International Journal of Obesity (2020) 44:1440–1449
https://doi.org/10.1038/s41366-020-0554-2

ARTICLE

Techniques and Methods

Murine genetic models of obesity: type I error rates and the power of commonly used analyses as assessed by plasmode-based simulation

Keisuke Ejima[1,2] · Andrew W. Brown[3] · Daniel L. Smith Jr[4,5,6] · Ufuk Beyaztas[7] · David B. Allison[1]

Received: 20 August 2019 / Revised: 6 February 2020 / Accepted: 12 February 2020 / Published online: 25 February 2020

Abstract

Background/Objectives Genetic contributors to obesity are frequently studied in murine models. However, the sample sizes of these studies are often small, and the data may violate assumptions of common statistical tests, such as normality of distributions. We examined whether, in these cases, type I error rates and power are affected by the choice of statistical test.

Subjects/Methods We conducted "plasmode"-based simulation using empirical data on body mass (weight) from murine genetic models of obesity. For the type I error simulation, the weight distributions were adjusted to ensure no difference in means between control and mutant groups. For the power simulation, the distributions of the mutant groups were shifted to ensure specific effect sizes. Three to twenty mice were resampled from the empirical distributions to create a plasmode. We then computed type I error rates and power for five common tests on the plasmodes: Student's *t* test, Welch's *t* test, Wilcoxon rank sum test (aka, Mann–Whitney *U* test), permutation test, and bootstrap test.

Results We observed type I error inflation for all tests, except the bootstrap test, with small samples (≤5). Type I error inflation decreased as sample size increased (≥8) but remained. The Wilcoxon test should be avoided because of heterogeneity of distributions. For power, a departure from the reference was observed with small samples for all tests. Compared with the other tests, the bootstrap test had less power with small samples.

Conclusions Overall, the bootstrap test is recommended for small samples to avoid type I error inflation, but this benefit comes at the cost of lower power. When sample size is large enough, Welch's *t* test is recommended because of high power with minimal type I error inflation.

Introduction

Murine (mouse and rat) models are widely used as preclinical experimental research models. Rodents offer the similarity of mammalian metabolic pathways, providing the opportunity to pursue research with human relevance where ethical and safety issues are involved. Researchers can control and manipulate the animal's genetic background and environment, enabling randomized, controlled experimentation, which is essential for drawing causal inference from experimental results. Murine models lay the preclinical foundation for translational research and provide potential mechanistic insights into basic biology.

Supplementary information The online version of this article (https://doi.org/10.1038/s41366-020-0554-2) contains supplementary material, which is available to authorized users.

✉ Keisuke Ejima
kejima@iu.edu

✉ David B. Allison
allison@iu.edu

[1] Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington, Bloomington, IN, USA

[2] Institute of Industrial Science, The University of Tokyo, Tokyo, Japan

[3] Department of Applied Health Science, Indiana University School of Public Health-Bloomington, Bloomington, IN, USA

[4] Department of Nutrition Sciences, University of Alabama at Birmingham, Birmingham, AL, USA

[5] Nutrition Obesity Research Center, University of Alabama at Birmingham, Birmingham, AL, USA

[6] Diabetes Research Center, University of Alabama at Birmingham, Birmingham, AL, USA

[7] Department of Mathematics, Bartin University, Bartin, Turkey

Table 1 Assumptions and hypotheses of five statistical tests commonly used to test for mean differences between groups.

Statistical tests	Assumptions	Null hypothesis
Student's *t* test	The distributions follow normal distribution. The variances are the same.	Means of the two groups are the same.
Welch's *t* test	The distributions follow normal distribution. The variances are not (necessarily) the same.	Means of the two groups are the same.
Wilcoxon test	No assumptions for the distributions.	Means of the two groups are the same given that all other characteristics of the distributions are identical.
Permutation test	The shape of the distributions (including variances) is identical.	Means of the two groups are the same.
Bootstrap test	No assumptions for the distributions.	Means of the two groups are the same.

