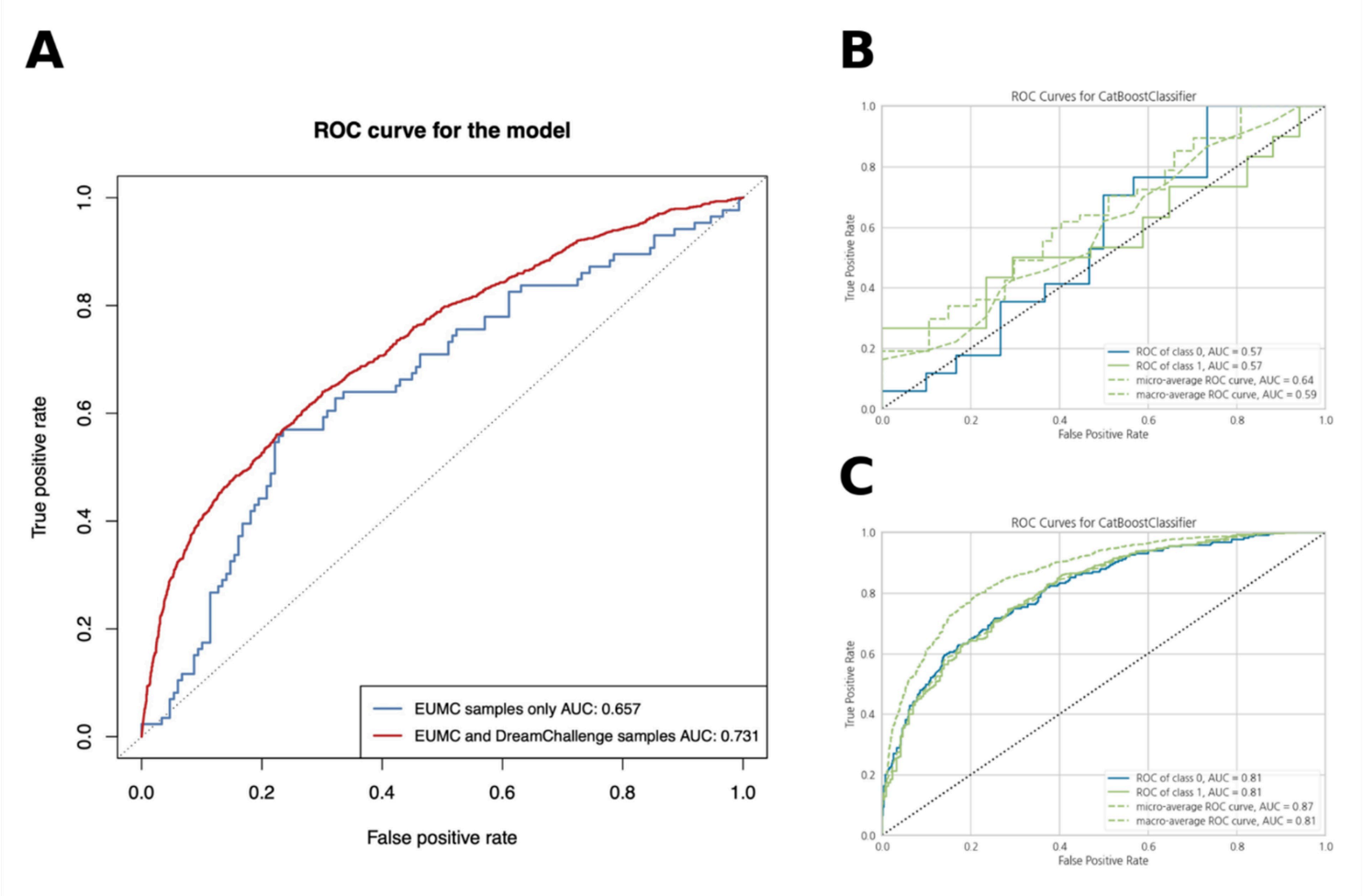
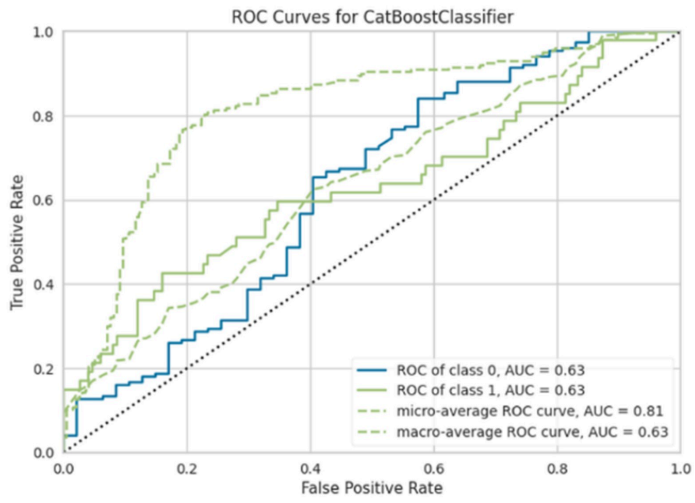


Figure 2

PTB prediction performance by algorithm and cohort integration. (A) ROC curves for the LASSO logistic regression model (SIAMCAT), EUMC alone (AUC=0.657) versus EUMC+DC (AUC=0.731). (B) ROC curves for the CatBoost model trained on EUMC alone (AUC=0.57). (C) ROC curves for the CatBoost model trained on the harmonized EUMC+DC dataset (AUC=0.81). In the CatBoost plots, the AUC reported in the text corresponds to the class-specific ROC curves (class 0 and class 1), not to the micro-average curve (see Methods).



