Research Square

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

First-Year Infections Requiring Penicillin and Gut Bacterial Features at One Year: A Secondary Analysis of the Public ABIS Birth Cohort Dataset

Caixi Hu
Guangdong Women and Children Hospital
Xuemei Deng
Guangdong Women and Children Hospital
Huachao Mai
853749142@qq.com
Guangdong Women and Children Hospital

Research Article

Keywords: ABIS birth cohort, gut bacterial features, infancy, penicillin-treated infections, infection burden, secondary analysis, 16S rRNA, Benjamini-Hochberg false discovery rate

Posted Date: November 1st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11102829/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Infections requiring antibiotics are common in infancy, and the first-year gut community is shaped by early exposures. Whether infection burden relates to it is unclear. We aimed to estimate the association between the ordinal burden of first-year infections requiring penicillin and the 230 publicly available gut bacterial features at one year, in a secondary analysis of the public ABIS dataset.

Results

Parent-reported first-year episodes of infection requiring penicillin (0, 1–2, >=3) were related to each of the 230 processed bacterial features in pre-specified, sequentially adjusted models (M0-M3). Per-feature linear regression used HC3 heteroskedasticity-robust standard errors, with one Benjamini-Hochberg false-discovery-rate (FDR) correction across all 230 features. Detection, sensitivity and multiple-imputation analyses (m = 20) were also performed. In the Model 3 complete-case analysis (n = 1,023), 3 of the 230 features reached q < 0.05, all negative: Coprococcus eutactus (beta = -0.283, 95% CI -0.403 to -0.164; q = 0.0008), Saccharimonadia unknown Species 103 (beta = -0.406; q = 0.0251) and Intestinimonas unknown Species 284 (beta = -0.191; q = 0.0486). Only the first two remained FDR-positive in the detection model; Intestinimonas unknown Species 284 did not and was least stable across sensitivity analyses. Imputation recovered the same three associations plus 21 exploratory signals.

Conclusions

A higher infection burden was associated with lower values for three of the 230 bacterial features at one year. The C. eutactus finding is concordant with a previously reported association in the same cohort, not independent. They do not support a causal role of penicillin or changes to prescribing.

Background

The bacterial community that becomes established in the gut during the first year of life is increasingly recognised as an important correlate of later health. Early-life microbial maturation follows a broadly reproducible trajectory that is shaped by delivery mode, feeding and environmental exposures, and departures from this trajectory have been associated with immune, metabolic and neurodevelopmental outcomes later in childhood [1–5]. Because this community is still malleable during infancy, exposures occurring in the first year are of particular interest.

Antibiotic exposure in early life is among the more consistently reported modifiers of the developing gut community. In children, antibiotic exposure has been associated with reduced microbiota richness and diversity and with characterisable shifts in composition, although the magnitude and persistence of these changes differ by antibiotic class [1, 6]. Macrolides have generally been associated with larger and longer-lasting changes, whereas penicillins and amoxicillin have been reported to leave a comparatively weaker imprint on community composition [7–10]. Against this class-specific background, prospective infant studies nevertheless indicate that a single course of amoxicillin or a macrolide, or a short course of systemic antibiotics, can perturb the developing microbiota, with some differences persisting for weeks to months [11, 12].

Separating the contribution of infection from that of its treatment is a long-standing design problem in this literature. Infections requiring antibiotic treatment are common in infancy, and many questionnaire-based measures of such events capture the infection and its treatment together, so that the two cannot always be distinguished within a single item. The indication and severity of an infection may influence the gut community and later health independently of the drug used, and in the same cohort antibiotic exposure and infection frequency have been shown to behave differently as exposures for clinical outcomes [13, 14]. An association estimated for a composite exposure of infection plus treatment therefore should not be attributed to the antibiotic alone [15], and designs that record infection and antibiotic dispensing separately can help to separate the two [16].

The All Babies in Southeast Sweden (ABIS) birth cohort offers an unusually detailed setting for this question. A one-year gut bacterial profile generated by 16S rRNA sequencing has been released as a public, processed dataset, and this resource has been used to examine future coeliac disease, type 1 diabetes and juvenile idiopathic arthritis [4, 17, 18]. In these studies the bacterial community has generally served as a correlate or predictor of a downstream disease outcome rather than as the outcome itself.

It is important to state that the same ABIS one-year bacterial resource has already been used to examine infection- and antibiotic-related exposures in relation to the bacterial community. In a multi-omics analysis of this cohort, Ahrens and colleagues dichotomised first-year exposure to “infection with antibiotics” and to otitis and performed differential-abundance testing across the full 13,197-ASV layer using DESeq2 with false-discovery control applied within each comparison; they reported that Coprococcus eutactus was more abundant both in children never exposed to antibiotics (FDR-adjusted $P = 1.7 \times 10^{-58}$) and in children who never had otitis (FDR-adjusted $P = 1.0 \times 10^{-38}$) [19]. The same group has also released the processed dataset analysed here and has used a closely related parent-reported exposure variable from the same public dataset, in that instance as a predictor of and matching covariate for future obesity, without modelling the bacterial features as an outcome [5]. An association between an infection- or antibiotic-related exposure and C. eutactus has therefore already been reported in the same cohort, and the present study does not provide a first description of this broader association.

What has not, to our knowledge within the sources we audited, been reported is an analysis that treats the ordinal burden of penicillin-treated infection as the exposure, treats the publicly released bacterial features themselves as the outcome, and applies a single multiplicity framework across all 230 features together with pre-specified presence/absence, sensitivity and missing-data analyses.

We therefore performed a secondary observational analysis of the publicly released ABIS dataset. Our aim was to estimate the association between the ordinal burden of infections requiring penicillin during the first year of life (0, 1–2 and ≥ 3 episodes) and each of the 230 publicly available gut bacterial features assessed at approximately one year of age. Associations were estimated in pre-specified sequentially adjusted models (M0-M3), with a single Benjamini-Hochberg false-discovery-rate correction applied across all 230 features, and were complemented by a pre-specified presence/absence analysis, several pre-specified sensitivity analyses and multiple imputation of the Model 3 covariates (m = 20). The analysis is intended to describe graded, transparently estimated associations; it is not designed to establish a causal effect of penicillin, and it does not constitute independent validation of previously published findings.

Methods

Study design and data source

This was a secondary observational analysis of the publicly released, processed dataset from the All Babies in Southeast Sweden (ABIS) birth cohort (Dryad DOI 10.5061/dryad.2v6wwq042) [5]. No new data were collected and no participant was contacted. The public dataset contains de-identified parent-reported questionnaire variables collected at birth and at one year of age, bacterial features derived from 16S rRNA gene sequencing of stool samples collected from diapers at approximately one year of age, and cord-blood proteomic features; only the questionnaire and bacterial variables were used here. The 230 bacterial features available in the public processed dataset were analysed as released and are referred to throughout as public processed bacterial features (bac_public). They are a processed layer and are not the full 16S dataset of the original study, which comprised 13,197 amplicon sequence variants (ASVs).

Participants

Of the 1,877 children in the public dataset, 149 were excluded because they had no usable microbiome data (all 230 bacterial features missing) and a further 389 children with microbiome data had no information on penicillin-treated infections, leaving 1,339 children in the Model 0 eligible set (Fig. 1). The complete-case and multiple-imputation analyses were conducted within this same eligible cohort but included different numbers of participants. The primary complete-case analysis included the 1,023 participants with complete Model 3 covariate data; successive exclusion of participants with incomplete Model 1, Model 2 and Model 3 covariate data left 1,282, 1,123 and 1,023 participants, respectively (Fig. 1). The multiple-imputation analysis retained all 1,339 eligible participants, an additional 316 participants relative to the primary analysis. No participant was excluded on the basis of outcome values. All analyses were performed at the participant level; the 230 bacterial features are a separate dimension, and the same participants contribute to all 230 feature models.

Exposure

The exposure was the parent-reported number of infections requiring penicillin during the first year of life, taken from the public variable meta_b_42c, labelled in the public dataset as "Infection requiring penicillin during the first year" (original categories: 1 = never, 2 = 1–2 times, 3 = 3–5 times, 4 = 6 times or more). Categories were recoded into an ordinal three-level burden variable (penicillin_inf3: 0, 1–2 and 3 or more episodes; the last category merges the original 3–5 and 6-or-more categories, because the highest-frequency categories were sparse) and were modelled with a single linear trend term (penicillin_trend: 0, 1, 2). Because the public dataset contains no dose, duration or dispensing record and the questionnaire item binds an infection to the treatment it required, the exposure is a parent-reported composite of infection and its treatment; infection and antibiotic treatment cannot be separated within this variable, and the variable is not a record of confirmed

penicillin use. Two binary contrasts (0 versus 1 or more episodes, and fewer than 3 versus 3 or more episodes) were pre-specified as sensitivity analyses and were not used to define the primary analysis.

Outcomes

The outcome was each of the 230 bacterial features available in the public processed dataset, analysed on the scale in which it was released (bac_public), with no further log or other transformation. The features are a processed relative-abundance layer; the public dataset contains no raw counts, no sequencing depth and no phylogenetic tree, and the 230 features do not represent the complete gut community. The primary outcome model was a continuous-outcome (abundance-value) model. A pre-specified detection (presence/absence) outcome was defined as a non-zero feature value (bac_public > 0).

Covariates

Covariates were pre-specified and were not selected on the basis of univariate P values; the P values in Table 1 are descriptive only. Model 0 was unadjusted. Model 1 added child sex, gestational age at delivery and delivery mode. Model 2 additionally added maternal pre-pregnancy BMI (continuous), maternal smoking during pregnancy (binary), exclusive breastfeeding duration (months; ordinal, 1–8 with 9 representing 9 months or more) and maternal education (categorical). Model 3 additionally included first-year common cold, gastroenteritis and influenza, each entered as an ordinal burden score (0–3) rather than as a count, so that the ordinal categories were not interpreted as true numbers of infections. Each general-infection score was computed as the original category minus one (1 to 0, 2 to 1, 3 to 2 and 4 to 3), and exclusive breastfeeding duration was entered as a numeric month variable (1–8, with 9 representing 9 months or more) rather than as a nine-level factor. Maternal infection during pregnancy was audited as a candidate covariate but was not adjusted for in the main model, because the item is a crude yes/no/do-not-know measure with ambiguous timing and adjustment risked over-adjustment. The alternative any/none coding of the three general-infection variables was a pre-specified sensitivity analysis. Model formulas and covariate parameters are listed in Additional file 1, Table S9.

