Brain, Behavior, and Immunity 112 (2023) 163–172

Contents lists available at ScienceDirect

Brain Behavior and Immunity

journal homepage: www.elsevier.com/locate/ybrbi

The power of effective study design in animal Experimentation: Exploring the statistical and ethical implications of asking multiple questions of a data set

R.A. Ankeny [a], A.L. Whittaker [b], M. Ryan [c,d], J. Boer [e], M. Plebanski [e], J. Tuke [c,d,1], S.J. Spencer [e,1,*]

[a] School of Humanities, University of Adelaide, Adelaide, South Australia 5005, Australia
[b] School of Animal and Veterinary Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia
[c] School of Computer and Mathematical Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia
[d] Australian Research Council Centre of Excellence for Mathematical and Statistical Frontiers, Australia
[e] School of Health and Biomedical Sciences, RMIT University, Melbourne, Victoria 3083, Australia

A B S T R A C T

One of the chief advantages of using highly standardised biological models including model organisms is that multiple variables can be precisely controlled so that the variable of interest is more easily studied. However, such an approach often obscures effects in sub-populations resulting from natural population heterogeneity. Efforts to expand our fundamental understanding of multiple sub-populations are in progress. However, such stratified or personalised approaches require fundamental modifications of our usual study designs that should be implemented in Brain, Behavior and Immunity (BBI) research going forward. Here we explore the statistical feasibility of asking multiple questions (including incorporating sex) within the same experimental cohort using statistical simulations of real data. We illustrate and discuss the large explosion in sample numbers necessary to detect effects with appropriate power for every additional question posed using the same data set. This exploration highlights the strong likelihood of type II errors (false negatives) for standard data and type I errors when dealing with complex genomic data, where studies are too under-powered to appropriately test these interactions. We show this power may differ for males and females in high throughput data sets such as RNA sequencing. We offer a rationale for the use of alternative experimental and statistical strategies based on interdisciplinary insights and discuss the real-world implications of increasing the complexities of our experimental designs, and the implications of not attempting to alter our experimental designs going forward.

1. Introduction

Multifactorial experimental designs are becoming commonplace as the complexity of scientific research increases and such important factors as biological sex are increasingly incorporated. However, investigators often fail to ensure a study is appropriately powered to statistically handle such complex analyses. One of the chief advantages of using highly standardised and well-understood animal models is that each variable can be precisely controlled so that the variable of interest can be studied. Use of model organisms, and rodents in particular, has become a cornerstone of contemporary biological practices in part for this reason (Ankeny and Leonelli, 2011, 2020). However, this approach is increasingly recognised as problematic. Given the cumulative developments of the past fifty years, we are now well past the 'low-hanging fruit' stage of biological discovery where it was assumed that a principle that applies to one biological organism or system is likely to apply generally or to humans in particular. The generally unrecognised limits of these models may be one of the many reasons why the bulk of our research on complex diseases such as Alzheimer's and Parkinson's have failed to generate a widely applicable cure. Natural heterogeneity in human populations appears to be obscuring potential beneficial effects of therapies in subsets of clinical populations. This blind spot is also likely to have its analogue at the pre-clinical level, with important implications for how we should conduct BBI-related research from its earliest laboratory stages: for instance, a 2015 analysis (Freedman et al., 2015) showed that more than 50% of preclinical animal studies were not reproducible, and estimated that US$28 billion was wasted per annum in the United States alone as a result.

Nowhere is this importance of population heterogeneity more evident than when biological sex is considered. As more research is

* Corresponding author.
E-mail address: Sarah.Spencer@rmit.edu.au (S.J. Spencer).
[1] Equal senior co-authors.

https://doi.org/10.1016/j.bbi.2023.06.012
Received 27 February 2023; Received in revised form 31 May 2023; Accepted 10 June 2023
Available online 12 June 2023

starting to incorporate sex as a variable, it is becoming obvious that females are not just "males with ovaries" (Barrientos et al., 2019). Many biological processes, from neuroimmune responses (Doust et al., 2021; Krukowski et al., 2018; Lynch, 2022) to how cognitive tasks are solved (Becegato and Silva, 2022; Bowman et al., 2022), are physiologically and mechanistically different between the sexes. Indeed, the immune system is critically different between males and females from birth, as evidenced by substantial differences in response to vaccination in children and neonates, well before major hormonal differences come into effect (Flanagan et al., 2017). In 2015, the U.S. National Institutes of Health (NIH) officially recognised the extreme paucity of research on female animal models and developed "Consideration of Sex as a Biological Variable" guidelines requiring that all funded grants incorporate both females and males into their analysis (where reasonable in the context of the disease being studied) or strongly justify not doing so (NIH, 2015). In acknowledgement of resource constraints in research, it was considered acceptable to meet this requirement using a mixed-sex design, pooling females and males into the same analysis without using sex as a covariate (NIH, 2015). (See Table 1 for definitions of statistical terms that are underlined in the text). However, this strategy risks overlooking a real effect when the direction of it is dissimilar between the sexes. As just one example, immunisation with the measles vaccine produces not just dissimilar, but opposite immune transcriptomics patterns (activation versus silencing) in male and female children (Noho-Konteh et al., 2016). On the other hand, fully examining both females and males in terms of all experimental measures, and statistically comparing them to reveal differences, involves immense resources, including increasingly complex statistical approaches.

Even supposing that we appropriately test the impact of biological sex on our treatment or mechanism of interest, there are still numerous important variables that are missing from our usual considerations in neuroimmune research. Age is a key one. As is well illustrated by the recent SARS-CoV2 pandemic, children are not merely miniature adults when it comes to neuroimmune function (Sominsky et al., 2020): they have fundamentally different ACE2 receptor distribution that likely makes their responses to SARS-CoV2 exposure very different from that of adults (Bunyavanich et al., 2020; Radzikowska et al., 2020). They also have very different neuroimmune and hypothalamic–pituitaryadrenal (HPA) axis responses to neuroimmune challenge as well as having developing brains that may be differentially impacted in the long term (Sominsky et al., 2018). In addition, age, which is sometimes erroneously represented as a single continuous random variable, can be acknowledged to incorporate as many as seven levels which may be relevant to differentiate depending on the question of interest: embryonic, foetal, neonatal, juvenile, adolescent, young adult (in females further including pre- and post-menopausal), and older adult. There are other factors such as genetic background, hormones, microbiome, diet, and stress present in early and adult environments that influence how we respond to challenges, and therefore how effective a treatment might be.

It is critical to note that these factors are not separate variables, but influence each other, requiring an understanding of the complex networks formed by them when designing a study. For example, dietary effects on the microbiome that affect immunity and inflammation (with likely consequences for neuroimmune interactions) can be specific to biological sex, a finding that has given rise to the concept of the 'microgenderome' (Vemuri et al., 2019) and calls to "embrace variation" in order to produce greater reproducibility (Witjes et al., 2020). Furthermore, when moving into even more complex biological spaces such as transcriptomic profiling, the high dimensionality and genetic complexity of these types of data sets produces additional confounders. These include the technical complexity of library preparation, biological and technical variability, and especially read count biases, all of which contribute to the reduction of both the sensitivity and precision of high-throughput experiments. For instance, since counting errors depend on gene expression level, the variation owing to the counting process dominates the variance in genes with low counts, whereas for genes with high counts, this effect becomes negligible (Wu et al., 2015). Since variation in gene expression measurements is driven not just by biological variation but also variation in sequencing counts, it is increasingly clear that greater attention must be paid to performing power assessments in more biologically complex studies.

