medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Factors shaping vaginal microbiota long-term community dynamics in young adult women

Tsukushi Kamiya[1,*], Nicolas Tessandier[1], Baptiste Elie[1,2], Claire Bernat[2,3], Vanina Boué[2], Sophie Grasset[2], Soraya Groc[2,4], Massilva Rahmoun[2], Christian Selinger[2,5], Michael S. Humphrys[6], Marine Bonneau[7], Christelle Graf[7], Vincent Foulongne[4], Jacques Reynes[8], Vincent Tribout[8], Michel Segondy[4], Nathalie Boulle[4], Jacques Ravel[6], Carmen Lía Murall[2,9], and Samuel Alizon[1,2,*]

[1]Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS, INSERM, Université PSL, Paris, France
[2]MIVEGEC, CNRS, IRD, Université de Montpellier, France
[3]Institut de Génomique Fonctionnelle, Université de Montpellier, CNRS, INSERM, Montpellier, France.
[4]PCCEI, Univ. Montpellier, Inserm, EFS, Montpellier, France
[5]Current address: Swiss Tropical and Public Health Institute, Basel, Switzerland
[6]Institute for Genomic Sciences, University of Baltimore, USA
[7]Department of Obstetrics and Gynaecology, Centre Hospitalier Universitaire de Montpellier, Montpellier, France
[8]Department of Infectious and Tropical Diseases, Centre Hospitalier Universitaire de Montpellier, Montpellier, France
[9]National Microbiology Laboratory (NML), Montreal Public Health Agency of Canada (PHAC), Canada
[*]Corresponding authors: tsukushi.kamiya@college-de-france.fr, samuel.alizon@college-de-france.fr

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Abstract

The vaginal microbiota is known to affect women's health. Yet, there is a notable paucity of high-resolution follow-up studies lasting several months, which would be required to interrogate the long-term dynamics and associations with demographic and behavioural covariates. Here, we present a high-resolution longitudinal cohort of 125 women followed for a median duration of 8.6 months, providing 11 samples per woman. Using a hierarchical Bayesian Markov model, we characterised the patterns of vaginal microbiota community persistence and transition, simultaneously estimated the impact of 16 covariates and quantified individual variability among women. We showed that 'optimal' (Community State Type (CST) I, II, and V) and 'sub-optimal' (CST III) communities are more stable over time than 'non-optimal' (CST IV) ones. Furthermore, we found that some covariates — most notably alcohol consumption — impacted the probability of shifting from one CST to another. We performed counterfactual simulations to confirm that alterations of key covariates, such as alcohol consumption, could shape the prevalence of different microbiota communities in the population. Finally, our analyses indicated that there is a relatively canalised pathway leading to the deterioration of vaginal microbiota communities, whereas the paths to recovery can be highly individualised among women. In addition to providing one of the first insights into vaginal microbiota dynamics over a year, our study showcases a novel application of a hierarchical Bayesian Markov model to clinical cohort data with many covariates. Our findings pave the way for an improved mechanistic understanding of

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

microbial dynamics in the vaginal environment and the development of novel preventative and therapeutic strategies to improve vaginal health.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Introduction

Epithelia of the human body are host to a diverse array of microorganisms. These microorganisms are collectively referred to as microbiota and their compositions are tightly associated with human health. In the human vaginal environment, the description of the microbiota dates back to Albert Döderlein in 1892. Its composition has been demonstrated to impact the acquisition risk of several sexually transmitted infections (STIs) [1], fertility (especially in medically-assisted procreation procedures) [2], and general well-being [3].

Vaginal microbiota communities comprise hundreds of species. To facilitate understanding, the variation in community composition is usually reduced to a handful of categories that capture key compositional signatures, such as the dominance of certain species or species evenness. This dimensionality reduction filters out noise in the data and facilitates the identification and visualisation of key patterns and relationships.

Potential drawbacks of reducing continuous variation include the risk of losing subtle but meaningful signals within the microbiota, as less dominant or rare taxa may be excluded despite their potential importance. Compared to the gut microbiota, however, vaginal microbiota communities tend to be highly structured and are often dominated by a small handful of species whose functional ecology is well-documented [4]. This contrasts with the highly diverse gut microbiota, where defining discrete community types, such as "enterotypes," remains contentious [5]. The high diversity and evenness in gut microbiota introduce continuous variations that can be oversimplified by strict categorical clustering.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

In contrast, vaginal microbiota composition aligns more naturally with categorical clustering, providing a robust understanding of key microbial patterns without significantly sacrificing interpretability.

One dimensionality reduction framework, i.e., community state types (CSTs), introduced by Ravel et al. [6], categorises vaginal microbial communities into five discrete state types. The CSTs considered 'optimal' for health are dominated by *Lactobacillus* species; *Lactobacillus crispatus*, *L. gasseri*, and *L. jensenii* for CST I, II, and V, respectively. Lactobacilli produce lactic acid and hydrogen peroxide, which create an acidic environment that helps to inhibit the growth of harmful pathogens [7]. On the other end of the spectrum, CST IV is the primary microbial context of bacterial vaginosis (BV), which elevates the risk of STI acquisition and spontaneous preterm birth, and is associated with symptoms such as malodor, discharge, and itching [4, 8]. This community is characterised by a diverse assemblage of anaerobic bacterial species from the *Gardnerella*, *Prevotella*, and *Fannyhessea* genera: recent classifications include sub-categories within CST IV (i.e., IV-A, IV-B, IV-C), each with a distinct microbial profile [9]. Finally, CST III, characterised by a dominance of *L. iners*, is considered 'sub-optimal' for women's health. While *L. iners* is a member of the *Lactobacillus* genus, it is less effective at producing lactic acid and hydrogen peroxide. As such, women with CST III tend to exhibit higher vaginal pH than those with CST I and are more prone to experiencing adverse health consequences, including vaginal infections [10].

The CST classification represents a snapshot of the microbiota community at the time of

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

sampling that facilitates the examination of clinically relevant microbiota variations across time and women. The development of the modern pipeline — through meta-barcoding sequencing of 16S DNA and clustering algorithms [9] — allows for CST-typing with enhanced efficiency and reduced observer bias compared to conventional microscopy-based methods of vaginal microbiota community typing (e.g., Nugent score).

The composition of vaginal microbiota is characteristically variable over both short and long timescales [11]. For instance, vaginal microbiota shifts throughout a woman's life, with prepubescent girls and postmenopausal women exhibiting lower levels of *Lactobacillus* dominance compared to women of reproductive age, though their bacterial communities are distinct from the CST IV typically seen during reproductive years [4]. On a short timescale, daily CST fluctuations are observed in some women of reproductive age, while others remain remarkably stable across menstrual cycles, suggesting that diverse factors influence the dynamics of vaginal microbiota communities [12]. For example, menstruation is a key driver of monthly dynamics, while clinical interventions such as antibiotics and probiotics can cause temporary perturbations [4].

A notable gap in the existing literature remains in the understanding of the long-term dynamics of vaginal microbiota in reproductive-aged women across several months. While some studies do follow this timespan, they focus on pregnancy-specific dynamics [13, 14], have large intervals between samples (often exceeding three months) [15], or involve modest sample sizes [16]. These limitations hinder our ability to fully understand the long-term patterns of CST stability and transitions in the general population of reproductive-aged

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

women, and the influence of clinically relevant factors such as demography, lifestyle, sexual practices, and medication.

In this study, we introduce an original follow-up cohort of 125 women in Montpellier, France. Our cohort presents a high-resolution longitudinal follow-up study with 2,103 microbial samples, spanning a median duration of over 8.6 months and a median of 11 samples per woman. We devise a hierarchical Bayesian Markov model to estimate transition probabilities between CSTs, associations between the transitions and 16 relevant covariates, and individual variability among women.

Materials and Methods

Longitudinal clinical data

The samples originated from the PAPCLEAR monocentric longitudinal cohort study, which followed 189 women longitudinally between 2016 and 2020. The participants were recruited through posters and leaflets circulated at the main sexually transmitted infection detection centre (CeGIDD) at the University Hospital of Montpellier (CHU) and at and around university campuses in the city. The inclusion criteria were to be between 18 and 25 years old, to be living in the area of Montpellier, France, to be in good health (no chronic disease), not to have a history of human papillomavirus (HPV) infection (e.g., genital warts or high-grade cervical lesion), and to report at least one new sexual partner over the last 12 months. Additional details about the protocol can be found elsewhere [17]. The longitudinal data

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

analysed in the present study are available at `https://doi.org/10.57745/FHQR9Z`.

The inclusion visit was performed by a gynaecologist or a midwife at the CeGIDD outside operating hours. After an interview, several samples were collected, including vaginal swabs with eSwabs (Coppan) in Amies preservation medium from which microbiota barcoding was later performed. The samples were aliquoted right after the visit and stored at -20°C, before being transferred to -70°C within a month. The participants also filled in a detailed questionnaire, which formed the basis of epidemiological covariates analysed in this study.

Subsequent on-site visits were scheduled every two or four months, depending on the HPV status. In between on-site visits, women were asked to perform eight self-samples at home with eSwabs in Amies medium and to keep them in their freezer. The self-samples were brought back in an isotherm bag at the next visit. These were then stored with the swab at -70°C until processing.

Microbiota metabarcoding and quantification

The microbiota metabarcoding was performed on 200μL of vaginal swabs specimen stored at -70° in Amies medium. The DNA extraction was performed using the MagAttract PowerMicrobiome DNA/RNA kit (Qiagen). Next-generation sequencing of the V3-V4 region of the 16S gene [18] was performed on an Illumina HiSeq 4000 platform (150 base pairs paired-end mode) at the Genomic Resource Center at the University of Maryland School of Medicine.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The taxonomic assignment was performed using the software package SpeciateIT (`https://github.com/Ravel-Laboratory/speciateIT`) and the CSTs were determined using the VALENCIA software package [9]. To examine longitudinal patterns, the present study included participants who contributed at least three samples: 125 women met the inclusion criterion, giving 2,103 samples in total.

Covariates

In the PAPCLEAR study, a questionnaire was given to each participant to record patient-level meta-data. We initially considered the following covariates based on previously proposed roles in influencing the vaginal milieu:

1st menstr. Number of years since the first menstruation: The morphology of the human vagina changes throughout life and the onset of puberty marks a key event that triggers cascading changes [19].

Alcohol Average number of glasses of alcoholic drinks consumed per week: Chronic presence of alcohol in the genital environment has been linked to a shift in the immune and microbiological conditions [20].

Antibio. Application of antibiotics during the study, either systemic (*Antibio. (Systemic)*) or genital (*Antibio. (Genital)*): The bacterial composition responds rapidly and transiently to antibiotic treatments that target bacteria either broadly or with a narrow taxonomic scale [21].

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

BMI Body mass index (BMI): Obesity has been implicated in elevating vaginal microbiota diversity and promoting *Prevotella* associated with BV [22].

Caucasian Identity as Caucasian ethnicity or other: Ethnicity has been linked to variation in vaginal microbiota compositions in several studies [6]. However, causal mechanisms remain an open question.

Cigarettes Cigarette smoking: Smoking has been implicated in the development of BV due to its anti-estrogenic effects and the presence of harmful substances such as benzo[a]pyrene diol epoxide (BPDE) [23].

