Risk of Bias Assessment Tool for Systematic Review and Metanalysis of the Gut Microbiome

Thomas Lampeter[1*], Charles Love[2], Trien Trey Tang[2], Aditi Marella[2], Hayden Young Lee[2], Armani Oganyan[2], Devin Moffat[2], Anisha Karim[2], Matthew Rusling DO[2*], Aubrey Massmann DO/MPH[2], Melanie Orr[1], Christian Bongiorno[2], Li-Lian Yuan PhD[2*]

[1]New York Institute of Technology College of Osteopathic Medicine, 101 Northern Blvd, Glen Head, NY 11545, USA

[2]Des Moines University College of Osteopathic Medicine, 3200 Grand Ave, Des Moines, IA 50312, USA

*corresponding authors: Lampeter.thomasm@gmail.com; matthewrrusling@gmail.com; lilian.yuan@dmu.edu

Author Contributions

M.R., Y.L., T.L., conceived and planned concept; A.M. consulted on methods in metanalysis and assisted in structuring the approach; T.L., M.R. drafted subdomains of interest with guidance from A.M.; T.L., C.L., T.T., A.SM., H.L., A.O., D.M. researched influencers on bias subdomains; T.L., C.L., T.T., A.SM., H.L., A.O., A.K., D.M. wrote body of manuscript describing subdomains of bias; T.L., C.L., T.T., A.SM., H.L., A.O., D.M. described summary criteria contained in table 1; L.Y., T.L., M.R. edited manuscript with input from all authors; M.O. consulted in manuscript editing; L.Y., guided manuscript formatting and editing; T.L. planned and performed validation test; M.R., L.Y., M.O. consulted on and reviewed validation test design, methods, and interpretation of results; T.L., created figures and table; L.Y. guided all phases of development; All authors made valuable intellectual contributions in discussion and creation of this manuscript.

10.1017/gmb.2023.12

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

Abstract:

Risk of bias assessment is a critical step of any metanalysis or systematic review. Given the low sample count of many microbiome studies, especially observational or cohort studies involving human subjects, many microbiome studies have low power. This increases the importance of performing metanalysis and systematic review for microbiome research in order to enhance the relevance and applicability of microbiome results. This work proposes a method based on the ROBINS-I tool to systematically consider sources of bias in microbiome research seeking to perform metanalysis or systematic review for microbiome studies.

Introduction:

The most common experimental design used to evaluate the effects of gut microbiome (GMB) genomic or taxonomic post-exposure remodeling has been cohort studies using either animal or human models. Randomized controlled trials (RCTs) for microbiome interventions are less common because we are still characterizing microbiome post-exposure remodeling to identify promising markers or targets for microbiome intervention that would warrant subsequent evaluation by RCTs. Therefore, results from a systematic review with quantitative or pooled metanalysis are essential in identifying candidates for RCTs.

A diligent risk of bias (ROB) assessment is a key step in systematic review or metanalysis to determine the likelihood that features of the study design or conduct of the study will give misleading results. GMB research is highly heterogeneous in its methods, reporting, and attempts to address bias. This manuscript and its associated rubric (**table 1**) are based on the Risk of Bias in Non-randomized Studies - of Interventions (ROBINS-I) tool, and are meant to be used as a GMB-specific adjunct to ROBINS-I. This manuscript and its associated rubric together form a tool that was developed to help standardize ROB assessment in metanalyses and systematic reviews of GMB studies. A small-scale validation test by first-time ROB assessors produced consistently similar ROB determinations, suggesting that this tool can successfully guide consistent ROB determinations. This tool may allow for improved ROB assessment when evaluating studies for metanalyses and systematic reviews of the GMB.

Using This Tool:

This manuscript and its associated rubric provide a framework for assessing ROB specific to GMB research. This tool strives to provide insight and reduce variability between individual researchers and groups conducting systematic reviews of the GMB. We do not seek to suggest best practices. Instead, we aim to indicate potential sources of bias that may significantly impact GMB studies and are thus vital when considering the strength of evidence for systematic review and metanalysis. The essential criteria in this manuscript are summarized in **table 1,** which was compiled to act as a rubric in guiding ROB determination.

Table 1, “the rubric,” guides the determination of low, moderate, or high ROB across seven domains. In each cell of the rubric, there are signaling statements to help guide low, moderate, or high ROB determination in that domain. Two additional ROB determinations are not included on the rubric as they are to be used at the judgement of the person assessing ROB in a study. They are “critical ROB” and “no information”. Critical ROB can be determined when a reviewer believes a study to be too problematic to

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

provide useful evidence on the effect of an intervention. As such, a study determined to be of critical ROB in any one domain should not be included in any synthesis. A determination of no information applies to domains where there is no clear evidence of a critical ROB *and* a lack of information to judge ROB otherwise.

1 – Confounding

1.1 Demographic Differences

Important demographic considerations in GMB studies are sex and age. Substantial differences in the gut microbiota are attributable to sex differences in mammals (Org *et al.* 2016, Kim *et al.* 2020). Because of this, any study which includes one sex in one arm and a different sex in another should be classified as having a high risk of bias. In addition to the risk of bias from sex, other demographic factors may also introduce confounding bias into the studies being examined. The GMB changes with age across numerous conditions, disease models, and species impacting microbial diversity and biome composition (Ticinesi *et al.* 2019, Liu *et al.* 2020). Therefore, age differences between cohorts and study arms should be assessed. If the study being examined uses organisms of one age in one arm and a different age in a second arm, it should be classified as having a high risk of bias. The age gap which introduces significant confounding bias, varies by organism. An example of an age gap that would introduce a high risk of bias is 8-week-old mice versus 1-year-old mice (Yoon *et al.* 2021).

1.2 Habitat Stability

The habitat in which organisms are kept substantially impacts their GMB (Singh *et al.* 2021). Mice, common subjects of microbiome research, are known to have highly variable microbiomes on arrival at a facility, likely because of transportation stress on the microbiome itself and the immune system and hormonal functions of the host organism (Lipinski *et al.* 2021, Montonye *et al*. 2018, Capdevila *et al.* 2007). Studies that do not allow for microbiome stabilization before research begins risk confounding bias due to a lack of habitat stability. Organisms should be acclimated to the study condition before baseline measurements or interventions are performed. However, an extensive acclimation period risks microbiome drift occurring due to the increasing age of the organism or other unknown factors, so habitat stabilization must be time limited (Hoy *et al.* 2015). Additional bias would also be introduced if the acclimation period is included in the interventional period of the research.

1.3 Genotype, Familial, & Source Differences:

Subject genotype, degree of familial relation, and in the case of animal models, the source can significantly impact GMB composition. Differences in the genotype of animal models have been found to impact the diversity and abundance of organisms (Campbell *et al.* 2012, McKnite *et al.* 2012, Leamy *et al.* 2014). For this reason, if the study being evaluated uses organisms of significantly different genotypes, such as the use of different strains of mice from the Collaborative Cross, where the effect of genotype difference is not the target of the study, it should be classified as having a high risk of bias. Suppose the study uses a similar genotype between treatment groups, such as the same strain of inbred animal model or monozygotic twin subjects. In that case, it should be considered a low risk of bias for confounding due to the genotype effect.

Regarding familial relation, genetically related subjects have been demonstrated to share a core of similar GMB for up to three generations in the female line (Turnbaugh *et al.* 2008, Valles-Colomer *et al.* 2021). With animal models, breeding within familial relations is often used to maintain genotypically and GMB homogeneity (Hufeldt *et al.* 2010). A caution regarding inbreeding is that while selective breeding between siblings can create a more stable and uniform GMB composition, the effects of genetic drift can also introduce confounders across multiple generations that may affect experimental reproducibility with subsequent generations (Laukens *et al.* 2016).

Additionally, with animal models, an organism's litter of origin impacts the gut microbiota (Vilson *et al.* 2018, Fujiwara *et al.* 2008). This may relate not only to parent genetics but also to the host of maternal factors that can affect the development of progeny GMB, including mode of delivery, maternal diet, maternal stress, and maternal antibiotic use (Friwell *et al.* 2010, Walker *et al.* 2017, Stokholm *et al.* 2014, Golubeva *et al.* 2015, Bailey *et al.* 2004, Zhang *et al.* 2021). For these reasons, if the study being examined utilizes organisms from differing litters (from separate mothers or separate deliveries from the same mother) that have not yet reached their mature adult development and are not randomly assorted between research arms, it should be classified as having a high risk of bias. Suppose a study uses organisms from the same mother and litter or randomly assorts progeny from different mothers and litters. In that case, it should be classified as having a low risk of bias.