So, is the best strategy to set up discovery studies that incorporate multiple natural variables (sex, age, diet, and so forth) and run our testing in large-cohort analyses from the earliest laboratory stages, using both female and male cell lines and organisms? Such an approach could reveal personalised responses to treatment or challenges from multimodal response distributions to the same experimental manipulation, ultimately giving us insights into what works, when, and for whom. However, appropriate experimental design for such a strategy is essential. Any study that asks too many questions from its data set risks failing to reject the null hypothesis if it is under-powered to test for all of the questions posed (Norrie, 2020). Without sufficient sample numbers in the set, these false negatives may mean that we consistently miss important differences and useful clinical effects within subpopulations.

Conversely, when dealing with more high-throughput data, the opposite is true, given that insufficient power in this context leads to very high numbers of false positives. The resource and ethical implications of these issues are considerable, since time, money, and most critically animal lives, are wasted on experiments that were never capable of revealing an effect, if one exists. Or else they may well be revealing effects where none are present. In this article, we examine the statistical feasibility of asking multiple questions within the same experimental cohort and discuss the ethical implications of this approach, as well as the consequences of not attempting it going forward. We use an interdisciplinary approach, combining biomedical sciences, statistics, bioinformatics, animal research ethics, and philosophy of the biomedical sciences, to discuss potential solutions to this critical problem in the BBI fields.

Table 1
Glossary of statistical terms (underlined in the main text).

Term	Definition
Categorical variables	A variable with a labelled classification rather than a number (e.g., no stress / mild stress / strong stress; sex; clusters of ages)
Continuous variable	Any variable that is measured (e.g., height, weight, temperature)
Coefficient	The β's in the model that describe the relationship between a variable and the outcome.
Covariate	A variable to be accounted for in the experiment that is not of primary interest.
Discrete variable	Any variable that is counted (e.g., the number of cases of COVID-19).
Full factorial	Examining all of the possible combinations of variables in the experiment.
Interaction	The combined effect of two or more variables.
Levels	Classifications of a categorical variable (e.g., an experiment incorporating no stress, mild stress, and strong stress has three levels).
Noise term	Random experimental noise that captures individual variability in results.
Outcome variable	The thing that we are trying to predict.
P-value	The probability of seeing results as extreme as the ones observed when no true effect exists.
Parameter	Anything to be estimated from the data such as effect size or a standard deviation.
Predictor	A variable used to understand the outcome variable. Predictors can be covariates, such as age, or variables of interest, such as treatment.
Reference level	The comparator group for a categorical variable, determined as the first category alphabetically.
Stratum (plural: strata)	A homogenous subgroup defined by a combination of variables under consideration.

2. Methods

All analysis was performed in R (R-Core-Team, 2022) using R (version 4.2.2) Studio (2022.12.0 + 353). We ran simulations using four scenarios to investigate the number of samples necessary to reveal an effect, should one be present, with a statistical power of 90% for standard data and 80% for high-throughput data as we increased the number of parameters contributing to our data set. To briefly recap the concept of statistical power, consider an experiment to detect the effect of two treatments on a given outcome. Where a true difference exists, we would conclude that there is a difference between the treatments 90% of the time if a test has a statistical power of 90%.

2.1. Experiment 1: Basic simulation

In the first scenario, we considered an experiment exploring the treatment effect of a drug B in comparison to the control A. We were interested in the effects of biological sex on the outcome, as well as whether the treatments had different effects on the outcome depending on the sex of the mouse. This experiment is described as a 'two-variable experiment' as we have two variables, treatment and sex, that may influence our continuous outcome variable y. An example of this type of study would be treating female and male mice with either the control or the drug and assessing circulating hormone levels. Treatment (Tx) is a categorical variable with two levels A and B (i.e., the control or the drug), and sex is a categorical variable with two levels F and M (i.e., female and male). See Table 1 for definitions of statistical terms that are underlined in the text.

We simulate from the model

$$y = \beta_0 + \beta_1 \mathrm{Tx} + \beta_2 \mathrm{sex} + \beta_3 \mathrm{Tx{:}sex} + \varepsilon\,, \tag{1}$$

where ε is standard normal noise.

In this model, we have the following relationships: '

- β_0 is the mean value of the outcome if a mouse was female and on treatment A.
- β_1 is the mean difference in the outcome if a mouse is on treatment B compared to treatment A.
- β_2 is the mean difference in the outcome of a male mouse compared to a female mouse. Note that we code the relationships alphabetically, in this case female becomes the reference level and the other – male – is the sex represented by β_2.
- β_3 is the mean difference in the outcome for treatment B compared to treatment A for male mice.

In our simulations, we considered three scenarios that could occur in this type of study:

- only a treatment effect: to simulate this, we set $\beta_1 = 2; \beta_2 = \beta_3 = 0$;
- a treatment and a sex effect, but no interaction effect: we achieved this by setting $\beta_1 = \beta_2 = 2$; $\beta_3 = 0$; and
- a treatment effect, a sex effect, and an interaction effect: this was achieved by setting: $\beta_1 = \beta_2 = \beta_3 = 2$.

The simulations were run with an arbitrarily selected fixed baseline effect of $\beta_0 = 10$. The values for β_1, β_2, and β_3 were chosen to represent an effect size of two standard deviations from the mean (in this case, a 20% difference between the groups).

Each experiment had a fully factorial design, that is, we assumed equal numbers of mice in each possible combination of treatment and sex. Thus, we had four possible combinations:

- female: Treatment A;
- female: Treatment B;
- male: Treatment A; and
- male: Treatment B.

We refer to the possible combinations as a stratum, so that in this case, we have four strata. For example, if we have an experiment that has four subjects in each stratum, there is a total of 16 subjects (Table 2). For each model, we considered 2 to 15 subjects in each stratum. A value of 2 is the lowest possible necessary for each stratum to allow us to fit an interaction model when needed. For each scenario and strata size, we simulated 200 data sets and fit the appropriate linear model (i.e. have estimated the values of the parameters in the model). We defined power for each coefficient as the proportion of times that the model returned a p-value of less than 0.05 in the 200 simulations.

2.2. Experiment 2: Simulation incorporating experimental data

To make our simulations applicable to a real data set, we next adapted existing published data (Di Natale et al., 2019) with noise added and numbers changed, and used the observed values of the coefficients in this analysis to provide realistic simulations. The original experiment examined whether the effect of chronic stress on ovarian follicle maturation was mediated by acylated ghrelin. We examined ovarian follicle-stimulating hormone receptor mRNA in rats that had experienced chronic stress (or no stress) and were given acylated ghrelin antagonist, D-Lys3 (or saline). The experiment consisted of 25 observations in a full factorial experiment with a treatment variable with two levels denoted C (control) and D (drug / antagonist) and a covariate with two levels denoted c (no stress) and d (stress).

A linear model of the form:

$$y \sim \mathrm{Tx} + \mathrm{covariate} + \mathrm{Tx{:}covariate} \tag{2}$$

was fitted using the lm() command in R. The observed coefficients were then used to simulate data of strata size 2 to 15 inclusively with treatment, covariate, and interaction effects given by the coefficients of the linear model. For each strata size, 200 data sets were simulated. Linear models of the form Equation (2) were then fitted to each simulated data set and the resultant p-values obtained. These observed p-values were used to calculate the power for each strata size.