Horm. contra. Use of hormonal contraception during the study: The vaginal hormonal landscape is affected by the use of hormonal contraceptives [24].

Lubricant Use of lubricant during the study: Personal lubricants contain various chemicals that differentially impact the growth of vaginal microbes in-vitro [25].

Menstr. cup Use of menstrual cups during the study: The vaginal microenvironment may be altered by the use of menstrual cups both physically and chemically. An elevated risk of fungal infections has been reported [26].

Partners Cumulative number of sexual partners: The genital microbiome can be transferred between sexual partners [27]. Such an external input could destabilise the resident community.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Red meat Average number of meals that include red meat consumption per week: Diet alters the vaginal environment for microbes. An unhealthy diet, linked to a high proportion of red meat consumption, has been linked to an elevated risk of BV [28].

Regular condom Regular use of condoms during sexual intercourse: Condom use can modify the vaginal microenvironment by altering the exchange of microbes between partners [29].

Regular sport Engaging in regular sporting activities, over 50% of the time: Physical activities influence immune responses, with leisure-time physical activity associated with a reduced risk of suspected bacterial infections compared to sedentary behaviour [30].

Stress Average stress level reported from 0 (min) to 3 (max): Stress hormones may disrupt vaginal flora, for instance, by inhibiting glycogen production, which is the primary fuel for lactobacilli [31].

Tampon Use of tampons during the study: The use of internal menstrual health products like tampons directly alters the vaginal environment, although negative effects from tampon use are seldom reported [32].

Vag. product Use of vaginal cream/tablet/capsule/gel/wipe during the study: Women frequently use over-the-counter vulvovaginal treatments that contain a variety of chemical components. However, the clinical effectiveness of these products in preventing BV is seldom systematically evaluated [33].

Chlamydia Tested positive for chlamydia.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Female/male affinity Affinity to female/male partner: Genital microbiome transfers during sexual activity are anticipated to vary based on the genders of the partners [34].

Pregnancy History of pregnancy: Pregnancy significantly changes the cervicovaginal environment, with increased estrogen from the ovaries and placenta leading to higher vaginal glycogen. This supports the growth of *Lactobacillus* species [35].

Spermicide Use of spermicide during the study: Spermicides are chemicals that prevent sperm from reaching an egg, but their use can change the vaginal microflora, potentially increasing the risk of genitourinary infections [36].

Vag. douching Use of vaginal douching during study: Vaginal douching, the practice of washing inside the vagina with a liquid solution, has been shown to increase the risk of disturbing the natural balance of vaginal flora [37].

Out of the covariates initially considered above, we excluded six (*Chlamydia*, *Female affinity*, *Male affinity*, *Pregnancy*, *Spermicide* and *Vag. douching*) as data were severely skewed towards the most common value ($> 90\%$ of data). During the study, any use of antibiotics was recorded with the date and we distinguished systemic (*Antibio. (Systemic)*) and genital topical (*Antibio. (Genital)*) applications, corresponding to 'Gynecological antiinfectives and antiseptics' ('G01' ATC codes), which consisted of metronidazole treatments, and 'Antibacterials for systemic use' ('J01' ATC codes), which were more diverse. Since the exact dates of treatment were recorded, *Antibio. (Systemic)* and *Antibio. (Genital)* were included as time-inhomogenous covariates in the model. All other covariates were

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

considered time-homogeneous meaning that the variation is among women, and static through time because the precise timing of changes in the covariate values was unknown.

To facilitate the comparison of covariate effects, we centred and scaled continuous variables [38] and deviation-coded binary variables. These transformations ensure that all covariates are modelled in a comparable scale and the intercept is located at a "representative reference value" of the modelled population: i.e., the population mean for continuous and the theoretical mid-point for binary values. Four continuous covariates (i.e., *Alcohol*, *BMI*, *Partners*, and *Red meat*) were log-transformed before scaling due to their right-skewed distribution. We found no strong correlations among the covariates included in the analysis (Supplementary Information S1).

Modelling

Markov model

Markov models are statistical models used to represent systems that transition between discrete states over time. These models are 'memoryless', meaning that the probability of transition to another state depends on the current state, but not its historical path. In clinical research, these models are often used to predict the transitions among health states (e.g., health, illness and remission), and the propensity to transition between these states is estimated from longitudinal follow-up data. Clinical follow-up data are typically modelled using the continuous-time Markov model [39], in which the probability of transition over

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

a given interval depends on the instantaneous transition intensity and the amount of time spent in the current state.

Vaginal microbiota state transitions are classically studied using continuous-time Markov models [13–15, 40, 41]. Our application of the continuous-time Markov model differs from those of the existing literature in its hierarchical Bayesian formulation, which allowed us to quantify individual variability among women (as unobserved heterogeneity, or random effects) and to estimate many covariate effects simultaneously (through the use of weakly informative priors).

Transition intensities

Transition intensities, q, refer to the instantaneous rate of moving from state i to state j in a participant p (e.g., CST I to CST IV), a process that may be affected by a vector of covariates, X. Taking the form of a proportional hazards model, these rates can be expressed as:

$$q_{p,i,j} = \mathrm{Exp}(\mu_{p,i,j} + \beta_{i,j}\ X), \tag{1}$$

where $\mu_{p,i,j}$ is the intercept and $\beta_{i,j}$ is the coefficient expressing the impact of a covariate(s). This intercept is further defined by the equation,

$$\mu_{p,i,j} = (\hat{\mu}_{i,j} + s_{p,i,j}) \cdot \mu_{sd} + \bar{\mu}, \tag{2}$$

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

where $\bar{\mu}$ and μ_{sd} are the prior mean and standard deviation of the intercept such that $\hat{\mu}_{i,j} \cdot \mu_{sd} + \bar{\mu}$ constitutes the non-centred parameterisation of the population-level intercept, $\mu_{i,j}$ and is assumed to be normally distributed, i.e., $\hat{\mu}_{i,j} \sim \mathcal{N}(0, 1)$.

Additionally, we allowed for unobserved heterogeneity in μ, i.e., $s_{p,i,j}$, where

$$s = \text{diag}(sd_s) \cdot L_s \cdot z_s. \tag{3}$$

We sampled from the corresponding weakly informative priors, namely $sd_s \sim t_4(0, 1)$, $L_s \sim \text{LKJCorrCholesky}(2)$ (which slightly favours correlations among unobserved heterogeneity closer to zero, reducing the likelihood of extreme positive or negative correlations), and $z_s \sim \mathcal{N}(0, 1)$, as recommended by the Stan development community [42, 43]. The multivariate normal density and the LKJ prior require the matrix parameters to be decomposed, which can be computationally intensive if done repeatedly. To ensure computational efficiency and numerical stability, the model was directly parameterised using the Cholesky factors of correlation matrices. This approach uses a multivariate version of the non-centred parameterisation.

For regression coefficients, the Student-t distributions with degrees of freedom 4 to 7 are recommended as generic, weakly informative, priors [43]: we sampled β from $\beta \sim t_4(0, 1)$, which places a comparatively wide tail within the recommendation. As all of our covariates have been proposed to impact vaginal microbiota communities *a priori* (see above), we did not strongly regularise the priors, for example, through the use of horseshoe priors [44].

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

We note that all covariates were modelled simultaneously, such that the interpretation of each coefficient is conditional upon other covariates included and accounts for the influence of other factors. We assumed that the covariates affect the transitions symmetrically (i.e., $\beta_{j,i} = -\beta_{i,j}$), meaning that the influence of a covariate on the affinity (or aversion) towards a particular CST is consistent, regardless of the direction of the transition.

Collectively, the transition intensities form the matrix, Q_p, in which the sum of intensities across a row, i.e., all transitions from a particular state, is defined to be zero, such that we have the following equation for the diagonal entries [39]:

$$q_{p,i,i} = -\sum_{j \neq i} q_{p,i,j}. \tag{4}$$

Transition probabilities and likelihood

Taking the matrix exponential of the Q_p matrix for each participant, p, we compute the matrix P_p such that:

$$P_p = \mathrm{Exp}\big((t_{k+1} - t_k)\ Q_p\big), \tag{5}$$

where k represents the sample identity for a given individual. The P_p matrix contains the transition probabilities between two observations (at k and $k+1$) and $t_{k+1} - t_k$ indicates the elapsed time between two observations.

Finally, the probability of observing a given state at the next sampling event (i.e., at

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

$k+1$) is modelled by the categorical distribution:

$$y_{k+1} \sim \text{Categorical}\big(P_p[y_k,]\big) \tag{6}$$

where $P_p[y_k,]$ is the ${y_k}^{\text{th}}$ row of the P_p matrix containing the probabilities of transition from the state observed at k.

Model fitting

We used a Bayesian approach to fit the above continuous-time Markov model to longitudinal data of vaginal microbiota CSTs. In total, the model consisted of 57 parameters and 12 hyper-parameters. Our model was written in Stan 2.26.1 and fitted through the RStan interface 2.32.3 [45]. The Stan programme is available at `https://doi.org/10.57745/FHQR9Z`.

One participant lacked information on the years since their initial menstruation. We imputed missing values using the mice package [46] and generated 10 imputed datasets to be fitted separately. For each imputed dataset, we fitted the model in parallel using four independent chains, each with $10,000$ sampled iterations and $1,000$ warm-up iterations. The MCMC samples from separate runs (i.e., based on differently imputed data) were subsequently combined for inference.

We confirmed over $1,000$ effective samples per imputed dataset and ensured convergence of independent chains ($\hat{R} < 1.01$) for all parameters [47]. We carried out a posterior predictive check by comparing the observed and predicted CST frequency. We also quanti-

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

fied the posterior z-score and posterior contraction to examine the accuracy and precision of posterior distributions and the relative strength of data to prior information [48] (Supplementary Information S2).

Counterfactual predictions

We took advantage of the parameterised model to simulate the population-level outcomes of each covariate, assuming that all covariates, but a focal one, are at the representative reference value (as described above) and then varying the focal parameter within the range of values observed in the studied cohort. The model predictions were generated by randomly drawing 100 samples from the posterior distributions and simulating the Markov model for each sampled parameter set. We focused on the CST frequency as the outcome of interest.