Regarding sourcing of animal models, subjects sourced from different vendors have substantial differences in GMB at baseline (Rasmussen *et al.* 2019, Long *et al.* 2021, Wolff *et al.* 2020). The microbiological or physiological basis of these effects is unknown but may be due to differential exposures to environmental or infectious factors between vendors (Mandal *et al.* 2020).

1.4 Extreme Diet

Dietary differences have been shown to alter the abundance of most gut microbes (Daniel *et al.* 2014, Ang *et al.* 2020, Li *et al.* 2021, Do *et al.* 2018). Because of this, maintaining the diet of interest is essential to avoid introducing confounding bias to the study. However, it may not always be possible to strictly control diet. This is especially relevant to clinical studies involving humans. In this situation, an evaluation of bias must note how a study documented these diet variations.

1.5 GMB Normalization

It is important to assure organisms being studied in research have similar baseline GMB. This allows for more definitive inference as to the effect of the intervention. Several strategies have been used to make the GMB as similar as possible over time. Removal of the entire GMB through the use of germ-free mice can allow for artificial seeding of a select group of organisms (Kennedy *et al.* 2018, Yi and Li 2012). However, the use of these mice necessarily limits the generalizability of a study. For this reason, research often uses organisms with populated GMBs and rely instead on antibiotics to homogenize the microbiome. The use of antibiotics introduces additional risks of bias which must be considered when evaluating a study (Theriot *et al.* 2016).The most significant risk of bias arises from beginning the intervention of interest before the gut microbiota has stabilized after normalization with antibiotics. The GMB continues to fluctuate unpredictably for long periods following antibiotic administration

(Merenstein *et al.* 2021). This variance has been found for at least a year after antibiotic usage in humans and for times ranging between one week and 16 weeks in mice depending on the length of the course of antibiotics used (Elvers *et al.* 2020, Rashid *et al.* 2015, Zhu *et al.* 2021). However, short, or single doses of antibiotics such as those often used to normalize the microbiome allow for substantial stabilization of the GMB within 7 days (Gu *et al*. 2020).

A third method used to standardize the GMB is to intermix the bedding of multiple cages and then redistribute it (Miyoshi *et al.* 2018). This method is less invasive than antibiotic usage and has a lower risk of long-term impact on the GMB than the use of antibiotics. The use of homogenization of the bedding allows for similar microbiomes to develop in more mice than can be practically housed in a single cage, where the organisms also share all of their bedding (McCafferty *et al.* 2013).

Because of the impact of different methods of GMB normalization, it is critical to note the method that was used to normalize the GMB and how long before the intervention this normalization was completed.

2 - Selection Bias

2.1 Extreme genotype

Host genotype shows a stable and heritable impact on GMB composition (Goodrich *et al.* 2016). In the context of GMB research, extreme genotype selection refers to the selection of GMB subjects with genotypes that vary significantly between subjects within a study. Selection of subjects with identical or similar genetic make-up limits genotype confounding effects. A subject with an established history of use along with maximized genetic correlation can be considered a low risk of selection bias. For example, while inbred Balb/C mice do have an extreme genotype, they also have a long-established history of use in immune modulation studies with their known Th2 immune response wherein they exhibit low IFNγ and high IL-4 production (Khan *et al.* 2022, Mills *et al.* 2000, Watanabe *et al.* 2004). Furthermore, prior literature has established the correlation between subject genetics and variation in the GMB population and subsequent disease states (Xu *et al.* 2020).

2.2 Randomization or Demographic Balancing Sufficiently Applied

Randomization is essential in ensuring subject-level differences between participants in the intervention and control groups can be attributed to chance alone. It is a standard method that attempts to create the necessary pre-intervention equivalence between groups, allowing for conclusions based on the effect of the intervention. In trials where randomization was not appropriately utilized, the outcome was overestimated by up to 40% compared to trials where randomization was utilized (Suresh, 2011). If randomization was not applied, implementing demographic balancing is an appropriate measure to ensure adequate control and intervention arms distribution. Any demographic balancing performed should be sufficiently described in the study. This method focuses on ensuring each group is demographically balanced at baseline to lessen the difference between groups and utilize randomization if no subject background information is available (Saint, 2015). Both randomization and demographic balancing can be applied to human and animal model studies. For example, in studies utilizing syngeneic

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

mice, randomization must be performed outside the scope of human intervention in that random number generators should assign mice numbers which can then correlate to intervention and control groups, hence this places randomization outside the scope of human influence, limiting bias to a maximum degree. In syngeneic animals, demographic balancing would have a limited impact on the bias, however, wherein studies utilize genetically unrelated animals, the need for implementation of both randomization and demographic balancing is necessary for limiting substantial bias (Hirst *et al.* 2014). Similar principles apply in human studies. Given a majority of human studies utilize genetically unrelated subjects, randomization is required to avoid high risk of bias. In human studies, a step beyond randomization should be taken, i.e., implementing blinded randomization with description of the randomization protocol to give the reader the ability to discern breaks in randomization or similar bias control methods within the study (Chalmers *et al.* 1983).

3 - Classification of Intervention

3.1 Intervention Bias

Bias in intervention can occur when interventions or outcomes are inappropriately selected for or measured. In non-differential misclassification, test subjects' exposures are misidentified, and they are categorized into the wrong group (McCoy, 2017). This misclassification can dilute the effect of the intervention causing effect estimates to favor the null (LaMorfe, 2016). The probability of non-differential misclassification is equal across all groups. Bias may be reduced by ensuring a proper background check on test subjects and equalizing any differences. On the other hand, differential misclassification occurs when misclassification of exposure or outcome is not equal between subjects and is less easily predictable in whether it will bias results towards or away from the null. Therefore, the probability of assigning subjects to the wrong group differs based on the individual. This may also introduce recall bias towards recalling specific exposures because the subject has the disease state versus a subject that does not. In GMB studies, this may present in the form of researchers explaining results that show a significant effect as attributed to specific causes but leaving out explanations for non-significant results. Because this type of misclassification is more applicable in case studies, it is less relevant for animal studies but can be prominent in human studies (Spencer *et al.* 2018).

3.2 Validation of Method

The establishment of an effective intervention is imperative for a successful study. Before the experiment, researchers must verify that their chosen intervention method will produce the intended effect. In studies where this is not done, the produced results may or may not be relied on because the protocol was never validated. Verification can be internal (tested and proved by the researchers) or external (via other established studies). If the study calls for a particular disease state to be expressed, it must be validated that the test subjects have the disease state. In studies that call for a specific procedure, there can be potential bias in how the readers know the procedure was correctly obtained if it is not reported. For example, in microbiome hypertension studies, animal subjects were tested based on blood pressure measurements by a well-established method, tail-cuff plethysmography (Marques *et al.* 2019). If a lesser-known and validated method was used, it could introduce a high risk of bias if researchers did not verify that their method was accurate. When testing for the effect of a disease state

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

as influenced by the microbiome, it is helpful to transplant the experimental group microbiome into a germ-free animal model to confirm the effect. This reduces an intermediate risk of bias by demonstrating that the effect of the intervention is associated with the levels of change in the microbiome (Gottfredson *et al.* 2015).

4 – Deviation from Intervention

It is well understood that experiments that deviate from their initial protocol have an increased potential for bias in their study should they decide to include data prior to the deviation. Therefore, all deviations from the protocol should be well documented with time stamps, and the data included in the study should also include the time at which it was collected—either post-protocol or pre-protocol addendum. Rationale and limitations should also be included should researchers decide to include data from any time the protocol was different.

5 - Missing Data

Missing data is prevalent in many academic disciplines, from the social to biomedical sciences, and may contribute to bias in any given study. GMB research likewise suffers from inadequate consideration of missing data and the statistical methods to address it. To begin, two types of missing data should be distinguished: missing data due to patient drop-out in clinical, longitudinal studies and missing data as a result of inadequate sequencing depth leading to “false zeroes” in the microbiome genetic data. Both have potential to increase ROB.