B

rank	Feature Id	Importances
1	*g__Gardnerella*	16.17
2	*s__Lactobacillus crispatus*	14.39
3	*g__Ureaplasma*	8.33
4	*g__Lactobacillus*	8.11
5	*g__Peptoniphilus*	7.60
6	simpson_index	4.04
7	*s__Ureaplasma parvum*	3.98
8	*s__Prevotella colorans*	2.13
9	*s__Lactobacillus hominis/gasseri*	2.09
10	*s__Lactobacillus acidophilus*	2.07
11	*g__Bifidobacterium*	2.04
12	*s__Atopobium vaginae*	2.04
13	*s__Lactobacillus paragasseri*	1.99
14	*s__Sneathia amnii*	1.88
15	*g__Finegoldia*	1.84
16	*s__Ureaplasma parvum/urealyticum*	1.75
17	*s__Staphylococcus haemolyticus*	1.52
18	*s__Lactobacillus jensenii*	1.45
19	*s__Lactobacillus gasseri/paragasseri*	1.42
20	*g__Staphylococcus*	1.36

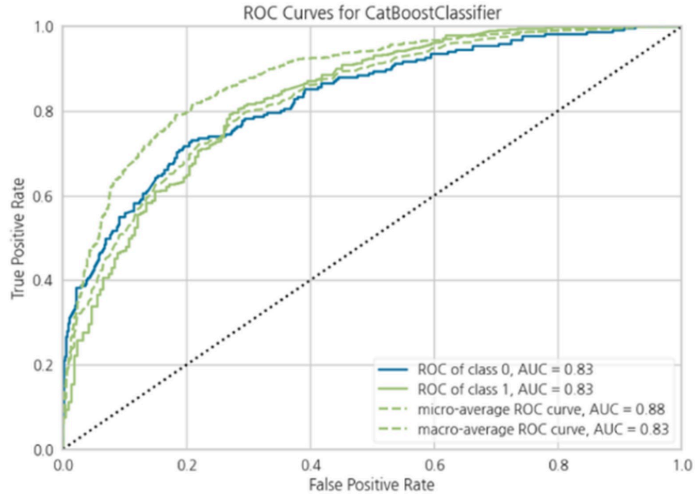
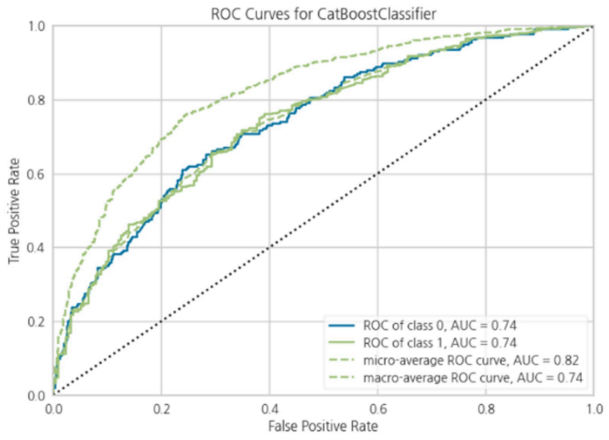


Figure 3

The short cervical length (sCL) classifier and its integration into the PTB prediction model. (A) ROC curves for the CatBoost sCL classifier trained on the harmonized EUMC+VMET dataset (n=983; AUC=0.63). (B) Top 20 features by importance for the sCL classifier; *Gardnerella* (16.17) and *Lactobacillus crispatus* (14.39) are the leading predictors. (C) ROC curves for the CatBoost PTB model trained on the harmonized EUMC+DC dataset after addition of the inferred sCL label (AUC=0.83; the model without the inferred sCL feature, AUC=0.81, is shown in Figure 2C).

A



B

rank	feature_name	feature_score
1	*g__Prevotella*	13.41
2	*s__Lactobacillus iners*	12.99
3	*g__Gardnerella*	12.93
4	shannon_index	12.04
5	*s__Lactobacillus crispatus*	11.96
6	*g__Lactobacillus*	11.96
7	*s__Lactobacillus jensenii*	11.52
8	*s__Lactobacillus reuteri*	6.74
9	*s__Lactobacillus coleohominis*	6.44

Figure 4

The nine-feature simplified model. (A) ROC curves for the CatBoost model restricted to the nine-feature panel, trained on the harmonized EUMC+DC dataset (AUC=0.74). (B) Feature importance scores for the nine-feature panel, ranked by score: *Prevotella*(13.41), *L. iners* (12.99), *Gardnerella* (12.93), Shannon index (12.04), *L. crispatus* (11.96), *Lactobacillus* (11.96), *L. jensenii* (11.52), *L. reuteri* (6.74), *L. coleohominis* (6.44).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PTBSupplementaryInformationdraftV2.3.docx