Results and Discussion

CSTs in the cohort

As is typical of vaginal microbiota communities, the microbial compositions sampled in PAPCLEAR were highly structured, and characterised by a relatively small number of operational taxonomic units (OTUs). The dominant species within these communities aligned closely with specific community state types (CSTs) as defined by Ravel et al. [6]. For example, CST I was primarily associated with *L. crispatus* and CST III with *L. iners*.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

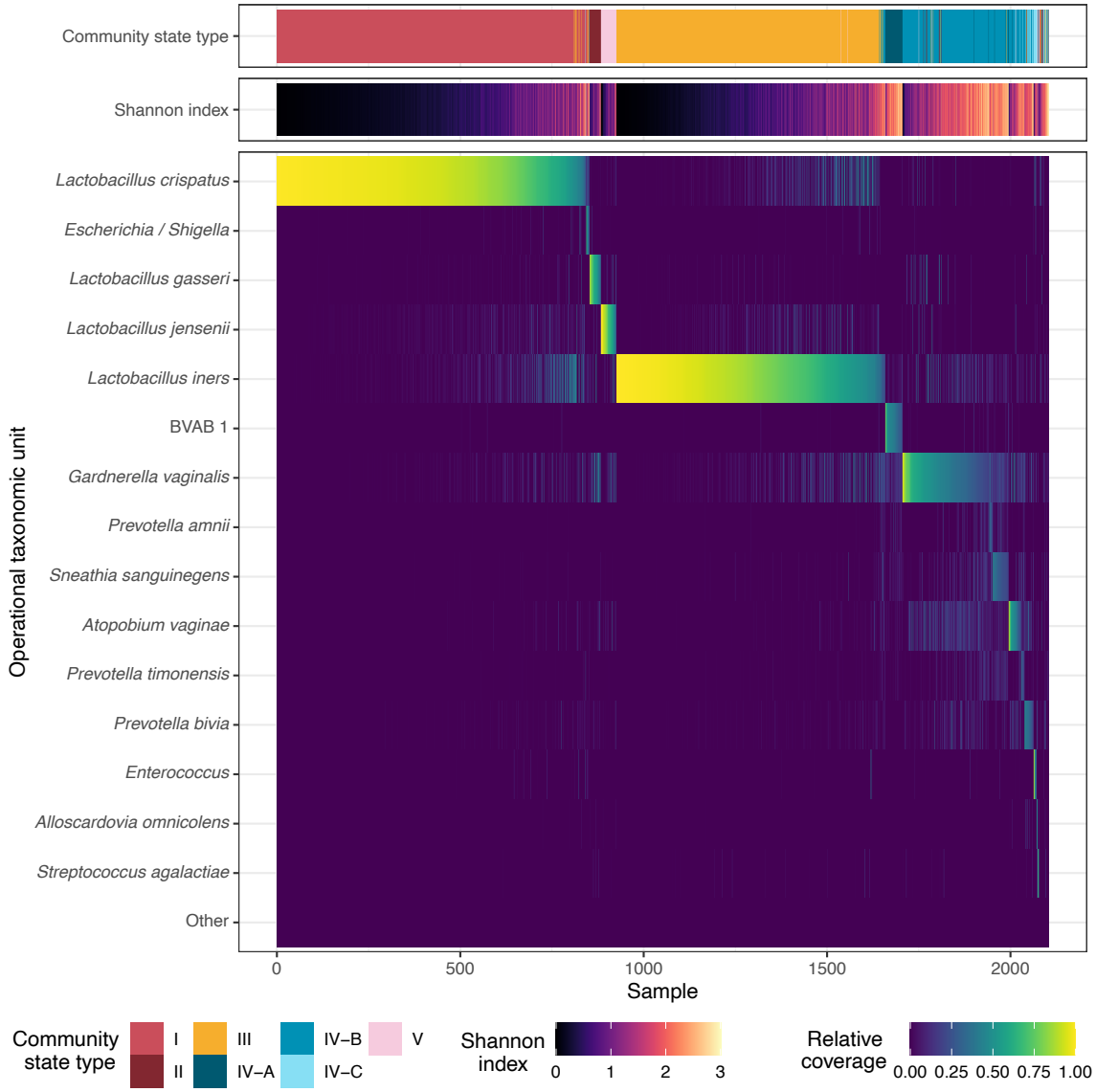

Figure 1: **Vaginal community state types (CSTs), diversity (Shannon Index), and relative coverage of the 15 most common taxonomic operational units (OTU) of 2,103 samples from the PAPCEAR cohort.** In over 98.5% of samples, a single of these 15 OTUs represented the most common OTU.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

In contrast, and as expected, CST IV communities exhibited a higher degree of microbial diversity compared to CSTs dominated by lactobacilli, reflecting a broader range of species typical of this community type (Fig. 1).

Our longitudinal dataset from the PAPCLEAR cohort represents one of the largest analysed to date in the context of the vaginal microbiota. Detailed participant characteristics are presented in Table 1. Briefly, the participants were between 18 and 25 years old and the majority of the 2,103 samples (73.7%) were self-collected at home, the rest being collected during on-site visits (Fig. 2a). The median follow-up duration was 8.64 months and the most common intervals between analysed samples were seven and 28 days (Fig. 2a & b). On average, each of the 125 participants contributed 11 samples (Fig. 2c).

The metabarcoding analysis on 16S RNA with the VALENCIA algorithm [9] was used to assign each sample to a CST. The vaginal microbiota communities were variable across women and over time (Fig. 2d). As CSTs I, II, and V are all dominated by lactobacilli and considered 'optimal' in terms of health, yet the latter two are rare (∼4% of all samples combined), we pooled the three optimal communities for further investigation. Overall, optimal communities were the most frequent, representing 44.5% of samples, followed by 'sub-optimal' (CST III) at 35.2% and 'non-optimal' communities (CST IV) at 20.4% (Fig. 2e and Table 1).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

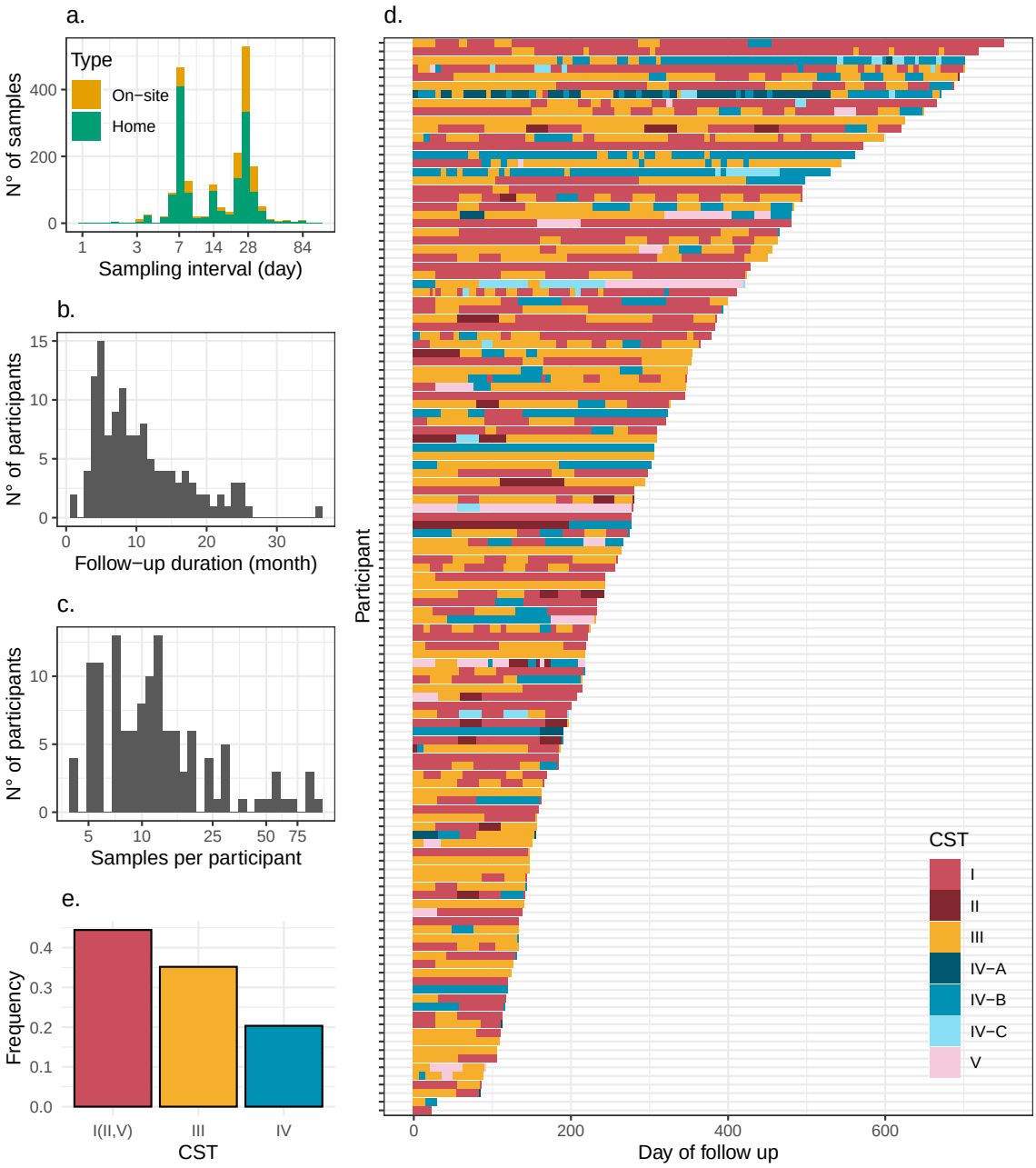

Figure 2: **Summary of vaginal microbiota samples analysed in the PAPCLEAR study.** a) Intervals between sampling events for clinical (i.e., on-site) and home samples. b) Follow-up duration per participant. c) Number of samples analysed per participant. d) Vaginal microbiota Community State Types (CST) over time in 125 participants. For visualisation, data are truncated at 750 days for a single individual whose duration exceeds this threshold. e) Frequency of the optimal (i.e., CSTs I, II, and V combined), sub-optimal (CST III) and non-optimal (CST IV) communities in all samples.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Probabilities of CST persistence

We implemented a continuous-time Markov model to capture the CST dynamics. Simulations based on the estimated parameters of our model (i.e., posterior predictive check) confirmed that it accurately captures the observed CST prevalence (Fig. 3a). The optimal,

Table 1: Summary profile of vaginal microbiota samples and covariates in the PAPCLEAR study. Q1 and Q3 refer to first (25%) and third (75%) quantiles. Level = 1 indicates the presence of a binary condition. See Materials and Methods for the covariate definitions.

	Level	Summary
Samples (Participants)		2103 (125)
CST (%)	I	847 (40.3)
	II	39 (1.85)
	III	740 (35.2)
	IV-A	54 (2.57)
	IV-B	342 (16.3)
	IV-C	32 (1.52)
	V	49 (2.33)
Sample type (%)	On-site	553 (26.3)
	Home	1550 (73.7)
Sampling interval (median (Q1,Q3))		21 (7, 28)
Follow-up duration (median (Q1,Q3))		8.64 (5.36, 14.0)
Samples per subject (median (Q1,Q3))		11 (7, 16)
Covariates		
Identifying as 'Caucasian' (%)	1	102 (81.6)
BMI (median (Q1,Q3))		21.19 (19.78, 23.46)
Alcohol (median (Q1,Q3))		3.14 (1.40, 5.07)
Smoker (%)	1	36 (28.8)
Stress level (from 0 to 3, median (Q1,Q3))		1.41 (1.00, 1.75)
Regular sport practice (%)	1	61 (48.8)
Red meat consumption (times per week, median (Q1,Q3))		0.50 (0.16, 1.00)
Years since 1st menstruation (median (Q1,Q3))		9 (7, 10)
Hormonal contraception (%)	1	32 (25.6)
Menstrual cup user (%)	1	46 (36.8)
Vaginal product user (%)	1	73 (58.4)
Tampon user (%)	1	89 (71.2)
Lifetime number of partners (median (Q1,Q3))		5 (3, 11)
Lubricant use (%)	1	58 (46.4)
Regular condom use by partner (%)	1	23 (18.4)
Male affinity (%)	1	124 (99.2)
Chlamydia infection at inclusion (%)	1	7 (5.6)
Pregnancy during follow-up (%)	1	4 (3.2)
Vaginal douching (%)	1	4 (3.2)
Spermicide user (%)	1	1 (0.8)
Female affinity (%)	1	10 (8.0)
Systemic antibiotic treatment (%)	1	65 (52.0)
Genital antibiotic treatment (%)	1	30 (24.0)