Table 1
Characteristics of participants by penicillin-treated infection burden during the first year

Characteristic	Category	Valid n	Overall	0 episodes	1–2 episodes	≥ 3 episodes	P
A. Pregnancy and birth							
Gestational age at delivery (weeks)		1307	40.0 (39.0–41.0)	40.0 (39.0–41.0)	40.0 (39.0–41.0)	40.0 (39.0–41.0)	0.778
Birth weight (g)		1317	3595.0 (3245.0-3940.0)	3580.0 (3220.0-3921.2)	3625.0 (3291.2–3970.0)	3630.0 (3237.5–4005.0)	0.257
Birth length (cm)		1311	51.0 (50.0–52.0)	51.0 (50.0–52.0)	51.0 (50.0–52.0)	51.0 (50.0–53.0)	0.275
Maternal pre-pregnancy BMI (kg/m2)		1240	24.0 (21.0–26.0)	24.0 (21.0–26.0)	23.0 (21.0–26.0)	25.0 (23.0–27.0)	0.0409
Delivery mode	Vaginal	1295	1062 (82.0%)	676 (82.7%)	330 (80.9%)	56 (80.0%)	0.843
	Cesarean	1295	147 (11.4%)	87 (10.6%)	50 (12.3%)	10 (14.3%)	
	Other	1295	86 (6.6%)	54 (6.6%)	28 (6.9%)	4 (5.7%)	
Parity	First parity	1307	568 (43.5%)	394 (48.2%)	162 (38.6%)	12 (17.1%)	< 0.001
	Previous parity	1307	739 (56.5%)	423 (51.8%)	258 (61.4%)	58 (82.9%)	
Maternal education	Low	1310	58 (4.4%)	29 (3.5%)	25 (6.0%)	4 (5.8%)	0.0103
	Medium	1310	736 (56.2%)	447 (54.3%)	241 (57.7%)	48 (69.6%)	
	High	1310	516 (39.4%)	347 (42.2%)	152 (36.4%)	17 (24.6%)	
Maternal smoking during pregnancy	No	1304	1195 (91.6%)	757 (92.3%)	381 (91.6%)	57 (83.8%)	0.0519
	Yes	1304	109 (8.4%)	63 (7.7%)	35 (8.4%)	11 (16.2%)	
Maternal infection during pregnancy	Yes	1278	356 (27.9%)	214 (26.7%)	116 (28.4%)	26 (38.8%)	0.133
	No	1278	898 (70.3%)	576 (71.8%)	282 (68.9%)	40 (59.7%)	
	Do not know	1278	24 (1.9%)	12 (1.5%)	11 (2.7%)	1 (1.5%)	
B. Infant and family							
Child sex	Male	1339	704 (52.6%)	417 (49.8%)	240 (55.9%)	47 (64.4%)	0.0137
	Female	1339	635 (47.4%)	420 (50.2%)	189 (44.1%)	26 (35.6%)	
Exclusive breastfeeding duration (months)		1266	4.0 (4.0–6.0)	4.0 (4.0–6.0)	4.0 (4.0–6.0)	4.0 (3.0–6.0)	0.415
Household smoking (first year)	Yes	1335	60 (4.5%)	30 (3.6%)	30 (7.0%)	0 (0.0%)	0.00359
	No	1335	1275 (95.5%)	804 (96.4%)	398 (93.0%)	73 (100.0%)	
C. Clinical history during the first year							
Common cold (first year)	Never	1325	24 (1.8%)	21 (2.5%)	2 (0.5%)	1 (1.4%)	< 0.001
	1–2 times	1325	410 (30.9%)	304 (36.6%)	97 (23.0%)	9 (12.5%)	
	3–5 times	1325	654 (49.4%)	392 (47.2%)	227 (53.8%)	35 (48.6%)	
	≥ 6 times	1325	237 (17.9%)	114 (13.7%)	96 (22.7%)	27 (37.5%)	
Gastroenteritis (first year)	Never	1267	857 (67.6%)	590 (71.3%)	229 (61.2%)	38 (58.5%)	0.00158
	1–2 times	1267	400 (31.6%)	234 (28.3%)	140 (37.4%)	26 (40.0%)	
	3–5 times	1267	10 (0.8%)	4 (0.5%)	5 (1.3%)	1 (1.5%)	
	≥ 6 times	1267	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	

Characteristic	Category	Valid n	Overall	0 episodes	1–2 episodes	≥ 3 episodes	P
A. Pregnancy and birth							
Influenza (first year)	Never	1243	1094 (88.0%)	755 (91.0%)	288 (81.8%)	51 (83.6%)	< 0.001
	1–2 times	1243	142 (11.4%)	74 (8.9%)	59 (16.8%)	9 (14.8%)	
	3–5 times	1243	7 (0.6%)	1 (0.1%)	5 (1.4%)	1 (1.6%)	
	≥ 6 times	1243	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Eczema (first year)	Never	1232	953 (77.4%)	654 (80.2%)	258 (73.3%)	41 (63.1%)	0.00255
	1–2 times	1232	170 (13.8%)	101 (12.4%)	51 (14.5%)	18 (27.7%)	
	3–5 times	1232	29 (2.4%)	18 (2.2%)	10 (2.8%)	1 (1.5%)	
	≥ 6 times	1232	80 (6.5%)	42 (5.2%)	33 (9.4%)	5 (7.7%)	

1. Model 0 sample, n = 1339 (0 episodes n = 837; 1–2 episodes n = 429; ≥3 episodes n = 73).

2. Values are n (%) for categorical variables and median (IQR) for continuous variables. Continuous variables were compared with the Kruskal-Wallis test; categorical variables with the chi-square test (Fisher exact test or simulated chi-square when an expected cell count was < 5).

3. Percentages for categorical variables use the variable-specific valid n as the denominator, not the total sample. Missing values were never treated as "no" or "not exposed". The "Valid n" column gives the number of non-missing observations for each variable within the Model 0 sample (n = 1339); the corresponding number of participants with missing data for that variable is 1339 minus its valid n. Per-variable completeness for the full public dataset (n = 1877) is reported in Additional file 1, Table S7, where the percentages use the 1877-participant denominator; those values are not the denominators used in this table.

4. Section A: characteristics at or before birth; Section B: infant and family characteristics during the first year; Section C: clinical history reported during the first year. The exposure (penicillin-treated infection burden) is itself a first-year clinical variable and is shown as the grouping variable.

5. P values are descriptive group comparisons only and were not used to select covariates; the covariate set was pre-specified.

6. The 2024 childhood obesity diagnosis (ICD-10) is a long-term follow-up outcome unrelated to the exposure-microbiome question; it is shown in a supplementary table rather than in the main Table 1.

7. ≥3 episodes is a merged category (original codes 3 = 3–5 times and 4 = ≥ 6 times), fixed before the analysis.

Statistical analysis

Each of the 230 bacterial features was modelled separately by ordinary least squares with HC3 heteroskedasticity-robust standard errors (confidence intervals were computed as the coefficient plus or minus 1.96 times the robust standard error):

bac_public ~ penicillin_trend + covariates

Regression coefficients represent the modelled difference in the public processed bacterial-feature value per one-step increase in the ordinal infection-burden scale (0, 1–2, 3 or more episodes). They are not the effect of one additional infection, not the effect of one additional penicillin course, not percentage changes in abundance and not log2 fold changes. For every model, the Benjamini-Hochberg false discovery rate (FDR) procedure was applied once across all 230 features within that model's family, and both raw P values and FDR q values were retained. The three-category model was also fitted for every feature, so that the ordinal spacing between categories was not assumed to be equally spaced. Trends and categorical contrasts were reported together.

Detection (presence/absence) analysis

In the pre-specified detection analysis, a logistic regression model was fitted for each feature with detection (bac_public > 0) as the outcome and penicillin_trend as the exposure, adjusted for the Model 3 covariate set, in the Model 3 complete-case sample (n = 1,023). The detection family for FDR control was again all 230 features; this is a separate family from the primary continuous-outcome trend model. Detection models were fitted by maximum-likelihood logistic regression without robust standard errors. Odds ratios are per one-step increase in the ordinal infection-burden scale and are not risk ratios or probabilities, and must not be read as percentage changes in detection probability.

Pre-specified sensitivity analyses

The pre-specified sensitivity analyses were: (S1) 0 versus 1 or more episodes; (S2) fewer than 3 versus 3 or more episodes; (S6) a rank inverse-normal transform of the outcome; a depth-adjusted analysis in which log reconstructed library size was added to each of the Model 0 to Model 3 covariate sets; a centred log-ratio (CLR) analysis of candidate reconstructed counts (counts reconstructed as round(exp(bac_public) − 1), zero values replaced by 0.5 before log transformation, and the log values then centred by the per-participant mean, which makes the result equivalent to the CLR of the corresponding composition); a negative-binomial model of candidate reconstructed counts; and a restriction of the FDR family to the 158 features detected in at least 10% of the Model 0 sample. Candidate reconstructed counts are not the original raw sequencing counts; the authors' exact transformation of the released values has not been published, and no DESeq2 analysis was performed. Each sensitivity analysis was evaluated within its own pre-specified FDR family, and direction consistency, coefficient magnitude, confidence-interval coverage and FDR status were reported as separate questions.

Missing data and multiple imputation

Complete-case analysis was the primary analysis. Missing Model 3 covariates were handled in a sensitivity analysis by multiple imputation with m = 20 imputations, applied to the Model 3 covariates only; the exposure and the microbiome outcomes were never imputed. Imputation was performed within the 1,339 Model 0-eligible participants. The imputation model included the nine Model 1 to Model 3 covariates together with auxiliary questionnaire variables and five principal components of the 230-feature matrix; the exposure and the outcomes were never imputed. Twenty imputations with 10 iterations each were generated, and estimates were pooled by Rubin's rules with a standard small-sample degrees-of-freedom approximation. Multiple-imputation results depend on the missing-at-random assumption, on the imputation model and on the choice of imputed variables, and were therefore not used as the primary analysis. Feature-level completeness in the full public dataset is reported in Additional file 1, Table S7.

Software and reproducibility

Analyses followed the pre-specified statistical analysis plan, which was finalised before the outcome models were fitted. All reported analyses were performed with the archived R scripts in R version 4.6.1 using the pre-specified random seed 20260918. Multiple imputation used the mice package (v3.19.0; predictive mean matching for continuous variables and logistic or polytomous regression for binary and categorical variables), and HC3 heteroskedasticity-robust standard errors were computed with the sandwich package (v3.1-3) through the lmtest package. The statistical analysis plan also permitted the use of FreeStatistics for routine procedures, but no result reported here was produced with FreeStatistics. Complete scripts, package versions, model formulas, per-step sample sizes, model warnings and session information were retained. All reported P values are two-sided. The use of artificial intelligence tools in the preparation of this manuscript is documented in the Declarations section ('Use of artificial intelligence').