2.3. Experiment 3: Simulation with additional parameter values

To make our simulations applicable to multiple experimental designs and explore the complexities of adding additional parameters to an analysis, we considered the number of parameters that we would need to model data with increasing numbers of variables beyond treatment and sex. The parameters need to account for any variable, the base level (intercept), any interaction terms, and the noise term.

The models considered were designed to slowly increase in complexity with the following predictors:

- treatment (e.g., drug): a categorical variable with two levels (i.e., control or drug);
- sex: a categorical variable with three levels (which allows us to consider cases where we may want to include more levels, for example an unspecified sex);
- age: a categorical variable with four levels; and
- weight: a continuous variable.

We added the variables in the order provided in the above list and considered three types of model:

Table 2
An explanation of equal number of subjects in each stratum. Having four subjects in each stratum makes 16 subjects in the total experiment.

	A	B
F	4	4
M	4	4

- main effects only;
- main effects plus two-way interaction terms; and
- main effects with two-way and three-way interaction terms.

As an example of a three-way interaction, consider the case where we would like to look at the effect of treatment X compared to treatment Z for older male mice compared to young female mice.

2.4. Experiment 4: Simulation incorporating larger data sets

Because gene dispersion is an important factor when evaluating whether a data set is appropriately powered, we analysed various high throughput mouse data sets from NCBI (Table 3). All three selected data sets originated from RNA sequencing (seq) experiments, since these are currently the most common genetic type of analysis. For power calculations of high throughput RNAseq data, we used count tables from moderate sized data sets from mouse, characterized either by treatment or sex. We used the PROPER R package designed by (Wu et al., 2015). to perform our calculations and visualize the differences. We performed 100 simulations on each of the three data sets (a total of 300 simulations) to appropriately evaluate power and error rates. In these simulations the concepts of stratified power by gene counts were highlighted.

3. Results

3.1. Experiment 1: Basic simulation

If we ask the question of whether our treatment (e.g., drug) affects our variable of interest (e.g., hormone concentrations), we see that 16 subjects in total, 4 per stratum, is sufficient to give us 90% power to detect a difference (Fig. 1A). The treatment-only model and the treatment plus sex model display similar power, so that if we ask whether sex affects hormone concentrations, we can also answer this question with 90% power with 16 subjects (Fig. 1B). However, as soon we want to ask the question of whether the effect of our drug treatment on hormone concentrations differs by sex, that is, whether there is an interaction between sex and treatment (Fig. 1C), it becomes much more costly to detect significance in any of our coefficients. We need at least 24 subjects (8 per stratum) to detect the main effects of treatment and sex with 90% power. Most notably, if we are interested in the interaction between the treatment and sex, we need at least 44 subjects (11 per stratum) to detect a significant coefficient with 90% power. Notice that this simulation demonstrates how an experiment designed to answer questions about the main effects of a treatment is significantly underpowered when we wish to consider possible interactions of the treatment with other variables. For example, a sample size required for 90% power when investigating main effects only returns 75% power in the interaction model (Fig. 1C).

3.2. Experiment 2: Simulation incorporating experimental data

To examine the statistical power needed for a treatment by sex interaction using real data, we used published gene expression data as the basis for a simulation to ask whether the effects of a life experience (in this case stress) on the expression of a gene of interest (in this case *Fshr* mRNA) would be influenced by changes to a hormonal pathway of interest (in this case disruption of the acylated ghrelin system with antagonist, D-Lys3; Fig. 2). We found a strong negative interaction effect in this experiment. That is, the drug or antagonist (treatment D) gives a higher mean mRNA (level of y) when the subjects are not stressed (i.e., for covariate level c), but a lower mean mRNA (level of y) when the subjects are stressed (i.e., for covariate level d).

Table 4 provides the observed coefficients from the data set used in our simulations. We see that the mean fold change is 1 for treatment C with covariate d the effect of treatment D compared to treatment c is 0.5, the effect of covariate d compared to covariate c is 0.6, and the effect of treatment D with covariate d is −1. In our simulation we used the following values (see the "estimate" column of Table 4):

$$\beta_0 = 1;\ \beta_1 = 0.5;\ \beta_2 = 0.6;\ \text{and}\ \beta_3 = -1$$

We next considered what the power would be for each of the coefficients in the two-way interaction model for simulations based on the coefficients observed in Table 4 for a range of total subject numbers (Fig. 3). In this experiment, we found that the power to detect a significant interaction term is about 85% with 25 subjects. We also observed that to achieve a power of 90% for all terms, we would need at least 50 subjects.

Table 3
NCBI accession numbers and data sets used for calculations.

Author	*GEO Acc. number*	*Study type*	*Animal numbers*	*Reference*
Bottomly et al.	GSE26024	Gene expression striatum	44	(Bottomly et al., 2011)
Lee et al.	GSE222450	SN38/PD1 treatment in head and neck squamous cell carcinoma	12	(Lee et al., 2023)
Wang et al.	GSE196121	Brain injury	20	(Wang et al., 2022)

3.3. Experiment 3: Simulation with additional parameter values

Increasingly the complexity of our experimental questions means that our study design must include consideration of multiple parameters including factors such as age, and health indices such as body weight. Here we illustrate the combinatorial explosion that occurs as we include extra predictors (Fig. 4). Note, the more levels that a categorical variable has, the greater the rate of combinatorial explosion will be. Observe that when we include two- and three-way interactions, we get a huge increase in the number of parameters in the final model. For instance, in a design where a treatment effect is observed and the number of parameters is three, the parameter number increases to 12 when weight, sex, and age are also assessed. For the same study design, the parameter number escalates to 42 and then 73 when two-way and three-way interactions are calculated. We should note that this calculation illustrates only the parameter number. For an experimental design to yield a valid statistical analysis, at least three samples per parameter would be required to allow calculation of a standard deviation, thus tripling the bare minimum sample number needed (to 120 in the case of this three-way interaction). These simulations highlight how sample size must be dramatically increased to achieve high power for interactions in complex designs.

3.4. Experiment 4: Simulation incorporating larger data sets

To expand our simulations to data sets with very large outputs we next assessed two different mouse data sets that were both stratified by the number of gene counts (Bottomly et al., 2011; Lee et al., 2023). Both data sets clearly show how the number of gene counts greatly influences whether the data set will provide sufficient power (Fig. 5). For the lower gene counts, even when using 10 samples per group both data sets are sitting well below the 80% desired power. The set from Bottomley et al. (Bottomly et al., 2011) reaches a power of less than 20% in this case (Fig. 5A). For genes that have counts between 10 and 20, Bottomley et al.'s data still require a sample size of at least five animals per group to marginally achieve the targeted power of 80%. However, for Lee et al.'s data set (Lee et al., 2023) only three biological replicates are required to achieve the optimal power for genes with this expression level (Fig. 5B). Most importantly, this simulation shows that different mouse data sets can exhibit extremely variable sequencing and biological replicate requirements.