5.1 Cause/Category of Missing Data

Missing data falls into multiple categories based on the mechanism of missingness: Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR) (Groenwold and Dekkers 2020). These categories apply assumptions to missing data based on the cause. MCAR assumes that data is missing due to a factor entirely unrelated for the study. MAR assumes data is missing due to observed variables relevant to the study. MNAR assumes data is missing based on unknown or not quantifiable variables to the authors. MAR and MNAR are most relevant to clinical research, specifically in regard to patient drop-out, including clinical GMB trials (Pugh *et al.* 2021). Sampling zeroes in microbiome data are a more generalized form of missing data but are primarily reminiscent of MAR (Kaul *et al.* 2017, Kaul *et al.* 2017). Each of these areas will be further discussed in the following sections. Under MAR, studies may utilize various statistical imputation techniques to replace missing data, though the most well-known and effective method is multiple imputations (Spineli *et al.* 2015). With MNAR, various statistical modeling techniques may address missing data. Such techniques are further discussed in relation to GMB studies in the section “Sequencing Depth and Sampling Zeroes.” The distinction between MAR and MNAR also indicates whether bias related to missing data is entirely removable in analysis - the former can, while the latter cannot (Mack *et al.* 2018). This should not be confused with the notion that MNAR assumptions immediately denote a study as biased. If the missingness in MNAR or MAR is independent of the outcome, then the study may be unbiased in regard to missing data. Thus, a study with MNAR data is not necessarily high ROB.

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

Notably, a significant number of studies do not clearly state the mechanism of missingness or adjust for missing data (Carpenter and Smuk, 2021). It is important that studies distinguish mechanism of missingness or explain relevant missing data. If a study does not acknowledge missingness in data or ensures the absence of missing data, the study may be considered high ROB. If a study acknowledges missing data but does not adequately address it through MAR/MNAR distinction and proper statistical techniques related to its missing data category, then the study may be considered intermediate ROB. If a study demonstrates all of this, it may be considered low ROB.

5.2 Subject Drop-out

Missing data in the form of patient drop-out has a marked effect on statistical power, type 1 error, and various outcome measures (Fiero *et al.* 2016, Cai *et al.* 2020, Thompson *et al.* 2011). In traditional clinical research, missing data has a clear effect on useful measures, such as relative risk and risk ratio calculations. Further, although researchers attempt to minimize drop-out and its statistical effects, drop-out ratios were reported to be greater than 40% depending on the study and the degree of unpleasantness in medical interventions to the patient (Schnicker *et al.* 2013, Li *et al.* 2021). Consequently, it has been proposed that a 20% drop-out ratio is reasonable (Furlan *et al.* 2009, Cramer *et al.* 2016). Interestingly, it has been shown that fecal sampling of patients in GMB studies has not been a significant reason for drop-out, suggesting typical sources of patient non-retention (Vandeputte *et al.* 2017). The effect of drop-out on statistical measures is expected to be the same in clinical GMB trials. Despite drop-out being common in clinical studies, its effect on outcome measures involving microbial compositional data (e.g., beta diversity) is not currently well described in clinical GMB studies. However, it is expected that such measurements relying on consistent analysis from a wide array of samples will be biased if there is inadequate sampling size.

The effect of bias comes into effect when there is interpretation between samples, in that missing data prevents consistent interpretation of genetic data through a larger body of samples. For example, microbiome samples stratified by disease state versus control should be held to higher statistical power, similar to traditional clinical studies. Yet, the complexity of GMB genetic analysis often prevents large sample sizes from being a practical implementation due to costs unless utilizing less-expensive protocols such as those involving qPCR to monitor microbial composition at high taxonomic levels (i.e., phyla) (Koliada *et al.* 2020). Some studies demonstrate shallow shotgun metagenomic sequencing as an alternative methodology for large, longitudinal GMB studies (Xu *et al.* 2021). Nonetheless, making interpretations in GMB data between samples stratified by host conditions may need to be more consistent and accurate when samples are unavailable from a patient drop-out. Based on the literature of other areas in clinical research as discussed, it is again reasonable to assert that drop-out will influence outcome measures if authors make interpretations across hosts of varying condition states.

Due to few clinical studies analyzing the effect of drop-out on GMB outcomes, it is reasonable to use a 20% patient drop-out ratio, as many clinical trials traditionally utilize. GMB studies that have a high patient dropout are considered high ROB. GMB studies that have low patient drop-out are considered low ROB.

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

5.3 Sequencing Depth and Sampling Zeroes

GMB researchers should consider sequencing depth as a contributor to missing data and subsequent bias. It is established that low-sequencing depth (2000 single-end reads per sample) can adequately predict the same diversity patterns as high-depth sequencing (on the scale of millions of reads per sample) (Caporaso *et al.* 2011, Lundin *et al.* 2012, Xiao *et al.* 2018). Experiments that quantify GMB outcome measures (like alpha and beta diversity) should utilize the same depth for all samples. Bias would be introduced if different sequencing depths are used for a set of samples. It should be noted, however, that false zeroes influence microbiome genetic data at both high and low depth. While true zeroes (or biological zeroes) represent true taxonomic absences, false zeroes (or sampling zeroes) represent a lack of sequencing depth to adequately detect certain microbial taxa. Notably, low sequencing depth, as is often the case of 16S rRNA sequencing, may not detect low abundance taxa or low taxa (subspecies) due to lower resolution. Though whole genome sequencing (WGS), such as shotgun metagenomic sequencing, utilizes high sequencing depth to sequence entire genomes, sampling zeroes still persist (Pereira-Marques *et al.* 2019).

At the time of writing, this issue of zero-inflation – or the excess of sampling zeroes at high and low depth – and the resulting bias in GMB genetic data is an active area of research. Interestingly, relatively few studies utilize any statistical modeling to correct for such missing data. Yet, various modeling techniques were recently developed to address zero-inflation (Deek and Li, 2021, Zhang *et al.* 2020, Ha *et al.* 2020). Similar to modeling techniques, imputation is a method traditionally used to address missing data in the form of patient drop out, but a promising imputation method is recently available to also deal with GMB sampling zeroes. Previous studies showed an increase of Pearson correlation from 0.59 (between 16S and WGS in non-corrected data) to 0.64 (between 16S and WGS in corrected data) (Jiang *et al. 2021)*. There were also marked differences in mean and standard deviation of abundances per taxon between corrected and non-corrected data. This suggests greater homogeneity of samples across sequencing methods if imputation is utilized to correct data. However, as our article focuses on the role of bias in GMB research, we do not yet place best-practice recommendations for a particular method of missing data correction.

As of date, few GMB studies utilize statistical techniques to correct for sampling zeroes. Furthermore, common bioinformatics pipelines (such as QIIME2) do not incorporate such techniques into data-correction programs.

As such, the available literature suggests future GMB studies that do not consider sampling zeroes and lack a statistical technique for missing data correction may be considered high ROB. Studies that utilize missing data correction may be considered low ROB. These data correction methods, once more, include various modeling techniques or imputation.

6 – Measurement of Outcomes

6.1 Sample collection

Currently, there is no standard method for sample collection for GMB studies. While biopsy of the lower intestine provides a controlled sampling site and an accurate microbiota account, it is expensive, time-consuming, and unsuitable for healthy control groups. In contrast, fecal collection is non-invasive and

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

cost-effective (Tang *et al.* 2020). Thus, it is a standard sampling method in both clinical and research applications. However, fecal collection introduces temporal inconsistency that is a risk of bias when unaccounted for.

Fecal samples collected at different times of the day are at risk for inaccurate representation of the absolute abundance of gut microbiota (Caporaso *et al.* 2011). Specifically for mouse studies, the snapshots of the microbiota provided by fecal samples is more accurate and consistent within treatment groups when collected in the morning due to the nocturnal feeding nature of mice (Jones *et al.* 2021). For studies involving subjects with unpredictable and inconsistent bowel movements, samples should be preserved immediately after defecation as oxidation of the outer layer can alter the microbiota (Pepper and Rosenfeld, 2012). Specifically, Firmicutes and Bifidobacteria Spp. are two known phylum that are unstable in the outer microenvironment when exposed to oxygen (Gorselak *et al.* 2015). Therefore, to minimize the differential errors, the methods of measurement must be consistent between control and intervention groups.

6.2 Blinding

In a GMB study, the primary outcome is based on definitive and objective genetic sequencing. Therefore, assessor bias is typically negligible, and a low risk of bias is expected (Higgins *et al.* 2022).