A growing recognition of appropriate ethical considerations for the design and implementation of preclinical research using animal models has encouraged the scientific research community to consider Replacement, Reduction, and Refinement of animal models (the so-called 3Rs [1]). This guidance includes the selection of the most appropriate model with the fewest animals required and consideration of alternatives to live animals when possible. In parallel with dissemination and adoption of the 3Rs, concerns over reproducibility of study results have increased in recent years in science generally [2, 3] and in preclinical and murine studies in particular [4–6]. For example, Kilkenny et al. surveyed "reporting, experimental design, and statistical analysis in published biomedical research using laboratory animals" and found that "[m]ost of the papers surveyed did not use randomization (87%) or blinding (86%), to reduce bias in animal selection and outcome assessment" [7]. Furthermore, Festing listed methodological issues related to poor study design in animal models, such as incorrect randomization, failure to blind, inadequate external validity, and incorrect statistical analysis [4]. To improve study design and minimize related errors, guidelines and checklists have been developed by various academic societies and organizations. For animal research, the Planning Research and Experimental Procedures on Animals: Recommendations for Excellence and the Animal Research: Reporting of In Vivo Experiments (ARRIVE) guidelines are utilized in planning animal experiments and reporting experimental results, respectively [8, 9].

Incorrect statistical analysis combined with small sample size is one major source of failure in reproducing results. The ARRIVE guidelines recommended checking three points regarding statistical methods: "(1) Provide details of the statistical methods used for each analysis. (2) Specify the unit of analysis for each dataset (e.g., single animal, group of animals, single neuron). (3) Describe any methods used to assess whether the data met the assumptions of the statistical approach" [8].

We need to "assess whether the data met the assumptions of the statistical approach" because if the data do not, we may report incorrect conclusions more or less frequently than expected. For example, genetic alterations that do not affect weight in reality might be wrongly reported as being *effective* (also called type I errors or false positives) more frequently if the assumptions are violated (see Supplemental Information for a more detailed explanation of the type I error rate).

There are many statistical tests for comparing means of groups; herein, we limit our discussion to those used to compare means of two groups. Table 1 summarizes five commonly used statistical tests with their assumptions and null hypotheses: (1) Student's *t* test [10], (2) Welch's *t* test [11], (3) Wilcoxon test [12, 13], (4) permutation test [14], and (5) bootstrap test [15]. Important differences in assumptions are that Student's and Welch's *t* tests assume normal distributions of each group. The permutation test and the bootstrap test are similar in computational processes; however, the permutation test assumes that the shape of the distributions is identical [16, 17]. Consequently, permutation tests assume equal variance as Student's *t* test. The null hypothesis of the Wilcoxon test is that the distributions of the two groups are the same, whereas that for the other four tests is that the means of the two groups are the same. Although the Wilcoxon test is not a test for mean or median, it has been frequently used to test central tendency.

Unfortunately, the assumptions behind each statistical test can themselves be difficult to test, especially when sample sizes are small, which is typical in many animal studies. For example, Student's *t* test is commonly used to test the difference in means between two groups. However, to use Student's *t* test, the data from two groups are assumed to follow a normal distribution with equal variance. The equality of variance can be tested, and if the test is significant (e.g., $p < 0.05$ for the test of equal variance), equal variance assumption is rejected, and Student's *t* test should not be used. However, even if the test of equal variance is not significant (e.g., $p > 0.05$), equal variance is not guaranteed. To conclude, determining that "equal variance" exists as opposed to "failing to reject

equal variance" when the test is not significant is an error of "accepting the null" [18, 19]. Because of the low power to test the equal variance assumption with small sample sizes, relying on such tests to determine whether Student's *t* test should be used may result in selecting inappropriate statistical methods.