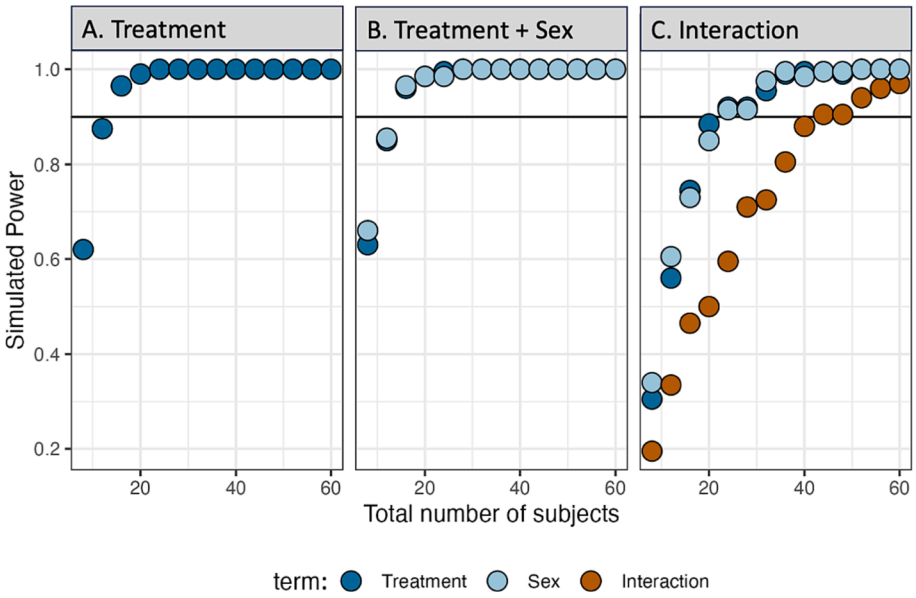

Fig. 1. Plot of power against full sample size for the three models: A) treatment only, B) treatment + sex, and C) the interaction model. The horizontal line indicates a power of 90%. Each dot represents the proportion of the simulations that gave a significant result for the considered coefficients.

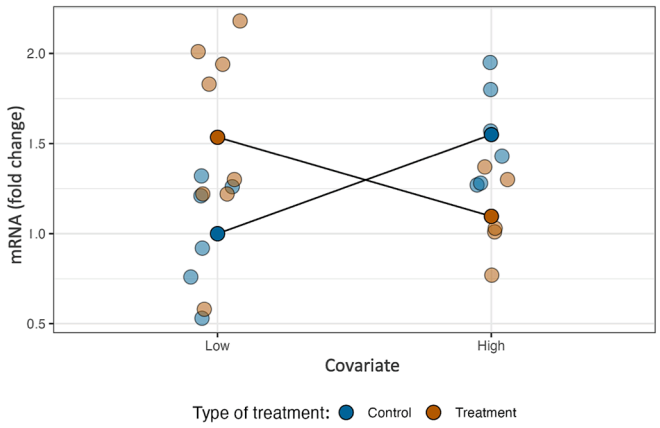

Fig. 2. Scatterplot of observed response variable (mRNA fold change) for the two levels of the covariate (c = non-stressed and d = stressed). The colour indicates the treatment level (C/blue = no antagonist [i.e., control]; D/brown = antagonist [i.e., drug]). The lines indicate the mean values for each combination of covariate and treatment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

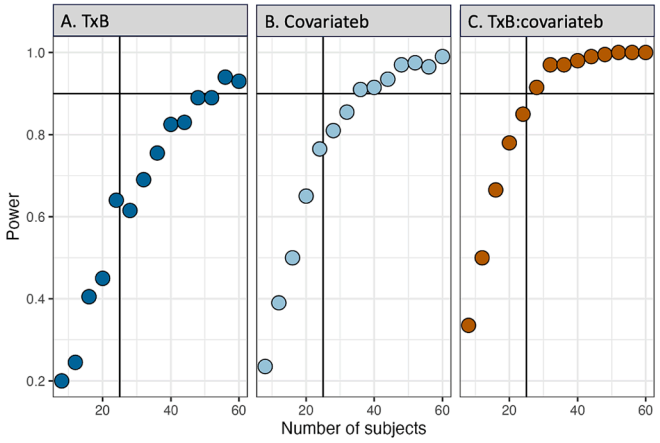

Fig. 3. Plot of power against total number of subjects for each of the coefficients in the two-way interaction model for simulations based on the observed coefficients for the experimental data simulation data set (Experiment 2). The horizontal line is for a power of 90%, while the vertical line is the 25 subjects included in the original experiment. To achieve a power of 90% for all terms (including (A) main effects [T × B] as well as (C) interactions [T × B: covariateb]), we would need at least 50 subjects in the study.

Table 4
Coefficients for the interaction linear model fitted to the experimental data simulation data set (Experiment 2).

Term	Estimate	Std Error	Statistic	P-value
(Intercept)	1.000	0.159	6.294	0.000
TxB	0.5	0.210	2.545	0.019
covariateb	0.6	0.225	2.448	0.023
TxB:covariateb	−1	0.316	−3.132	0.005

We next expanded our analysis to a data set comparing female and male mice (Wang et al., 2022). Notably, our simulations indicate gene dispersion may vary considerably between the sexes. In the female cohort, a sample size of 7 sufficed to provide the desired statistical power of 80% for genes with low average read counts (~10 read counts; Fig. 6). In males, however, a sample size of 7 would not provide a power of 80% for genes with average read counts lower than 320; a sample size of 10 would be required to confidently reach this. These data indicate wide variation in the gene dispersion for females versus males and further highlight the need for careful study design and power analyses in every case.

4. Discussion

Interactions between sex and treatment require increased sample sizes to maintain power

In our basic simulation (Experiment 1), we calculated the power cost of asking increasingly complex questions using laboratory experimental designs. We have shown that if an experimenter is only interested in main effects (such as treatment, i.e., control versus drug), there is a relatively manageable pattern of power required to detect these effects. However, once we begin to investigate interactions between sex and

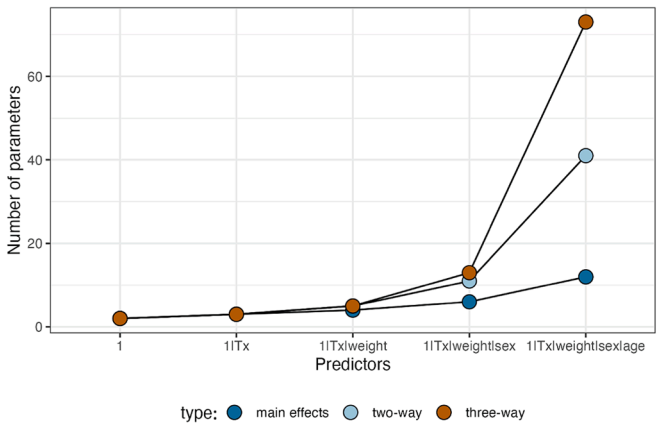

Fig. 4. Plot of the number of parameters needed for each considered model. The x-axis gives the predictors considered in each model, while the colour indicates the type of interactions considered in the model. 1 indicates the intercept. Note how parameter number (and therefore necessary sample size) is dramatically increased when considering interactions in complex designs.

treatment, we require much larger sample sizes to maintain power. A similar pattern is seen in simulations based on observed treatment, covariate, and interaction effects from a real experiment (Di Natale et al., 2019). We note that the incorporation of covariates such as sex or experimental conditions results in a large increase in sample size to maintain power that also can be observed in high-throughput data sets as seen in Experiment 4 (Li et al., 2021).

Incorporating additional variables dramatically increases the sample sizes required to maintain power

This problem is exacerbated with every variable added. In our simulation incorporating additional variables into the data set (Experiment 3), we observe the huge increase in the number of parameters in the final model as we add predictors and interaction terms. To allow such tests to be performed, we would need at least one subject per parameter. For example, in the full three-way interaction model seen in Fig. 4, we would need at least 43 subjects. However, we need at least three samples per parameter to calculate a standard deviation, and, as our power analysis shows us, we would need many more subjects per parameter to ensure good estimation of those parameters (i.e., $43 \times 5 = 213$ subjects).

