

Assessing the adoption of Causal Language and Methodology in Human Microbiome Studies: A Methodological Review

Albina Tskhay

McGill University

Alibek Moldakozhayev

Harvard Medical School

Cristina Longo

University of Montreal

Roxana Behruzi

McGill University

Celia Greenwood

McGill University

Stan Kubow

McGill University

Vadim N Gladyshev

Harvard Medical School

Tibor Schuster

`tibor.schuster@mcgill.ca`

McGill University

Research Article

Keywords: microbiome, methodological review, analysis strategies, causal inference, observational studies, confounding control, study design, effect estimation

Posted Date: February 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8681336/v1>

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

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Background

Microbiome research seeks to clarify how microbial communities influence human health. Although etiological research paradigms are evolving in the bio-medical sciences, many microbiome studies continue to rely on association-based methods that detect statistical patterns but are limited identifying causal mechanisms needed to inform clinical or public health interventions. This methodological review evaluates the extent to which modern causal inference approaches have been adopted in human microbiome studies and identifies persistent challenges to their broader implementation.

Methods

We systematically reviewed human microbiome studies published between 2019 and 2024 that examined links between the microbiome and health outcomes, or between exposures/interventions and microbiome composition, across ten high-impact journals identified using the Scimago Journal and Country Ranking. Eligible studies were retrieved from PubMed using a predefined search strategy. Two reviewers independently screened titles, abstracts and full texts and extracted data on study design, sampling, analytical framework, confounding control, effect size reporting, and the use of causal language. Analyses were performed using standardized extraction templates.

Results

Across 205 included studies, adoption of causal inference approaches in microbiome research remains limited. Only 15% of studies used designs or analytical strategies capable of approximating causal effects—12% were randomized controlled trials and 3% were observational studies employing formal causal inference methods. Longitudinal designs were common (45%). However, 30% of studies did not address confounding, and more than 40% did not report intervention-relevant or clinically actionable effect sizes. Studies making stronger causal claims were also more likely to propose intervention-relevant recommendations, regardless of the underlying study design.

Conclusion

The limited use of rigorous causal inference approaches remains a key barrier to producing actionable evidence in microbiome research. Greater adoption of principled confounding control, improved use of mediation and effect-modification frameworks, and more consistent reporting of interpretable effect sizes are necessary to strengthen causal claims. Advancing methodological standards and promoting interdisciplinary collaboration will be essential for translating microbiome findings into clinically meaningful insights.

BACKGROUND

Over the last two decades, microbiome research has gained increasing attention due to its potential to reveal complex interactions between microbial communities and human health (1–7). Despite advances in statistical and computational methods, challenges in addressing etiological questions remain pervasive in microbiome research (8). Randomized controlled trials are often infeasible, so the field relies heavily on observational data analyzed through association-based approaches that require additional assumptions to support causal interpretation (9). Central among these is the need for an adequate representation of the data-generating mechanisms induced by the study design, research design, sampling frameworks, and measurement approaches (8, 10).

Microbiome composition is influenced by a complex web of host and environmental factors, including diet, environment, genetics, and host physiology (11, 12). The interdependence of these variables creates challenges indisentangling causal relationships from spurious associations using observational studies (8). For example, distinguishing the effect of a specific microbial taxon from that of underlying host characteristics requires that key identifiability assumptions hold. Modern causal inference theory formalizes these assumptions and provides guidance on the conditions under which causal effects can be estimated from observational data (13–16). In particular, three assumptions are fundamental: consistency, which links observed outcomes to potential outcomes under the exposure actually received; exchangeability, which requires that, conditional on measured covariates, exposure assignment is independent of potential outcomes; and positivity, which ensures that all relevant exposure levels occur with nonzero probability within strata of the covariates.

Intrinsic properties of microbiome data can lead to violations of these key assumptions. In particular, high dimensionality and underdetermination make it difficult to measure or adjust for all relevant confounders, threatening exchangeability. Overdispersion and sparsity, including a large proportion of zeros, can complicate the definition and consistent measurement of treatment or exposures, affecting consistency. Time-varying confounding further challenges positivity and exchangeability, as confounders that are themselves affected by prior treatment or outcomes can change over time and influence subsequent treatment assignment (17–19).

These challenges are compounded by limitations in study design and measurement. Many microbiome studies are cross-sectional, limiting temporal inferences and, consequently, causal interpretation - unless strong external assumptions are met (10, 20). Even longitudinal studies are affected by heterogeneity in sampling, sequencing protocols, and batch effects, which threaten consistency (21). The microbiome's dynamic nature, changing over time and in response to environmental stimuli, further complicates the establishment of well-defined and generalizable causal relationships, thereby increasing the need for more complex study designs to capture these dynamics.

In contrast, related fields such as epidemiology and pharmacoepidemiology, have successfully adopted causal inference frameworks, emphasizing rigorous study design, careful confounding control, well-defined estimands, and transparent reporting (22). The relative underutilization of such frameworks in

microbiome research limits the ability to generate actionable insights and leaves critical questions about whether interventions on the microbiome could improve health outcomes.

Motivated by these limitations, this review systematically assesses the adoption of causal inference methods in human microbiome research. By analysing recent observational studies using 16S rRNA or metagenomic sequencing examining links between the microbiome and health outcomes, or between exposures/interventions and microbiome composition and published in leading journals, we aim to provide a comprehensive overview of the current state of causal inference in the field and outline directions for future methodological development.

METHODS

The study's objectives, design, and methods were registered with the Open Science Framework (OSF) <https://osf.io/bfn2q> (23).

Search

To identify publications that best represent microbiome research domain, we selected nine journals with the highest publication output in Microbiology Domain between 2019 and 2024 (i.e., excluding journals publishing only literature reviews) were identified from Scimago Journal and Country Ranking, which ranks journals based on the weighted number of citations received by a journal's publications (24). Scientific Reports was selected as tenth journal, as a non-microbiome-specific journal identified as having the highest number of microbiome-related publications in the same reference period (25).

Relevant studies published between 2019 and 2024 from each journal were identified through a PubMed search using a pre-specified search strategy (Supplementary material). Only studies that primarily used human gut microbiome data obtained by 16S rRNA or metagenomic shotgun sequencing and examining links between the microbiome and health outcomes, or between exposures/interventions and microbiome composition were included. These platforms were selected because they are the most widely used and methodologically mature approaches for characterizing gut microbial communities, providing sufficient taxonomic and/or functional resolution to support comparative analyses and causal inference-oriented study designs (26). Literature reviews, descriptive studies, pooled analysis of published studies, studies in animals, and studies that investigated viruses or eukaryotic organisms were excluded. The date range was chosen to cover studies appearing since the cost of 16S rRNA sequencing and newer technologies made it more possible to study larger sample sizes and hence associations with phenotypes (27).

Screening

Two reviewers (A.T. and A.M.) independently screened titles and abstracts, followed by full-text assessment when necessary. Screening was conducted separately within each journal with the aim of

identifying up to 25 eligible articles per journal for data extraction, for a target total of 250 articles. We deemed this number sufficient to quantify the proportion of articles that would exhibit certain characteristics: with 250 articles, would allow us to obtain a 95%CI half-width of less than 7 percentage points. For journals with more than 100 eligible records, title and abstract screening was capped at 100 articles, whereas for journals with fewer records, all identified articles were screened. Articles meeting the predefined eligibility criteria were advanced to full-text review until the journal-specific cap was reached.

Data extraction

Two rounds of data extraction were conducted on the included studies to minimize the risk of errors. Data was extracted on the following pre-defined aspects: study design, inferential framework, sample characteristics, data characteristics, statistical methods, mediation analysis and claims, confounding control and justification, number of citations (in Google Scholar), and reporting and action recommendation sentences. A detailed summary of all extraction variables can be found in Table S1.

Relevant data were systematically identified by analyzing the title, abstract, and conclusion sections of each article. Special attention was given to the outcomes and results emphasized by the authors, reflecting the central focus of each report. For each study, we then identified domains most pertinent to the reported findings and systematically extracted relevant details. This process included the following study specifics:

General Information

We recorded general information for each study, including the name(s) of the author(s), the study title, and the journal or report from which the data were extracted. For each study, the publication year and citation count were recorded. Citation data were retrieved from Google Scholar between October and December 2024.