7 – Reporting of Results

7.1 Selection of Reported Results

Selective reporting of results can lead to biased interpretations of significance and or non-significance via particular selection of results from multiple outcome measures in estimating outcome effect. Bias in selection of reported results can be difficult to detect without access to a protocol from which one can compare pre-specified intended outcomes of interest to the outcomes analyzed in the published paper (Heneghan *et al.* 2019). Often, results are selected for significance, omitted for non-significance, or omitted for adverse effect of intervention (Dwan *et al.* 2013, Hedin *et al.* 2016, Van der Steen *et al.* 2019).

Validation Test

Four medical students with no prior experience in ROB assessment were recruited to test this tool by using it to independently assess ROB on three selected studies of similar length in a predetermined sequence (Wu *et al.* 2017, Mohammed *et al.* 2020, Saunders *et al.* 2020). Subjects were provided with the manuscript and ROB rubric. They were asked to track time to completion per study and complete the ROB rubric for each study. Subjects assessed ROB in an average of 44.75 minutes per study with time to completion generally decreasing from the first study assessed to the last study assessed.

Inter-rater variability was assessed by assigning values of 1, 2, and 3 to low, medium, and high ROB in order to construct visual representations of rater scores in each sub-domain of bias and to compare summed ROB scores between raters for each study. **Figures 1.1-1.3** demonstrate variability within a

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

study in each subdomain of bias assessed by this tool between raters. The figures demonstrate similar ROB judgements between at least three of four raters in the majority of subdomains across the three studies assessed.

Figure 2 demonstrates variation in summed ROB score by rater for each of the three studies. It shows the decreasing magnitude of difference between raters' summed ROB scores with each subsequent use of the tool from a max-score min-score difference of six points in study1 and study3, and of four points in study2 out of 45 possible points. One way ANOVA test of rater subdomain scores across all subdomains for each study returned p-values of 0.554, 0.568, and 0.399 for study1, study2, and study3 respectively indicating no significant difference between overall ROB assessment scores between raters of the same study. First time ROB assessors using this tool showed a relatively high degree of concordance in ROB determination at the subdomain level and in magnitude of summed ROB score.

Conclusion

Risk of bias assessment is a crucial step in systematic review and metanalysis to assess quality of information being collected. By outlining common sources of bias that can impact GMB research following the structure of the ROBINS-I tool, this tool can serve as an adjunct to improve and standardize ROB assessment of GMB studies. A standardized ROB assessment for GMB studies will improve accuracy of risk assessment, improve reproducibility between researchers, and promote the inclusion of high-quality information in systematic reviews and metanalyses of the GMB.

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

Financial Support

This work was supported by the Iowa Osteopathic Education and Research (IOER) Foundation.

Conflicts of Interest declarations in manuscripts

Authors declare no conflicts of interest.

Research Transparency and Reproducibility

Following the journal's policy for supporting research transparency and reproducibility, we will make all data and protocols available to readers.

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

Figure Captions:

Figure 1.1 - Inter-rater variability in ROB determinations by subdomain for validation test study 1, *"Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug"* by Wu *et al*. 2017, where "1" on the y-axis indicates that the rater determined the study to be at low ROB for the subdomain indicated on the x-axis; "2" indicates medium ROB and "3" indicates a high ROB determination by the individual rater.

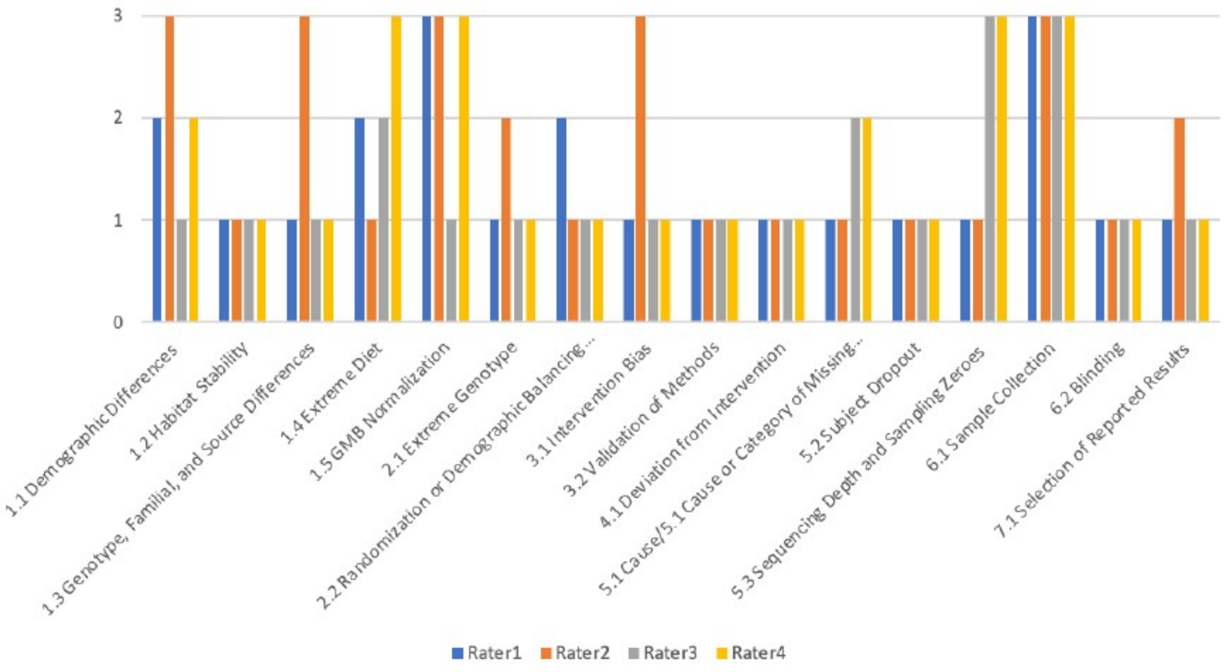
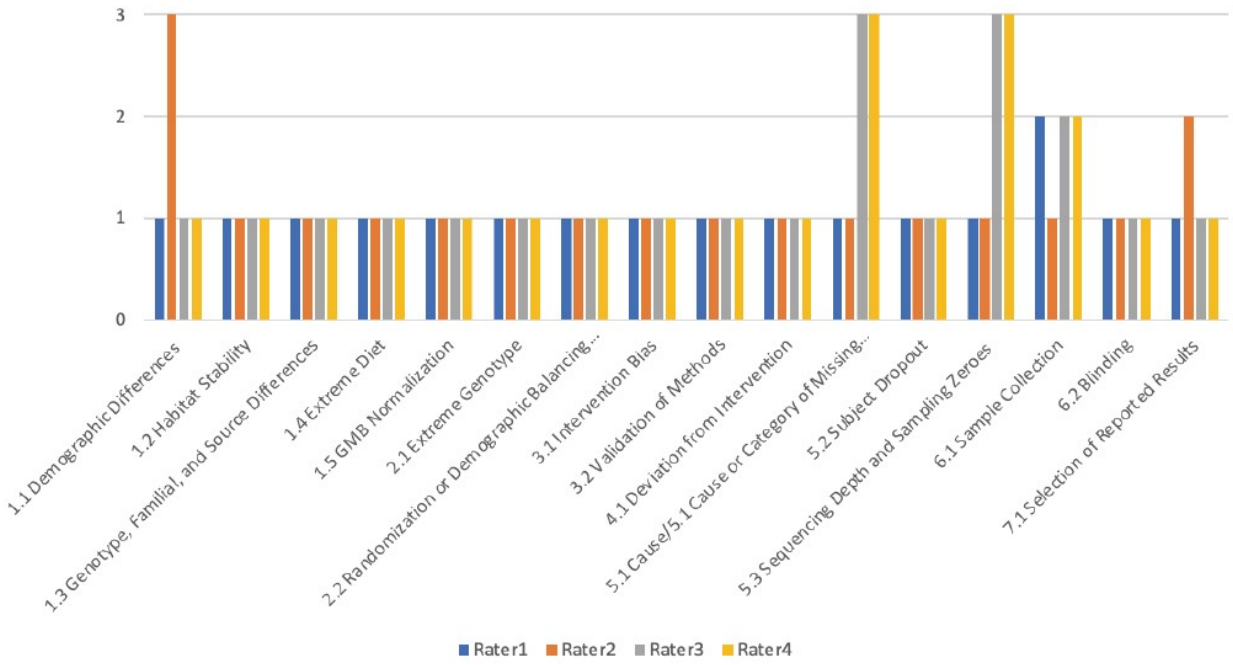


Figure 1.2 - Inter-rater variability in ROB determinations by subdomain for validation test on study 2, *"Protective effects of Δ9-tetrahydrocannabinol against enterotoxin-induced acute respiratory distress syndrome are mediated by modulation of microbiota"* by Mohammed *et al.* 2020, where "1" on the y-axis indicates that the rater determined the study to be at low ROB for the subdomain indicated on the x-axis; "2" indicates medium ROB and "3" indicates a high ROB determination by the individual rater.



https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

Figure 1.3 - Inter-rater variability in ROB determinations by subdomain for validation test on study 3, *"Gut microbiota manipulation during the prepubertal period shapes behavioral abnormalities in a mouse neurodevelopmental disorder model"* by Saunders *et al.* 2020, where "1" on the y-axis indicates that the rater determined the study to be at low ROB for the subdomain indicated on the x-axis; "2" indicates medium ROB and "3" indicates a high ROB determination by the individual rater.