High throughput genomic data from female mice is less dispersive than male data sets, influencing the number of samples required to achieve the desired power of analysis

Further complications arise when using complex genomic data sets to simulate required sample size (Experiment 4). Aside from the classical biological variability issues inherent in all data sets, high-throughput data sets also contain technical variabilities and considerable read count variability. Read counts are typically characterised by unequal variabilities (Law et al., 2014) and emerge primarily as a consequence of gene dispersion (Yoon and Nam, 2017). Count data applied to multiple types of functional genomics applications is usually modelled with a Poisson distribution where the mean and variance of read count are equivalent. However, RNAseq read count data are typically characterised by overdispersion, meaning that they notoriously exhibit variance of gene counts that are much larger than the mean. Therefore, the negative binomial distribution has become the popular choice for modelling RNAseq data, since it can capture gene dispersion where the variance of counts typically increases with the mean (Robinson and Smyth, 2008). Many different R-based procedures have been published that could be helpful with sample size calculations to reveal the required experimental power. For example, sample sizeRNAseq (Bi and Liu, 2016) is an extension from the data packages that originated from microarrays, while others provide the source code to simulate power calculations for either single or multiple genes and for sample size calculations in the presence of confounding covariates (Li et al., 2021). More recent studies have shown that negative binomial mixed models when estimated using the maximum likelihood test could be used for longitudinal data (Tsonaka et al., 2020).

Unfortunately, the majority of these R packages are lacking on many different fronts. For instance, covariates often are accounted for, but the modelling does not allow for tailoring gene dispersion values to personal data sets, instead relying on simulated or standardised data sets built into the package. Admittedly, it is incredibly difficult to design packages or provide source code for data simulation that can account for the many biological heterogeneities found in complex genomic data sets. For instance, it is important to consider that differentially expressed genes may reflect changes in the transcriptome that are disease-induced rather than disease-causing (Porcu et al., 2021); common genetic patterns may exist regardless of disease (Crow et al., 2019). Studies have shown that numerous genes are characterised by dominant isoforms which account for far more of the total expression of a particular gene than any of the remaining isoforms (Law et al., 2014). Finally, a study from Cote et al. (Cote et al., 2022) suggests that the choice of covariate adjustment can have considerable effects on the structure and accuracy of the resulting co-expression gene network. It is notable then that our results show that high throughput data from female mice is less dispersive than male data

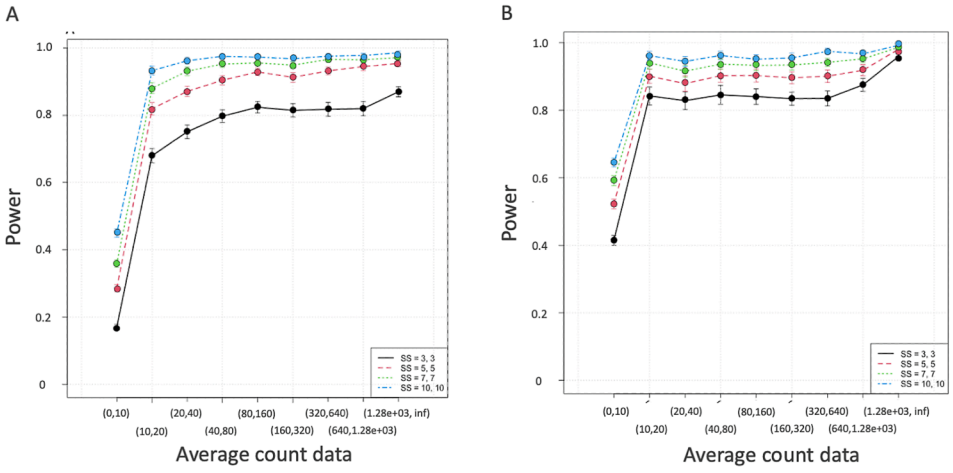

Fig. 5. Gene count stratification and experimental power. Mouse data sets from A) Bottomley et al. (Bottomly et al., 2011) and B) Lee et al. (Lee et al., 2023) that have been stratified by gene counts. The X-axes show average gene counts, while the Y-axes show the experimental power reached. Lower gene counts may not allow the achievement of appropriate power thresholds (80%) in any of the data sets. SS: sample size.

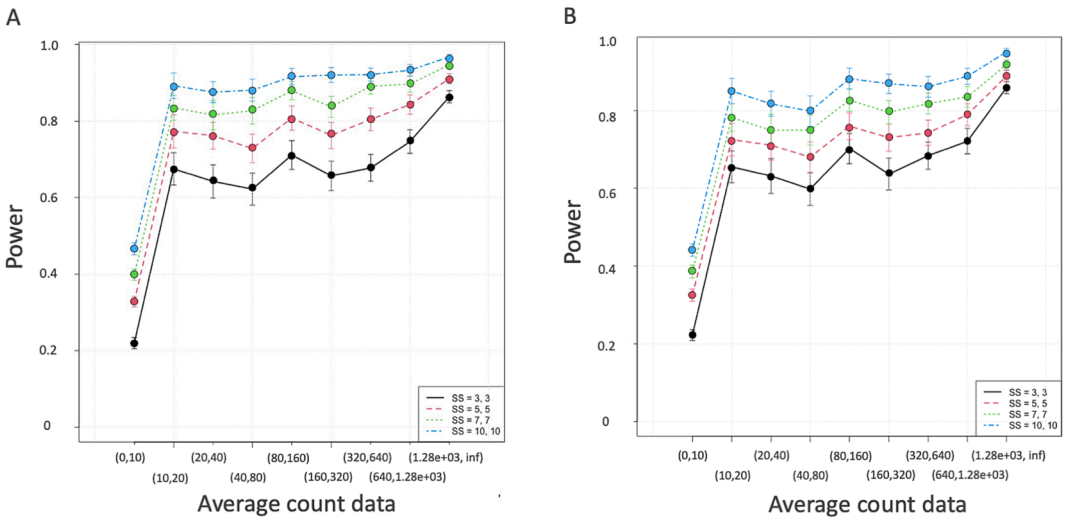

Fig. 6. Gene count stratification and experimental power for females and males. Mouse data sets from A) females and B) males from Wang et al. (Wang et al., 2022) that have been stratified by gene counts. The X-axes show average gene counts, while the Y-axes show the experimental power reached. In females, appropriate power thresholds (80%) were reached with a sample size of 7 for the lower gene counts (~10) but for males a sample size of 10 is needed. SS: sample size.

sets and therefore requires lower samples size for each group compared to the male counterpart. This is interesting and in line with previous findings in which extensive *meta*-analysis showed that male gene expression proves to be on average slightly more variable than female gene expression (Itoh and Arnold, 2015). Although the results in this study came from extensive analysis of microarray probes, the authors have investigated over 5 million of these. Together with our data these findings strongly suggest that although small, this variability can influence the number of samples required to achieve the desired power of analysis.

Multifactorial analyses have substantial statistical implications that need to be understood

Our findings together with the considerations discussed above illustrate the substantial statistical complexity of asking multiple questions of the same data set while appropriately integrating the complex biological background driving and underlying the data sets. They highlight how a type II error (false negative) is very likely to result from highly complex studies that are not appropriately powered, and that type I errors (false positive) become an even bigger problem when dealing with data sets where thousands of genes will be compared to each other. They also reinforce the fact that power calculations themselves are complex, and that a sophisticated understanding of the mathematics and biology behind these calculations is required, rather than simply using R packages off the shelf. So, what does this mean for our studies? Clearly there are strong ethical implications associated with including large numbers of animals within a study, particularly when end-points are terminal or painful. But there are also ethical implications associated with not doing this, as key effects could go unnoticed due to underpowered studies.