Characteristics of Included Studies

The study design was classified into the following categories: randomized controlled trial, non-randomized trial, prospective cohort study, retrospective cohort study, case-control study, cross-sectional study, or nested case-control study. Based on the research questions, methodology, and conclusions, the inferential framework employed by each study was categorized as one of the following: statistical (hypothesis) testing, Bayesian inference, causal inference, supervised machine learning (e.g., prediction modeling or classification), or unsupervised learning (e.g., cluster analysis). The inferential framework was classified as causal inference if the study met at least one of the following predefined criteria: (i) the study was a randomized controlled trial; (ii) the authors explicitly stated a causal objective (i.e., to determine the causal effect of X on Y) or identified the analysis as causal; or (iii) the study was

observational but employed established causal inference methods aimed at satisfying identifiability assumptions, such as explicit control for confounding using propensity score–based approaches (e.g., matching, weighting, stratification), causal mediation analysis, or other counterfactual-based estimation strategies.

We extracted the total number of participants in each study, recorded the number of groups along with their respective sample sizes. The sampling method was recorded as either random or non-random sampling.

Exposure and Outcome Variables

The type of primary outcome was classified as binary, continuous, or categorical. Similarly, the type of primary exposure was categorized as binary, continuous, or categorical. Microbiome-related variables were further categorized according to their role in the presumed causal structure - as exposure, outcome, mediator, or moderator.

Microbiome Measures

We recorded the primary microbiome measures used to generate the main conclusions, including alpha diversity, beta diversity, relative abundance, or absolute abundance.

Statistical Models

We documented the statistical models used in each study to perform inferential analysis. Possible methods included nonparametric tests (e.g., Wilcoxon signed-rank test, Mann-Whitney U test, Kruskal-Wallis test), correlation measures (e.g., Spearman's or Pearson's correlation coefficients), hypothesis testing (e.g., t-tests, analysis of variance), regression models (e.g., logistic regression, linear regression, multinomial logistic regression, mixed-effects models), machine learning methods (e.g., random forests, support vector machines, deep learning), and microbiome-specific models (e.g., Linear Discriminant Analysis Effect Size), ANCOM (Analysis of Composition of Microbiomes), MaAsLin (Multivariate Association with Linear Models), PLS-DA (Partial Least Squares Discriminant Analysis))

Confounding Control

We examined whether confounding control was implemented and, if so, documented the strategy used. Strategies included covariate adjustment, inverse probability weighting (IPW), matching, restriction, standardization, or stratification. If no confounding control strategy was specified, this was noted. We also assessed whether the choice of confounding control was justified, categorizing the rationale as data-driven (e.g., difference testing or other methods), based on directed acyclic graphs (DAGs), informed by expert input, grounded in literature, or not addressed.

Multiple Comparisons

We assessed whether studies addressed the issue of multiple comparisons, which refers to the inflation of type I error resulting from conducting multiple statistical hypothesis tests. Specifically, we recorded whether any correction for multiple testing was reported (e.g., Bonferroni adjustment, false discovery rate control) and documented this as “yes” if any such method was applied, or “no” if the issue was not addressed.

Mediation and Modification Analyses

For each study, we examined whether mediation mechanisms were claimed and noted the corresponding sentence or paragraph and mediation analysis conducted. Similarly, we recorded whether effect modification was reported and type of analyses were conducted.

Assessment of Causal Language

We extracted the effect size measures reported in the studies, along with sentences used to report and interpret these measures. Where action recommendations or future research recommendations were provided, we recorded these as “yes” or “no” and captured the corresponding sentence(s). Statements were selected in accordance with the criteria described in reference (28).

We evaluated the use of causal versus associational language using two complementary approaches. First, a human reviewer (AT) assessed the strength of causal claims using the framework in (23), which defines causal language, action recommendations, and their grading (none, weak, moderate, strong). Second, a Large Language Model (LLM), ChatGPT based on the GPT-4 architecture (29), was used to assess both the strength of causal claims and the strength of practical recommendations. A standardized query was applied to all sentences (Supplementary material), and the model was blinded to other characteristics of the studies to minimize bias. This analysis was conducted in January 2025.

Software

For abstract and full-text screening and data extraction, we used Covidence, a web-based platform that streamlines systematic review workflows (30). The extraction templates within Covidence were customized to include the sections described above to ensure consistency and comprehensiveness in data collection. Descriptive graphs and tables were generated using R statistical software (version 4.3.0), which facilitated the analysis and visualization of the data (31).

RESULTS

Study selection and included sample

A total of 1,864 records were identified across the ten journals included in this review (Fig. 1 Standard systematic review guidelines such as PRISMA were not fully applicable to this methodological review (32); therefore, Fig. 1 presents a PRISMA-style flow diagram to transparently document the article

selection process. Following title and abstract screening, 679 records were screened, and 259 articles were assessed at the full-text level. Of these, 205 studies met the predefined methodological inclusion criteria and were included in the final analysis. As screening was conducted using a journal-stratified sampling strategy with a maximum of 25 included studies per journal, the number of screened and included articles varied across journals, reflecting differences in publication volume and eligibility rather than differential exclusion at later stages of review.

This figure summarizes the study selection process across ten journals. Screening was conducted sequentially within each journal until either the prespecified target of 25 eligible publications was reached or no additional publications were available for screening. Consequently, the number of records screened and studies included varies by journal. In total, data were extracted from 205 publications. Because the objective of this review was to characterize the landscape of methodological approaches rather than to exhaustively synthesize all available evidence, a journal-stratified sampling strategy was applied with a maximum of 25 included studies per journal. The flow diagram therefore represents a purposeful screening process rather than complete enumeration.

Study designs, sample sizes, and group structure

The publication and methodological characteristics of the 205 included studies are summarized in Table 1. The included studies comprised primarily prospective cohort studies (28.8%, n = 59), cross-sectional studies (26.8%, n = 55), and case-control studies (23.9%, n = 49), with fewer randomized controlled trials (12.2%, n = 25), pre-post studies (4.4%, n = 9), nested case-control studies (2.4%, n = 5), and retrospective cohort studies (1.5%, n = 3).

Table 1
Distribution of study characteristics across 205 included studies.

Study characteristics		Number (out of 205)	Percentage (%)
Journal	<i>Cell Host & Microbe</i>	25	12.2
	<i>Gut Microbes</i>	25	12.2
	<i>Microbiome</i>	25	12.2
	<i>npj Biofilms and Microbiomes</i>	25	12.2
	<i>Scientific Reports</i>	25	12.2
	<i>mSystems</i>	24	11.7
	<i>mSphere</i>	18	8.8
	<i>mBio</i>	18	8.8
	<i>Nature Microbiology</i>	15	7.3
	<i>ISME</i>	5	2.4
Year of publication	2019	16	7.8
	2020	36	17.6
	2021	39	19.1
	2022	43	21.1
	2023	39	19.1
	2024	31	15.2
Study design	Prospective cohort study	59	28.8
	Cross-sectional study	55	26.8
	Case-control study	49	23.9
	Randomized controlled trial	25	12.2
	Pre-post study	9	4.4
	Nested case-control study	5	2.4

This table summarizes key methodological and reporting features of the included studies, including publication year, study design, analytical approach, role of the microbiome in the causal structure, confounding control strategies, outcome modelling practices, and reporting of effect sizes. Counts and percentages reflect the number of studies exhibiting each characteristic out of the total sample (N = 205).

Study characteristics		Number (out of 205)	Percentage (%)
	Retrospective cohort study	3	1.5
Inferential framework	Statistical (hypothesis) testing	160	78
	Causal inference	31	15.1
	Supervised machine learning	8	3.9
	Unsupervised machine learning	4	2
	Bayesian inference	2	1
Role of microbiome studied	Exposure	77	37.6
	Outcome	75	36.6
	Mediator	46	22.4
	Moderator/ Effect modifier	7	3.4
Confounding control strategy	Adjustment	101	49.3
	None	71	34.6
	Group-level matching	27	13.2
	Stratification	4	2
	Restriction	2	1
Confounding control justification	None	71	34.6
	Difference testing	46	22.4
	Data-driven	10	4.9
	Literature-driven	10	4.9
	DAG	2	1
Multiple comparison control	Yes	161	78.5
Mediation mechanism claimed	Yes	115	56.1

This table summarizes key methodological and reporting features of the included studies, including publication year, study design, analytical approach, role of the microbiome in the causal structure, confounding control strategies, outcome modelling practices, and reporting of effect sizes. Counts and percentages reflect the number of studies exhibiting each characteristic out of the total sample (N = 205).

Study characteristics		Number (out of 205)	Percentage (%)
Moderation mechanism claimed	Yes	50	24.4
Effect size measure	p-value	51	24.9
	Correlation coefficient	44	21.5
	Regression coefficient	24	11.7
	LEfSe	21	10.2
	Log-fold change	20	9.8
	Coefficient of determination	14	6.8
	Area under the curve	11	5.4
	Odds ratio	9	4.4
	Other	11	5.4
Future research recommended	Yes	178	86.8
Practical recommendations given	Yes	105	51.2
This table summarizes key methodological and reporting features of the included studies, including publication year, study design, analytical approach, role of the microbiome in the causal structure, confounding control strategies, outcome modelling practices, and reporting of effect sizes. Counts and percentages reflect the number of studies exhibiting each characteristic out of the total sample (N = 205).			