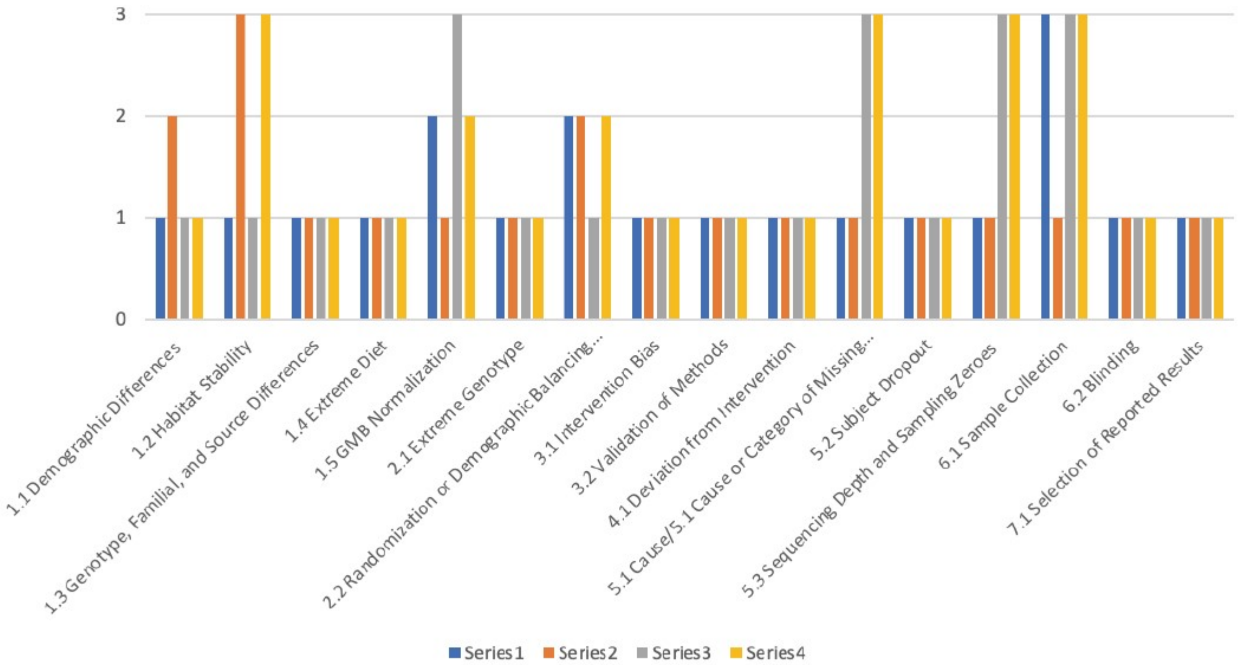
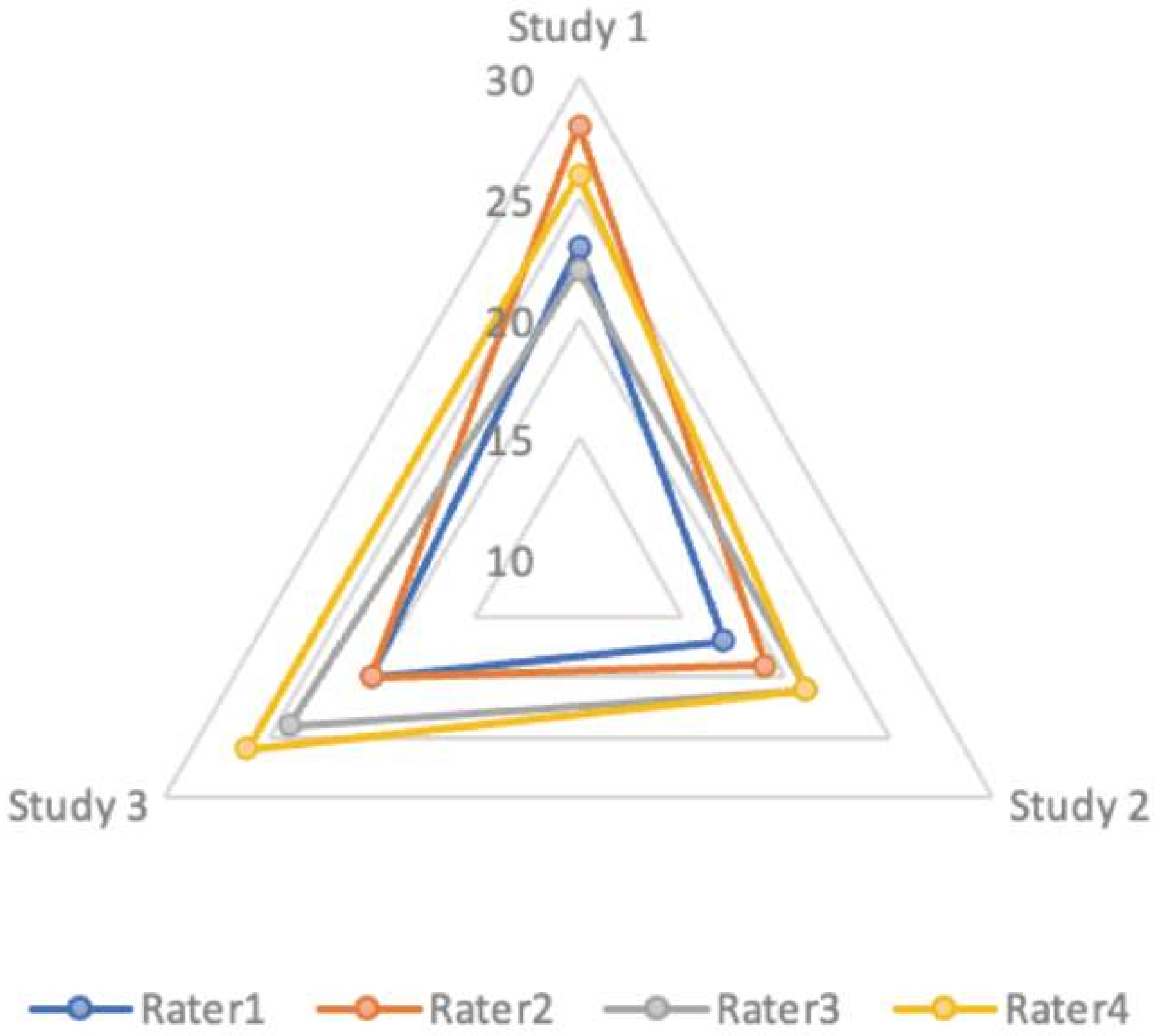


Figure 2 – Visual representation comparing summed ROB score (as determined by assigning point values of 1, 2, and 3 to low, medium, and high ROB respectively) by rater for each of the three studies assessed in the validation test where each increasingly large concentric triangle indicates an increase of 5 points.



https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

References:

Org *et al.* (2016). Sex differences and hormonal effects on gut microbiota composition in mice. *Gut microbes*, *7*(4), 313–322.

Kim *et al.* (2020). Sex Differences in Gut Microbiota. *The world journal of men's health*, *38*(1), 48–60. https://doi.org/10.5534/wjmh.190009

Ticinesi *et al.* (2019). Gut Microbiota, Muscle Mass and Function in Aging: A Focus on Physical Frailty and Sarcopenia. *Nutrients*, *11*(7), 1633.

Liu *et al.* (2020). Aging Increases the Severity of Colitis and the Related Changes to the Gut Barrier and Gut Microbiota in Humans and Mice. *The journals of gerontology. Series A, Biological sciences and medical sciences*, *75*(7), 1284–1292.