Results

Study population and participant characteristics

Of the 1,877 children in the public ABIS dataset, 149 were excluded because they had no usable microbiome data (all 230 bacterial features missing), leaving 1,728 children. A further 389 children with microbiome data had no information on penicillin-treated infections, leaving 1,339 children in the Model 0 eligible set (Fig. 1).

The complete-case and multiple-imputation analyses were conducted within the same eligible cohort but included different numbers of participants. The multiple-imputation (MI) analysis included all 1,339 eligible participants and imputed the Model 3 covariates (m = 20); the exposure and the microbiome outcomes were never imputed. The primary complete-case analysis included the 1,023 participants with complete Model 3 covariate data; an additional 316 participants were included in the MI analysis. Successive exclusion of participants with incomplete Model 1, Model 2 and Model 3 covariate data left 1,282, 1,123 and 1,023 participants, respectively (Fig. 1).

Within the Model 0 sample, 837 children (62.5%) had no penicillin-treated infection, 429 (32.0%) had 1–2 episodes and 73 (5.5%) had ≥ 3 episodes. Table 1 summarises participant characteristics by infection-burden group. Just over half the children were male (704/1,339, 52.6%); median gestational age at delivery was 40.0 weeks (IQR 39.0–41.0), most deliveries were vaginal (1,062/1,295, 82.0%), median birth weight was 3,595 g (IQR 3,245-3,940), and median exclusive breastfeeding duration was 4.0 months (IQR 4.0–6.0). Representative group differences in first-year history included a rise in ≥ 6 common-cold episodes from 13.7% (0 episodes) to 37.5% (≥ 3 episodes) and a fall in the proportion reporting no gastroenteritis from 71.3% to 58.5%; parity and maternal education also differed across groups.

The P values in Table 1 describe between-group differences only and were not used to select covariates, which were pre-specified. The exposure and the first-year clinical history were recorded during the same period and are not all pre-exposure variables. Variable-level completeness in the full public dataset (n = 1,877) is reported in Additional file 1, Table S7; the reduction from 1,339 to 1,023 participants in the complete-case pathway reflects incomplete Model 3 covariate data (Fig. 1).

Associations between penicillin-treated infection burden and bacterial features

In the pre-specified Model 3 complete-case analysis (n = 1,023), the ordinal association between penicillin-treated infection burden and each of the 230 bacterial features was tested, with Benjamini-Hochberg (BH) false discovery rate (FDR) correction applied once across all 230 features (Additional file 1, Table S1). Three features reached q < 0.05 (Table 2, Fig. 2), and all three associations were negative (Fig. 3).

Table 2
Bacterial features associated with penicillin-treated infection burden (Model 3 primary analysis)

Bacterial feature	Unique taxon ID	n	beta (95% CI)	P	FDR q
Coprococcus eutactus	bac_Coprococcus.eutactus	1023	-0.283 (-0.403, -0.164)	< 0.001	0.0008
Saccharimonadia unknown Species 103	bac_Saccharimonadia.unknown.Species.103	1023	-0.406 (-0.621, -0.192)	< 0.001	0.0251
Intestinimonas unknown Species 284	bac_Intestinimonas.unknown.Species.284	1023	-0.191 (-0.300, -0.082)	< 0.001	0.0486
1. Pre-specified primary analysis: Model 3 complete-case sample, n = 1023.					
2. Model: bac_public ~ penicillin_trend (0 = 0, 1 = 1–2, 2 = ≥ 3 episodes) + child sex + gestational age + delivery mode + maternal pre-pregnancy BMI + maternal smoking + breastfeeding duration + maternal education + ordinal burden scores for common cold, gastroenteritis and influenza; per-feature ordinary least squares with HC3 robust standard errors.					
3. beta is the modelled difference in the public processed bacterial-feature value per one-step increase in the ordinal infection-burden scale. It is not the effect of one additional infection or of one additional penicillin course, and it is not a causal effect.					
4. bac_public was used directly, with no further log transformation. beta cannot be converted to a percentage change: beta = − 0.283 must NOT be reported as a 28.3% decrease in abundance.					
5. Benjamini-Hochberg FDR was applied once across all 230 bacterial features within the Model 3 family; raw P values and FDR q values are both shown.					
6. Intestinimonas unknown Species 284 reaches FDR q = 0.0486 in this primary analysis; it does not reach FDR significance in several sensitivity analyses (Table 5), so it must not be described as significant in all analyses.					
7. Display names omit the programmatic bac_ prefix; the unique taxon ID maps to the taxonomy audit (Additional file 1, Table S8) and to Additional file 1, Table S1.					

Coprococcus eutactus had the smallest q value (β = -0.283, 95% CI -0.403 to -0.164; P < 0.001; q = 0.0008). Saccharimonadia unknown Species 103 (β = -0.406, 95% CI -0.621 to -0.192; P < 0.001; q = 0.0251) and Intestinimonas unknown Species 284 (β = -0.191, 95% CI -0.300 to -0.082; P < 0.001; q = 0.0486) were also FDR-positive. Coprococcus eutactus is a formally named species; Saccharimonadia unknown Species 103 and Intestinimonas unknown Species 284 were not resolved to species level in the public feature labels and are reported as unidentified taxa (Additional file 1, Table S8).

Regression coefficients represent the estimated difference in the publicly processed bacterial-feature value per one-category increase in penicillin-treated infection burden (0, 1–2 and ≥ 3 episodes), after adjustment for the Model 3 covariate set. Coefficients are on the public processed scale used directly (bac_public), with no further log transformation; they are not percentage changes in abundance and not log2 fold changes.

Figure 4 shows unadjusted observed means of the three features by infection-burden group in the Model 0 sample (n = 1,339). In the ≥ 3-episode group, all 73 observed values of Coprococcus eutactus were zero, so the group mean and its conventional confidence interval degenerate to 0 (0 to 0).

Associations across sequentially adjusted models

Adding covariates left the direction of all three associations unchanged, but the magnitude did not change uniformly (Table 3, Additional file 2, Supplementary Figure S1). Coprococcus eutactus had β = -0.244 at M0 (n = 1,339), -0.260 at M1 (n = 1,282), -0.246 at M2 (n = 1,123) and − 0.283 at M3 (n = 1,023); Saccharimonadia unknown Species 103 had β = -0.311, -0.318, -0.380 and − 0.406 across M0-M3. For these two features the absolute coefficient was larger in M3 than in M0. Intestinimonas unknown Species 284 was essentially unchanged (-0.191 at M0 and − 0.191 at M3; -0.192 at M1 and − 0.189 at M2). All three features reached $q < 0.05$ within each model-specific FDR family.

Table 3
Stepwise adjusted models (M0-M3) for the FDR-confirmed bacterial features

Bacterial feature	Unique taxon ID	Model	n	beta (95% CI)	P	FDR q
Coprococcus eutactus	bac_Coprococcus.eutactus	M0	1339	-0.244 (-0.339, -0.149)	< 0.001	0.0001
		M1	1282	-0.260 (-0.359, -0.162)	< 0.001	< 0.0001
		M2	1123	-0.246 (-0.355, -0.137)	< 0.001	0.0022
		M3 (PRIMARY)	1023	-0.283 (-0.403, -0.164)	< 0.001	0.0008
Saccharimonadia unknown Species 103	bac_Saccharimonadia.unknown.Species.103	M0	1339	-0.311 (-0.475, -0.147)	< 0.001	0.0099
		M1	1282	-0.318 (-0.490, -0.147)	< 0.001	0.0095
		M2	1123	-0.380 (-0.569, -0.192)	< 0.001	0.0045
		M3 (PRIMARY)	1023	-0.406 (-0.621, -0.192)	< 0.001	0.0251
Intestinimonas unknown Species 284	bac_Intestinimonas.unknown.Species.284	M0	1339	-0.191 (-0.275, -0.108)	< 0.001	0.0006
		M1	1282	-0.192 (-0.278, -0.105)	< 0.001	0.0011
		M2	1123	-0.189 (-0.281, -0.097)	< 0.001	0.0045
		M3 (PRIMARY)	1023	-0.191 (-0.300, -0.082)	< 0.001	0.0486

1. Models are fitted independently; the rows for one feature form one display block.

2. Model definitions: M0 unadjusted; M1 + child sex, gestational age, delivery mode; M2 + maternal pre-pregnancy BMI, maternal smoking, breastfeeding duration, maternal education; M3 + ordinal burden scores (0–3) for common cold, gastroenteritis and influenza.

3. Sample sizes differed across models because of missing covariate data (M0 n = 1339, M1 n = 1282, M2 n = 1123, M3 n = 1023). Consequently, changes in regression coefficients across models may reflect both covariate adjustment and changes in the analytic sample; they cannot be attributed to the added covariates alone.

4. FDR q values are computed within each model-specific family of 230 features and are therefore not comparable across models; M3 is the pre-specified primary analysis.

5. The direction of the adjusted association is stable across models, but the magnitude does not decrease uniformly with adjustment: for Coprococcus eutactus and Saccharimonadia unknown Species 103 the absolute coefficient is larger in M3 than in M0 (|beta| 0.244 -> 0.283 and 0.311 -> 0.406 respectively), while for Intestinimonas unknown Species 284 it is essentially unchanged (0.191 -> 0.191). The phrase "stepwise attenuation" is therefore not appropriate and has been removed.

6. beta is on the public processed bacterial-feature scale (bac_public used directly, no further transformation); see the Table 2 footnotes for interpretation limits.

Because model sample sizes differed, changes in β may reflect both covariate adjustment and the changed analytic sample.

Presence/absence analysis of the primary bacterial features

In the pre-specified presence/absence sensitivity analysis (Table 4), detection was a non-zero feature value (bac_public > 0), and logistic regression estimated the odds ratio (OR) per one-category increase in infection burden in the Model 3 complete-case sample (n = 1,023).