The average number of human participants across studies was 2331 (range 8–422417), with a median sample size of 91. Case-control studies showed the greatest heterogeneity in sample size (median = 98; range = 20–422417), driven by a small number of very large studies. Prospective cohort and cross-sectional studies were moderately sized (medians = 81 and 216, respectively), while randomized controlled trials and pre-post studies tended to include smaller cohorts (medians = 56 and 20 respectively).

More than half of the studies (53.2%, n = 109) compared two study groups. An additional 15.1% of studies (n = 31) focused on a single study group, typically within longitudinal designs or subgroup analyses. Among single-group studies, the median sample size was 43 (range: 8–1054), whereas two-group studies had a median sample size of 82 (range: 9 – 3890). One study analyzed 16 distinct cohorts, representing the largest number of groups included in a single analysis.

Conceptual role of the microbiome in study designs

Across the included studies, the microbiome was most frequently conceptualized as an exposure variable (37.6%), followed closely by its role as an outcome (36.6%) (Table 1). In 22.0% and 3.4% of studies, the microbiome was analyzed as a mediating variable and effect modifier, respectively.

Variable types and operationalization across microbiome roles

To examine how conceptual roles translated into analytical practice, we assessed the types of variables used to represent exposures and outcomes when the microbiome was studied as an exposure, outcome, mediator, or effect modifier (Fig. 2).

When the microbiome was analyzed as an exposure ($n = 77$), continuous microbiome measures dominated (74/77, 96.1%). The most common configuration paired binary outcomes with continuous microbiome exposures (38/77, 49.4%), followed by continuous outcomes with categorical microbiome exposures (27/77, 35.1%) and continuous outcomes with continuous microbiome exposures (9/77, 11.7%). Binary or discrete microbiome exposures were rare, together accounting for only 3.9% of studies.

In studies treating the microbiome as an outcome ($n = 75$), two main patterns emerged. Continuous microbiome outcomes were most frequently paired with binary exposures (38/75, 50.7%) or discrete exposures (32/75, 42.7%). Continuous exposures were uncommon (5/75, 6.7%), mostly appearing in continuous–continuous configurations (4/75, 5.3%).

When the microbiome functioned as a mediating variable ($n = 46$), the most frequent configurations involved continuous outcomes. Continuous outcomes paired with binary exposures accounted for 30.4% of studies (14/46), while continuous outcomes paired with discrete exposures represented 23.9% (11/46). Other combinations were less common and more evenly distributed, with no single alternative pattern exceeding 13% of studies.

Finally, in the smaller subset of studies examining the microbiome as a moderator ($n = 7$), patterns were heterogeneous. Discrete or binary exposures predominated (5/7, 71.4%), most commonly paired with categorical or continuous outcomes (each 2/7, 28.6%). All other exposure–outcome combinations occurred once or not at all.

Each heatmap depicts the frequency of combinations of exposure and outcome variable types (binary, categorical, continuous) used when the microbiome is studied as an exposure, outcome, mediator, or effect modifier.

Statistical methods used in microbiome analyses

The overall landscape of statistical methods employed across the included studies is shown in Fig. 3a. In total, 51 distinct analytical methods were identified and grouped into ten broad categories, including non-parametric tests, parametric tests, compositional analyses, correlation-based methods, linear and

generalized linear models, machine learning approaches, causal inference methods, survival analysis, Bayesian methods, and other approaches.

When grouped by microbiome role and metric, methodological heterogeneity varied considerably. In studies with the microbiome as an exposure (Fig. 3b), alpha diversity analyses ($n = 10$) were dominated by non-parametric tests (7/10, 70%), with the remaining analyses using a mix of parametric, compositional, and survival methods. Beta diversity was assessed exclusively with non-parametric methods (2/2). Relative abundance studies ($n = 125$) employed a wide range of approaches: non-parametric tests were most frequent (28/125, 22.4%), followed by compositional analyses (24/125, 19.2%), parametric tests (15/125, 12.0%), generalized linear models and machine learning (each 12/125, 9.6%), correlation analyses (11/125, 8.8%), causal inference methods (2/125, 1.6%), survival analysis (1/125, 0.8%), and other methods (20/125, 16.0%).

Panels summarize the number of studies applying each method type when analyzing exposures ($n = 77$), outcomes ($n = 75$), alpha diversity ($n = 41$), beta diversity ($n = 16$), and relative abundance ($n = 314$). Bars represent counts of studies using non-parametric, compositional, parametric, generalized linear, machine-learning, correlation-based, causal-inference, survival, Bayesian, and other analytical approaches.

In studies with the microbiome as an outcome variable (Fig. 3c), alpha diversity analyses ($n = 20$) were primarily non-parametric (11/20, 55%), with parametric tests (4/20, 20%), generalized linear models (2/20, 10%), and other methods (3/20, 15%). Beta diversity comparisons ($n = 11$) relied mostly on non-parametric approaches (9/11, 81.8%), with the remaining studies using parametric and other methods (2/11, 18.2%). Relative abundance studies ($n = 106$) showed high heterogeneity: non-parametric tests were most common (35/106, 33.0%), followed by compositional analyses (21/106, 19.8%), parametric tests (13/106, 12.3%), generalized linear models (12/106, 11.3%), other methods (14/106, 13.2%), machine learning (6/106, 5.7%), and correlation analyses (5/106, 4.7%).

Reported effect size measures and uncertainty

Across the included studies, reporting practices showed the predominance of difference-testing approaches. The p-value alone or along with other measures was the most commonly reported measure overall, appearing in 25% of studies (Fig. 4). Correlation coefficients (21.5%) and regression coefficients (11.7%) were the next most frequently reported effect size measures. Only 10.2% of studies ($n = 21$) reported confidence intervals for effect size estimates, limiting the interpretability and comparability of reported findings.

This faceted bar chart illustrates the distribution of commonly reported effect size measures across different microbiome roles.

The choice of effect measures varied by the conceptual role of the microbiome. In microbiome-as-exposure studies, correlation coefficients were most frequently reported (16/77, 20.8%), followed by a Linear discriminant analysis effect size (LEfSe) (11/77, 14.3%), area under the curve or AUC (9/77,

11.7%), regression coefficients (9/77, 11.7%), and log fold-change (7/77, 9.1%), while hazard ratios were rarely reported (2/77, XX %). In microbiome-as-outcome studies, correlation coefficients (16/75, 21.3%) and regression coefficients (10/75, 13.3%) were most commonly estimated, followed by the coefficient of determination (7/75, 9.3%), log fold-change (6/75, 8.0%), and LEfSe (5/75, 6.7%), with fewer instances of the odds ratio (2/75), fold-change (1/75), or incidence ratio (1/75). In microbiome-as-mediators studies, correlation coefficients (11/46, 23.9%), the log fold-change (6/46, 13.0%), regression coefficients (5/46, 10.9%), and LEfSe (5/46, 10.9%) were primarily reported, with rarer instances for the F-statistic (1/46), hazard ratio (1/46), and Mendelian randomization effect size (1/46). Moderator studies were less consistent, with only a few reporting AUC (2/7), coefficient of determination (2/7), correlation coefficient (1/7), LEfSe (1/7), and no studies reporting regression or risk ratios.

Confounding Control in Observational Studies

Among the 171 observational studies (excluding randomized controlled trials and pre-post designs), 32.0% (n = 54) did not apply any form of confounding control (Fig. 5). Of these, 70.0% (n = 38) provided no justification for the absence of confounding control, while 27.8% (n = 15) relied on difference testing to assess covariate balance between groups.

Among studies that implemented confounding control, adjustment was the most common strategy (52.0%, n = 89), followed by group-level matching (14.0%, n = 24). Restriction and stratification were rarely used, each appearing in 1.2% of studies (not shown in the Figure). Among studies that adjusted for confounding, 52.8% (n = 47) did not justify their selection of covariates, and 31.5% (n = 28) relied on significance testing to guide covariate inclusion.

Mediation and Mechanistic Claims

More than half of the included studies (53.2%, n = 109) reported conducting mediation analyses, and 56.1% (n = 115) explicitly claimed a mediation mechanism. Overall, 55% of studies treating the microbiome as an exposure and 33.3% of studies where microbiome was an outcome made claims about mediation mechanisms. Mediation claims were most common in studies examining the microbiome as a mediator (38.3%, n = 44) or as an exposure (37.4%, n = 43).

The most frequently used method to support mediation claims was Spearman's correlation (39.1%), followed by Pearson's correlation (7.0%). A small proportion of studies supported mediation claims using additional laboratory/animal experiments (6.1%), while an equal proportion (6.1%) claimed causal mechanisms without providing analytical evidence.