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

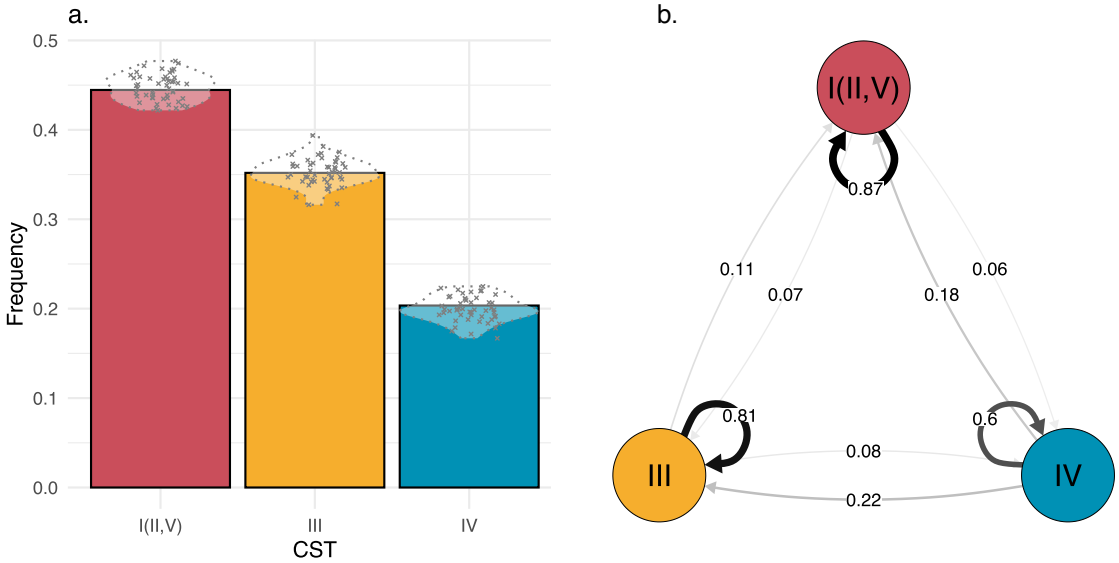

Figure 3: **Prevalence and transition probabilities among vaginal microbiota community state types (CSTs).** a) Observed (bars) and predicted prevalence (crosses) of CSTs I (II, V), III and IV. The model predictions were generated by drawing 100 random samples from the posterior distributions and simulating the Markov model for each sampled parameter set. b) Mean estimated weekly transition probabilities of CSTs I (II, V), III and IV. The arrow thickness indicates the persistence or transition probability.

CST I (II, V), and sub-optimal, CST III, communities showed a high degree of stability, with weekly probabilities to remain in the current state estimated at 87% (95% credibility interval (95CrI): 78 - 93%) and 81% (95CrI: 68 - 90%), respectively (Fig. 3b). In contrast, the weekly persistence probability of the non-optimal CST IV was 60% (95CrI: 35 - 80%, Fig. 3b). These transition probabilities translate into sojourn times (i.e., the expected time spent in a given state before moving to another) in CST I (II, V), III and IV of 6.9 days (95CrI: 2.9 - 13.6 days), 4.23 days (95CrI: 1.8 - 8.4 days) and 1.6 days (95CrI: 0.58 - 3.8 days), respectively.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

The reported persistence and transition probabilities in the literature vary widely based on the cohort characteristics. For example, focusing on women during pregnancy, DiGiulio et al. [13] estimated that the four *Lactobacillus*-dominated CSTs (CSTs I, II, III, and V) were more stable than CST IV. Notably, both CST I and II showed 98% probability of weekly persistence. The enhanced persistence of *Lactobacillus*-dominated communities during pregnancy owes itself to specific vaginal conditions during pregnancy including the up-regulation of oestrogen and progesterone that facilitates lactobacilli [13, 49].

In addition, the temporal dynamics of vaginal microbiota are notably different in women with BV. In contrast to pregnant women, those experiencing symptomatic BV generally exhibit less stable vaginal microbiota communities. In the cohort of Ravel et al. [50], which focused on women with symptomatic BV, Brooks et al. [40] found significantly lower stability across all CSTs. The probability of these CSTs persisting ranged from 38% to 48%, with CST I persisting only 46% of the time over a week.

Among studies that focused on non-pregnant, healthy young women — with no particular emphasis on BV — the analysis by Brooks et al. [40] of the Chaban et al. cohort [16] (N = 27; Canada) estimated weekly persistence probabilities of 75% for CST I, 78% for III, 60% for IV-A, and 88% for V. In the Gajer et al. dataset [12] (N = 32; USA), analysed again by Brooks et al., [40], CST I, II and III demonstrated 72%, 84% and 77% weekly persistence probabilities, respectively. In this dataset, CST IV sub-categories showed markedly different stability with CST IV-A with weekly persistence of 38% and CST IV-B with persistence of 82%. A third study, Munoz et al. [15] (N = 88; South Africa), reported

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

the stability of vaginal microbiota in women in a three-month time frame using a different microbiota classification system consisting of four categories predominantly associated with: *L. crispatus* (similar to CST I), *L. iners* (similar to CST III), *G. vaginalis* (similar to CST IV), or *Prevotella* spp. (similar to CST IV). They found similar persistence for CST I and CST IV-like communities ranging from 51 to 53% over three months while the CST III-like community was more stable at 62% over the same period. Recasting in the three-month time scale, our estimates show the same extent of stability for CST I(II, V) at 51% (95% CrI: 29-72%) while CST III (38%, 95% CrI: 19-61%) and CST IV (15%, 95% CrI: 5-34%) were less stable. Taken together, our estimates of vaginal microbiota community stability are within the range of values reported in other cohorts. However, the dynamics of vaginal microbiota communities are likely geographically variable even among healthy young women.

Covariate effects on transitions

The Bayesian approach, which can accommodate vaguely informative priors on the covariate effects, allows for the simultaneous inclusion of many covariates (as hazard ratios; Eq. 1) which would otherwise prove difficult in Markov models [39]. We identified 16 covariates based on previously proposed roles in influencing the vaginal milieu and assumed that covariates have a symmetrical effect on CST transitions: e.g., the magnitude of a given covariate effect on the transition from CST I to III is identical to that on the transition from CST III to I. We identified alcohol consumption as the strongest and most consistent

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

effect while several other covariates were identified as possible drivers of CST transitions.

Alcohol consumption

The estimated hazard ratios on community transitions indicate that alcohol consumption favoured the sub-optimal (CST III) community over optimal (CST I(II, V)) with 97% probability (Fig. 4). Because of our symmetry assumption, this can mean that alcohol consumption increases the pace of transition from CST I(II, V) to CST III or reduces that in the opposite direction by the same magnitude. Alcohol consumption also tended to favour CST IV over CST III, although with a lower credibility level (with 73% probability of the hazard ratio $\neq 1$, Fig. 4).

To examine how these effects translate to the population level, we carried out counterfactual simulations in which all participant characteristics were set to the representative value observed in the studied cohort, except for alcohol consumption, which ranged from non-drinking to the level of the heaviest drinking observed in our cohort (19 drinks per week). The simulations demonstrated that the expected prevalence of the optimal (CST I (II, V)) community was 18% (95% CrI of 9 to 27%) higher in a hypothetical population of non-drinkers compared to that of average-level drinkers who consumed three drinks per week (Fig. 5; Supplementary Information S3). In turn, the prevalence of the optimal community was 19% (95% CrI of 10 to 29%) higher in the population of average-level drinkers than in the heaviest drinkers. As the optimal community declined with alcohol consumption, the prevalence of the non-optimal (CST IV) community was found to be 9% (95% CrI

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

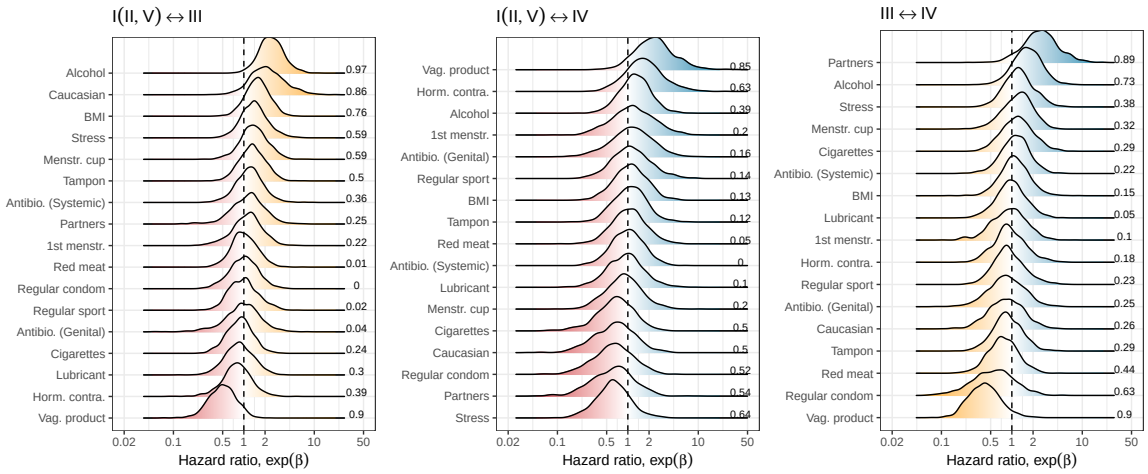

Figure 4: **Estimated covariate effects on community transition rates.** With the symmetry assumption, there are three sets of covariate effects on transitions. The impact of covariates on community transition rates was estimated for a given set of community states as the log hazard ratio, β. The figure shows the posterior distributions of $\exp(\beta)$, the hazard ratio for the three sets of transition sets, and the corresponding 16 covariates. The numbers on the right-hand side of each panel indicate the probability that the estimated effect is different from the hazard ratio of 1 (i.e., the proportion of posterior distributions sampled on the dominant side of the effect). For example, alcohol consumption was estimated to favour CST III over CST I (II,V) at a credibility level of 97%.

of 2 to 15%) higher among average drinkers compared to non-drinkers. Therefore, while the strongest impact of alcohol on community transitions appears to be between the optimal (CST I (II, V)) and sub-optimal (CST III) communities, an additional, non-zero impact on the sub-optimal to non-optimal (CST IV) transition means that alcohol consumption ultimately promotes non-optimal communities at the expense of optimal ones. As the effects of covariates are estimated simultaneously, potential confounding factors, including the number of partners, condom use and smoking, are controlled for in our findings.

