nature portfolio

Corresponding author(s): Lars Engstrand, Ms number: NMED-A145552

Last updated by author(s): Sep 20, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Questionnaire data was collected on QuestBack. Biological samples were registered in an in-house SQL database.
Data analysis	We used our in-house Snakemake (v4.8.1) workflow StaG-mwc (v0.5.1) [44] [45]. The workflow used fastp (v0.23.2) s for adapter removal and quality trimming . Kraken2 (v2.1.2) annotation against the GRCh38 human genome was used for host DNA removal. Taxonomic annotation was performed with MetaPhlAn 4.0. Taxonomic tables were cleaned using the R decontam package. All statistical analyses were performed in R (2022.02.4+500). All plots were created using the R package ggplot2 v3.4.2. Matching was performed using the R package MatchIt. centred-log ratio (CLR) transformation using the R package compositions v2.0-8. Statistical analysis was performed using Vegan package v2.6-4. R package Scutr v0.1.2 was used for subsampling and randomization. ANCOM-BC2 v2.0.1 for differential abundance analysis. R Caret package v6.0-94 alongside randomForest package v4.7-1,1 and R package net v7.3.18 were used for machine learning. ROC curves and plots were generated using pROC v1.18.4.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The main code used in the work is available on the CTMR GitHub page, which is accessible at https://github.com/ctmrbio/ML_swemami. The metagenomic sequences have been submitted to ENA (project number PRJEB81814).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	Gender of the mothers was not recorded in the dataset, but for consistency we refer to the biological sex of pregnant individuals as "women/mothers" throughout this manuscript.
Reporting on race, ethnicity, or other socially relevant groupings	Race and ethnicity data were not collected. Participants place of birth is reported as "Sweden/other". An aggregated socio-economic score was estimated based on co-habitation with a co-parent, full-time employment and 3 years or more of post-secondary education.
Population characteristics	The population characteristics reported are: age, pre-pregnancy BMI, Swedish born (y/n), University education (y/n), socio-economic score (described above), nulliparity, co-habitation, conception mode, previous obstetric history (IUFD, PTB, RPL, subfertility), regular menses, PCOS, endometriosis, smoking and other tobacco usage, diet characteristics (daily fiber intake, probiotics, iron supplementation), medication (antihistaminics and other allergy drugs, neuroactive drugs, GI drugs, diabetes drugs, antibiotics, antihypertensives) and psychometrics (EPDS, PSS-4).
Recruitment	All pregnant women (before gestational week 20) residing in Sweden with a personal identification number and understanding Swedish or English were eligible to participate, since no healthcare visits were required . Pregnant women were recruited through online advertisement on social media and in pregnancy-related mobile applications as well as posters on campus and in antenatal clinics.
Ethics oversight	The study protocol was approved by the Regional Board of Ethics, Stockholm, Sweden (2017/1118-31).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Cases were selected as the 132 women who gave birth spontaneously before 37 completed weeks (preterm) and had both filled in the first questionnaire and sent in the first fecal samples. Each case was matched with two controls, defined as women who gave birth after 38 completed weeks of pregnancy. The variables included in the matching were age (continuous), pre-pregnancy BMI (continuous), pregnancy week at sampling (continuous), nulliparity (yes/no) and diagnosed PCOS (yes/no).
Data exclusions	Twin pregnancies, miscarriages and early-term pregnancies were excluded
Replication	Repetition in this study was one on the machine learning models which were evaluated with leave-one-out validation, 10-fold cross-validation, and 10,000 bootstrap iterations. 10 Iteration of random subsampling was also performed to assure with reproducible internal seed values.
Randomization	In the machine learning of section of the analysis to address imbalance between cases and controls, random subsampling of controls (scutr v0.1.2) was performed, with 10 iterations of training (80%) and test (20%) sets. Performance was assessed on all random test sets using AUROC, sensitivity, specificity, positive predictive value, and negative predictive value (pROC v1.18.4), with confidence intervals derived from caret confusion matrices.

Blinding	Blinding was not relevant to this study, since the ultimate goal was to train a machine learning model on cases and controls.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*