Strength of Causal Claims and Practical Recommendations

In the assessment of the strength of causal claims, overall agreement between the human's and LLM's assessments was high, with the LLM tending to assign slightly more conservative ratings, (Figs. 6a and 6b). The human reviewer identified 34 studies as making strong causal claims, whereas GPT identified

18. Conversely, 115 studies were rated as making no or weak causal claims by the human reviewer, compared with 136 by GPT.

Regarding practical recommendations, 51.2% of studies (n = 105) provided recommendations for implementation. Case-control studies reported causal recommendations in 25 of 49 studies (51.0%). Randomized controlled trials provided recommendations in 15 of 25 studies (60%) and prospective cohort studies did so in 33 of 59 (55.9%). Nested case-control studies provided recommendations in 4 of 5 studies. Pre-post studies reported recommendations in 4 of 9 studies. Cross-sectional studies were less likely to lead to such recommendations, with 23 of 55 reporting recommendations (51.0%). and retrospective cohort studies rarely did so, with only 1 of 3 reporting recommendations.

Among studies that provided recommendations, most issued weak or moderate recommendations, and no study issued a strong recommendation (Figs. 6c and 6d). Studies with weaker causal claims tended to issue weaker recommendations, although a small number of studies issued recommendations despite making no explicit causal claims.

Finally, the strength of causal claims varied by the conceptual role of the microbiome (Fig. 6e). Studies treating the microbiome as an exposure most commonly made no or weak causal claims, whereas moderate and strong claims were more frequent when the microbiome was examined as a mediator or outcome. Studies considering the microbiome as an effect modifier tended toward weak or moderate claims, although the small number of such studies limits interpretation.

DISCUSSION

This methodological review demonstrates that, despite rapid growth and methodological innovation in microbiome research, rigorous causal inference approaches remain infrequently applied. Most studies continue to rely on association-based analyses, with limited attention to confounding control, counterfactual reasoning, and clinically interpretable effect estimation.

Study design and temporal structure

One notable pattern observed in the reviewed literature is a shift toward more temporally structured study designs. In a prior systematic review by Bardenhorst et al. (10), covering studies published in 2018–2019, only 27% employed longitudinal designs. In contrast, 45% of studies in the present review used longitudinal or prospective designs. This change is consistent with increased recognition of temporality as a key component of causal interpretation. However, while improved temporal ordering strengthens causal plausibility, it does not, on its own, address other core challenges related to confounding, measurement, or effect interpretation (33, 34).

Microbiome role, variable representation, and causal consistency

A particularly informative finding concerns how variable types are distributed according to the microbiome's conceptual role in the causal model. When treated as an exposure, microbiome features were overwhelmingly represented as continuous measures, such as relative abundances or diversity indices. Although ecologically intuitive, this dominance of broad, continuously defined exposures complicates causal interpretation. Continuous microbiome measures often lack a clear intervention analogue, making it difficult to define well-specified causal estimands or to satisfy the consistency assumption, which requires that exposure levels correspond to coherent interventions.(35) Moreover, the limited use of causal frameworks explicitly designed to handle continuous, high-dimensional exposures further constrains interpretability. As a result, even when statistical associations are robust, the corresponding causal questions often remain unclear, and effect estimates may be difficult to compare across studies or translate into meaningful biological or clinical interpretations.

Confounding control and variable selection

A recurring limitation across studies was the inconsistent handling and reporting of confounding. Although demographic, behavioral, and clinical covariates were frequently collected, few studies clearly articulated the rationale for confounder selection or employed formal tools such as directed acyclic graphs (DAGs) to justify adjustment strategies. Design-based approaches, including matching or restriction, were also rarely reported.

In many cases, confounder selection appeared to rely on difference testing or data-driven significance thresholds as recommended by some current guidelines including the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) (36). This approach is problematic, as it may fail to identify variables that strongly influence the outcome but are weakly associated with the exposure, leaving residual confounding unaddressed (37). It is not expected that a fixed predefined set of confounders would be accounted for in microbiome studies, because ideally the confounder set is to be tailored to the specific research question, design, and population. However, as shown by Vujkovic-Cvijin et al. (38), factors such as sex, BMI, diet, alcohol consumption, and bowel movement quality explain substantial variation in gut microbiota composition and may confound disease associations if not appropriately accounted for. The inconsistent treatment of such variables reflects a broader challenge in aligning statistical practices with causal reasoning in microbiome research.

Effect and uncertainty measures reporting and interpretability

Another prominent pattern concerns the limited reporting of effect and uncertainty measures. Many studies relied primarily on p-values alone to convey results, while fewer reported effect estimates, and only a small fraction provided confidence intervals. This practice restricts interpretability, as statistical significance alone does not convey the magnitude, direction, or precision of an effect.

As noted in the clinical and epidemiologic literature (39, 40) small and clinically irrelevant effects may achieve statistical significance in large samples, whereas meaningful effects may go undetected in smaller studies. Without effect sizes and confidence intervals, it is difficult to assess the practical

relevance of reported findings or to compare results across studies. The observed variability in reported effect measures further limits synthesis, particularly when microbiome features are analyzed on different scales or treated differently across causal roles.

Mediation analyses and reliance on correlations

Mediation analyses represent another area where current practices impose substantial limitations on causal interpretation. Spearman's correlation was frequently used to explore relationships among exposures, microbiome features, and outcomes, often as a proxy for mediation. While correlation analyses can reveal associations suggestive of potential pathways, they do not establish causal mechanisms.

Several challenges arise in this context. First, correlation measures do not distinguish cause from effect or account for exposure–mediator interactions. Second, microbiome-wide correlation analyses implicitly assume independence among microbial features, an assumption violated by compositional constraints and microbial interdependence. Third, correlation-based approaches typically do not control for confounders influencing both the mediator and the outcome, increasing the risk of spurious associations (41, 42).

As a result, while correlation analyses may be useful for exploratory purposes, they offer limited insight into causal mediation. More formal causal mediation frameworks, such as those developed by VanderWeele (43, 44) explicitly define direct and indirect effects under counterfactual assumptions and can, in principle, accommodate treatment–microbiome interactions. The limited adoption of such frameworks in the reviewed literature underscores a gap between the complexity of the scientific questions being posed and the analytical tools commonly applied.

Strength of causal claims and reported recommendations

Importantly, this review highlights that relatively few microbiome studies make strong explicit causal claims or offer practical recommendations. Studies that did do so were more commonly those based on stronger designs, such as randomized controlled trials, pre–post interventions, or longitudinal cohort studies. This pattern suggests that, despite substantial analytical effort across the field, much of the published microbiome literature does not yet translate into intervention-relevant conclusions.

The limited prevalence of causal claims and actionable recommendations underscores the need for broader adoption of causal inference frameworks in microbiome research. Without clearly articulated causal questions and corresponding analytical strategies, large volumes of data may yield primarily associative findings with restricted downstream utility. More systematic use of causal inference techniques has the potential to improve research efficiency by enabling clearer interpretation of results, better prioritization of follow-up studies, and more direct relevance to clinical or public health decision-making.

Implications for the field

Taken together, these findings highlight slow adoption of the application and reporting of causal inference in microbiome research. These limitations, ranging from variable definition and confounding control to effect measure and size reporting and mediation analysis, do not diminish the value of existing studies but rather delineate the boundaries of what current practices can support. Causal inference frameworks offer a conceptual structure that may help address these challenges, but their potential lies in clarifying assumptions, estimands, and interpretability rather than in providing prescriptive solutions.

This is the first methodological review to summarize the methodological landscape, informing future efforts to align research questions, analytical strategies, and causal claims in microbiome science.

Limitations

This study has several limitations. First, our review was restricted to 205 research articles published in 10 journals. While these journals are highly regarded in the microbiome field, the selection includes only one broad-topic journal, potentially limiting the generalizability of our findings. Standards for causal inference and reporting practices may vary across journals, especially in broader biomedical or interdisciplinary domains. By focusing on the top journals within the microbiome domain, our aim was not to provide an exhaustive survey but to identify prevailing trends, commonly used frameworks, and key methodological challenges in current causal inference practices. Second, stratifying by microbiome role or microbiome measures further reduced the number of studies available for analysis. This limitation is particularly relevant for studies examining the microbiome as a moderator, which were limited to only seven studies, preventing strong conclusions about trends in this specific subgroup. Third, many microbiome studies employ multiple analytical approaches rather than a single method. To facilitate mapping of current methodologies, we extracted data on up to two main analyses per study. Nonetheless, some additional analyses were not included, meaning the overview provided here does not fully capture the complete methodological landscape. Fourth, microbiome research is inherently complex, with study questions, microbiome measures, and analytical techniques varying widely. In this review, we stratified studies only by the microbiome's conceptual role and the type of microbiome measure. Further stratifications could provide a more nuanced understanding of patterns and trends in current practices but were beyond the scope of this analysis. Finally, the assessment of causal language and the strength of recommendations involved a degree of subjectivity. Although we employed a standardized rubric and resolved discrepancies through consensus discussions, some level of interpretive ambiguity is unavoidable.