Alcohol consumption may impact the vaginal microbiota through a variety of mecha-

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

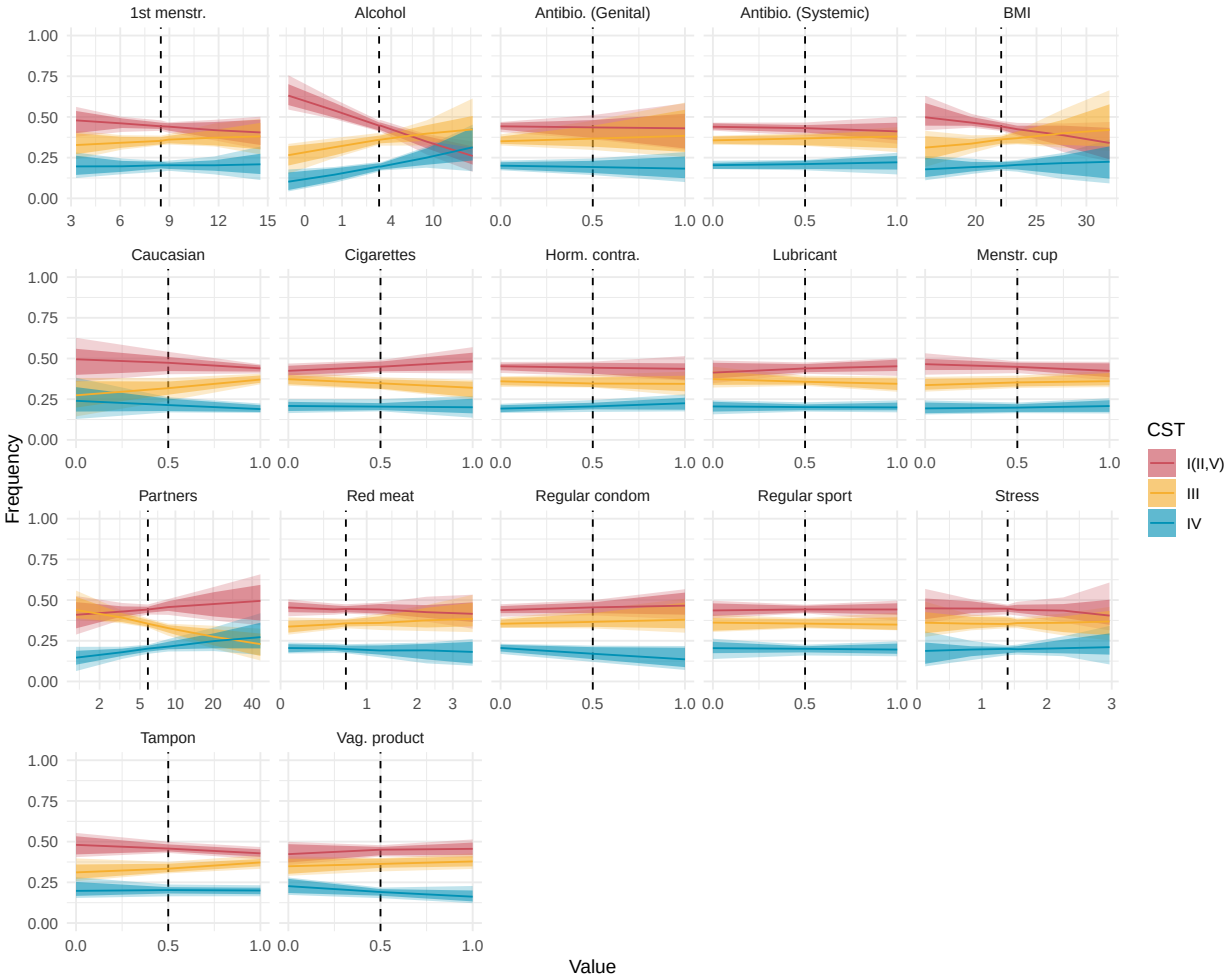

Figure 5: **Counter-factual simulations predict population-level consequences of covariates.** Based on estimated hazard ratios (Fig. 4), the population-level impact (i.e., the prevalence of CST I (II,V), III and IV) was simulated for each covariate. The vertical dashed lines indicate the intercept used in estimation: i.e., the population mean for continuous and midpoint for binary variables. For continuous variables, the range of values explored was determined by the minimum and maximum values reported in the PAP-CLEAR study.

nisms. Physiologically, the chronic presence of alcohol in the genital environment has been linked to a shift in immune and microbiological conditions [20]. In addition, alcohol is a

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

known modifier of sexual behaviour, which in turn has been demonstrated to increase the risk of BV, linked to CST IV [51]. Finally, alcohol alters the microbial profile in other body parts, which in turn could cross over to the vaginal milieu. For example, *Prevotella*, a genus commonly found in CST IV communities, is enriched in the oral microbiota of drinkers [52]. Similarly, others postulate the effect of alcohol on the gut microbiota may have a concurrent influence on the vaginal microbiota [53].

While there remains a lack of consensus among existing studies (briefly reviewed by Froehle et al. [53]), cohort and cross-sectional studies from diverse geographical contexts (namely, Australia, Denmark, Sweden, Thailand, Tanzania, Uganda and USA) have previously reported an association between alcohol consumption and BV [53–61]. In addition to corroborating these findings, our Markov model offers a novel insight into the ecology of microbial communities underlying these observations: alcohol consumption destabilises the optimal (CST I (II, V)) communities towards sub-optimal (CST III), which opens the gate for the deterioration towards non-optimal (CST IV), associated with BV. To the authors' knowledge, there have been no alcohol cessation studies reporting its impact on vaginal microbiota. Such studies are necessary to establish causal links, similar to those conducted on the effects of smoking [23], douching [62], and antibiotics [21] on vaginal microbiota compositions.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Potential effects of other covariates

Other factors with possible effects on transitions (i.e., with more than 80% probability of hazard ratio $\neq 1$) included the use of vaginal intimate hygiene products, number of sexual partners and self-reported 'Caucasian' identity.

Vaginal hygiene products: The use of vaginal hygiene products, defined broadly here to include vaginal cream, tablet, capsule, gel and wipe, appeared to have multifaceted effects. Between CST I (II, V) and CST III, their use was positively linked to maintaining or transitioning to CST I (II, V) with 90% probability (Fig. 4). For the CST I (II, V) and CST IV pair, it tended to favour a shift towards CST IV, with 85% probability. Finally, between CST III and CST IV, their use was more likely to support the persistence or a shift towards CST III, also with 90% probability. The circular effects suggest that women may experience different effects of the products marketed for 'vaginal intimate hygiene' depending on the predisposition with certain CSTs. Nonetheless, the circular effects on community transitions meant that there was no noticeable impact at the population level in our counterfactual simulations (Fig. 5).

Number of sexual partners: A higher number of sexual partners was also found to potentially favour CST IV over CST III, increasing the risk of maintaining (or transitioning to) CST IV with 89% probability of the hazard ratio $\neq 1$. The association between CST IV and the lifetime number of partners is consistent with the hypothesis that external importation of microbes could alter the dynamics of vaginal microbiota and is in line with

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

earlier work [63, 64]. Population-level simulations predict that an increasing number of sexual partners tends to reduce the prevalence of the sub-optimal (CST III) community. For example, CST III was 13% (95% CrI of 2 to 21%) less common in a hypothetical population with the highest number of partners than one conforming to the average number. The decrease was accompanied by a tendency for the other CSTs to increase, although the trend was less clear for CST I(II,V) and CST IV, individually (Fig. 5).

Causasian identity: It is worth noting that our cohort was not designed to achieve comprehensive coverage of self-reported ethnic identity, with over 80% identifying as Caucasians (Table 1). Nonetheless, identifying oneself as a 'Caucasian' tended to favour CST III over CST I(II, V) with 86% probability. European studies focusing on the role of ethnicity are rare. However, a North American study has observed a qualitatively opposite trend: CST III communities are comparably rare in women who identify themselves as Caucasian compared to those identifying as Asian, Black and Hispanic (26.8 versus 42.7, 31.4 and 36.1%, respectively [6]). While previous studies have revealed differences in vaginal microbiota compositions among ethnic groups, the relative importance of biological, societal, and environmental factors remains an open question [6, 65–67].

Antibiotics: Notably, we found little association between antibiotic consumption and CST transitions, neither for local treatment for BV (genital application of metronidazole) nor systemic treatment (antibiotic treatment via oral intake). Such a lack of effect in our study may be because the changes in the vaginal microbiota compositions following an antibiotic treatment take place in a shorter time scale than our sampling intervals: the most

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

common sampling intervals were either 7 or 28 days (Table 1). In comparison, Brooks et al. [40] found rapid CST transition following BV medication in the cohort of Ravel et al. [50], which involved daily sampling. On a longer time scale, the re-emergence of BV-associated communities following treatments is a well-documented clinical challenge [68–70].

Unobserved individual variability in community transition

While we incorporated 16 covariates into our Markov model, some variations among women remain unaccounted for. To quantify these, we estimated the extent of individual variability (i.e., unobserved heterogeneity, or random effects) in community transitions for each transition pair using a hierarchical Bayesian approach (Eq.2 & 3).

The highest variability was observed among women in the transitions involving 'recovery' to an optimal (CST I (II,V)) from a non-optimal (CST IV) state (Fig. 6). On the other hand, inverse transitions from optimal to non-optimal exhibited some of the lowest individual variability. The same is true, although to a lesser extent, for the shifts from sub-optimal (CST III) to optimal. These findings suggest that there are relatively limited pathways leading to the deterioration of vaginal microbiota communities, whereas the routes to recovery can be more individualised and the source of this variation remains to be fully elucidated.

The presence of individual-level random effects indicates that a considerable part of the variability remains unaccounted for by the 16 covariates in this study. One possible cause is that our study left out key drivers of the vaginal milieu. For example, while menstrual cycles

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

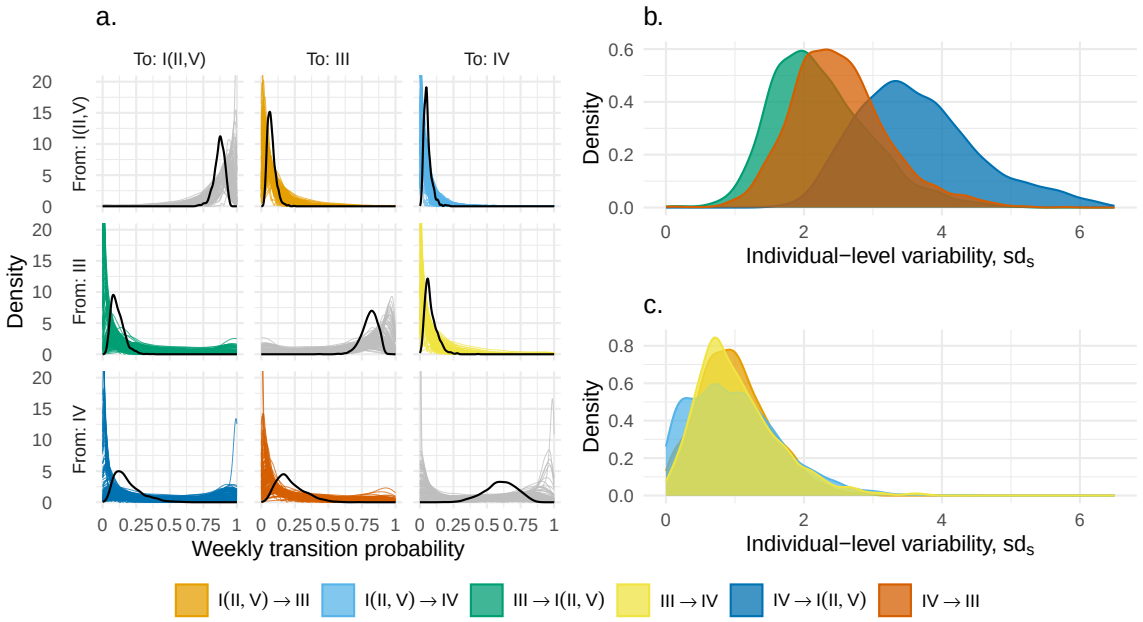

Figure 6: **Individual-level variability in vaginal microbiota community state type (CST) transitions.** a) The population average (thick black) and individual (thinner colours) weekly transition probabilities. Between-women individual variation for transitions to b) a more optimal and c) a less optimal state. Colours indicate the type of transition between CSTs.