Table 4
Detection (presence/absence) logistic-regression sensitivity analysis

Bacterial feature	Unique taxon ID	n	Detected n	Detection %	OR (95% CI)	P	FDR q	FDR < 0.05
Coprococcus eutactus	bac_Coprococcus.eutactus	1023	81	7.9	0.38 (0.22, 0.67)	< 0.001	0.0399	yes
Saccharimonadia unknown Species 103	bac_Saccharimonadia.unknown.Species.103	1023	606	59.2	0.58 (0.46, 0.72)	< 0.001	0.0005	yes
Intestinimonas unknown Species 284	bac_Intestinimonas.unknown.Species.284	1023	153	15.0	0.62 (0.43, 0.88)	0.008	0.2092	no
1. Pre-specified presence/absence sensitivity analysis: logistic regression with Y = detection of the bacterial feature (bac_public > 0) and X = penicillin_trend (0/1/2 adjusted for the Model 3 covariate set); complete-case n = 1023. No model showed separation or zero variation (status "ok" for all three features).								
2. Detection % uses the whole Model 3 sample as the denominator (n = 1023). Detected n: Coprococcus eutactus 81, Saccharimonadia unknown Species 103 606, Intestinimonas unknown Species 284 153. Coprococcus eutactus has few detection events (81); the model converged normally with no separation and no alternative estimator was substituted.								
3. Only Coprococcus eutactus (q = 0.0399) and Saccharimonadia unknown Species 103 (q = 0.0005) reach BH-FDR q < 0.05 in this sensitivity analysis. Intestinimonas unknown Species 284 has OR 0.62 (0.43, 0.88) but q = 0.2092 and is therefore NOT FDR-significant here; this does not change its FDR-positive status in the complete-case primary abundance analysis. The three features must not be described as all being FDR-confirmed in the detection model.								
4. The odds ratio is the multiplicative change in the odds of detection per one-step increase in the ordinal infection-burden scale. OR = 0.38 means the odds of detection are estimated to be 0.38 times those at the next lower burden level; it is not a risk ratio, not a probability, and not the effect of an actual number of infections, and must not be read as a percentage decrease in detection probability.								
5. Detected n (81 events) is reported because small event counts affect the precision of the odds ratio; it is not by itself a reason to consider the model invalid.								
6. Display names omit the bac_ prefix; unique taxon IDs map to the taxonomy audit (Additional file 1, Table S8).								

All three features had ORs below 1 (Additional file 2, Supplementary Figure S4, Additional file 1, Table S4). Coprococcus eutactus: OR 0.38 (95% CI 0.22–0.67; P < 0.001; q = 0.0399), detected in 81 of 1,023 participants (7.9%). Saccharimonadia unknown Species 103: OR 0.58 (95% CI 0.46–0.72; P < 0.001; q = 0.0005), detected in 606 participants (59.2%). Intestinimonas unknown Species 284: OR 0.62 (95% CI 0.43–0.88; P = 0.008; q = 0.2092), detected in 153 participants (15.0%). Coprococcus eutactus and Saccharimonadia unknown Species 103 therefore remained FDR-positive in the detection model, whereas Intestinimonas unknown Species 284 did not; this does not change its FDR-positive status in the continuous-outcome primary analysis. The OR is the multiplicative change in the odds of detection per one-category increase in infection-burden category.

Sensitivity analyses and multiple imputation

Across the pre-specified sensitivity analyses (Table 5, Additional file 1, Table S5), the three features were not uniformly FDR-significant. Coprococcus eutactus remained negative and FDR-significant in every analysis in which it was tested; it was not included in the prevalence ≥ 10% analysis because its detection rate (7.9%) fell below that threshold, which is a family-definition restriction rather than a negative result. Saccharimonadia unknown Species 103 was FDR-positive in most analyses but not in the high-burden binary contrast (< 3 vs ≥ 3 episodes; β = -0.453, 95% CI -1.023 to 0.118; q = 0.5876) or in the reconstructed-count CLR analysis (q = 0.0544). Intestinimonas unknown Species 284 was not FDR-positive in the high-burden binary contrast (q = 0.5005), the reconstructed-count CLR analysis (q = 0.2392), the reconstructed-count negative-binomial analysis (q = 0.1425) or the prevalence ≥ 10% analysis (q = 0.0501, just above 0.05 and not rounded down).

Table 5
Trend and pre-specified sensitivity analyses for the FDR-confirmed features

Bacterial feature	Analysis	n	beta (95% CI)	P	FDR q	FDR < 0.05	Model; exposure coding; outcome scale
Coprococcus eutactus	Primary (M3 complete case)	1023	-0.283 (-0.403, -0.164)	< 0.001	0.0008	yes	OLS-HC3; trend 0/1/2; bac_public
	S1: 0 vs ≥ 1 episodes	1023	-0.323 (-0.486, -0.160)	< 0.001	0.0136	yes	OLS-HC3; binary 0 vs ≥ 1; bac_public
	S2: <3 vs ≥ 3 episodes	1023	-0.471 (-0.610, -0.332)	< 0.001	< 0.0001	yes	OLS-HC3; binary < 3 vs ≥ 3; bac_public
	Prevalence ≥ 10% restriction	not tested	NA	NA	NA	NA	OLS-HC3; trend 0/1/2; bac_public
	S6: rank inverse-normal	1023	-0.112 (-0.160, -0.065)	< 0.001	0.0010	yes	OLS-HC3; trend 0/1/2; inverse-normal score
	Sequencing-depth adjusted	1023	-0.285 (-0.404, -0.165)	< 0.001	0.0009	yes	OLS-HC3; trend 0/1/2; bac_public + depth
	Reconstructed counts, CLR	1023	-0.252 (-0.383, -0.120)	< 0.001	0.0211	yes	OLS-HC3; trend 0/1/2; CLR (reconstructed)
	Reconstructed counts, negative binomial	1023	-1.340 (-2.025, -0.655)	< 0.001	0.0041	yes	Neg. binomial; trend 0/1/2; recon. counts
	Multiple imputation (m = 20)	1339	-0.266 (-0.373, -0.159)	< 0.001	0.0003	yes	MI pooling; trend 0/1/2; bac_public
Saccharimonadia unknown Species 103	Primary (M3 complete case)	1023	-0.406 (-0.621, -0.192)	< 0.001	0.0251	yes	OLS-HC3; trend 0/1/2; bac_public
	S1: 0 vs ≥ 1 episodes	1023	-0.511 (-0.770, -0.252)	< 0.001	0.0136	yes	OLS-HC3; binary 0 vs ≥ 1; bac_public
	S2: <3 vs ≥ 3 episodes	1023	-0.453 (-1.023, 0.118)	0.120	0.5876	no	OLS-HC3; binary < 3 vs ≥ 3; bac_public
	Prevalence ≥ 10% restriction	1023	-0.406 (-0.621, -0.192)	< 0.001	0.0344	yes	OLS-HC3; trend 0/1/2; bac_public
	S6: rank inverse-normal	1023	-0.191 (-0.295, -0.086)	< 0.001	0.0273	yes	OLS-HC3; trend 0/1/2; inverse-normal score
	Sequencing-depth adjusted	1023	-0.415 (-0.628, -0.202)	< 0.001	0.0159	yes	OLS-HC3; trend 0/1/2; bac_public + depth
	Reconstructed counts, CLR	1023	-0.423 (-0.667, -0.179)	< 0.001	0.0544	no	OLS-HC3; trend 0/1/2; CLR (reconstructed)
	Reconstructed counts, negative binomial	1023	-0.230 (-0.361, -0.098)	< 0.001	0.0154	yes	Neg. binomial; trend 0/1/2; recon. counts
	Multiple imputation (m = 20)	1339	-0.288 (-0.463, -0.114)	0.001	0.0234	yes	MI pooling; trend 0/1/2; bac_public
Intestinimonas unknown Species 284	Primary (M3 complete case)	1023	-0.191 (-0.300,	< 0.001	0.0486	yes	OLS-HC3; trend 0/1/2; bac_public

Bacterial feature	Analysis	n	beta (95% CI)	P	FDR q	FDR < 0.05	Model; exposure coding; outcome scale
			-0.082)				
	S1: 0 vs ≥ 1 episodes	1023	-0.237 (-0.376, -0.098)	< 0.001	0.0393	yes	OLS-HC3; binary 0 vs ≥ 1; bac_public
	S2: <3 vs ≥ 3 episodes	1023	-0.225 (-0.457, 0.008)	0.058	0.5005	no	OLS-HC3; binary < 3 vs ≥ 3; bac_public
	Prevalence ≥ 10% restriction	1023	-0.191 (-0.300, -0.082)	< 0.001	0.0501	no	OLS-HC3; trend 0/1/2; bac_public
	S6: rank inverse-normal	1023	-0.110 (-0.175, -0.045)	< 0.001	0.0402	yes	OLS-HC3; trend 0/1/2; inverse-normal score
	Sequencing-depth adjusted	1023	-0.191 (-0.301, -0.082)	< 0.001	0.0475	yes	OLS-HC3; trend 0/1/2; bac_public + depth
	Reconstructed counts, CLR	1023	-0.157 (-0.281, -0.033)	0.013	0.2392	no	OLS-HC3; trend 0/1/2; CLR (reconstructed)
	Reconstructed counts, negative binomial	1023	-0.496 (-0.894, -0.097)	0.015	0.1425	no	Neg. binomial; trend 0/1/2; recon. counts
	Multiple imputation (m = 20)	1339	-0.206 (-0.299, -0.113)	< 0.001	0.0008	yes	MI pooling; trend 0/1/2; bac_public

1. All analyses use the frozen Model 3 covariate set unless stated; the primary complete-case result is the first row of each block. n = 1023 for complete-case analyses and n = 1339 for multiple imputation (all Model 0-eligible participants; missing covariates were imputed; exposure and outcomes were never imputed).

2. The last column gives “model; exposure coding; outcome scale”; these three concepts must not be conflated. Full definitions of the column codes and of the rounding conventions for P and q values are given in Additional file 3.

3. The prevalence ≥ 10% restriction limits the BH-FDR family to the 158 features detected in at least 10% of the Model 0 sample. Coprococcus eutactus (detection 7.9%) is outside that family and is reported as “not tested”; this is a family-definition effect, not a failed result.

4. Multiple-imputation estimates are pooled by Rubin's rules from 20 imputations of the Model 3 covariates; the pooled beta and 95% CI are taken from the pooled imputation estimates, not the complete-case results. MI is a sensitivity analysis only and depends on the MAR assumption and the imputation model.

5. Significance is judged by each analysis's pre-specified BH-FDR family; direction consistency, magnitude, CI coverage and FDR status are reported separately. Coprococcus eutactus is FDR-significant in every analysis in which it was tested. Saccharimonadia unknown Species 103 is not FDR-significant in S2 or the reconstructed-count CLR analysis. Intestinimonas unknown Species 284 is not FDR-significant in S2, reconstructed-count CLR, reconstructed-count negative binomial, or the prevalence restriction (q = 0.0501, just above 0.05 and not rounded down).

6. Outcome scales differ across analyses (public processed feature scale, inverse-normal scores, CLR, reconstructed counts, negative binomial counts); beta values must NOT be compared across rows with different outcome scales. Reconstructed counts are round(exp(bac_public) – 1), not original raw sequencing counts; no DESeq2 analysis was performed.

The three primary bacterial features retained negative associations and remained FDR-significant in the multiple-imputation analysis (n = 1,339; m = 20): Coprococcus eutactus β = -0.266 (95% CI -0.373 to -0.159; q = 0.0003), Saccharimonadia unknown Species 103 β = -0.288 (95% CI -0.463 to -0.114; q = 0.0234) and Intestinimonas unknown Species 284 β = -0.206 (95% CI -0.299 to -0.113; q = 0.0008) (Additional file 1, Table S6a). MI depends on the missing-at-random assumption and on the imputation model.