Murine genetic models of obesity pose such challenges with assumptions. As presented herein, there is frequently less variance in the distribution of body weight in control animals than in their genetic mutant counterparts. Furthermore, sex differences may be present, and weights are not necessarily normally distributed. Studies and statistical theory have established how selection of statistical methods and sample size affect type I error and power from theoretical perspectives or by using simulated data in which population distributions are explicitly specified (normal distribution is frequently used) [20–22]. However, whether and how the sample size and statistical test influence type I error rates and power has not been demonstrated using empirical data for murine genetic models of obesity. Such tests can provide important insights into the idiosyncrasies that animal researchers will face. This is particularly important for research where outcomes measured with high reliability and validity are considered for translational impact (e.g., body weight in nutrition and obesity research). Given that we cannot know the true population distribution from which the empirical data are created, and that the empirical data are not necessarily created from commonly used probability distributions (e.g., normal distribution), assessing type I error rates and power using empirical data is more realistic and practical.

In this study, therefore, we investigated the influence of sample size and choice of statistical tests on the testing of mean differences in body mass between control and mutant murine groups. We assessed type I error rates and power for "plasmodes," which are simulated experiments created from empirical mouse data.

Materials/Subjects and methods

What is a plasmode?

The plasmode was proposed by Cattell and Jaspers, defined as "a set of numerical values fitting a mathematico-theoretical model" [23]. In the age of bulky, complex, high-dimensional data (such as sequence data), Mehta et al. concisely defined the term as "a real dataset whose true structure is known" [24]. On the other hand, classical simulation data (as defined herein) are drawn from specific distributions with specific parameters (i.e., mean, variance-covariance). Compared with simulation data, a plasmode has an advantage because it can preserve data structure, without assuming distributions and interdependency of variables: "Plasmodes may represent actual experimental data sets better than simulations do" [24]. The plasmode approach has been used to evaluate various statistical methods [25–27]. Plasmodes can be constructed by conducting experiments or resampling from empirical data. In this study, we used empirical data from a mouse genetic study of obesity, described in the next section, and resampled from the data to create a plasmode. Although obesity can be defined in multiple ways, we focused on body weight.

Data source

We used the data of Bouchard et al., who investigated the relationship between obesity and cholesterol cholelithiasis using three polygenic and five monogenic mouse models of obesity. We focused on monogenic strains and available controls. We included two monogenic mouse models of obesity (carboxypeptidase E [Cpe^{fat}], leptin receptor [$Lepr^{db}$]) and their control with the same genetic background (C57BLKS/J) and four monogenic mouse models (agouti yellow [A^{y}], tubby [*tub*], leptin [Lep^{ob}], leptin receptor [$Lepr^{db}$]) and their control with the same genetic background (C57BL/6J). All animals were reported to be treated in the same condition: "Animals were provided free access to Rodent Laboratory Chow (Purina Mills, Richmond, VA) and acidified water (adjusted with HCl to a pH of 2.8–3.2) to retard microbial growth. Mice were housed in a temperature (22–23 °C) controlled room with alternating 14:10 h light–dark cycles of regular diurnal periodicity. When 10 weeks old, the mice were weighed and fed a lithogenic diet containing 15% butter fat, 1% cholesterol, and 0.5% cholic acid for 8 weeks" [28]. We investigated the influence of genetic differences on body mass (weight). To avoid the effect of the different diets after baseline assessment (the diets before and after the baseline assessment were not the same), we used weight at baseline (10 weeks old). The data were downloaded from the Mouse Phenome Database at The Jackson Laboratory [29]. In the following analysis, each monogenic mouse model was compared against its corresponding control.

Plasmode simulation

We assume in the plasmode simulation that the empirical data represent a whole population. Therefore, we treated each sex × genotype combination as representing the whole population of those animals. An overview of the simulation process to compute type I error rates and power is summarized in Fig. 1.