have been demonstrated to influence daily and weekly transitions [12], they were omitted from our analysis because the timing of menstruation was ambiguous in the PAPCLEAR study. Furthermore, while large-scale longitudinal studies present logistical challenges, a citizen-science-based approach offers the potential for expanding the cohort size, thereby enhancing the statistical power needed to examine additional covariates [71]. Secondly, further resolution on individual variability may be gained by incorporating time-varying covariates, which could accommodate changes in participant behaviours during the follow-up. In continuous-time Markov models, time-varying covariates are assumed piecewise

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

constant, meaning they are constant between sampling events [39]. Such an assumption is convenient as covariate values are rarely known between sampling events. Without precise knowledge of the timing of the covariate changes, however, it is unclear whether the previous covariate value (at $t-1$) or the new covariate value (at t) should influence the transition. Consequently, our analysis focused on static covariates, with the exception of antibiotic treatments for which the exact application dates were known. Aggregating participant behaviours as static covariates eliminated the uncertainty of covariate dynamics, albeit at a potentially lost opportunity for further precision.

Limitations and opportunities

A potential limitation from a clinical methodological perspective is that the majority of samples were collected at home during the PAPCLEAR study. While home sampling could introduce variability, the participants were provided with detailed instructions to minimise the difference in swabbing techniques between on-site and home samples, and we verified consistency in sampling dates by having participants fill out online questionnaires during sampling.

Another possible limitation of our study is the resolution of microbiota community classification. We focused on three CST groups with varying health implications: optimal (CST I (II,V)), sub-optimal (III), and non-optimal (IV). This decision stemmed from the fact that detailed classifications in a Markov model would increase the number of possible transitions, and it would be difficult to estimate transitions between rare types. However,

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

significant functional differences may exist within these CSTs. For instance, the VALENCIA algorithm classifies subcategories within some CSTs [9], and Brooks et al. demonstrated that CST IV-B is more stable than CST IV-A [40]. We also note that there are several clustering algorithms of microbial communities besides the CST framework [71,72], which may offer differing insights on community transitions. Furthermore, the centroid distance computed by VALENCIA for CST assignment may also be leveraged to develop a quantitative, multi-dimensional perspective of the vaginal microbiota communities. Such a quantitive perspective may enhance our understanding of within-CST variabilities — although we are unaware of an existing approach that accommodates the temporal patterns in such data. Finally, the metagenomics approach holds the promise to uncover within-species diversities: e.g., metagenomics CSTs (MgCSTs) have identified with 25 distinct communities [73]. Such an approach helps to identify lineage replacements in women with stable CSTs and investigating the impact of antibiotic treatments on the prevalence of resistance genes could yield insights into the within-species dynamics of vaginal microbes.

A promising direction for future research is the joint analysis of CST dynamics and sexually transmitted infections such as HPV. Previous studies have found a weak association between CST IV and HPV detection risk [74]. However, these studies tested the CST effect after estimating transition rates and pooled all high-risk and low-risk HPVs, making it difficult to identify coinfections or reinfections. The PAPCLEAR cohort, with genotype-specific follow-ups, could provide new insights into the link between CST and HPV infection, potentially identifying causal relationships.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Conclusion

We showcased a novel application of a hierarchical Bayesian Markov model to original clinical cohort data of vaginal microbiota dynamics. Our approach facilitated the simultaneous estimation of several covariate effects on community transitions and the identification of unobserved variability in these transitions. Our work paves the way for an improved ecological understanding of microbial dynamics within the vaginal environment and indicates lifestyle alterations (such as reduced alcohol consumption) that may promote vaginal health.

Ethics

This study has been approved by the Comité de Protection des Personnes (CPP) Sud Méditerranée I (reference number 2016-A00712-49); by the Comité Consultatif sur le Traitement de l'Information en matière de Recherche dans le domaine de la Santé (reference number 16.504); by the Commission Nationale Informatique et Libertés (reference number MMS/ABD/ AR1612278, decision number DR-2016–488), by the Agence Nationale de Sécurité du Médicament et des Produits de Santé (reference 20160072000007), and is registered at ClinicalTrials.gov under the ID NCT02946346.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Funding

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 648963, to SA). The authors acknowledge further support from the Fondation pour la Recherche Medicale (to TK), the Agence Nationale de la Recherche contre le SIDA (ANRS-MIE, to NT), and the MemoLife Labex (to BE).

Acknowledgements

The authors thank Olivier Supplisson for his helpful feedback. We acknowledge the ISO 9001 certified IRD i-Trop HPC (member of the South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported within this article (bioinfo.ird.fr and www.southgreen.fr). DNA extracts were (partly) performed through the genotyping and sequencing facilities of ISEM (Institut des Sciences de l'Evolution-Montpellier) and Labex Centre Méditerranéen Environnement Biodiversité.

References

[1] van de Wijgert JHHM. The vaginal microbiome and sexually transmitted infections are interlinked: Consequences for treatment and prevention. PLOS Medicine. 2017;14(12):e1002478.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[2] Haahr T, Zacho J, Bräuner M, Shathmigha K, Skov Jensen J, Humaidan P. Reproductive outcome of patients undergoing in vitro fertilisation treatment and diagnosed with bacterial vaginosis or abnormal vaginal microbiota: a systematic PRISMA review and meta-analysis. BJOG. 2019;126(2):200–207.

[3] Bilardi JE, Walker S, Temple-Smith M, McNair R, Mooney-Somers J, Bellhouse C, et al. The burden of bacterial vaginosis: women's experience of the physical, emotional, sexual and social impact of living with recurrent bacterial vaginosis. PLOS ONE. 2013;8(9):e74378.

[4] France M, Alizadeh M, Brown S, Ma B, Ravel J. Towards a deeper understanding of the vaginal microbiota. Nature Microbiology. 2022;7(3):367–378.

[5] Cheng M, Ning K. Stereotypes about enterotype: the old and new ideas. Genomics, Proteomics and Bioinformatics. 2019;17(1):4–12.

[6] Ravel J, Gajer P, Abdo Z, Schneider GM, Koenig SS, McCulle SL, et al. Vaginal microbiome of reproductive-age women. Proceedings of the National Academy of Sciences. 2011;108:4680–4687.

[7] France MT, Fu L, Rutt L, Yang H, Humphrys MS, Narina S, et al. Insight into the ecology of vaginal bacteria through integrative analyses of metagenomic and metatranscriptomic data. Genome Biology. 2022;23(1):66.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[8] Coudray MS, Madhivanan P. Bacterial vaginosis—A brief synopsis of the literature. European Journal of Obstetrics & Gynecology and Reproductive Biology. 2020 Feb;245:143–148.

[9] France MT, Ma B, Gajer P, Brown S, Humphrys MS, Holm JB, et al. VALENCIA: a nearest centroid classification method for vaginal microbial communities based on composition. Microbiome. 2020;8(1):166.

[10] Petrova MI, Reid G, Vaneechoutte M, Lebeer S. *Lactobacillus iners*: Friend or Foe? Trends in Microbiology. 2017;25(3):182–191.

[11] Cancelo-Hidalgo MJ, Coello LB. Genitourinary syndrome of the menopause: vaginal health and microbiota. In: Cano A, editor. Menopause: a comprehensive approach. Cham; 2017. p. 91–107.

[12] Gajer P, Brotman RM, Bai G, Sakamoto J, Schütte UM, Zhong X, et al. Temporal dynamics of the human vaginal microbiota. Science Translational Medicine. 2012;4(132):132ra52–132ra52.

[13] DiGiulio DB, Callahan BJ, McMurdie PJ, Costello EK, Lyell DJ, Robaczewska A, et al. Temporal and spatial variation of the human microbiota during pregnancy. Proceedings of the National Academy of Sciences. 2015;112(35):11060–11065.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[14] Serrano MG, Parikh HI, Brooks JP, Edwards DJ, Arodz TJ, Edupuganti L, et al. Racioethnic diversity in the dynamics of the vaginal microbiome during pregnancy. Nature Medicine. 2019;25(6):1001–1011.

[15] Munoz A, Hayward MR, Bloom SM, Rocafort M, Ngcapu S, Mafunda NA, et al. Modeling the temporal dynamics of cervicovaginal microbiota identifies targets that may promote reproductive health. Microbiome. 2021;9(1):1–12.

[16] Chaban B, Links MG, Jayaprakash TP, Wagner EC, Bourque DK, Lohn Z, et al. Characterization of the vaginal microbiota of healthy Canadian women through the menstrual cycle. Microbiome. 2014;2:1–12.

[17] Murall CL, Rahmoun M, Selinger C, Baldellou M, Bernat C, Bonneau M, et al. Natural history, dynamics, and ecology of human papillomaviruses in genital infections of young women: protocol of the PAPCLEAR cohort study. BMJ Open. 2019;9(6):e025129.

[18] Frank JA, Reich CI, Sharma S, Weisbaum JS, Wilson BA, Olsen GJ. Critical evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. Applied and Environmental Microbiology. 2008;74(8):2461.

[19] Farage M, Maibach H. Lifetime changes in the vulva and vagina. Archives of Gynecology and Obstetrics. 2006;273:195–202.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[20] Loganantharaj N, Nichols WA, Bagby GJ, Volaufova J, Dufour J, Martin DH, et al. The effects of chronic binge alcohol on the genital microenvironment of simian immunodeficiency virus-infected female rhesus macaques. AIDS Research and Human Retroviruses. 2014;30(8):783–791.

[21] Mayer BT, Srinivasan S, Fiedler TL, Marrazzo JM, Fredricks DN, Schiffer JT. Rapid and profound shifts in the vaginal microbiota following antibiotic treatment for bacterial vaginosis. Journal of Infectious Diseases. 2015;212(5):793–802.

[22] Si J, You HJ, Yu J, Sung J, Ko G. *Prevotella* as a hub for vaginal microbiota under the influence of host genetics and their association with obesity. Cell Host & Microbe. 2017;21(1):97–105.

[23] Brotman RM, He X, Gajer P, Fadrosh D, Sharma E, Mongodin EF, et al. Association between cigarette smoking and the vaginal microbiota: a pilot study. BMC Infectious Diseases. 2014;14(1):1–11.

[24] Achilles SL, Austin MN, Meyn LA, Mhlanga F, Chirenje ZM, Hillier SL. Impact of contraceptive initiation on vaginal microbiota. American Journal of Obstetrics and Gynecology. 2018;218(6):622–e1.

[25] Laniewski P, Owen KA, Khnanisho M, Brotman RM, Herbst-Kralovetz MM. Clinical and personal lubricants impact the growth of vaginal lactobacillus species and colo-

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

nization of vaginal epithelial cells: an in vitro study. Sexually Transmitted Diseases. 2021;48(1):63–70.