However, multiple imputation identified 21 additional FDR-positive features that did not meet the FDR threshold in the complete-case analysis (Additional file 1, Table S6b); with the 3 primary features, 24 features were FDR-positive under MI. These 21 additional findings were considered exploratory and were not included among the primary findings. Across all 230 features, 3 were complete-case FDR-positive, 21 MI-only, 179 not FDR-positive in either analysis with a consistent direction, and 27 with opposite signs under the two analyses (Additional file 2, Supplementary Figure S2); the 27 direction-inconsistent features were not FDR-positive in either analysis.

Discussion

In this secondary analysis of the publicly released ABIS dataset, we examined the ordinal association between the burden of first-year infections requiring penicillin and 230 publicly available gut bacterial features at approximately one year of age. In the pre-specified Model 3 complete-case analysis (n = 1,023), three of the 230 features reached $q < 0.05$ and all three associations were negative: Coprococcus eutactus (β = -0.283, 95% CI -0.403 to -0.164; q = 0.0008), Saccharimonadia unknown Species 103 (β = -0.406, 95% CI -0.621 to -0.192; q = 0.0251) and Intestinimonas unknown Species 284 (β = -0.191, 95% CI -0.300 to -0.082; q = 0.0486). The three features are not equivalent: C. eutactus is a formally named species, whereas the other two are unresolved taxa identified in the public feature labels only to class and genus level, respectively. In the pre-specified presence/absence analysis, C. eutactus (OR 0.38, q = 0.0399) and Saccharimonadia unknown Species 103 (OR 0.58, q = 0.0005) remained FDR-positive, whereas Intestinimonas unknown Species 284 (OR 0.62, q = 0.2092) did not. Multiple imputation (n = 1,339; m = 20) recovered the same three negative associations and identified a further 21 FDR-positive features; these additional signals were treated as exploratory and are not part of the primary findings reported here.

The most direct point of comparison is a previous analysis of the same ABIS one-year bacterial resource [19]. In that study the exposure was dichotomised, the outcome layer comprised all 13,197 ASVs analysed with DESeq2, and false-discovery control was applied within each comparison. Coprococcus eutactus was reported to be more abundant in children never exposed to antibiotics (FDR-adjusted $P = 1.7 \times 10^{-58}$) and in children without otitis (FDR-adjusted $P = 1.0 \times 10^{-38}$). Our finding for C. eutactus is directionally concordant with that report. Because it is derived from the same cohort and the same one-year bacterial resource, however, it represents concordance with earlier work in the same cohort and not independent external validation. The analyses differ in exposure coding (ordinal versus binary), outcome layer (the 230 publicly released features versus 13,197 ASVs), statistical model (linear regression of the processed bacterial-feature value on the ordinal exposure as a single linear trend term, with heteroskedasticity-robust standard errors, versus DESeq2) and multiplicity framework (one Benjamini-Hochberg family across all 230 features versus within-comparison control). A later analysis by the same group used a closely related exposure variable from the same public dataset, but modelled future obesity with the microbiome and cord proteome as predictors and did not report a model in which a bacterial feature was the outcome [5]. More broadly, our findings are directionally consistent with the general observation that early-life antibiotic and infection exposures accompany altered gut community composition [1, 6, 11]. At the same time, the human evidence on antibiotic-class specificity indicates that penicillins leave a comparatively weak imprint on community composition [7–10]. One explanation that remains hypothesis-generating is that the exposure analysed here is a composite of infection and its treatment rather than penicillin alone, so that infection-related factors, treatment, or their combination could each contribute; differences in age window and in taxonomic resolution may also contribute. The present design cannot distinguish among these possibilities.

C. eutactus is a formally named species in the family Lachnospiraceae and has species-level functional evidence from non-epidemiological settings. A mouse study reported that administration of C. eutactus alleviated colitis through an acetate-mediated IgA response and microbiota restoration [20]; in vitro genomic work showed that C. eutactus and related strains use β-glucan as a growth substrate [21]; the relative abundance of C. eutactus was selected as a predictor of language development in 3-year-old children in a rural Ugandan cohort [22]; and an isolate attenuated stress-related behavioural changes in mice [23]. Genus-level reviews describe Coprococcus as a butyrate-producing genus of potential biomarker interest [24]; that butyrate evidence is genus-level and does not transfer to the species, whereas the acetate link was demonstrated specifically for C. eutactus in mice [20]. Taken together, these findings are compatible with C. eutactus behaving as a fibre-utilising commensal whose lower abundance and lower probability of detection in children with a higher infection burden could reflect loss of a potentially beneficial community member. The supporting evidence nevertheless remains indirect, deriving from animals, in vitro systems or other cohorts, and it does not establish a mechanism in this cohort. Apart from the same-cohort report discussed above, we identified no direct human evidence linking penicillin or antibiotic exposure to C. eutactus abundance at species level.

Saccharimonadia unknown Species 103 was not resolved below class in the public feature labels and belongs to the phylum Patescibacteria. We found no direct evidence on this specific taxon. Background information on the Saccharibacteria (TM7) group derives mainly from the oral cavity and from episymbiotic lifestyles [25], and a separate one-year infant cohort reported Saccharibacteria taxa in relation to neurodevelopment without examining this exposure [26]. Because these sources concern a different taxon, a different anatomical site or a different design, we do not attribute any specific metabolic function to this feature.

Intestinimonas unknown Species 284 was resolved to the genus Intestinimonas (family Oscillospiraceae) but not to species. Here too we found no direct evidence on the specific taxon. The available evidence concerns the genus or a different species: Intestinimonas butyriciproducens and its fructoselysine-to-butyrate fermentation have been associated with favourable cardiometabolic markers and improved host metabolic health [27], and the genus and its type species have been characterised taxonomically and physiologically [28, 29]. This is genus-level or other-species evidence and must not be assigned to Intestinimonas unknown Species 284; we therefore do not describe this feature as a known butyrate producer. Its limited robustness (see below) should also constrain interpretation.

Across the pre-specified sensitivity analyses the three features did not behave uniformly. C. eutactus remained negative and FDR-significant in every analysis in which it was tested; it was not included in the prevalence ≥ 10% restriction because its detection rate (7.9%) fell below the threshold defining that family, which is a family-definition restriction rather than a negative result. Saccharimonadia unknown Species 103 was FDR-positive in most analyses but not in the high-burden binary contrast (< 3 versus ≥ 3 episodes; β = -0.453, 95% CI -1.023 to 0.118; q = 0.5876) or in the reconstructed-count centred-log-ratio analysis (q = 0.0544). Intestinimonas unknown Species 284 was not FDR-positive in the high-burden binary contrast (q = 0.5005), the reconstructed-count centred-log-ratio analysis (q = 0.2392), the reconstructed-count negative-binomial analysis (q = 0.1425) or the prevalence ≥ 10% analysis (q = 0.0501, just above the threshold and not rounded down); less weight should therefore be placed on this feature. Under multiple imputation all three features remained negative and FDR-significant, and 21 additional features became FDR-positive; the latter depend on the missing-at-random assumption and on the imputation model, are exploratory only, and are not part of the primary findings. Direction consistency, coefficient magnitude, confidence-interval coverage and FDR status are separate questions that we report separately; notably, the 27 features whose coefficients changed sign between the complete-case and imputed analyses were not FDR-positive in either analysis.

Because the exposure combines infection and penicillin treatment, several non-exclusive processes could contribute to the observed associations: the infection itself and its indication or severity; an effect of penicillin treatment on the gut community; and confounding by factors related to both infection frequency and microbiome composition, such as host genetics, dietary detail, other antibiotic classes and household or social factors. Existing human evidence on penicillin specifically is limited and does not point consistently to a strong compositional effect. In a large prescription-linked metagenomic study, penicillin V and extended-spectrum penicillins were associated with only a few species, whereas clindamycin, fluoroquinolones and flucloxacillin accounted for most of the associations [9]; a prospective observational study of adults receiving penicillin V for pharyngotonsillitis, which used culture-based faecal sampling of specific bacterial groups and resistance rather than whole-community metagenomic sequencing, found an increase in ampicillin-resistant Enterobacterales with limited evidence of a compositional shift [10]; and a small prospective infant cohort showed that a single course of amoxicillin or a macrolide perturbs the microbiota [11]. Interpretation of the present associations as a direct effect of penicillin is therefore not supported by the available evidence, and we frame the possible mechanisms as hypotheses that the present design cannot test.

These findings add transparent and reproducible observational evidence on the relationship between the burden of early-life infection requiring treatment and gut bacterial features at one year. They may be useful for framing and powering future studies of early-life antibiotic exposure and microbiome development, and for pre-specifying which features are worth re-examining in independent cohorts. They do not, however, justify any change to antibiotic prescribing, to the management of infections, or to paediatric care. The exposure was reported by parents, without dose or duration; infection and treatment could not be separated; and the associations are observational. It is also relevant that, in the same cohort, early-life infections and antibiotic use have shown limited or outcome-specific clinical associations [14, 30].

The strengths of this analysis are its pre-specified statistical analysis plan, the use of an ordinal exposure entered as a single linear trend term rather than a binary exposure, testing of all 230 publicly released features within a single Benjamini-Hochberg family, a pre-specified presence/absence model, several pre-specified sensitivity analyses, multiple imputation of the Model 3 covariates, and complete reporting of all 230 features including null results, from a publicly released dataset that permits reproduction. Several limitations qualify the interpretation. First, this is a secondary analysis of publicly released processed data; the 230 features are a processed relative-abundance layer without raw counts, sequencing depth or a phylogenetic tree, and they do not represent the complete gut community. Second, the data and research content overlap substantially with previous work in the same cohort [5, 19], and in particular the association with C. eutactus has already been reported there. Third, the exposure is a parent-reported composite measure and may be affected by recall bias; dose and duration were not recorded, the category boundaries are coarse, and misclassification cannot be excluded. Fourth, infection and actual antibiotic treatment cannot be distinguished. Fifth, the exposure and the microbiome sampling window overlap in time, so the temporal ordering of exposure and outcome is not established and reverse causation cannot be excluded. Sixth, residual confounding remains possible, including by host genetics, dietary detail and antibiotic classes other than penicillin, which were not fully modelled. Seventh, interpretation is constrained by differing taxonomic levels and by the two unresolved taxa, which cannot be assigned species identities or functions; the numbers 103 and 284 are internal feature identifiers and carry no taxonomic meaning. Eighth, the complete-case and multiple-imputation analyses rest on different assumptions: the primary complete-case analysis included 1,023 of the 1,339 eligible participants, which may introduce selection, while the imputation analysis depends on the missing-at-random assumption and on the imputation model and generated 21 additional exploratory signals. Finally, we identified no independent external validation; the findings are limited to the ABIS population and to this published feature layer.