CONCLUSION

As microbiome research matures, there is increasing interest in moving beyond descriptive associations toward robust, interpretable causal claims. This review highlights both encouraging developments and areas of concern in the field's current methodological practices. The rise in longitudinal and

interventional designs is a promising step toward satisfying the fundamental requirement of temporality. Yet, this shift has not been consistently matched by appropriate analytic frameworks or transparent reporting. Moreover, meaningful effect estimation remains limited.

By identifying current patterns and gaps, this review highlights the pressing need for methodological innovations and interdisciplinary collaborations between microbiome researchers, statisticians, and epidemiologists to strengthen causal inference practices in the field. Addressing these challenges is essential to unlocking the full potential of microbiome studies to inform precision medicine, public health interventions, and our broader understanding of human health and disease.

Declarations

Conflicts of interest

No conflicts of interest declared

Funding

We acknowledge the Fonds de Recherche du Québec-Santé (FRQ-S) for the doctoral award to Albina Tskhay and Alibek Moldakozhayev. FRQ-S, did not play any role in the design, analysis or interpretation of results.

Authors' contributions

Albina Tskhay designed the data extraction template; screened abstracts and full-text articles; extracted and analysed the data; interpreted the results; and drafted the manuscript. Alibek Moldakozhayev screened abstracts and full-text articles, extracted data, interpreted the results, and critically revised the manuscript for important intellectual content. Cristina Longo co-supervised Albina Tskhay, contributed to the design of the data extraction template, interpreted the results, and critically revised the manuscript for important intellectual content. Roxana Behruzi, Celia Greenwood, Stan Kubow contributed to the interpretation of the results and critically revised the manuscript for important intellectual content. Vadim N. Gladyshev supervised Alibek Moldakozhayev and critically revised the manuscript for important intellectual content. Tibor Schusted supervised Albina Tskhay, contributed to the design of the data extraction template, interpreted the results, and critically revised the manuscript for important intellectual content.

References

1. Helmink BA, Khan MAW, Hermann A, Gopalakrishnan V, Wargo JA (2019) The microbiome, cancer, and cancer therapy. *Nat Med* 25(3):377–388
2. Hsiao EY, McBride SW, Hsien S, Sharon G, Hyde ER, McCue T et al (2013) Microbiota modulate behavioral and physiological abnormalities associated with neurodevelopmental disorders. *Cell* 155(7):1451–1463
3. Li B, Selmi C, Tang R, Gershwin ME, Ma X (2018) The microbiome and autoimmunity: a paradigm from the gut–liver axis. *Cell Mol Immunol* 15(6):595–609
4. MacQueen G, Surette M, Moayyedi P (2017) The gut microbiota and psychiatric illness. *J Psychiatry Neurosci* 42(2):75–77
5. Roy Sarkar S, Banerjee S (2019) Gut microbiota in neurodegenerative disorders. *J Neuroimmunol* 328:98–104
6. Wang P-X, Deng X-R, Zhang C-H, Yuan H-J (2020) Gut microbiota and metabolic syndrome. *Chin Med J* 133(7):808–816
7. Alibek K, Farmer S, Tskhay A, Moldakozhayev A, Alibek K, Isakov T (2019) The Role of Infection, Inflammation and Genetic Alterations in ASD Etiopathogenesis: A Review. *J Neurol Psychiatr Disord* 2:105
8. Relman DA (2020) Thinking about the microbiome as a causal factor in human health and disease: philosophical and experimental considerations. *Curr Opin Microbiol* 54:119–126
9. Lin H, Peddada SD (2020) Analysis of microbial compositions: a review of normalization and differential abundance analysis. *npj Biofilms Microbiomes* 6(1):60
10. Bardenhorst SK, Berger T, Klawonn F, Vital M, Karch A, Rübsamen N (2021) Syst Rev Curr Pract mSystems 6(1). 10.1128/msystems.01154-20. Data Analysis Strategies for Microbiome Studies in Human Populations—a
11. Xia Y, Sun J, Chen D-G (2018) What Are Microbiome Data? Springer Singapore, pp 29–41
12. Lutz KC, Jiang S, Neugent ML, De Nisco NJ, Zhan X, Li Q (2022) A Survey of Statistical Methods for Microbiome Data Analysis. *Front Appl Math Stat.* ;8
13. Nearing JT, Douglas GM, Hayes MG, MacDonald J, Desai DK, Allward N et al (2022) Microbiome differential abundance methods produce different results across 38 datasets. *Nat Commun* 13(1):342
14. VanderWeele TJ, Hernán MA (2013) Causal Inference Under Multiple Versions of Treatment. *J Causal Inference* 1(1):1–20
15. Lanza ST, Moore JE, Butera NM (2013) Drawing causal inferences using propensity scores: a practical guide for community psychologists. *Am J Community Psychol* 52(3–4):380–392
16. Lousdal ML (2018) An introduction to instrumental variable assumptions, validation and estimation. *Emerg Themes Epidemiol* 15(1):1
17. Xia Y, Sun J, Chen D-G (2018) What are microbiome data? Statistical analysis of microbiome data with R: Springer; pp. 29–41

18. Lutz KC, Jiang S, Neugent ML, De Nisco NJ, Zhan X, Li Q (2022) A survey of statistical methods for microbiome data analysis. *Front Appl Math Stat* 8:884810
19. Wozniak H, Gaïa N, Lazarevic V, Le Terrier C, Beckmann TS, Balzani E et al (2024) Early reduction in gut microbiota diversity in critically ill patients is associated with mortality. *Ann Intensiv Care* 14(1):174
20. Barnett TA, Koushik A, Schuster T (2023) Invited Commentary: Cross-Sectional Studies and Causal Inference-It's Complicated. *Am J Epidemiol* 192(4):517–519
21. Debelius J, Song SJ, Vazquez-Baeza Y, Xu ZZ, Gonzalez A, Knight R (2016) Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome Biol* 17(1):217
22. Fischbach MA (2018) Microbiome: focus on causation and mechanism. *Cell* 174(4):785–790
23. Tskhay A, Moldakozhayev A, Longo C, Schuster T (2024) Methodological Evaluation in Microbiome Research: Approaches Employed to Study Microbiome as an Exposure and Outcome Variable. *Systematic Review Pre-Registration*
24. SCImago SJR – SCImago Journal & Country Rank [Portal] n.d [Available from: <http://www.scimagojr.com>]
25. Huang Z, Liu K, Ma W, Li D, Mo T, Liu Q (2022) The gut microbiome in human health and disease—Where are we and where are we going? A bibliometric analysis. *Front Microbiol.* ;13
26. Knight R, Vrbancac A, Taylor BC, Aksenov A, Callewaert C, Debelius J et al (2018) Best practices for analysing microbiomes. *Nat Rev Microbiol* 16(7):410–422
27. Minich JJ, Humphrey G, Benitez RAS, Sanders J, Swafford A, Allen EE, Knight R (2018) High-Throughput Miniaturized 16S rRNA Amplicon Library Preparation Reduces Costs while Preserving Microbiome Integrity. *mSystems* 3(6). 10.1128/msystems.00166 – 18
28. Haber NA, Wieten SE, Rohrer JM, Arah OA, Tennant PWG, Stuart EA et al (2022) Causal and Associational Language in Observational Health Research: A Systematic Evaluation. *Am J Epidemiol* 191(12):2084–2097
29. OpenAI GPT May 13 ed2024
30. Innovation VH Covidence systematic review software Melbourne, Australia
31. de Micheaux PL, Drouilhet R, Liqueur B (2013) The R software. *Fundamentals of Programming and Statistical Analysis.* :978-1
32. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD et al (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* 372:n71
33. Haine D, Dohoo I, Dufour S (2018) Selection and misclassification biases in longitudinal studies. *Front veterinary Sci* 5:99
34. Laird NM (1988) Missing data in longitudinal studies. *Stat Med* 7(1–2):305–315
35. Chatton A, Rohrer JM (2024) The Causal Cookbook: Recipes for Propensity Scores, G-Computation, and Doubly Robust Standardization. *Adv Methods Practices Psychol Sci* 7(1):25152459241236149

36. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP (2007) The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Ann Intern Med* 147(8):573–577
37. Sourial N, Vedel I, Le Berre M, Schuster T (2019) Testing group differences for confounder selection in nonrandomized studies: flawed practice. *CMAJ* 191(43):E1189–e93
38. Vujkovic-Cvijin I, Sklar J, Jiang L, Natarajan L, Knight R, Belkaid Y (2020) Host variables confound gut microbiota studies of human disease. *Nature* 587(7834):448–454
39. Sullivan GM, Feinn R (2012) Using Effect Size—or Why the P Value Is Not Enough. *J Graduate Med Educ* 4(3):279–282
40. Gardner MJ, Altman DG (1986) Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J (Clin Res Ed)* 292(6522):746–750
41. Janse RJ, Hoekstra T, Jager KJ, Zoccali C, Tripepi G, Dekker FW, van Diepen M (2021) Conducting correlation analysis: important limitations and pitfalls. *Clin Kidney J* 14(11):2332–2337
42. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ (2017) Microbiome Datasets Are Compositional: And This Is Not Optional. *Front Microbiol* 8:2224
43. VanderWeele TJ (2013) A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology* 24(2):224–232
44. VanderWeele TJ (2014) A unification of mediation and interaction: a 4-way decomposition. *Epidemiology* 25(5):749–761

Figures

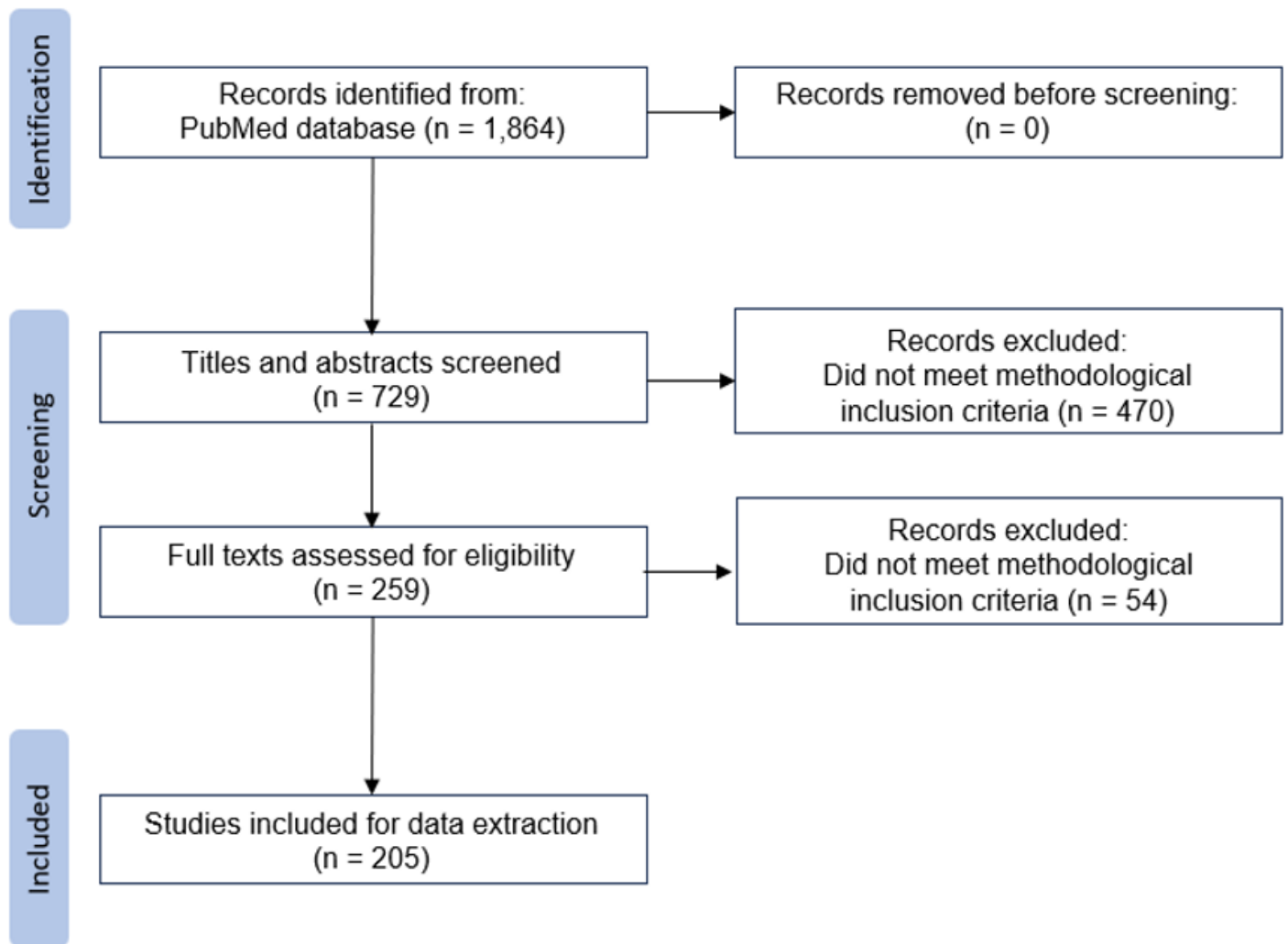


Figure 1

PRISMA flow diagram of study identification, screening, and inclusion

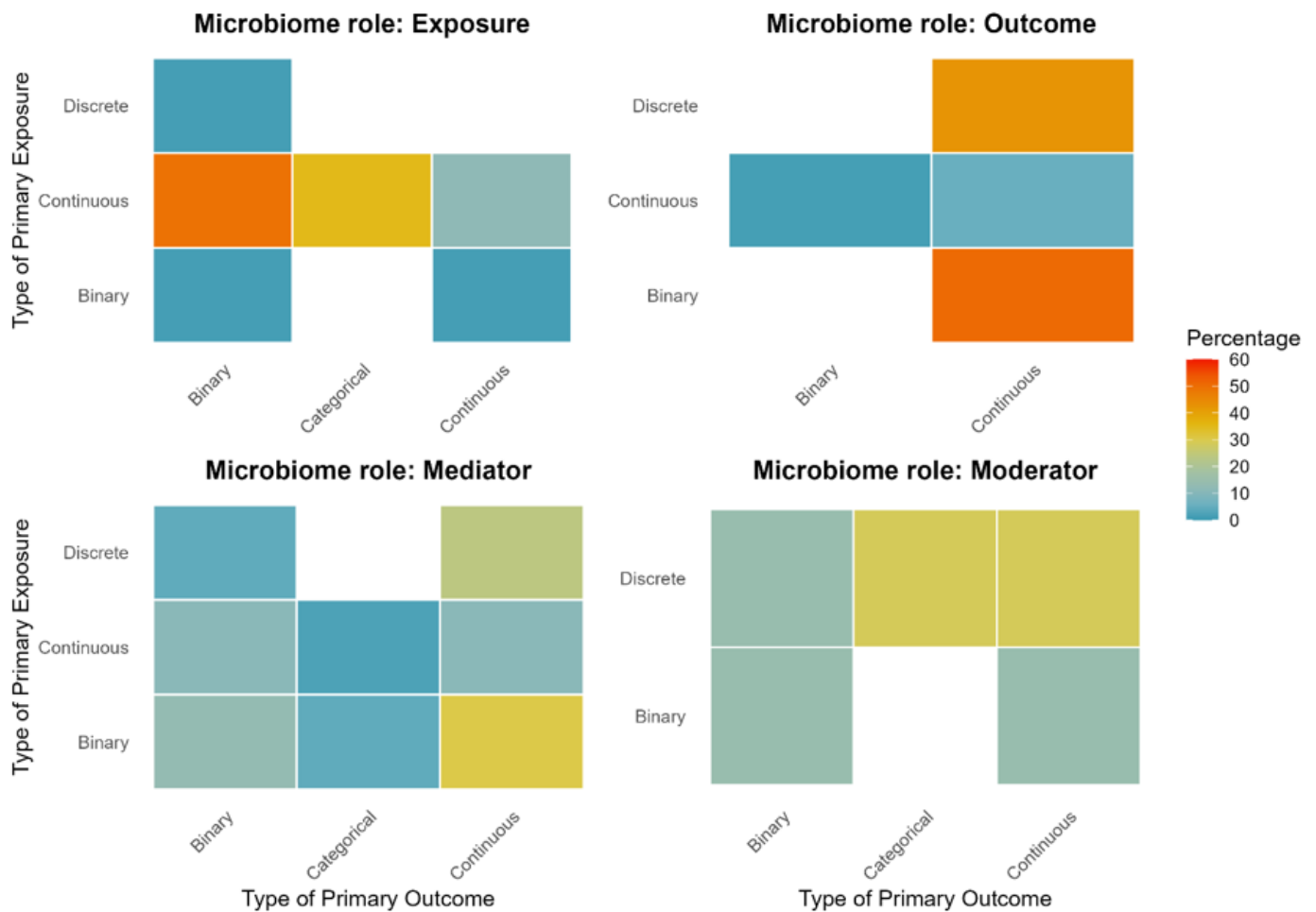


Figure 2

Variable Type Combinations in Microbiome Research Across Causal Roles.