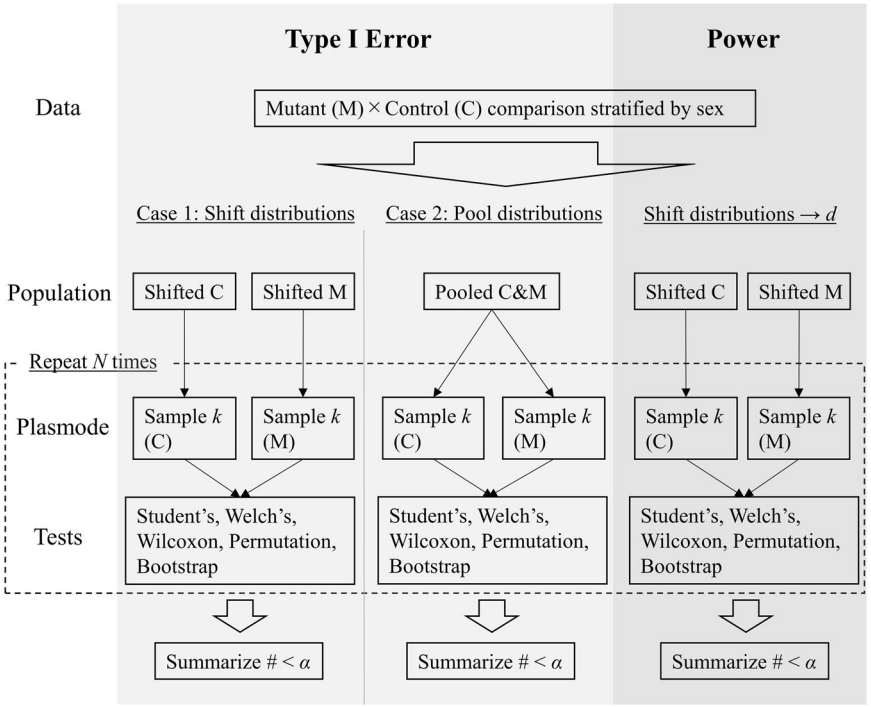

Fig. 1 Summary of the simulation protocol. The data were extracted from a genetically modified murine model published in Bouchard et al. [28] and stratified by sex. After the creation of populations of control and mutant animals, N plasmodes (each plasmode consists of k control animals and k mutant animals) were created by resampling from the populations. The five different tests were implemented for each of the plasmodes, and the p values obtained were summarized to compute type I error rates or power.

To compute type I error rates using the empirical data, we reformed the empirical weight distributions to realize the null hypothesis (i.e., there is no mean difference between the groups) in two ways. In the first case (Case 1), we added a constant value to each animal's body weight in the control group such that the means of the control and the mutant animals were the same. In the second case (Case 2), we combined the data from the control and the mutant groups into one pooled distribution and assumed that both the control and the mutant data were generated from the combined distribution. To compute power (or type II error rates) using the empirical data, we reformed the empirical distributions to realize a nominal Cohen's d (1.0, 1.5, 3.0) by adding a constant value to each animal's body weight in the control group.

Because "[a]nalysis of a single plasmode is minimally compelling," we created numerous plasmodes and analyzed each plasmode independently. The results obtained from the analyses were combined and used to assess the quality of each statistical test (described in the next section). Each plasmode is composed of k_1 mice of the mutant group and k_2 mice of the control group of the same sex. These mice are resampled from the empirical populations with replacement. For simplicity, we enforced equal sample sizes between groups: $k_1 = k_2 = k$. We repeated the entire process separately for different k (3–5, 8, 12, 16, and 20). The simulation was stratified by sex within mutant-control comparisons.

Five different statistical tests on each plasmode

Because murine studies frequently focus on the mean difference between treatment and control groups, we compared mean differences in body weight between mutant groups and their corresponding control groups stratified by sex using five common statistical tests: (1) Student's t test [10], (2) Welch's t test [11], (3) Wilcoxon test [12, 13], (4) permutation test [14], and (5) bootstrap test [15]. These statistical tests were performed on each plasmode.