Conclusions

In this secondary analysis of the publicly released ABIS dataset, a higher burden of infections requiring penicillin during the first year of life was associated with lower abundance values for three of the 230 bacterial features and with a lower probability of detection for two of them

at approximately one year of age. In the abundance (continuous-value) models, C. eutactus, Saccharimonadia unknown Species 103 and Intestinimonas unknown Species 284 were all FDR-positive with negative coefficients; in the detection (presence/absence) model, C. eutactus and Saccharimonadia unknown Species 103 remained FDR-positive whereas Intestinimonas unknown Species 284 did not, and this feature was also the least stable across the sensitivity analyses. The three features are not taxonomically equivalent, and the finding for C. eutactus is concordant with an association previously reported in the same cohort [19] and is therefore not independent. These associations should be understood as signals that require evaluation in independent settings; they do not establish a causal effect of penicillin, and statistical significance should not be equated with clinical importance.

Abbreviations

ABIS
All Babies in Southeast Sweden
ASV
amplicon sequence variant
BH
Benjamini-Hochberg
BMI
body mass index
CI
confidence interval
CLR
centred log-ratio
FDR
false discovery rate
HC3
heteroskedasticity-consistent covariance estimator of type 3
IQR
interquartile range
MI
multiple imputation
OLS
ordinary least squares
OR
odds ratio
SAP
statistical analysis plan
16S rRNA
16S ribosomal RNA
TM7
candidate phylum TM7 (Saccharibacteria).

Declarations

Ethics approval and consent to participate

This study was a secondary observational analysis of fully de-identified data that had already been released publicly. No new data were collected, no participant was contacted, and no restricted or individually identifiable data were accessed. The public dataset record states that the ABIS study includes participant consent for data sharing and publication of de-identified data, and that the deposited data underwent a formal de-identification process before release; the name of the original ethics committee, the approval number and the date of approval are not recorded in the public dataset documentation and are therefore reported here as not available from the public record. The authors' institution, Guangdong Women and Children Hospital, did not require this secondary analysis of publicly available, de-identified data to undergo a separate institutional ethics review or to obtain a separate exemption determination; accordingly, no separate approval or exemption number is reported, and none has been invented.

Consent for publication

The public dataset record states that the ABIS study includes participant consent for data sharing and publication of de-identified data, which covers the de-identified dataset analysed and reported in this article. No individual-level identifiable data are presented.

Competing interests

The authors declare that they have no competing interests.

Use of artificial intelligence

Artificial intelligence tools (DeepSeek) were used, under human supervision, to assist with manuscript drafting, with translation from Chinese into English, and with English language editing and polishing. Generative artificial intelligence was also used to assist with selected graphical elements in the supplementary materials. The scientific data presented in the figures were derived from the study's statistical analyses and were not created or altered by artificial intelligence. The AI tools did not select the analysis variables, did not execute the statistical analyses and did not generate any of the numerical results reported. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the final manuscript.

Funding

The publication costs of this article will be covered by the Xuemei Deng Inheritance Studio Construction Project for National Famous Traditional Chinese Medicine Experts in Women's and Children's Health (Chinese original: [illegible]), hosted by Guangdong Women and Children Hospital, Guangzhou, China. The studio is established under the national Inheritance Studio Construction Project for National Famous Traditional Chinese Medicine Experts in Women's and Children's Health, organised by the National Administration of Traditional Chinese Medicine and the National Health Commission of the People's Republic of China. No separate research grant was awarded for this secondary analysis; no grant number is available for this support and no funding amount is reported, and none has been invented. The English rendering of the project name above is a descriptive translation of the official Chinese name and is not an official English name.

Author Contribution

All authors made substantive contributions to this work, read and approved the final manuscript, and agree to be accountable for all aspects of the work. CH: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Writing - original draft. XD: Supervision, Funding acquisition. HM: Supervision, Writing - review & editing.

Acknowledgements

The authors thank the ABIS study team and the participating families, and the Dryad Digital Repository for the public release of the de-identified dataset. No other individual or organisation contributed to this analysis in a way that would warrant acknowledgement.

Data Availability

Availability of data and materialsThe dataset supporting the conclusions of this article is available in the Dryad Digital Repository, https://doi.org/10.5061/dryad.2v6wwq042 [31]. No restricted or individually identifiable data were accessed. The analysis scripts and the derived tables and figures are available from the corresponding author on reasonable request; the scripts are not deposited in a public repository. This is a secondary analysis of a publicly released dataset, and the request route is provided so that editors, reviewers and readers can access the analysis scripts and the derived outputs.

References

1. Bokulich NA, Chung J, Battaglia T, Henderson N, Jay M, Li H et al. Antibiotics, birth mode, and diet shape microbiome maturation during early life. Sci Transl Med. 2016. 10.1126/scitranslmed.aad7121. PMID: 27306664.

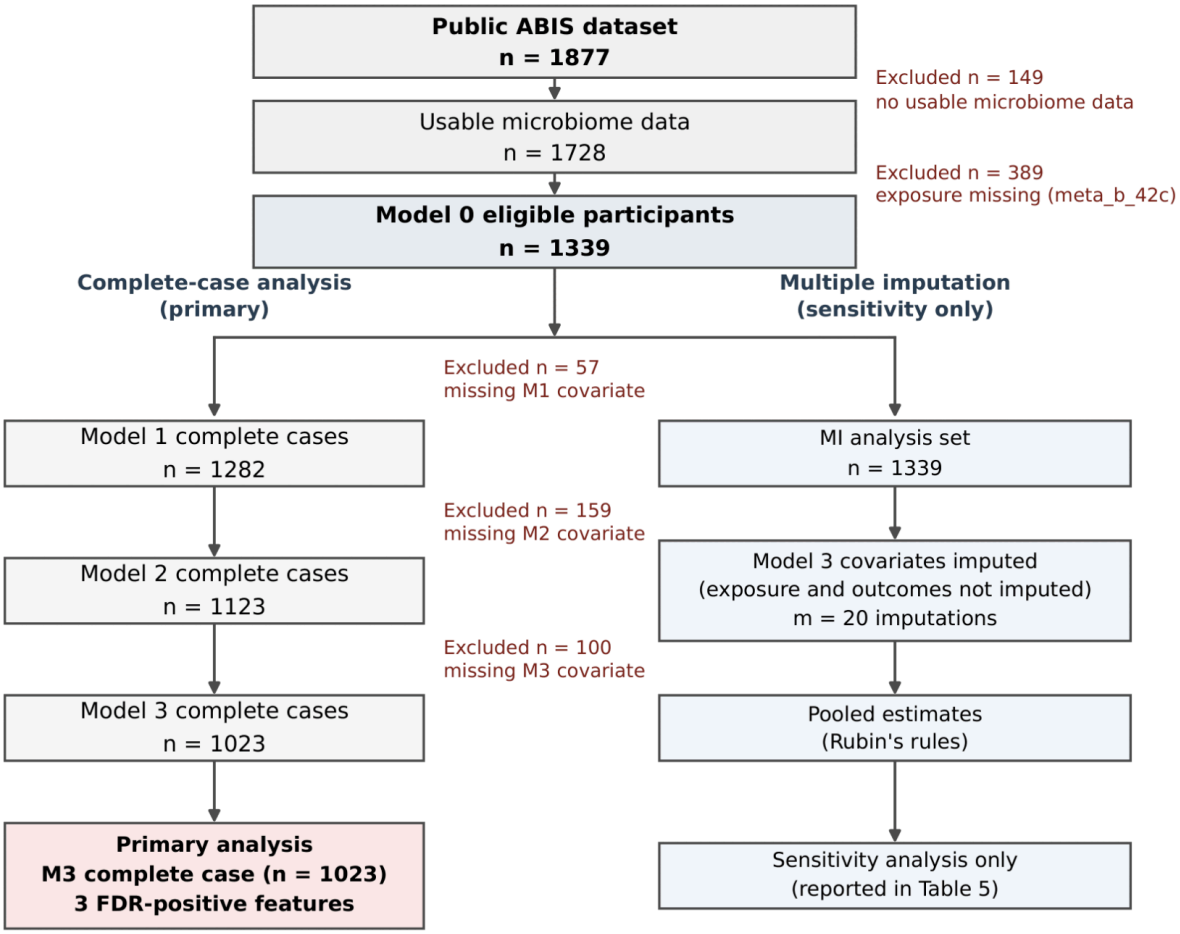
2. Grech A, Collins CE, Holmes A, Lal R, Duncanson K, Taylor R et al. Maternal exposures and the infant gut microbiome: a systematic review with meta-analysis. Gut Microbes. 2021. 10.1080/19490976.2021.1897210. PMID: 33978558.
3. Russell JT, Roesch LFW, Ördberg M, Ilonen J, Atkinson MA, Schatz DA et al. Genetic risk for autoimmunity is associated with distinct changes in the human gut microbiome. Nat Commun. 2019. 10.1038/s41467-019-11460-x. PMID: 31399563.
4. Bélteky M, Milletich PL, Ahrens AP, Triplett EW, Ludvigsson J. Infant gut microbiome composition correlated with type 1 diabetes acquisition in the general population: the ABIS study. Diabetologia. 2023. 10.1007/s00125-023-05895-7. PMID: 36964264.
5. Ahrens AP, Dias R, Hyötyläinen T, Andersson White P, Orešič M, Triplett EW et al. Early-life proteomic and microbiome features signal obesity risk across 26 years of follow-up. mSystems. 2026;11(6):e0142425. 10.1128/msystems.01424-25. PMID: 42206849.
6. McDonnell L, Gilkes A, Ashworth M, Rowland V, Harries TH, Armstrong D et al. Association between antibiotics and gut microbiome dysbiosis in children: systematic review and meta-analysis. Gut Microbes. 2021. 10.1080/19490976.2020.1870402. PMID: 33651651.
7. Korpela K, Salonen A, Virta LJ, Kekkonen RA, Forslund K, Bork P et al. Intestinal microbiome is related to lifetime antibiotic use in Finnish pre-school children. Nat Commun. 2016. 10.1038/ncomms10410. PMID: 26811868.
8. Elvers KT, Wilson VJ, Hammond A, Duncan L, Huntley AL, Hay AD et al. Antibiotic-induced changes in the human gut microbiota for the most commonly prescribed antibiotics in primary care in the UK: a systematic review. BMJ Open. 2020. 10.1136/bmjopen-2019-035677. PMID: 32958481.
9. Baldanzi G, Larsson A, Sayols-Baixeras S, Dekkers KF, Hammar U, Nguyen D et al. Antibiotic use and gut microbiome composition links from individual-level prescription data of 14,979 individuals. Nat Med. 2026. 10.1038/s41591-026-04284-y. PMID: 41814006.
10. Rystedt K, Edquist P, Giske CG, Hedin K, Tyrstrup M, Ståhlgren GS et al. Effects of penicillin V on the faecal microbiota in patients with pharyngotonsillitis - an observational study. JAC Antimicrob Resist. 2023. 10.1093/jacamr/dlad006. PMID: 36816747.
11. Korpela K, Salonen A, Saxen H, Nikkonen A, Peltola V, Jaakkola T et al. Antibiotics in early life associate with specific gut microbiota signatures in a prospective longitudinal infant cohort. Pediatr Res. 2020. 10.1038/s41390-020-0761-5. PMID: 31954376.
12. Kwon Y, Cho YS, Lee YM, Kim SJ, Bae J, Jeong SJ. Changes to gut microbiota following systemic antibiotic administration in infants. Antibiotics (Basel). 2022. 10.3390/antibiotics11040470. PMID: 35453221.
13. Kindgren E, Ludvigsson J. Infections and antibiotics during fetal life and childhood and their relationship to juvenile idiopathic arthritis: a prospective cohort study. Pediatr Rheumatol Online J. 2021. 10.1186/s12969-021-00611-4. PMID: 34530851.
14. Mårild K, Lerchova T, Östensson M, Imberg H, Størdal K, Ludvigsson J et al. Early-life infections, antibiotics and later risk of childhood and early adult-onset inflammatory bowel disease: pooled analysis of two Scandinavian birth cohorts. Aliment Pharmacol Ther. 2025. 10.1111/apt.18358. PMID: 39450871.
15. Li P, Chang X, Chen X, Wang C, Shang Y, Zheng D et al. Early-life antibiotic exposure increases the risk of childhood overweight and obesity in relation to dysbiosis of gut microbiota: a birth cohort study. Ann Clin Microbiol Antimicrob. 2022. 10.1186/s12941-022-00535-1. PMID: 36329476.
16. Li DK, Chen H, Ferber J, Odouli R. Infection and antibiotic use in infancy and risk of childhood obesity: a longitudinal birth cohort study. Lancet Diabetes Endocrinol. 2017. 10.1016/S2213-8587(16)30281-9. PMID: 27815091.
17. Milletich PL, Ahrens AP, Russell JT, Petrone JR, Berryman MA, Agardh D et al. Gut microbiome markers in subgroups of HLA class II genotyped infants signal future celiac disease in the general population: ABIS study. Front Cell Infect Microbiol. 2022. 10.3389/fcimb.2022.920735. PMID: 35959362.
18. Kindgren E, Ahrens AP, Triplett EW, Ludvigsson J. Infant gut microbiota and environment associate with juvenile idiopathic arthritis many years prior to disease onset, especially in genetically vulnerable children. EBioMedicine. 2023. 10.1016/j.ebiom.2023.104654. PMID: 37329576.
19. Ahrens AP, Hyötyläinen T, Petrone JR, Igelström K, George CD, Garrett TJ et al. Infant microbes and metabolites point to childhood neurodevelopmental disorders. Cell. 2024;187(8):1853–1873.e15. 10.1016/j.cell.2024.02.035. PMID: 38574728.
20. Yang R, Shan S, Shi J, Li H, An N, Li S, et al. Coprococcus eutactus, a potent probiotic, alleviates colitis via acetate-mediated IgA response and microbiota restoration. J Agric Food Chem. 2023. 10.1021/acs.jafc.2c06697. . PMID: 36786768.
21. Alessi AM, Gray V, Farquharson FM, Flores-López A, Shaw S, Stead D et al. β-Glucan is a major growth substrate for human gut bacteria related to Coprococcus eutactus. Environ Microbiol. 2020. 10.1111/1462-2920.14977. PMID: 32141148.
22. Kort R, Schlösser J, Vazquez AR, Atukunda P, Muhoozi GKM, Wacoo AP et al. Model selection reveals the butyrate-producing gut bacterium Coprococcus eutactus as predictor for language development in 3-year-old rural Ugandan children. Front Microbiol. 2021. 10.3389/fmicb.2021.681485. PMID: 34149673.
23. Xu L, Wang S, Wu L, Cao H, Fan Y, Wang X et al. Coprococcus eutactus screened from healthy adolescent attenuates chronic restraint stress-induced depression-like changes in adolescent mice: potential roles in the microbiome and neurotransmitter modulation. J