Each heatmap depicts the frequency of combinations of exposure and outcome variable types (binary, categorical, continuous) used when the microbiome is studied as an exposure, outcome, mediator, or effect modifier.

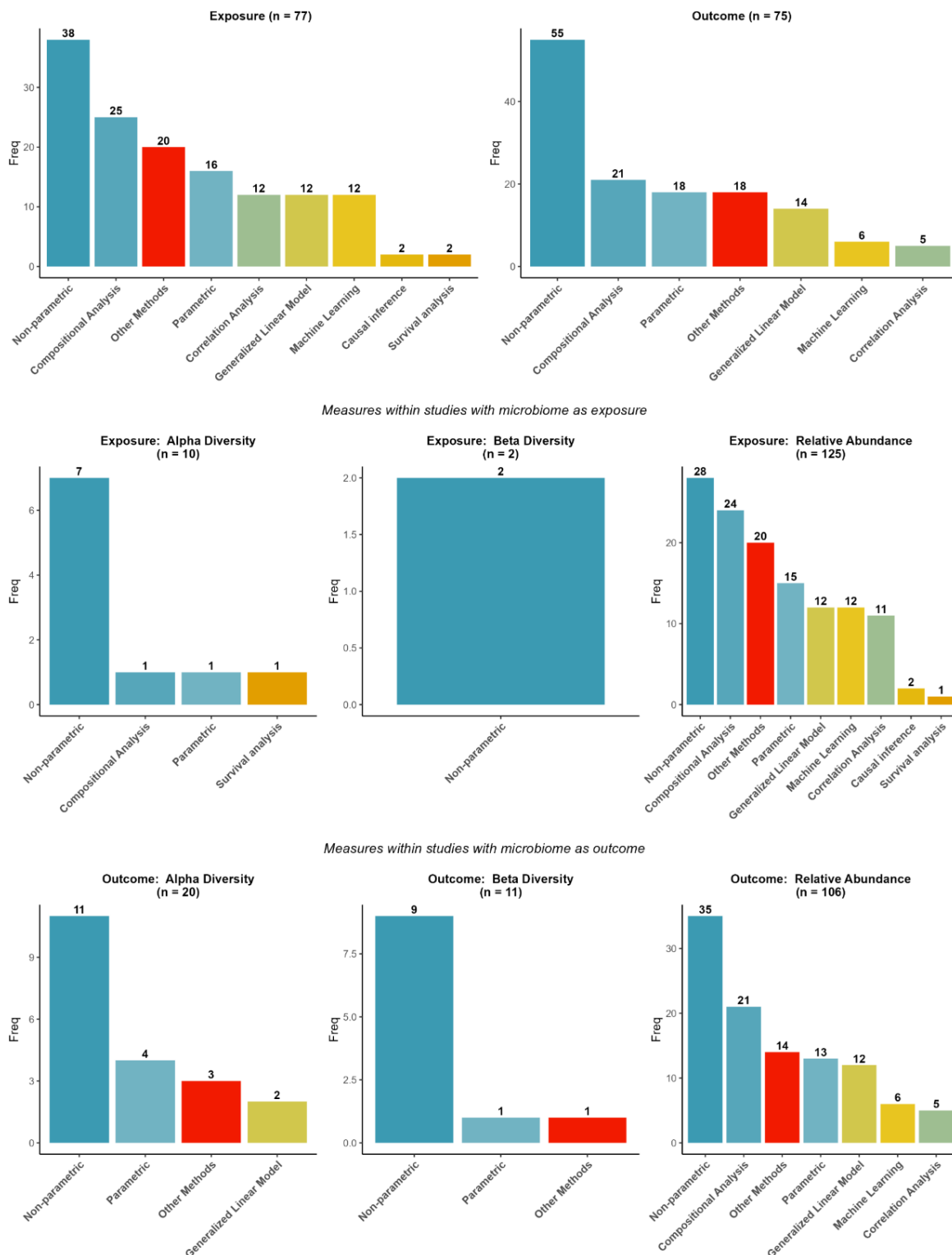


Figure 3

Frequency of Each Method Category Used in Studies to Arrive to The Main Conclusions Grouped by (A) Category and (B) Primary Microbiome Measures.

Panels summarize the number of studies applying each method type when analyzing exposures (n = 77), outcomes (n = 75), alpha diversity (n = 41), beta diversity (n = 16), and relative abundance (n = 314). Bars

represent counts of studies using non-parametric, compositional, parametric, generalized linear, machine-learning, correlation-based, causal-inference, survival, Bayesian, and other analytical approaches.

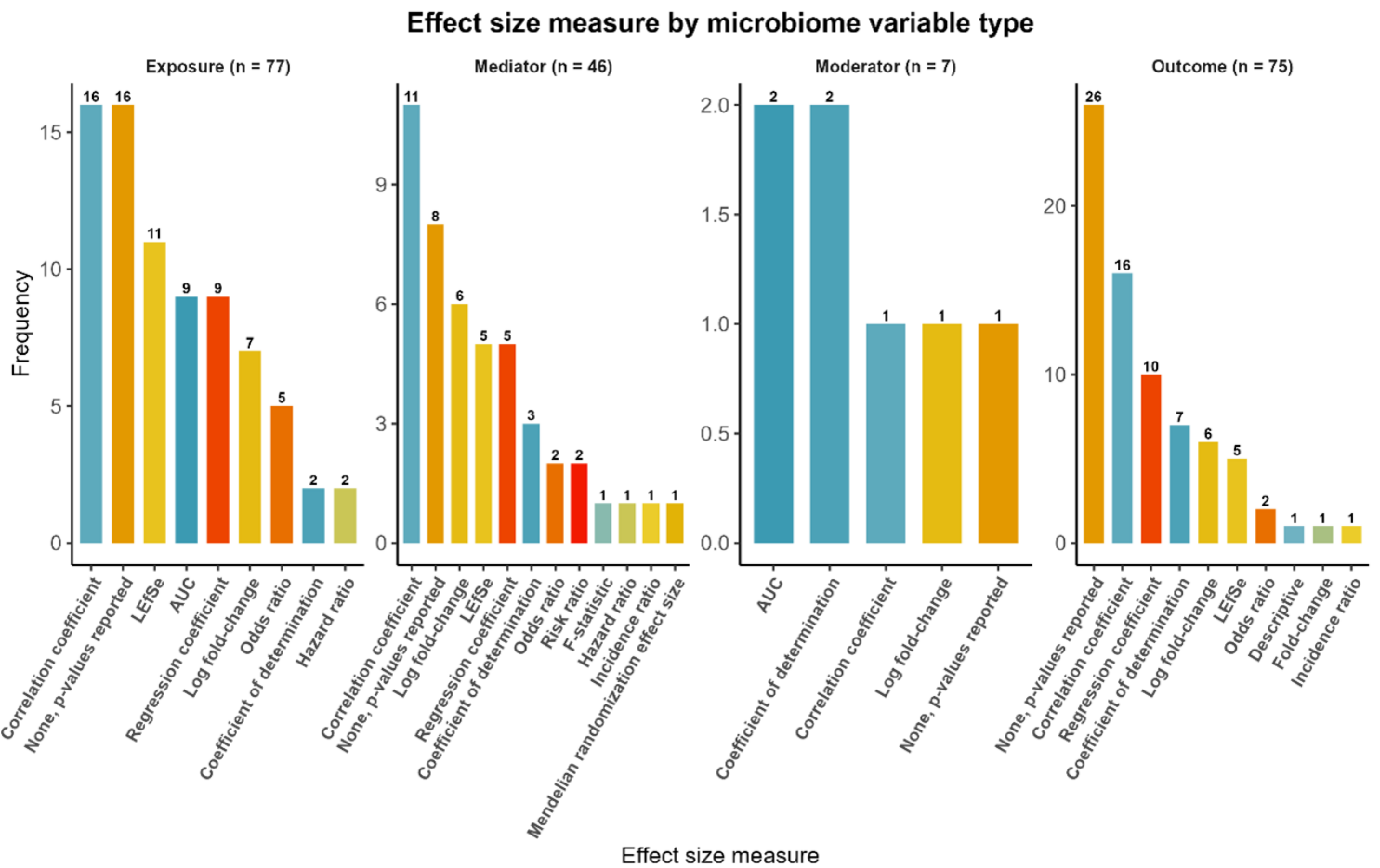


Figure 4

Effect Size Measures Used in Microbiome Research by Variable Role.

This faceted bar chart illustrates the distribution of commonly reported effect size measures across different microbiome roles.