[26] Tessandier N, Uysal IB, Elie B, Selinger C, Bernat C, Boué V, et al. Does exposure to different menstrual products affect the vaginal environment? Molecular Ecology. 2023;32(10):2592–2601.

[27] Vodstrcil LA, Twin J, Garland SM, Fairley CK, Hocking JS, Law MG, et al. The influence of sexual activity on the vaginal microbiota and *Gardnerella vaginalis* clade diversity in young women. PLOS ONE. 2017;12(2):e0171856.

[28] Noormohammadi M, Eslamian G, Kazemi SN, Rashidkhani B. Association between dietary patterns and bacterial vaginosis: a case–control study. Scientific Reports. 2022;12(1):12199.

[29] Hutchinson KB, Kip KE, Ness RB. Condom use and its association with bacterial vaginosis and bacterial vaginosis-associated vaginal microflora. Epidemiology. 2007;p. 702–708.

[30] Pape K, Ryttergaard L, Rotevatn TA, Nielsen BJ, Torp-Pedersen C, Overgaard C, et al. Leisure-time physical activity and the risk of suspected bacterial infections. Medicine and Science in Sports and Exercise. 2016;48(9):1737–1744.

[31] Amabebe E, Anumba DO. Psychosocial stress, cortisol levels, and maintenance of vaginal health. Frontiers in Endocrinology. 2018;p. 568.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[32] Klebanoff MA, Nansel TR, Brotman RM, Zhang J, Yu KF, Schwebke JR, et al. Personal hygienic behaviors and bacterial vaginosis. Sexually Transmitted Diseases. 2010;37(2):94.

[33] Van Kessel K, Assefi N, Marrazzo J, Eckert L. Common complementary and alternative therapies for yeast vaginitis and bacterial vaginosis: a systematic review. Obstetrical & Gynecological Survey. 2003;58(5):351–358.

[34] Ma ZS. Microbiome transmission during sexual intercourse appears stochastic and supports the red queen hypothesis. Frontiers in Microbiology. 2022;12:789983.

[35] Juliana NC, Peters RP, Al-Nasiry S, Budding AE, Morré SA, Ambrosino E. Composition of the vaginal microbiota during pregnancy in women living in sub-Saharan Africa: a PRISMA-compliant review. BMC Pregnancy and Childbirth. 2021;21(1):1–15.

[36] Gupta K, Hillier SL, Hooton TM, Roberts PL, Stamm WE. Effects of contraceptive method on the vaginal microbial flora: a prospective evaluation. Journal of Infectious Diseases. 2000;181(2):595–601.

[37] Brotman RM, Klebanoff MA, Nansel TR, Andrews WW, Schwebke JR, Zhang J, et al. A longitudinal study of vaginal douching and bacterial vaginosis—a marginal structural modeling analysis. American Journal of Epidemiology. 2008;168(2):188–196.

[38] Gelman A. Scaling regression inputs by dividing by two standard deviations. Statistics in Medicine. 2008;27(15):2865–2873.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[39] Christopher H Jackson. Multi-State Models for Panel Data: The msm Package for R. Journal of Statistical Software. 2011;38(8):1–29.

[40] Brooks JP, Buck GA, Chen G, Diao L, Edwards DJ, Fettweis JM, et al. Changes in vaginal community state types reflect major shifts in the microbiome. Microbial Ecology in Health and Disease. 2017;28(1):1303265.

[41] Fettweis JM, Serrano MG, Brooks JP, Edwards DJ, Girerd PH, Parikh HI, et al. The vaginal microbiome and preterm birth. Nature Medicine. 2019;25(6):1012–1021.

[42] Team SD. Stan Functions Reference; 2024. Accessed: 2024-02-05. `https://mc-stan.org/docs/functions-reference`.

[43] Stan Development Team. Prior Choice Recommendations; 2023. Accessed: 2024-02-15. `https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations`.

[44] Piironen J, Vehtari A. Sparsity information and regularization in the horseshoe and other shrinkage priors. Electronic Journal of Statistics. 2017;11(2):5018–5051.

[45] Stan Development Team. RStan: the R interface to Stan; 2023. R package version 2.32.3. Available from: `https://mc-stan.org/`.

[46] van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of Statistical Software. 2011;45(3):1–67.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[47] Stan Development Team. The Stan Core Library; 2018. Version 2.18.0. Available from: `http://mc-stan.org/17`.

[48] Betancourt M. Towards a principled Bayesian workflow; 2020. Available from: `https://betanalpha.github.io/assets/case_studies/principled_bayesian_workflow.html`.

[49] Odogwu NM, Onebunne CA, Chen J, Ayeni FA, Walther-Antonio MR, Olayemi OO, et al. *Lactobacillus crispatus* thrives in pregnancy hormonal milieu in a Nigerian patient cohort. Scientific Reports. 2021;11(1):18152.

[50] Ravel J, Brotman RM, Gajer P, Ma B, Nandy M, Fadrosh DW, et al. Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. Microbiome. 2013;1(1):1–6.

[51] Fethers KA, Fairley CK, Hocking JS, Gurrin LC, Bradshaw CS. Sexual risk factors and bacterial vaginosis: a systematic review and meta-analysis. Clinical Infectious Diseases. 2008;47(11):1426–1435.

[52] Liao Y, Tong XT, Jia YJ, Liu QY, Wu YX, Xue WQ, et al. The effects of alcohol drinking on oral microbiota in the Chinese population. International Journal of Environmental Research and Public Health. 2022;19(9):5729.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[53] Froehle L, Ghanem KG, Page K, Hutton HE, Chander G, Hamill MM, et al. Bacterial vaginosis and alcohol consumption: a cross-sectional retrospective study in Baltimore, Maryland. Sexually Transmitted Diseases. 2021;48(12):986–990.

[54] Smart S, Singal A, Mindel A. Social and sexual risk factors for bacterial vaginosis. Sexually Transmitted Infections. 2004;80(1):58–62.

[55] Thorsen P, Vogel I, Molsted K, Jacobsson B, Arpi M, Møller BR, et al. Risk factors for bacterial vaginosis in pregnancy: a population-based study on Danish women. Acta Obstetricia et Gynecologica Scandinavica. 2006;85(8):906–911.

[56] Shoubnikova M, Hellberg D, Nilsson S, Mårdh PA. Contraceptive use in women with bacterial vaginosis. Contraception. 1997;55(6):355–358.

[57] Rugpao S, Sriplienchan S, Rungruengthanakit K, Lamlertkittikul S, Pinjareon S, Werawatakul Y, et al. Risk factors for bacterial vaginosis incidence in young adult Thai women. Sexually Transmitted Diseases. 2008;35(7):643–648.

[58] Baisley K, Changalucha J, Weiss HA, Mugeye K, Everett D, Hambleton I, et al. Bacterial vaginosis in female facility workers in north-western Tanzania: prevalence and risk factors. Sexually Transmitted Infections. 2009;85(5):370–375.

[59] Francis SC, Looker C, Vandepitte J, Bukenya J, Mayanja Y, Nakubulwa S, et al. Bacterial vaginosis among women at high risk for HIV in Uganda: high rate of recurrent diagnosis despite treatment. Sexually Transmitted Infections. 2016;92(2):142–148.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[60] Cu-Uvin S, Hogan JW, Warren D, Klein RS, Peipert J, Schuman P, et al. Prevalence of lower genital tract infections among human immunodeficiency virus (HIV)—seropositive and high-risk HIV-seronegative women. Clinical Infectious Diseases. 1999;29(5):1145–1150.

[61] French AL, Adeyemi OM, Agniel DM, Evans CT, Yin MT, Anastos K, et al. The association of HIV status with bacterial vaginosis and vitamin D in the United States. Journal of Women's Health. 2011;20(10):1497–1503.

[62] Brotman RM, Ghanem KG, Klebanoff MA, Taha TE, Scharfstein DO, Zenilman JM. The effect of vaginal douching cessation on bacterial vaginosis: a pilot study. American Journal of Obstetrics and Gynecology. 2008 Jun;198(6):628.e1–628.e7.

[63] Sobel JD, Sobel R. Current and emerging pharmacotherapy for recurrent bacterial vaginosis. Expert Opinion on Pharmacotherapy. 2021;22(12):1593–1600.

[64] Morsli M, Gimenez E, Magnan C, Salipante F, Huberlant S, Letouzey V, et al. The association between lifestyle factors and the composition of the vaginal microbiota: a review. European Journal of Clinical Microbiology & Infectious Diseases. 2024;43(10):1869–1881.

[65] Zhou X, Brown CJ, Abdo Z, Davis CC, Hansmann MA, Joyce P, et al. Differences in the composition of vaginal microbial communities found in healthy Caucasian and black women. ISME J. 2007;1(2):121–133.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[66] Fettweis JM, Brooks JP, Serrano MG, Sheth NU, Girerd PH, Edwards DJ, et al. Differences in vaginal microbiome in African American women versus women of European ancestry. Microbiology. 2014;160(Pt 10):2272.

[67] Borgdorff H, Veer Cvd, Houdt Rv, Alberts CJ, Vries HJd, Bruisten SM, et al. The association between ethnicity and vaginal microbiota composition in Amsterdam, the Netherlands. PLOS ONE. 2017;12(7):e0181135.

[68] Lambert JA, John S, Sobel JD, Akins RA. Longitudinal analysis of vaginal microbiome dynamics in women with recurrent bacterial vaginosis: recognition of the conversion process. PLOS ONE. 2013;8(12):e82599.

[69] Srinivasan S, Liu C, Mitchell CM, Fiedler TL, Thomas KK, Agnew KJ, et al. Temporal variability of human vaginal bacteria and relationship with bacterial vaginosis. PLOS ONE. 2010;5(4):e10197.

[70] Armstrong E, Hemmerling A, Joag V, Huibner S, Kulikova M, Crawford E, et al. Treatment success following standard antibiotic treatment for bacterial vaginosis is not associated with pre-treatment genital immune or microbial parameters. Open Forum Infectious Diseases. 2023 Jan;p. ofad007.

[71] Lebeer S, Ahannach S, Gehrmann T, Wittouck S, Eilers T, Oerlemans E, et al. A citizen-science-enabled catalogue of the vaginal microbiome and associated factors. Nature Microbiology. 2023;8(11):2183–2195.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

[72] Symul L, Jeganathan P, Costello EK, France M, Bloom SM, Kwon DS, et al. Sub-communities of the vaginal microbiota in pregnant and non-pregnant women. Proceedings of the Royal Society B. 2023;290(2011):20231461.

[73] Holm JB, France MT, Gajer P, Ma B, Brotman RM, Shardell M, et al. Integrating compositional and functional content to describe vaginal microbiomes in health and disease. Microbiome. 2023;11(1):259.

[74] Brotman RM, Shardell MD, Gajer P, Tracy JK, Zenilman JM, Ravel J, et al. Interplay between the temporal dynamics of the vaginal microbiota and human papillomavirus detection. Journal of Infectious Diseases. 2014;210(11):1723–1733.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

Supplementary Information

Competing interests

SAlizon is a recommender for PCI Evolutionary Biology and PCI Ecology.