Affect Disord. 2024. 10.1016/j.jad.2024.04.050. PMID: 38649105.

24. Notting F, Pirovano W, Sybesma W, Kort R. The butyrate-producing and spore-forming bacterial genus Coprococcus as a potential biomarker for neurological disorders. Gut Microbiome (Camb). 2023. 10.1017/gmb.2023.14. PMID: 39295905.
25. Bor B, Bedree JK, Shi W, McLean JS, He X. Saccharibacteria (TM7) in the human oral microbiome. J Dent Res. 2019. 10.1177/0022034519831671. PMID: 30894042.
26. Davias A, Lyon-Caen S, Iszatt N, Monot C, Rayah Y, Ilhan ZE et al. Associations between the gut microbiota at one-year and neurodevelopment in children from the SEPAGES cohort. Brain Behav Immun Health. 2025. 10.1016/j.bbih.2025.101063. PMID: 40791964.
27. Rampanelli E, Romp N, Troise AD, Ananthasabesan J, Wu H, Gül IS et al. Gut bacterium Intestinimonas butyriciproducens improves host metabolic health: evidence from cohort and animal intervention studies. Microbiome. 2025. 10.1186/s40168-024-02002-9. PMID: 39833973.
28. Bui TP, Shetty SA, Lagkouvardos I, Ritari J, Chamlagain B, Douillard FP et al. Comparative genomics and physiology of the butyrate-producing bacterium Intestinimonas butyriciproducens. Environ Microbiol Rep. 2016. 10.1111/1758-2229.12483. PMID: 27717172.
29. Kläring K, Hanske L, Bui N, Charrier C, Blaut M, Haller D et al. Intestinimonas butyriciproducens gen. nov., sp. nov., a butyrate-producing bacterium from the mouse intestine. Int J Syst Evol Microbiol. 2013. 10.1099/ijs.0.051441-0. PMID: 23918795.
30. Bélteky M, Wahlberg J, Ludvigsson J. Infections and antibiotic use in early childhood have limited importance in developing manifest type 1 diabetes - the ABIS cohort study. Front Endocrinol (Lausanne). 2025. 10.3389/fendo.2025.1529447. PMID: 40060384.
31. Ahrens AP, Dias R, Hyotylainen T, Andersson White P, Oresic M, Triplett EW et al. Data from: Early life proteomic and microbiome features signal obesity risk across 26 years of follow-up. Dryad Digital Repository. 2026. https://doi.org/10.5061/dryad.2v6wwq042.

Figures

Figure 1. Flow of participants from the public ABIS dataset to the primary and sensitivity analyses



Analysis unit: one participant per row (1877 unique ObesityID; no repeated samples).
The 230 bacterial features are a separate dimension and are not participants: the same 1023 participants contribute to all 230 feature models. The two branches are alternative analyses of the SAME 1339 participants and must never be added together.

Figure 1

Flow of participants from the public ABIS dataset to the primary and sensitivity analyses.

Arrows indicate successive selection steps from the public dataset downwards. Of 1877 children, 149 were excluded because they had no usable microbiome data (all 230 bacterial features missing; no child had a partially missing microbiome profile). A further 389 children had microbiome data but no information on penicillin-treated infections (meta_b_42c), leaving 1339 children in the Model 0 eligible set. The 83 children with exposure data but no microbiome data were excluded at the first step, so the two steps do not double-count. The 1339 participants are then analysed along one of two alternative pathways. In the complete-case pathway, 57 participants were excluded for a missing Model 1 covariate (leaving 1282), a further 159 for a missing Model 2 covariate (leaving 1123), and a further 100 for a missing Model 3 covariate (leaving 1023); this Model 3 complete-case analysis is the pre-specified primary analysis. In the multiple-imputation pathway the same 1339 participants are retained and the Model 3 covariates are imputed (m = 20 imputations); the exposure and the microbiome outcomes were never imputed, and this pathway is a sensitivity analysis only. The two pathways are alternative analyses of the same participants and are never added together. The unit of analysis is one participant per row; the 230 bacterial features are a separate dimension, and the same participants contribute to all 230 feature models. Because a participant may be missing more than one covariate,

per-variable missingness does not sum to the number excluded at each step; variable-level detail is in the missing-data report (Additional file 1, Table S7).