Yoon *et al.* (2021). Association between aging-dependent gut microbiome dysbiosis and dry eye severity in C57BL/6 male mouse model: a pilot study. *BMC microbiology*, *21*(1), 106.

Singh *et al.* (2021). Cage and maternal effects on the bacterial communities of the murine gut. *Scientific reports*, *11*(1), 9841. https://doi.org/10.1038/s41598-021-89185-5

Lipinski *et al.* (2021). Cage Environment Regulates Gut Microbiota Independent of Toll-Like Receptors. *Infection and immunity*, *89*(9), e0018721. https://doi.org/10.1128/IAI.00187-21

Montonye *et al.* (2018). Acclimation and Institutionalization of the Mouse Microbiota Following Transportation. *Frontiers in microbiology*, *9*, 1085. https://doi.org/10.3389/fmicb.2018.01085

Capdevila *et al.* (2007). Acclimatization of rats after ground transportation to a new animal facility. *Laboratory animals*, *41*(2), 255–261. https://doi.org/10.1258/002367707780378096

Hoy *et al.* (2015). Variation in Taxonomic Composition of the Fecal Microbiota in an Inbred Mouse Strain across Individuals and Time. *PloS one*, *10*(11), e0142825. https://doi.org/10.1371/journal.pone.0142825

Campbell *et al.* (2012). Host genetic and environmental effects on mouse intestinal microbiota. *The ISME journal*, *6*(11), 2033–2044. https://doi.org/10.1038/ismej.2012.54

McKnite *et al.* (2012). Murine gut microbiota is defined by host genetics and modulates variation of metabolic traits. *PloS one*, *7*(6), e39191. https://doi.org/10.1371/journal.pone.0039191

Leamy *et al.* (2014). Host genetics and diet, but not immunoglobulin A expression, converge to shape compositional features of the gut microbiome in an advanced intercross population of mice. *Genome biology*, *15*(12), 552. https://doi.org/10.1186/s13059-014-0552-6

Turnbaugh *et al.* (2008). A core gut microbiome in obese and lean twins. *Nature*, *457*(7228), 480–484. https://doi.org/10.1038/nature07540

Valles-Colomer *et al.* (2021). Variation and transmission of the human gut microbiota across multiple familial generations. *Nature Microbiology*, *7*(1), 87–96. https://doi.org/10.1038/s41564-021-01021-8

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

Hufeldt *et al.* (2010). Family relationship of female breeders reduce the systematic inter-individual variation in the gut microbiota of inbred laboratory mice. *Laboratory Animals, 44*(4), 283–9. https://doi.org/10.1258/la.2010.010058

Laukens *et al.* (2016). Heterogeneity of the gut microbiome in mice: guidelines for optimizing experimental design. *Fems Microbiology Reviews, 40*(1), 117–32. https://doi.org/10.1093/femsre/fuv036

Vilson *et al.* (2018). Disentangling factors that shape the gut microbiota in German Shepherd dogs. *PloS one*, *13*(3), e0193507. https://doi.org/10.1371/journal.pone.0193507

Fujiwara *et al.* (2008). Assessing changes in composition of intestinal microbiota in neonatal BALB/c mice through cluster analysis of molecular markers. *The British journal of nutrition*, *99*(6), 1174–1177. https://doi.org/10.1017/S0007114507862349

Friswell *et al.* (2010). Site and strain-specific variation in gut microbiota profiles and metabolism in experimental mice. *Plos One, 5*(1), 8584. https://doi.org/10.1371/journal.pone.0008584

Walker *et al.* (2017). The prenatal gut microbiome: are we colonized with bacteria in utero ?: colonization of the gut microbiome in utero. *Pediatric Obesity, 12*, 3–17. https://doi.org/10.1111/ijpo.12217

Stokholm *et al.* (2014). Antibiotic use during pregnancy alters the commensal vaginal microbiota. *Clinical Microbiology and Infection : The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases, 20*(7), 629–35. https://doi.org/10.1111/1469-0691.12411

Golubeva *et al.* (2015). Prenatal stress-induced alterations in major physiological systems correlate with gut microbiota composition in adulthood. *Psychoneuroendocrinology, 60*, 58–74. https://doi.org/10.1016/j.psyneuen.2015.06.002

Bailey *et al.* (2004). Prenatal Stress Alters Bacterial Colonization of the Gut in Infant Monkeys. *Journal of Pediatric Gastroenterology and Nutrition*, *38*(4), 414–421. https://doi.org/10.1097/00005176-200404000-00009

Zhang *et al.* (2021). Prenatal stress leads to deficits in brain development, mood related behaviors and gut microbiota in offspring. *Neurobiology of Stress*, *15*, 100333. https://doi.org/10.1016/j.ynstr.2021.100333

Rasmussen *et al.* (2019). Mouse Vendor Influence on the Bacterial and Viral Gut Composition Exceeds the Effect of Diet. *Viruses*, *11*(5), 435. https://doi.org/10.3390/v11050435

Long *et al.* (2021). Shared and distinctive features of the gut microbiome of C57BL/6 mice from different vendors and production sites, and in response to a new vivarium. *Lab animal*, *50*(7), 185–195. https://doi.org/10.1038/s41684-021-00777-0

Wolff *et al.* (2020). Vendor effects on murine gut microbiota and its influence on lipopolysaccharide-induced lung inflammation and Gram-negative pneumonia. *Intensive care medicine experimental*, *8*(1), 47. https://doi.org/10.1186/s40635-020-00336-w

Mandal *et al*. (2020). Temporospatial shifts within commercial laboratory mouse gut microbiota impact experimental reproducibility. *BMC biology*, *18*(1), 83. https://doi.org/10.1186/s12915-020-00810-7

Daniel *et al.* (2014). High-fat diet alters gut microbiota physiology in mice. *The ISME journal*, *8*(2), 295–308. https://doi.org/10.1038/ismej.2013.155

Ang *et al.* (2020). Ketogenic Diets Alter the Gut Microbiome Resulting in Decreased Intestinal Th17 Cells. *Cell*, *181*(6), 1263–1275.e16. https://doi.org/10.1016/j.cell.2020.04.027

Li *et al.* (2021). Ketogenic Diets Induced Glucose Intolerance and Lipid Accumulation in Mice with Alterations in Gut Microbiota and Metabolites. *mBio*, *12*(2), e03601-20. https://doi.org/10.1128/mBio.03601-20

Do *et al.* (2018). High-Glucose or -Fructose Diet Cause Changes of the Gut Microbiota and Metabolic Disorders in Mice without Body Weight Change. *Nutrients*, *10*(6), 761. https://doi.org/10.3390/nu10060761

Kennedy *et al.* (2018). Mouse Microbiota Models: Comparing Germ-Free Mice and Antibiotics Treatment as Tools for Modifying Gut Bacteria. *Frontiers in physiology*, *9*, 1534. https://doi.org/10.3389/fphys.2018.01534

Yi and Li (2012). The germfree murine animal: an important animal model for research on the relationship between gut microbiota and the host. *Veterinary microbiology*, *157*(1-2), 1–7. https://doi.org/10.1016/j.vetmic.2011.10.024

Theriot *et al.* (2016). Antibiotic-Induced Alterations of the Gut Microbiota Alter Secondary Bile Acid Production and Allow for Clostridium difficile Spore Germination and Outgrowth in the Large Intestine. *mSphere*, *1*(1), e00045-15. https://doi.org/10.1128/mSphere.00045-15

Merenstein *et al.* (2021). *Bifidobacterium animalis* subsp. *lactis* BB-12 Protects against Antibiotic-Induced Functional and Compositional Changes in Human Fecal Microbiome. *Nutrients*, *13*(8), 2814. https://doi.org/10.3390/nu13082814

Elvers *et al.* (2020). 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*, *10*(9), e035677. https://doi.org/10.1136/bmjopen-2019-035677

Rashid *et al.* (2015). Determining the Long-term Effect of Antibiotic Administration on the Human Normal Intestinal Microbiota Using Culture and Pyrosequencing Methods. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America*, *60 Suppl 2*, S77–S84. https://doi.org/10.1093/cid/civ137

Zhu *et al.* (2021). Effects of long-term antibiotic treatment on mice urinary aromatic amino acid profiles. Bioscience reports, 41(1), BSR20203498. https://doi.org/10.1042/BSR20203498