The ethics and practices of "reduction": Is reduction the most ethical strategy?

From the practical standpoint of obtaining approval for research involving non-human animals, many jurisdictions approve studies locally via an institutionally based animal ethics committee focused on humane care and use of animals for scientific purposes and tend to do so in accordance with the principles of the 3Rs (articulated originally by Russell and Burch in 1959 and subsequently modified and reinterpreted). The 3Rs - replacement, reduction, and refinement - are taken to be fundamental to the activities performed by investigators and animal carers as well as ethics committees and research institutions, but are not always easy to evaluate or implement. There are tensions associated with reduction in relation to the findings from the simulations detailed above, since incorporating consideration of multiple natural variables into animal-based studies and the associated requirements of increasing sample size can be viewed as potentially at odds with the goal of reduction.

In the original treatise by Russell and Burch (1959), the control of variation by making animals more standardised was central to reduction efforts that sought to use fewer animals while still obtaining comparable levels of information or obtaining more information from the same number of animals. Sound experimental design and robust statistical analysis, combined with the increasingly widespread availability of standardised animal strains, led researchers to use a reduced number of animals to achieve their scientific aims. The drive to control variation between individuals through standardisation became central to biomedical experimentation in the 20th century, giving rise to the thousands of inbred, almost genetically homogenous rodent strains commonly used in biomedical research today (Rader, 2004).

These forms of standardisation led to use of animals with a narrow range of characteristics such as age and sex, and together with tight control over husbandry and protocols to standardise the environment within which experiments are performed, were utilised to maximize sensitivity (i.e., internal validity) while minimising the numbers of animals used. However, efforts to reduce variation within experiments may limit the inferences that can be made to other laboratories with different experimental conditions, as well as producing misleading results (Richter et al., 2010; Richter et al., 2009) and contributing to the so-called 'reproducibility crisis' (Baker, 2015). Results may be limited as they relate to highly standardised mice strains in their local context, which cannot be replicated precisely across laboratories, leading to a form of limited external validity. Some argue that these failings can be rectified by applying processes of standardisation more rigorously and publishing details of protocols.

External validity is much more difficult to assess or achieve than internal validity. In the case of biomedical research, external validity is also critical for efficient translation from laboratory settings to the clinic and requires solid evidence about underlying similarities between the organisms used for research and those onto which they are being projected, and the relevance of these similarities to the functions or processes under study (Ankeny and Leonelli, 2011, 2020). Genetic approaches to similarity dominated in the second half of the 20th century, together with assumptions that conservation of genotype ensured effective translation. However, we are increasingly aware that other factors such as phenotypic plasticity (the degree to which living

organisms are highly responsive to their environment) operate even within organismal types and highly standardised environments, with phenotypes in control mice found to fluctuate unexpectedly between batches in the same laboratory (Karp et al., 2014).

4.1. Changing our approach to experimental design

The evidence presented here points to the need to make experimental conditions in BBI research much more heterogeneous and to track a greater number of potentially relevant factors to ensure effective translation of results to our highly heterogenous human population. To do so will require shifting away from various long-held institutional and disciplinary norms about animal experimentation to make it acceptable, and even necessary, to use more animals where clearly required. This shift will need to be accompanied by appropriate funding to accommodate these changes.

Assuming the merit of our proposal based on the evidence provided, what are the barriers for researchers and others involved in animal ethics to adopting heterogenisation strategies? The first is surely educative. Any laboratory animal science textbook describes the primary goal of animal research management as standardisation. Adoption of alternate experimental and methodological strategies requires considerable shifts in underlying assumptions and understandings, and will only succeed with focused training, ongoing research, and clear demonstration of benefit. More training for biologists in design and analysis will be necessary as has been noted by other authors (Weissgerber et al., 2016). This training must include more attention to the framing of research questions to permit robust answers with particular sample sizes, as well as greater understanding of the trade-offs associated with power inherent in particular types of study design (see e.g. (Lazic et al., 2018)).

Ethics committees must participate in this transition: even those committee members who are scientists rarely have advanced training in statistics, and few jurisdictions mandate that a statistician should be involved in ethics committee deliberations. Increased use of more diverse and complex forms of study design, and the need to balance oversimplified imperatives to reduce experimental animal numbers with methodological requirements to increase their numbers, will require refined guidelines and additional training. We propose that inclusion of a biostatistician or bioinformatician should be mandatory for ethics review processes, as it would offer an agile solution to the current knowledge gap present in most ethics committees and related bodies. As experimental scientists are able to gather more and more complex data, and as issues related to population heterogeneity continue to be acknowledged and incorporated into study design, this inclusion becomes more of an imperative. In turn, these requirements will make increased training in these fields necessary.

We also concur with those (e.g., (Cheleuitte-Nieves, 2019)) who argue that achieving reproducible and reliable preclinical research results should be viewed as a joint responsibility of various participants in animal research practices, including not only researchers and animal care technicians but also those involved in ethics reviews, journal publication, and grant funding. It will be necessary to monitor and manage a wider range of factors during study design and execution, and to document and report on these factors. The PREPARE (Planning Research and Experimental Procedures on Animals: Recommendations for Excellence: (Smith et al., 2018) and ARRIVE (Animal Research: Reporting of In Vivo Experiments: (Percie du Sert et al., 2020) guidelines provide useful advice on how to manage and report on intrinsic and extrinsic factors in the processes of animal experimentation in order to improve reproducibility and reliability. We encourage greater attention in such guidelines to more detailed reporting of the power calculations and other statistics associated with study design and execution.

The recent shift toward inclusion of both biological sexes in experimental animal studies presents clear challenges to researchers, ethics committee members, and others involved in animal experimentation and ethics. This concept appears to have been embraced by many, following increasing support from large funding bodies such as the NIH (Garcia-Sifuentes and Maney, 2021). However, in the authors' personal experience, there often appears to be a (likely erroneous) assumption that this requirement necessitates a simple doubling in animal numbers, which many seem willing to accept. This observation leads us to wonder whether they would understand the detailed statistical justifications for a tripling or quadrupling of numbers if needed to achieve statistical power, and change the status quo based on this, or be able to gauge proposed study methods in terms of power?

Most importantly, not to recognise the implications of the power requirements as detailed here for BBI research and beyond would be unethical and create potential harms for humans in clinical settings, particularly those who deviate from the standard norms associated with the 'standardised' biology that has been largely assumed to date. Although there has been some attention over the last two to three decades to injustices created by solely using young, white men as research subjects and excluding women of reproductive age and others altogether particularly from pharmaceutical testing (Barrientos et al., 2019; Ravindran et al., 2020), there has been limited focus on the earliest stages of preclinical research such as those explored here.