Summarize plasmode results

For computation of the type I error rate, we created plasmodes N (=15,352), and obtained p values for each of the five tests. The point estimates of the type I error rates were computed using nominal significance levels (α; the value is one of {0.05, 0.01, 0.005, 0.001}): $\hat{\alpha} = \frac{\sum_{i=1}^{N} I(p_i<\alpha)}{N}$. The 95% CI of the type I error rates were calculated using a single proportion: $\hat{\alpha} \pm 1.96\sqrt{\frac{\hat{\alpha}(1-\hat{\alpha})}{N}}$. Using $N = 15{,}352$ provides a 95% CI of ±0.0005 for an α of 0.001, and thus is within rounding error for estimating the 95% CI for that α value. For the power computation, we created 1000 plasmodes. To obtain reference values of power for each test, we created 1000 samples randomly sampled from normal distribution $N(0, 1)$ and $N(d, 1)$, where d (the value is one of {1.0, 1.5, 3.0}) corresponds to nominal effect size, and performed each test on the samples.

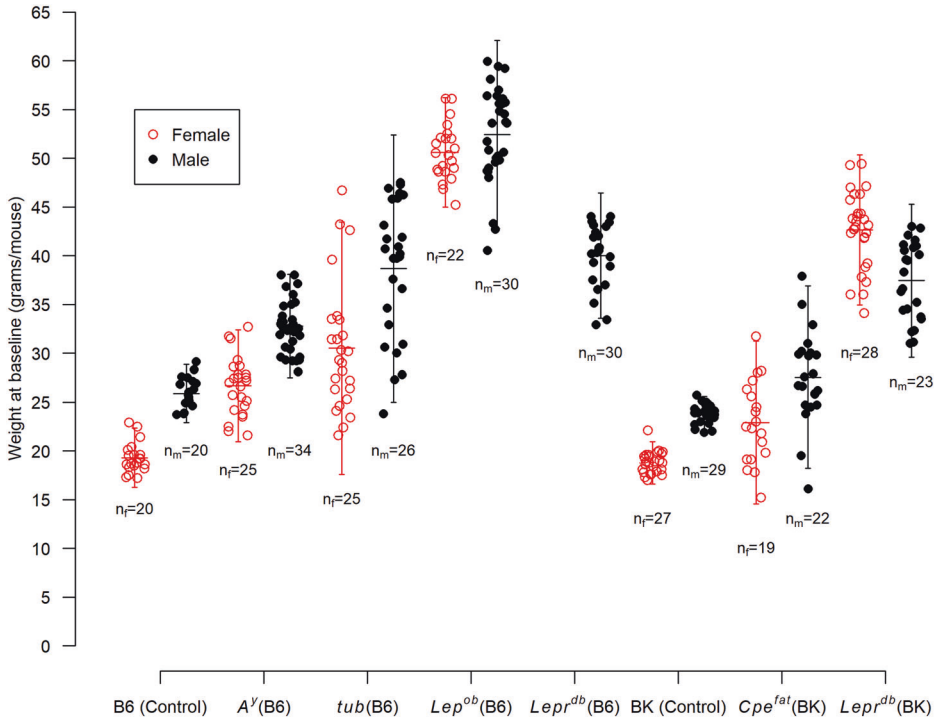

Fig. 2 Summary of baseline body mass (weight). Baseline body masses (weights) were used from four monogenic murine models of obesity (agouti yellow [A^y], tubby [*tub*], leptin [Lep^{ob}], and leptin receptor [$Lepr^{db}$]) and their control with the same genetic background (C57BL/6J; B6) and two monogenic murine models of obesity (carboxypeptidase E [Cpe^{fat}] and leptin receptor [$Lepr^{db}$]) and their control with the same genetic background (C57BLKS/J; BK). The open red circle and closed black circles correspond to female and male mice, respectively. The error bars correspond to the 95% confidence intervals of each group (female mice for $Lepr^{db}$ (B6) were not available).