Gu *et al.* (2020). Effect of the Short-Term Use of Fluoroquinolone and β-Lactam Antibiotics on Mouse Gut Microbiota. *Infection and drug resistance*, *13*, 4547–4558. https://doi.org/10.2147/IDR.S281274

Miyoshi *et al.* (2018). Minimizing confounders and increasing data quality in murine models for studies of the gut microbiome. *PeerJ*, *6*, e5166. https://doi.org/10.7717/peerj.5166

McCafferty *et al.* (2013). Stochastic changes over time and not founder effects drive cage effects in microbial community assembly in a mouse model. *The ISME journal*, *7*(11), 2116–2125. https://doi.org/10.1038/ismej.2013.106

Goodrich *et al.* (2016). Genetic determinants of the gut microbiome in UK twins. Cell Host & Microbe, 19(5), 731–743. https://doi.org/10.1016/j.chom.2016.04.017

Khan *et al.* (2022). Effects of Shrimp Peptide Hydrolysate on Intestinal Microbiota Restoration and Immune Modulation in Cyclophosphamide-Treated Mice. Molecules, 27(5), 1720. https://doi.org/10.3390/molecules27051720

Mills *et al.* (2000). M-1/M-2 macrophages and the Th1/Th2 paradigm. Journal of immunology, 164(12), 6166–6173. https://doi.org/10.4049/jimmunol.164.12.6166

Watanabe *et al.* (2004). Innate immune response in Th1- and Th2-dominant mouse strains. Shock, 22(5), 460–466. https://doi.org/10.1097/01.shk.0000142249.08135.e9

Xu *et al.* (2020). The interplay between host genetics and the gut microbiome reveals common and distinct microbiome features for complex human diseases. Microbiome, 8(1), 145. https://doi.org/10.1186/s40168-020-00923-9

Suresh (2011). An overview of randomization techniques: An unbiased assessment of outcome in clinical research. Journal of human reproductive sciences , 4 (1), 8 – 11. https://doi.org/10.4103/0974 - 1208.82352

Saint (2015). Randomization Does Not Help Much, Comparability Does. PloS one, 10 (7), e0132102. https://doi.org/10.1371/journal.pone.0132102

Hirst *et al.* (2014) The Need for Randomization in Animal Trials: An Overview of Systematic Reviews. PLOS ONE 9(6): e98856. https://doi.org/10.1371/journal.pone.0098856

Chalmers *et al.* (1983). Bias in treatment assignment in controlled clinical trials. *The New England journal of medicine*, *309*(22), 1358–1361. https://doi.org/10.1056/NEJM198312013092204

McCoy (2017). Understanding the Intention-to-treat Principle in Randomized Controlled Trials. *The western journal of emergency medicine*, *18*(6), 1075–1078. https://doi.org/10.5811/westjem.2017.8.35985

LaMorfe (2016). Information Bias (Observation Bias). Information bias(observation bias). Retrieved March 26, 2022, fromhttps://sphweb.bumc.bu.edu/otlt/mph-modules/ep/ep713_bias/EP713_Bias4.html

Spencer *et al.* **Misclassification bias.** In Catalogue Of Bias 2018. http://www.catalogueofbiases.org/biases/misclassificationbias

Marques *et al.* (2019). Guidelines for transparency on GutMicrobiome Studies in essential and experimental hypertension. Hypertension,74(6),1279–1293.https://doi.org/10.1161/hypertensionaha.119.13079

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

Gottfredson *et al.* Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research in Prevention Science: Next Generation. Prev Sci. 2015 Oct;16(7):893-926. doi: 10.1007/s11121-015-0555-x. PMID:25846268; PMCID: PMC4579256.

Groenwold and Dekkers (2020). Missing data: the impact of what is not there. *European journal of endocrinology*, *183*(4), E7–E9. https://doi.org/10.1530/EJE-20-0732

Pugh *et al.* (2021). Missing repeated measures data in clinical trials. *Neuro-oncology practice*, *9*(1), 35–42. https://doi.org/10.1093/nop/npab043

Kaul *et al.* (2017). Structural zeros in high-dimensional data with applications to microbiome studies. *Biostatistics (Oxford, England)*, *18*(3), 422–433. https://doi.org/10.1093/biostatistics/kxw053

Kaul *et al.* (2017). Analysis of Microbiome Data in the Presence of Excess Zeros. *Frontiers in microbiology*, *8*, 2114. https://doi.org/10.3389/fmicb.2017.02114

Spineli *et al.* (2015). Addressing missing participant outcome data in dental clinical trials. *Journal of dentistry*, *43*(6), 605–618. https://doi.org/10.1016/j.jdent.2015.03.007

Mack *et al.* Managing Missing Data in Patient Registries: Addendum to Registries for Evaluating Patient Outcomes: A User’s Guide, Third Edition [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2018 Feb. Types of Missing Data. Available from: https://www.ncbi.nlm.nih.gov/books/NBK493614/

Carpenter and Smuk (2021). Missing data: A statistical framework for practice. *Biometrical journal. Biometrische Zeitschrift*, *63*(5), 915–947. https://doi.org/10.1002/bimj.202000196

Fiero *et al.* (2016). Statistical analysis and handling of missing data in cluster randomized trials: a systematic review. *Trials*, *17*, 72. https://doi.org/10.1186/s13063-016-1201-z

Cai *et al.* (2020). Estimands and missing data in clinical trials of chronic pain treatments: advances in design and analysis. *Pain*, *161*(10), 2308–2320. https://doi.org/10.1097/j.pain.0000000000001937

Thompson *et al.* (2011). Who's left? Symptoms of schizophrenia that predict clinical trial dropout. *Human psychopharmacology*, *26*(8), 609–613. https://doi.org/10.1002/hup.1253,

Schnicker *et al.* (2013). Drop-out and treatment outcome of outpatient cognitive-behavioral therapy for anorexia nervosa and bulimia nervosa. *Comprehensive psychiatry*, *54*(7), 812–823. https://doi.org/10.1016/j.comppsych.2013.02.007

Li *et al.* (2021). Analysis of risk factors and construction of prediction model of drop out from peritoneal dialysis. *Medicine*, *100*(3), e24195. https://doi.org/10.1097/MD.0000000000024195

Furlan *et al.* (2009). 2009 updated method guidelines for systematic reviews in the Cochrane Back Review Group. *Spine*, *34*(18), 1929–1941. https://doi.org/10.1097/BRS.0b013e3181b1c99f

Cramer *et al.* (2016). A Systematic Review and Meta-Analysis Estimating the Expected Dropout Rates in Randomized Controlled Trials on Yoga Interventions. *Evidence-based complementary and alternative medicine : eCAM*, *2016*, 5859729. https://doi.org/10.1155/2016/5859729

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

Vandeputte *et al.* Practical considerations for large-scale gut microbiome studies, *FEMS Microbiology Reviews*, Volume 41, Issue Supplement_1, August 2017, Pages S154–S167, https://doi.org/10.1093/femsre/fux027

Koliada *et al.* (2020). Seasonal variation in gut microbiota composition: cross-sectional evidence from Ukrainian population. *BMC microbiology*, *20*(1), 100. https://doi.org/10.1186/s12866-020-01786-8

Xu *et al.* (2021). Characterization of Shallow Whole-Metagenome Shotgun Sequencing as a High-Accuracy and Low-Cost Method by Complicated Mock Microbiomes. *Frontiers in microbiology*, *12*, 678319. https://doi.org/10.3389/fmicb.2021.678319

Caporaso *et al.* (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*, *108 Suppl 1*(Suppl 1), 4516–4522. https://doi.org/10.1073/pnas.1000080107

Lundin *et al.* (2012). Which sequencing depth is sufficient to describe patterns in bacterial α- and β-diversity?. *Environmental microbiology reports*, *4*(3), 367–372. https://doi.org/10.1111/j.1758-2229.2012.00345.x

Xiao *et al.* (2018). Necessary Sequencing Depth and Clustering Method to Obtain Relatively Stable Diversity Patterns in Studying Fish Gut Microbiota. *Current microbiology*, *75*(9), 1240–1246. https://doi.org/10.1007/s00284-018-1516-y

Pereira-Marques *et al.* (2019). Impact of Host DNA and Sequencing Depth on the Taxonomic Resolution of Whole Metagenome Sequencing for Microbiome Analysis. *Frontiers in microbiology*, *10*, 1277. https://doi.org/10.3389/fmicb.2019.01277