4.2. Considerations for study design and reporting existing underpowered data

It is therefore increasingly clear that we must not only pay much more attention to whether our data sets are appropriately powered, but also to understanding the underlying biology associated with experiments in order to design research questions and studies that maximise outcomes for humans while reducing unnecessary harms to non-human animals. In their published policy on inclusion of sex as a biological variable, the NIH (NIH, 2015) published four "C" factors to incorporate into study design and we believe these are also useful when considering multifactorial analysis (Fig. 7). These are "consider", "collect", "characterise", and "communicate". For our purposes, "consider" refers to appropriate planning of the questions we really want to ask in our study design. "Collect" refers to the acquisition of all relevant and accessible samples and data, including from multiple sexes and ages insofar as it is sensible and feasible within resource allocation. Strong study design and data collection can allow for future studies and collaborations without the sacrifice of repeated full cohorts of animals. "Characterise" refers to the reporting of characteristics of individual data points where possible. For instance, even if a study is too underpowered to test sex effects, reporting the data points in different colours or symbols can allow other researchers to determine if a future, fully powered, investigation into sex differences might be worth pursuing. "Communicate" stresses that full reporting of statistical power and data characteristics will help frame the data and interpretations in the right context and help the reader appreciate the probable strength of any conclusions.

To this list, we add "calculate", "collaborate", and "compensate". "Calculate" refers to the need to run power calculations and simulations of the sort presented here before a study commences. "Collaborate" may allow us the resources to run experiments that are appropriately powered to incorporate multifactorial designs in the pooling of financial and personnel resources. Finally, via "compensate", we recommend choosing non-standard sex, ages, or other backgrounds when designing a study when resources allow for restricted choices. In addition to certain advantages in choosing females over males, such as the reduced gene dispersion in females as shown here, this strategy will help even out the publication record over time and contribute to a balanced understanding of physiology on an organism-wide basis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence

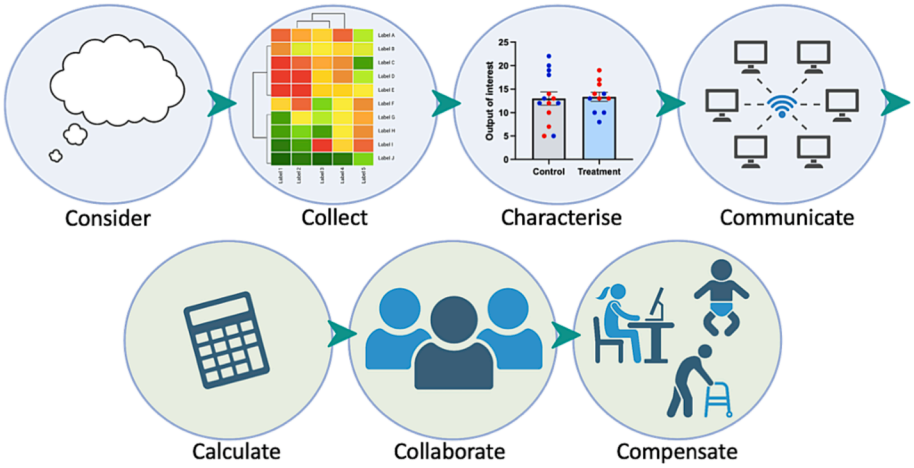

Fig. 7. Considerations for study design and reporting. As outlined by the NIH with respect to inclusion of sex as a biological variable (NIH, 2015), we advocate that when incorporating multifactorial designs one should consider, collect, characterise, and communicate one's strategy. Consider design before commencing experimentation. Collect data on crucial parameters such as sex wherever possible. Characterise data – even if the study is underpowered, merely illustrating the data points may encourage other groups to pursue the finding in more detail. Communicate all findings, including those that are underpowered (with due discussion of the data limitations). In addition to the NIH considerations, we also suggest to calculate the sample sizes needed to achieve the desired power before commencing experimentation, including running appropriate simulations as we have done here. Collaborate, with the idea that multiple groups may be better resourced to include appropriate sample sizes in return for investigation of different questions in the same subjects. Collaborating with biomedical statisticians and bioinformaticians will also allow for better study design and interpretation and will reduce the production and output of unsalvageable data. And compensate, in that if all other considerations are equal and the resources only allow for limited targeted questions, choosing to study females or non-standard ages may balance the body of literature and may also lead to more promising findings than choosing to study young-adult males (again). Top panels adapted from NIH recommendations (NIH, 2015). Figure created with BioRender.com; Toronto, Canada.

the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

MP is supported by a National Health and Medical Research Council (NHMRC) Senior Research Fellowship (SRFB) 1154850. SJS is supported by funding from an European Union (EU) Joint Program on Neurodegenerative Disease (JPND) Grant: (SOLID JPND2021-650-233); an NHMRC Ideas Grant (2019196) and an Australian Research Council Discovery Project (ARC; DP230101331). RAA's work relating to this paper was supported by an Australian Research Council Discovery Project (DP230101331).