We mainly argue the case with $\alpha = 0.05$ given that it is a standard threshold used in many domains, though we recognize 0.05 has been challenged as the appropriate threshold (cf. [30]), and others have argued against using bright line significance testing altogether (cf. [31]). We used the Karst high-throughput computing cluster at Indiana University (Bloomington, IN) and the statistical computing software R 3.4.1 (R Development Core Team) for all simulations and calculations. All statistical tests are two-tailed. The simulation protocol and the analytic procedures including the parameters (i.e., sample size, significance level, and effect size) were internally prespecified. The codes used in this study will be available online (https://doi.org/10.5281/zenodo.1488359).

Results

Characterization of weight distributions of mutant and control animals

Table S1 and Fig. 2 summarize the data. A number of animals of each strain (stratified by sex) ranged from 19 to 34. Mean weights of all the mutant groups were heavier than those of the corresponding control groups, and this difference was statistically significant by use of any of the four tests testing mean difference. The distributions are also statistically significantly different by use of the Wilcoxon test. The effect size was assessed by Cohen's *d*, which ranged from 1.10 to 13.57. In light of the rules of thumb for effect size proposed by Sawilowsky, the weight difference between the controls and the mutants was considered to be "very large" or "huge" (0.01 = "very small," 0.2 = "small," 0.5 = "medium," 0.8 = "large," 1.2 = "very large," 2.0 = "huge") [32]. We note that the equal variance assumption was rejected for all the comparisons ($p < 0.05$; *F*-test).

Type I error rates

The computed type I error rates with $\alpha = 0.05$ for 11 sex × strain combinations are shown in Fig. 3. For Student's *t* test, Welch's *t* test, and the permutation test, substantial type I error inflation (i.e., the lower bound of 95% CI was over 0.05; inflated) was observed for small sample sizes ($n = 3$ or 4) for Case 1, whereas the bias in the type I error rate was relatively smaller for Case 2. The magnitude of type I error inflation was mitigated as the sample size increased regardless of sex or strain for all tests except the bootstrap test, which rarely had inflated type I error rates (see Fig. 3). When type I error rate inflation was observed for the bootstrap test, the magnitude of inflation was relatively small compared with the other tests, even with small sample sizes. However, bootstrap tests were frequently conservative (i.e., type I error rates lower than significance levels), particularly at low sample sizes. The proportions of strain × sex combinations for each sample size with inflated or conservative type I error rates from Fig. 3 are summarized in Fig. 4. For Case 1, inflated error rates were reduced by increasing sample size for Student's *t* test, Welch's *t* test,