Deek and Li (2021). A Zero-Inflated Latent Dirichlet Allocation Model for Microbiome Studies. *Frontiers in genetics*, *11*, 602594. https://doi.org/10.3389/fgene.2020.602594

Zhang *et al.* (2020). Zero-Inflated gaussian mixed models for analyzing longitudinal microbiome data. *PloS one*, *15*(11), e0242073. https://doi.org/10.1371/journal.pone.0242073

Ha *et al.* (2020). Compositional zero-inflated network estimation for microbiome data. *BMC bioinformatics*, *21*(Suppl 21), 581. https://doi.org/10.1186/s12859-020-03911-w

Jiang *et al.* (2021). mbImpute: an accurate and robust imputation method for microbiome data. *Genome biology*, *22*(1), 192. https://doi.org/10.1186/s13059-021-02400-4

Tang *et al.* (2020). Current Sampling Methods for Gut Microbiota: A Call for More Precise Devices. *Frontiers in cellular and infection microbiology*, *10*, 151. https://doi.org/10.3389/fcimb.2020.00151

Caporaso *et al.* (2011). Moving pictures of the human microbiome. Genome biology, 12(5), R50. https://doi.org/10.1186/gb-2011-12-5-r50

Jones *et al.* (2021). Fecal sample collection methods and time of day impact microbiome composition and short chain fatty acid concentrations. *Scientific reports*, *11*(1), 13964. https://doi.org/10.1038/s41598-021-93031-z

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

Pepper and Rosenfeld (2012). The emerging medical ecology of the human gut microbiome. *Trends in ecology & evolution*, *27*(7), 381–384. https://doi.org/10.1016/j.tree.2012.03.002

Gorzelak *et al.* (2015). Methods for Improving Human Gut Microbiome Data by Reducing Variability through Sample Processing and Storage of Stool. *PloS one*, *10*(8), e0134802. https://doi.org/10.1371/journal.pone.0134802

Higgins *et al. Cochrane Handbook for Systematic Reviews of Interventions* version 6.3 (updated February 2022). Cochrane, 2022. Available from www.training.cochrane.org/handbook.

Heneghan (2019). *Outcome reporting bias*. Catalog of Bias. Retrieved September 6, 2022, from https://catalogofbias.org/biases/outcome-reporting-bias/

Dwan *et al.* (2013). Systematic Review of the Empirical Evidence of Study Publication Bias and Outcome Reporting Bias — An Updated Review. *PLoS ONE*, *8*(7), e66844. https://doi.org/10.1371/journal.pone.0066844

Hedin *et al.* (2016). Publication Bias and Nonreporting Found in Majority of Systematic Reviews and Meta-analyses in Anesthesiology Journals. *Anesthesia & Analgesia, 123* (4), 1018-1025. doi: 10.1213/ANE.0000000000001452.

Van der Steen *et al.* (2019). Causes of reporting bias: a theoretical framework. *F1000Research*, *8*, 280. https://doi.org/10.12688/f1000research.18310.2

Wu *et al.* (2017). Metformin alters the gut microbiome of individuals with treatment-naive type 2 diabetes, contributing to the therapeutic effects of the drug. *Nature Medicine*, *23*(7), 850–858. https://doi.org/10.1038/nm.4345

Mohammed *et al.* (2020). Protective effects of Δ9-tetrahydrocannabinol against enterotoxin-induced acute respiratory distress syndrome are mediated by modulation of microbiota. *British Journal of Pharmacology*, *177*(22), 5078–5095. https://doi.org/10.1111/bph.15226

Saunders *et al.* (2020). Gut microbiota manipulation during the prepubertal period shapes behavioral abnormalities in a mouse neurodevelopmental disorder model. *Scientific Reports*, *10*(1). https://doi.org/10.1038/s41598-020-61635-6

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

678

Domain	High ROB	Moderate ROB	Low ROB
1 - Confounding			
1.1 Demographic Differences	- Age: consistently different between study arms - Sex: consistently different between study arms	- Age: mixed ages within a study arm, but equal in distribution between study arms - Sex: mixed sexes within a study arm, but equal in distribution between study arms	- Age: consistently similar between study arms - Sex: consistently similar between study arms
1.2 Habitat Stability	- No acclimation period, or acclimation period <2 days - Acclimation period included in the interventional period	- Acclimation period ≥2 days but <5 days	- Acclimation period ≥5 days and <9 weeks
1.3 Genotype, Familial, and Source Differences	- Significantly different subject genotypes between study arms (where genotype effect is not the target of investigation) - Non-matured animal models from different litters and/or mothers without random assortment into study arms - Comparison of animal subjects from different source or vendor between study arms	- Animal subjects from same vendor, but from separate and temporally spaced orders without random assortment into study arms	- Adequately similar genotypes used between study arms (where host genotype effect is not the target of study) - Animal subjects from same litter - Animal subjects from same vendor and same order - Adult animal subjects from different litters/mothers/vendors randomly assorted into study arms

https://doi.org/10.1017/gmb.2023.12 Published online by Cambridge University Press

1.4 Extreme Diet	- No statement of dietary standards or documentation of dietary variation - Major deviations from stated diet	- Study uses human subjects outside of a highly controlled environment (for example an inpatient healthcare setting)	- Use of identical diet between study arms where diet is not the target of study
1.5 GMB Normalization	- No documented means of verified GMB normalization methods employed prior to intervention - Use of different normalization methods between study arms or use of non-validated technique	- Antibiotic normalization employed < 7 days prior to intervention	- Antibiotic normalization employed ≥7 days prior to intervention - Validated technique of GMB normalization employed
2 – Selection Bias			
2.1 Extreme Genotype	- Subjects of known extremely different genotypes - Subjects with no established history of use	- Syngeneic subjects with limited established history of use	- Syngeneic subjects with established history of use
2.2 Randomization or Demographic Balancing Sufficiently Applied	- Absence of both RCT and implementation of consistent host demographic across study	- Utilization of RCT or implementation of consistent host demographics across study	- Utilization of RCT and implementation of consistent host demographics across study
3 - Classification of Intervention			
3.1 Intervention Bias	- Differential misclassification of intervention or test subject based on exposures present or suspected	- n/a	- Differential misclassification of intervention or test subject based on exposures absent or not suspected
3.2 Validation of Method	- No validation that treatment method produces intended effect	- n/a	- Documented use of validated methods

	- Use of new method without internal validation		- Use of a new method with adequate internal validation
4 – Deviation from Intervention			
4.1 Deviation from Intervention	- Large deviations to protocol without adequate time stamps, rationale, and limitations noted	- Slight deviations to protocol with adequate time stamps, rationale, and limitations noted	- Intervention successfully carried out without protocol deviation
5 - Missing Data			
5.1 Cause or Category of Missing Data	- Does not address missing data qualitatively or quantitatively - Or, does not ensure to readers the absence of missing data	- Acknowledges missing data qualitatively or quantitatively - Inadequate MAR/MNAR distinction or proper statistical correction	- Addresses missing data qualitatively or quantitatively, or ensures absence of missing data. - Adequate MAR/MNAR distinction or proper statistical correction
5.2 Subject Dropout	- Subject drop-out exceeds 20%	n/a	- Subject drop-out is equal to or less than 20%
5.3 Sequencing Depth and Sampling Zeroes	- Does not address sampling zeroes with statistical correction	n/a	- Addresses sampling zeroes with statistical correction
6 – Measurement of Outcomes			
6.1 Sample Collection	- Inconsistent collection time	- Animal models: Collected at same time, not in the morning - Human models: Inconsistent collection time, but preserved immediately after defecation	- Animal models: Collected at same time, in the morning - Human models: Consistent collection time & preserved immediately after defecation

6.2 Blinding	- No double blinding when the primary measurement is subjective	n/a	- Primary outcome is objective measure such as genetic sequencing not subject to bias by the subject or investigator - Primary outcome is subjective and double or greater blinding employed
7 – Reporting of Results			
7.1 Selection of Reported Results	Any of: - Omission of stated outcomes that are unfavorable or statistically insignificant - Addition of outcomes not in initial protocol - Results reported are only on a subset of data - Changing outcome(s) of interest	- Any of the above, but with valid and satisfactory explanation provided	- Inclusion of relevant null and significant findings as stated in protocol

Table 1 – Rubric of domains and subdomains of bias with signaling statements to guide risk of bias assessment of gut microbiome studies.