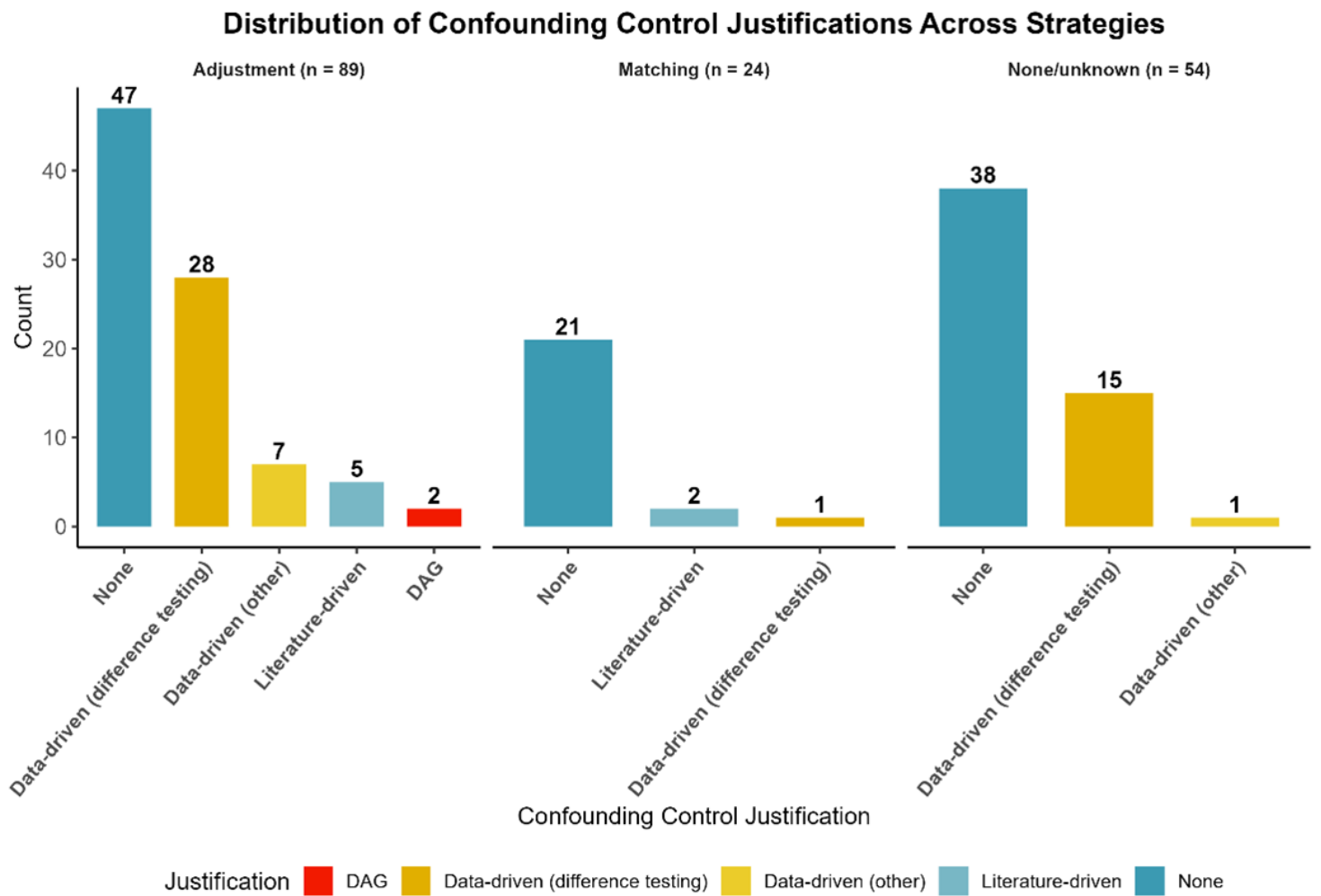


Figure 5

Distribution of Justifications for Confounding Control Methods in Epidemiologic Studies. The percentages shown on the y-axis are calculated relative to each confounding control method on the x-axis. Dots represent the upper and lower confidence intervals (CIs) for each justification category within a confounding control method.

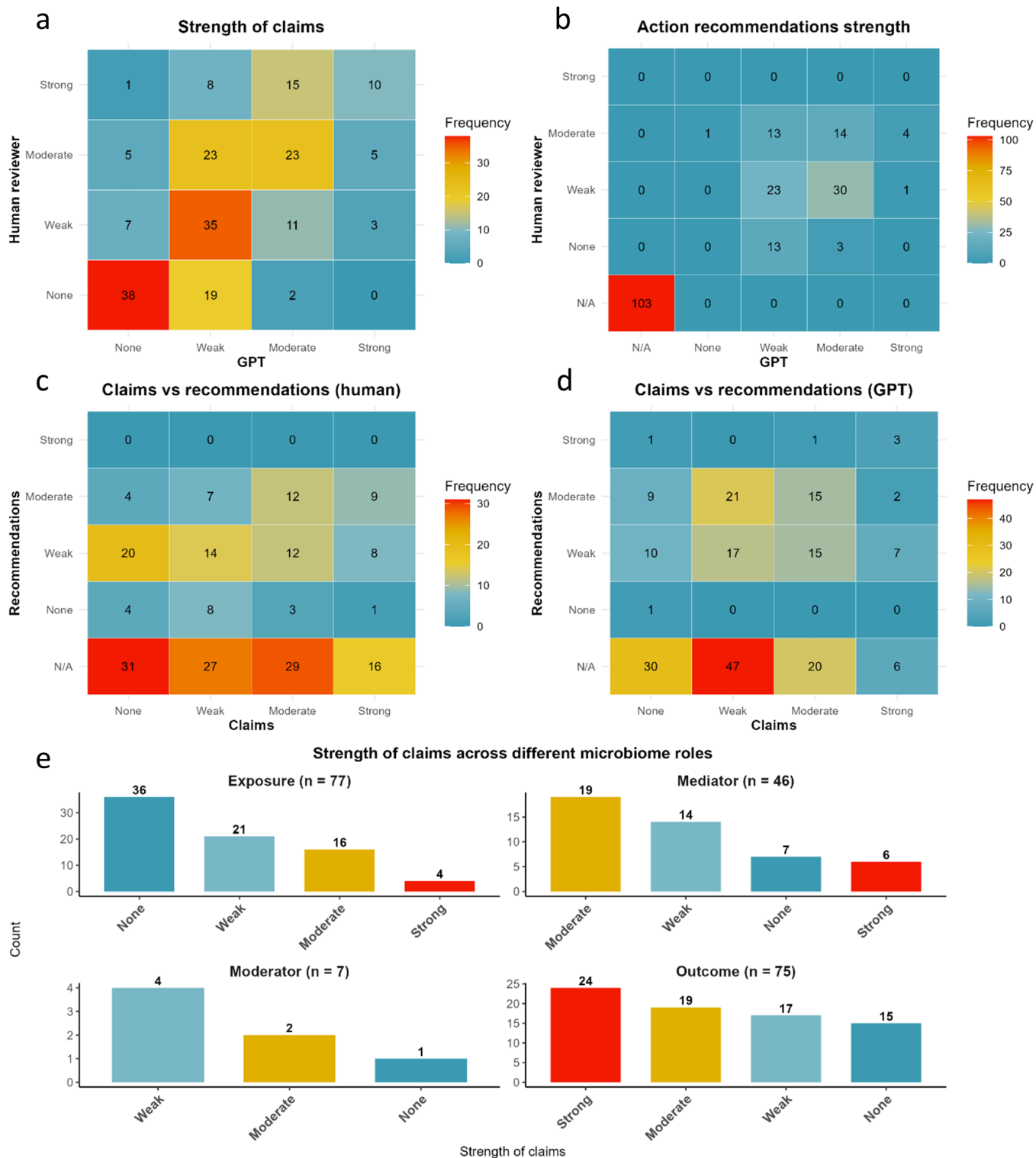


Figure 6

Analysis of Reporting Claims' and Practical Recommendations' Strength. (A-B) Comparison of the strength of causal claims and action recommendations as assessed by a human reviewer and GPT. Overall agreement was high, with the LLM ratings being slightly more conservative. Most discrepancies were within one level on the 4-point scale (none, weak, moderate, strong). (C-D) Strength of Causal Claims compared against Strength of Recommendations for gradings done by human reviewer and GPT.

"N/A" indicates no recommendations were provided. Other categories include "none" (recommendations not based on causal reasoning), and "weak," "moderate," or "strong" based on their implied causal strength. (E) Strength of Causal Claims by Microbiome Role. Each facet represents a different microbiome role (exposure, mediator, outcome, or effect modifier), with within-category percentages showing the distribution of causal claim strengths.