References

Ankeny, R.A., Leonelli, S., 2020. Model Organisms. Cambridge University Press, Cambridge.
Ankeny, R.A., Leonelli, S., 2011. What's so special about model organisms? Studies Hist Phil Sci 42 (2), 313–323.
Baker, M., 2015. Reproducibility crisis: Blame it on the antibodies. Nature 521 (7552), 274–276.
Barrientos, R.M., Brunton, P.J., Lenz, K.M., Pyter, L., Spencer, S.J., 2019. Neuroimmunology of the female brain across the lifespan: Plasticity to psychopathology. Brain Behav Immun 79, 39–55.
Becegato, M., Silva, R.H., 2022. Object recognition tasks in rats: Does sex matter? Front Behav Neurosci 16, 970452.
Bi, R., Liu, P., 2016. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. BMC bioinformatics 17, 146.
Bottomly, D., Walter, N.A.R., Hunter, J.E., Darakjian, P., Kawane, S., Buck, K.J., Searles, R.P., Mooney, M., McWeeney, S.K., Hitzemann, R., Zhuang, X., 2011. Evaluating gene expression in C57BL/6J and DBA/2J mouse striatum using RNA-Seq and microarrays. PloS one 6 (3), e17820.
Bowman, R., Frankfurt, M., Luine, V., 2022. Sex differences in cognition following variations in endocrine status. Learn Mem 29 (9), 234–245.
Bunyavanich, S., Do, A., Vicencio, A., 2020. Nasal Gene Expression of Angiotensin-Converting Enzyme 2 in Children and Adults. JAMA 323, 2427–2429.
Cheleuitte-Nieves, C., 2019. Enrichment and outcomes in female lab mice. Lab Anim (NY) 48 (2), 53–54.
Cote, A.C., Young, H.E., Huckins, L.M., 2022. Comparison of confound adjustment methods in the construction of gene co-expression networks. Genome Biol 23, 44.
Crow, M., Lim, N., Ballouz, S., Pavlidis, P., Gillis, J., 2019. Predictability of human differential gene expression. Proc. Natl. Acad. Sci. U.S.A. 116 (13), 6491–6500.
Di Natale, M.R., Soch, A., Ziko, I., De Luca, S.N., Spencer, S.J., Sominsky, L., 2019. Chronic predator stress in female mice reduces primordial follicle numbers: implications for the role of ghrelin. The Journal of endocrinology 241, 201–219.
Doust, Y.V., King, A.E., Ziebell, J.M., 2021. Implications for microglial sex differences in tau-related neurodegenerative diseases. Neurobiology of aging 105, 340–348.
Flanagan, K.L., Fink, A.L., Plebanski, M., Klein, S.L., 2017. Sex and Gender Differences in the Outcomes of Vaccination over the Life Course. Annu Rev Cell Dev Biol 33 (1), 577–599.
Freedman, L.P., Cockburn, I.M., Simcoe, T.S., 2015. The Economics of Reproducibility in Preclinical Research. PLoS Biol 13 (6), e1002165.
Garcia-Sifuentes, Y., Maney, D.L., 2021. Reporting and misreporting of sex differences in the biological sciences. Elife 10.
Itoh, Y., Arnold, A.P., 2015. Are females more variable than males in gene expression? Meta-analysis of microarray datasets. Biol Sex Differ 6, 18.
Karp, N.A., Speak, A.O., White, J.K., Adams, D.J., Hrabé de Angelis, M., Hérault, Y., Mott, R.F., Gkoutos, G.V., 2014. Impact of temporal variation on design and analysis of mouse knockout phenotyping studies. PloS one 9 (10), e111239.
Krukowski, K., Grue, K., Frias, E.S., Pietrykowski, J., Jones, T., Nelson, G., Rosi, S., 2018. Female mice are protected from space radiation-induced maladaptive responses. Brain Behav Immun 74, 106–120.
Law, C.W., Chen, Y., Shi, W., Smyth, G.K., 2014. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome biology 15 (2), R29.
Lazic, S.E., Clarke-Williams, C.J., Munafò, M.R., 2018. Training in experimental design and statistics is essential: Response to Jordan. PLoS Biol 16 (10), e3000022.
Lee, Y.M., Chen, Y.H., Ou, D.L., Hsu, C.L., Liu, J.H., Ko, J.Y., Hu, M.C., Tan, C.T., 2023. SN-38, an active metabolite of irinotecan, enhances anti-PD-1 treatment efficacy in head and neck squamous cell carcinoma. J Pathol.
Li, X., Rai, S.N., Rouchka, E.C., O'Toole, T.E., Cooper, N.G.F., 2021. Adjusted Sample Size Calculation for RNA-seq Data in the Presence of Confounding Covariates. BioMedInformatics 1, 47–63.
Lynch, M.A., 2022. Exploring Sex-Related Differences in Microglia May Be a Game-Changer in Precision Medicine. Frontiers in aging neuroscience 14, 868448.
NIH, 2015. Consideration of Sex as a Biological Variable in NIH-funded Research. NOT-OD-15-102.
Noho-Konteh, F., Adetifa, J.U., Cox, M., Hossin, S., Reynolds, J., Le, M.T., Sanyang, L.C., Drammeh, A., Plebanski, M., Forster, T., Dickinson, P., Ghazal, P., Whittle, H., Rowland-Jones, S.L., Sutherland, J.S., Flanagan, K.L., 2016. Sex-Differential Non-Vaccine-Specific Immunological Effects of Diphtheria-Tetanus-Pertussis and Measles Vaccination. Clin Infect Dis 63, 1213–1226.
Norrie, J.D., 2020. Remdesivir for COVID-19: challenges of underpowered studies. Lancet 395 (10236), 1525–1527.
Percie du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M.T., Baker, M., Browne, W. J., Clark, A., Cuthill, I.C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S.T., Howells, D.W., Karp, N.A., Lazic, S.E., Lidster, K., MacCallum, C.J., Macleod, M., Pearl, E.J., Petersen, O.H., Rawle, F., Reynolds, P., Rooney, K., Sena, E.S., Silberberg, S.D., Steckler, T., Wurbel, H., 2020. The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. PLoS Biol 18, e3000410.
Porcu, E., Sadler, M.C., Lepik, K., Auwerx, C., Wood, A.R., Weihs, A., Sleiman, M.S.B., Ribeiro, D.M., Bandinelli, S., Tanaka, T., Nauck, M., Volker, U., Delaneau, O., Metspalu, A., Teumer, A., Frayling, T., Santoni, F.A., Reymond, A., Kutalik, Z., 2021. Differentially expressed genes reflect disease-induced rather than disease-causing changes in the transcriptome. Nat Commun 12, 5647.

Rader, K., 2004. Making Mice: Standardizing Animals for American Biomedical Research. Princeton University Press, Princeton.
Radzikowska, U., Ding, M., Tan, G.e., Zhakparov, D., Peng, Y., Wawrzyniak, P., Wang, M., Li, S., Morita, H., Altunbulakli, C., Reiger, M., Neumann, A.U., Lunjani, N., Traidl-Hoffmann, C., Nadeau, K.C., O'Mahony, L., Akdis, C., Sokolowska, M., 2020. Distribution of ACE2, CD147, CD26, and other SARS-CoV-2 associated molecules in tissues and immune cells in health and in asthma, COPD, obesity, hypertension, and COVID-19 risk factors. Allergy 75 (11), 2829–2845.
Ravindran, T.S., Teerawattananon, Y., Tannenbaum, C., Vijayasingham, L., 2020. Making pharmaceutical research and regulation work for women. BMJ 371, m3808.
R-Core-Team, 2022. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Richter, S.H., Garner, J.P., Würbel, H., 2009. Environmental standardization: cure or cause of poor reproducibility in animal experiments? Nat Methods 6 (4), 257–261.
Richter, S.H., Garner, J.P., Auer, C., Kunert, J., Würbel, H., 2010. Systematic variation improves reproducibility of animal experiments. Nat Methods 7 (3), 167–168.
Robinson, M.D., Smyth, G.K., 2008. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics 9, 321–332.
Smith, A.J., Clutton, R.E., Lilley, E., Hansen, K.E.A., Brattelid, T., 2018. PREPARE: guidelines for planning animal research and testing. Laboratory animals 52 (2), 135–141.
Sominsky, L., Jasoni, C.L., Twigg, H.R., Spencer, S.J., 2018. Hormonal and nutritional regulation of postnatal hypothalamic development. The Journal of endocrinology 237, R47–R64.
Sominsky, L., Walker, D.W., Spencer, S.J., 2020. One size does not fit all - Patterns of vulnerability and resilience in the COVID-19 pandemic and why heterogeneity of disease matters. Brain Behav Immun 87, 1–3.
Tsonaka, R., Signorelli, M., Sabir, E., Seyer, A., Hettne, K., Aartsma-Rus, A., Spitali, P., 2020. Longitudinal metabolomic analysis of plasma enables modeling disease progression in Duchenne muscular dystrophy mouse models. Hum Mol Genet 29, 745–755.
Vemuri, R., Sylvia, K.E., Klein, S.L., Forster, S.C., Plebanski, M., Eri, R., Flanagan, K.L., 2019. The microgenderome revealed: sex differences in bidirectional interactions between the microbiota, hormones, immunity and disease susceptibility. Semin Immunopathol 41 (2), 265–275.
Wang, Y., Wernersbach, I., Strehle, J., Li, S., Appel, D., Klein, M., Ritter, K., Hummel, R., Tegeder, I., Schafer, M.K.E., 2022. Early posttraumatic CSF1R inhibition via PLX3397 leads to time- and sex-dependent effects on inflammation and neuronal maintenance after traumatic brain injury in mice. Brain Behav Immun 106, 49–66.
Weissgerber, T.L., Garovic, V.D., Milin-Lazovic, J.S., Winham, S.J., Obradovic, Z., Trzeciakowski, J.P., Milic, N.M., 2016. Reinventing Biostatistics Education for Basic Scientists. PLoS Biol 14 (4), e1002430.
Witjes, V.M., Boleij, A., Halffman, W., 2020. Reducing versus Embracing Variation as Strategies for Reproducibility: The Microbiome of Laboratory Mice. Animals (Basel) 10.
Wu, H., Wang, C., Wu, Z., 2015. PROPER: comprehensive power evaluation for differential expression using RNA-seq. Bioinformatics 31, 233–241.
Yoon, S., Nam, D., 2017. Gene dispersion is the key determinant of the read count bias in differential expression analysis of RNA-seq data. BMC genomics 18, 408.