JReynes reports personal fees from Gilead (payment or honoraria for lectures, presentations, or educational events) Merck (payment or honoraria for lectures, presentations, or educational events), Moderna (honoraria for lectures, or manuscript writing), Shionogi (honoraria for presentation) and ViiV Healthcare (consulting and payment or honoraria for lectures, presentations, or educational events) and support for attending meetings and/or travel from Gilead and Pfizer, all outside of the submitted work.

JRavel is co-founder of LUCA Biologics, a biotechnology company focusing on translating microbiome research into live biotherapeutics drugs for women's health. He is Editor-in-Chief at *Microbiome*.

None of the other authors report any conflict of interest.

S1: Pairwise correlations between covariates

There were no strong correlations among covariates, with the strongest correlation found between BMI and stress ($r = 0.41$; Fig. S1).

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

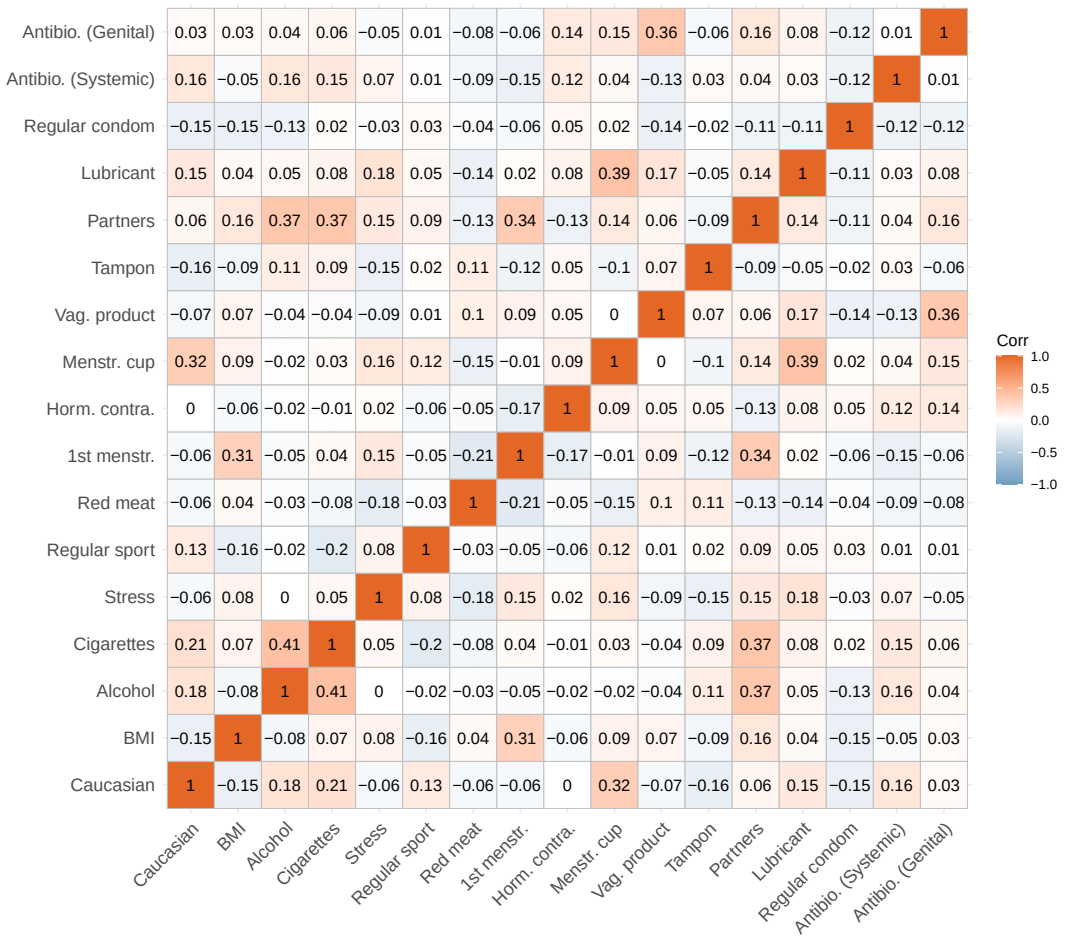

Figure S1: Correlation between covariates. Pairwise Pearson's correlation coefficients between covariates. Parameter descriptions are found in Materials and Methods.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

S2: Assessment of posterior accuracy, precision and prior contraction

We leveraged the properties of posterior distributions to identify potential model fitting problems that might manifest from our model assumptions. To examine the accuracy and precision of posterior distributions, we first generated simulated observations based on the estimated posterior mean parameters. We then refitted our model to the simulated observations (i.e., secondary fitting) to compute the posterior z-score for each parameter, which measures how closely the posterior recovers the parameters of the data generating process [48]:

$$z = \frac{\mathbb{E}_{\text{sim}} - \mathbb{E}_{\text{post}}}{\sigma_{\text{sim}}},$$

where $\mathbb{E}_{\text{post}}$ denotes the posterior mean of the fit to the actual data that we consider the 'true' parameter. $\mathbb{E}_{\text{sim}}$ and σ_{sim} denote the mean and standard deviation of the posterior distribution of the secondary fitting. The smaller the z-score, the closer the bulk of the posterior is to the true parameter [48]. In contrast, large z-values may be indicative of overfitting and, or poor prior specifications [48].

To examine the influence of the likelihood function in relation to prior information, we computed the posterior contraction, k:

$$k = 1 - \frac{\sigma^2_{\text{post}}}{\sigma^2_{\text{prior}}}$$

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

where σ^2_{post} and σ^2_{prior} correspond to the variance of posterior and prior distributions, respectively. The k values close to zero indicate that data contain little information (i.e., rendering priors strongly informative). Conversely, values close to 1 indicate that data are much more informative than the prior [48].

We found that most of our model parameters and hyperparameters — were estimated with accuracy, precision, and identifiability, with the absolute posterior z-scores below three (Fig. S2). The posterior distributions for covariate coefficients, β, contracted by 86% on average, and at least 75%, compared to the prior distribution, meaning that the covariate coefficients were well-identified from data (Fig. S2). Although we used generic priors recommended by Stan [42], the L_s parameters that define correlations among between-woman variation showed limited posterior contraction (i.e., $\leq\sim 0.25$), indicating that these parameters are poorly informed by data. As such, we refrain from making biological inferences about these correlations.

S3: Predicted difference in community state type (CST) prevalence at various counterfactual scenarios.

Our counterfactual simulations predicted that alcohol consumption and the number of partners are factors that impact the population-level outcome in terms of the prevalence of different community state types. The full list of comparisons is available in Fig. S3.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

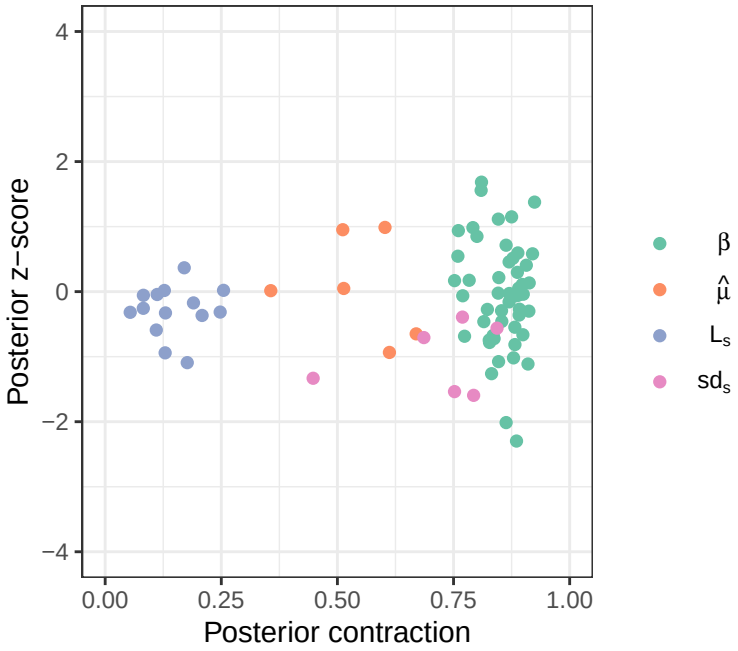

Figure S2: Accuracy, precision and identifiability of estimated parameters. Posterior z-score (y-axis) measures how closely the posterior recovers the parameters of the true data-generating process and posterior contraction (x-axis) evaluates the influence of the likelihood function over the prior, respectively. Smaller absolute posterior z-scores indicate that the posterior accurately recovers the parameters of the data-generating process: the absolute value beyond three to four may indicate substantial bias [48]. The posterior contraction values close to one indicate that data are much more informative than the prior. The estimated parameters are represented by a filled dot.

medRxiv preprint doi: https://doi.org/10.1101/2024.04.08.24305448; this version posted December 1, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. All rights reserved. No reuse allowed without permission.

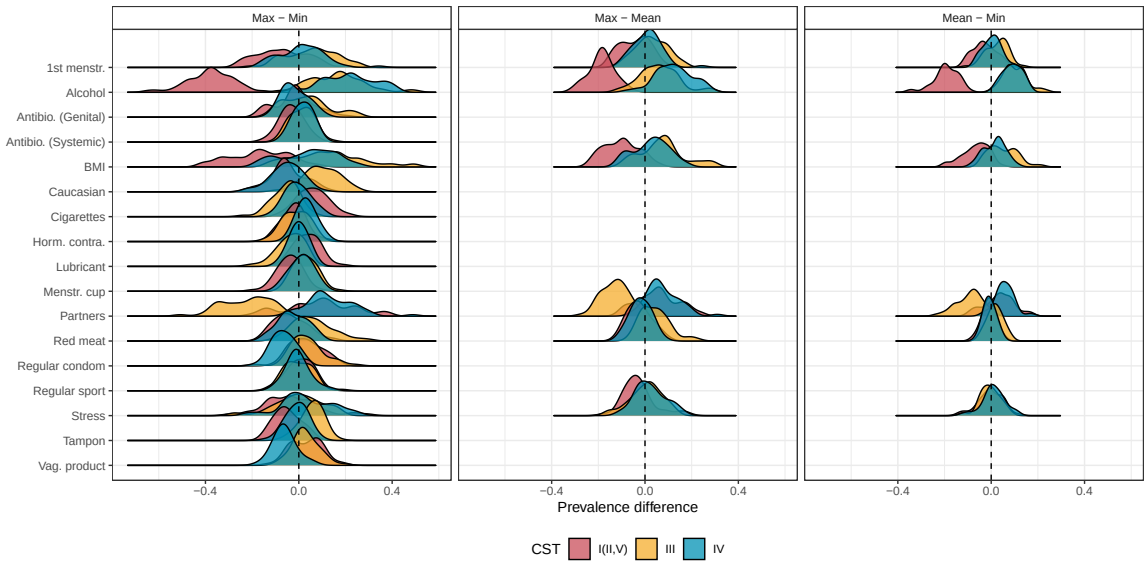

Figure S3: Difference in community state type (CST) prevalence at predicted various counterfactual scenarios. The differences were calculated from posterior samples simulated at 0 and 1 for binary variables and at the population maximum and minimum values recorded by the PAPCLEAR for continuous variables (left panel). Additional differences were computed between the population maximum and mean (middle panel) and the population mean and minimum for continuous variables (right panel). Parameter descriptions are found in Materials and Methods.