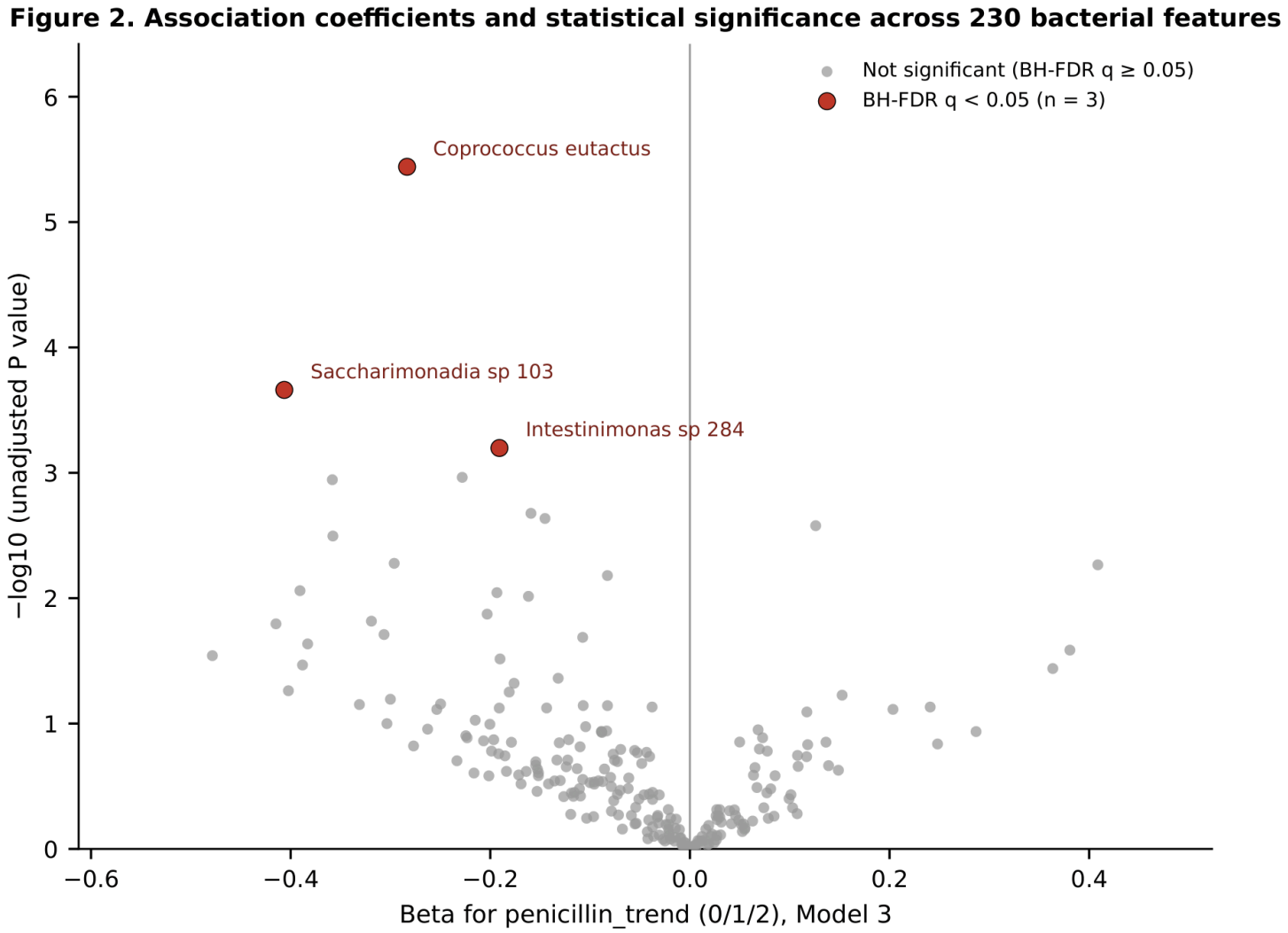


Figure 2

Association coefficients and statistical significance across 230 bacterial features.

Each point is one of the 230 bacterial features available in the public processed ABIS dataset. The x-axis is the regression coefficient (beta) for penicillin_trend (0 = 0 episodes, 1 = 1-2 episodes, 2 = >=3 episodes) in the Model 3 complete-case analysis (n = 1023); the y-axis is the unadjusted P value on a -log10 scale. The model was fitted per feature by ordinary least squares with HC3 heteroskedasticity-robust standard errors, adjusted for child sex, gestational age, delivery mode, maternal pre-pregnancy BMI, maternal smoking, breastfeeding duration, maternal education, and ordinal burden scores (0-3) for common cold, gastroenteritis and influenza. Benjamini-Hochberg FDR was applied once across all 230 features within the Model 3 family; points in red (n = 3) reach q < 0.05. The outcome was the public processed bacterial-feature value (bac_public) used directly with no further log transformation, so beta is not a log2 fold change and cannot be converted to a percentage change. This is a coefficient-and-significance plot for a continuous outcome, not a classical differential-abundance volcano plot.

Figure 3. Effect estimates for the FDR-confirmed bacterial features (Model 3)

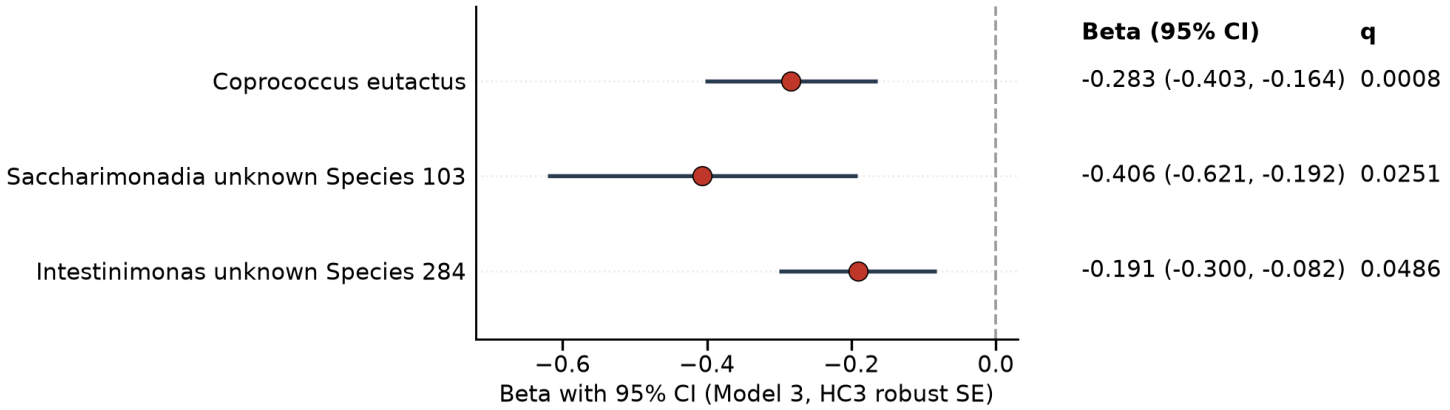


Figure 3

Associations between penicillin-treated infection burden and bacterial features at one year.

Points are regression coefficients (beta) and horizontal lines are 95% confidence intervals based on HC3 heteroskedasticity-robust standard errors, from the pre-specified primary Model 3 complete-case analysis (n = 1,023). Penicillin-treated infection burden was coded as 0, 1 and 2, corresponding to 0, 1-2 and >=3 parent-reported episodes of infection treated with penicillin during the first year. Model 3 included child sex, gestational age, delivery mode, maternal pre-pregnancy BMI, maternal smoking during pregnancy, exclusive breastfeeding duration, maternal education, and first-year common cold, gastroenteritis and influenza (ordinal burden scores 0-3). Benjamini-Hochberg false discovery rate correction was applied once across all 230 bacterial features available in the public processed ABIS dataset, and only features reaching q < 0.05 are displayed (n = 3). The dashed vertical line marks beta = 0. Beta is the modelled difference in the public processed feature value per one-step increase in the ordinal infection-burden scale; values were used as released (bac_public) with no further transformation, so beta is not the effect of one additional infection or penicillin course, cannot be converted to a percentage change, and no causal effect is implied. Display names omit the bac_ prefix; full identifiers are given in Additional file 1, Tables S1 and S8. Saccharimonadia unknown Species 103 belongs to the Saccharibacteria/TM7 (Patescibacteria) lineage; the other members of that lineage in the 230-feature set were directionally consistent but did not pass FDR (TM7x sp. 44: beta = -0.358, q = 0.082), and each feature was tested as a separate association. The 230 features are those available in the public processed dataset and are not the complete 16S dataset of the original study.

Figure 4. Observed bacterial-feature values across penicillin-treated infection-burden groups

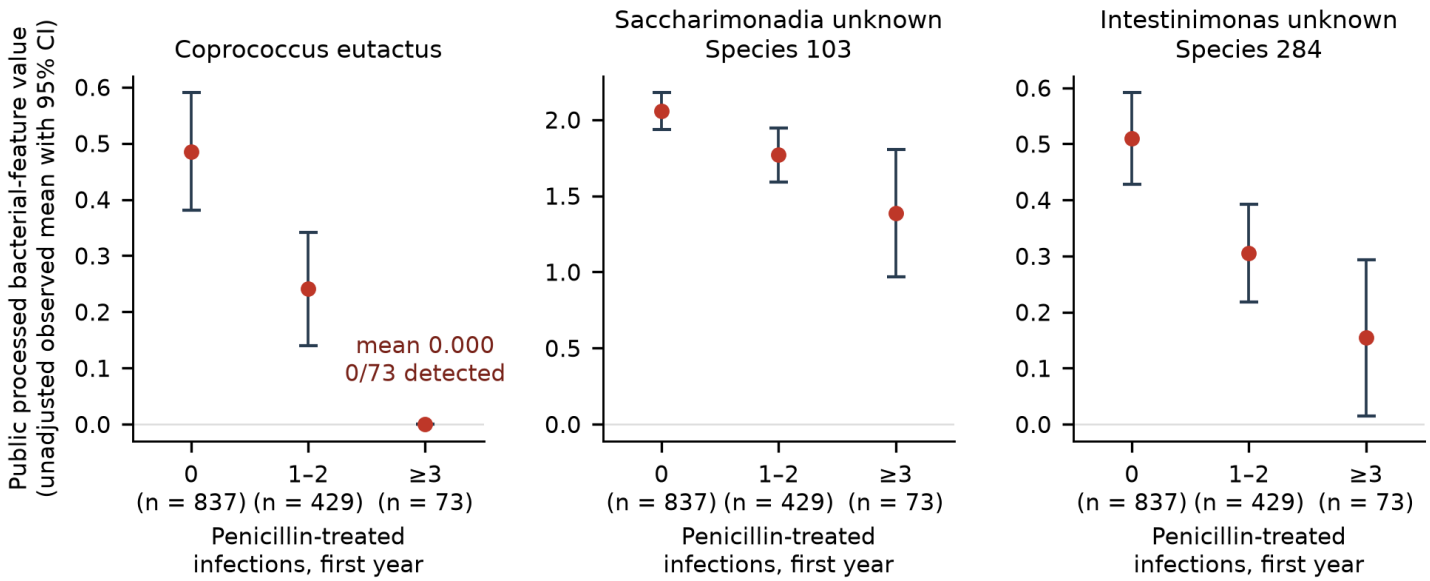


Figure 4

Observed bacterial-feature values across penicillin-treated infection-burden groups.

Unadjusted observed means (points) with 95% confidence intervals (error bars) of the public processed bacterial-feature values by penicillin-treated infection-burden group in the Model 0 sample (n = 1339; 0 episodes n = 837, 1-2 episodes n = 429, >=3 episodes n = 73). These are descriptive observed values and are not covariate-adjusted Model 3 predictions; the group means must not be interpreted as the M3 regression coefficients. The x-axis is an ordinal infection-burden classification, not a penicillin dose, a cumulative number of treatment days, or an equally spaced count of infections; the statistical evidence for a graded association comes from the pre-specified Model 3 trend model, not from the visual slope across groups. Each panel uses its own y-axis scale, so visual slopes should not be compared between panels. In the >=3 group, all 73 observed values of Coprococcus eutactus were zero (0 of 73 detected), so under the conventional mean confidence-interval method used throughout this figure the sample standard deviation, standard error and confidence interval all degenerate to zero and the interval is displayed as 0 to 0 with zero width. This degenerate interval reflects the observed sample and the method used; it should not be read as evidence that the population mean is known without uncertainty.

Supplementary Files

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

- Additionalfile1SupplementaryTables.xlsx
- Additionalfile2SupplementaryFigures.pdf
- Additionalfile3SupplementaryLegends.docx