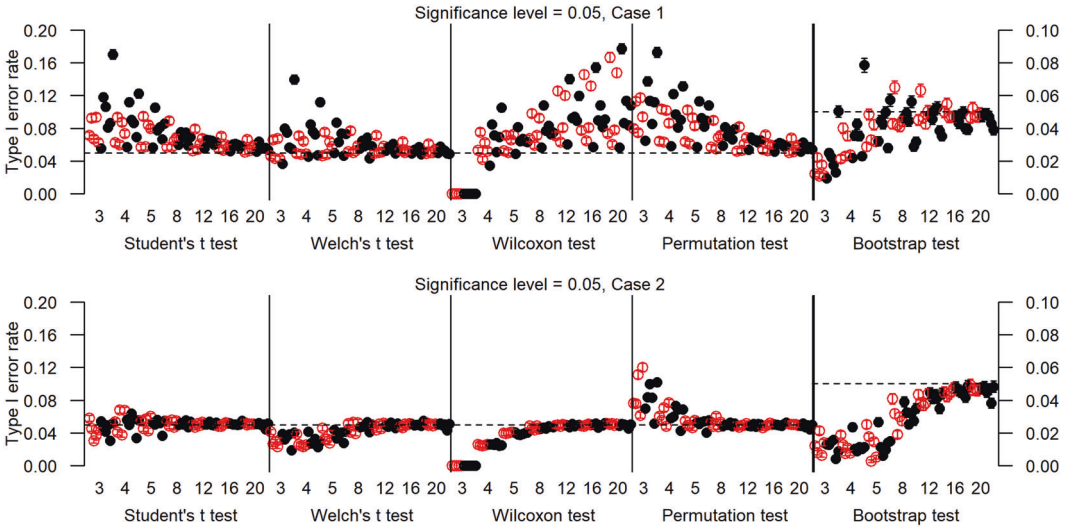

Fig. 3 Estimated type I error rate from the plasmode-based simulation (significance level = 0.05). Estimated type I error rates for each case (Case 1: both distributions are centered on the same mean, maintaining other characteristics of the individual-group distributions including variances; Case 2: the two distributions are pooled into a single distribution) with nominal significance level, 0.05. Open red and closed black circles with bars are type I error rates and the 95% CI of female and male mice for different mutants (five and six strains for each sex), respectively. A horizontal dotted line corresponds to the significance level. Note that the scale of the *y*-axis is different only for the bootstrap test and is shown on the right.

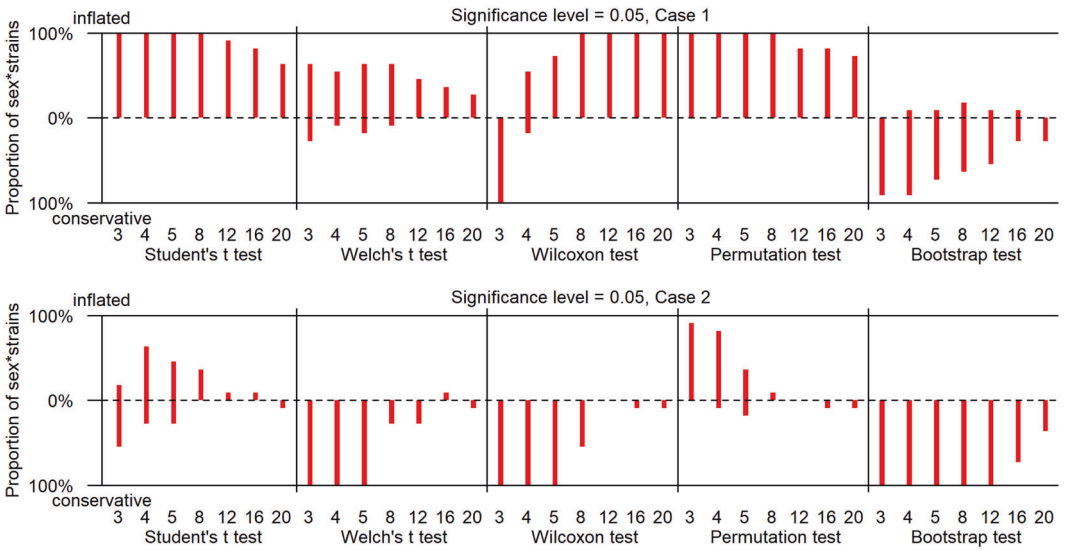

Fig. 4 Summary of type I error rate for each sample size (significance level = 0.05). The length of the red bar over and under the dashed line is the proportion of strain × sex combinations for each sample size with significant inflated (the lower bound of 95% CI of type I error rate is above the significance level) or conservative (the upper bound of 95% CI of type I error rate is below the significance level) type I error rates for each sample size.

and permutation test, but some inflation still remained with sample sizes of 20. The magnitude of reduction was larger for Welch's *t* test, which is reasonable because difference in variances is accounted for.

The interpretation of the type I error rate for the Wilcoxon test is complicated because it tests the difference in distribution rather than only the central tendency. For Case 1, the two plasmode groups were resampled from different distributions with different characteristics but with the same mean. As the sample size increased, the type I error rate increased for differences between the two because the increased sample allowed increased power to detect differences between the two distributions, not just differences in means. However, as the plasmodes were resampled from a pooled population for Case 2, this type I error was not inflated because the plasmodes were drawn from the same distribution.

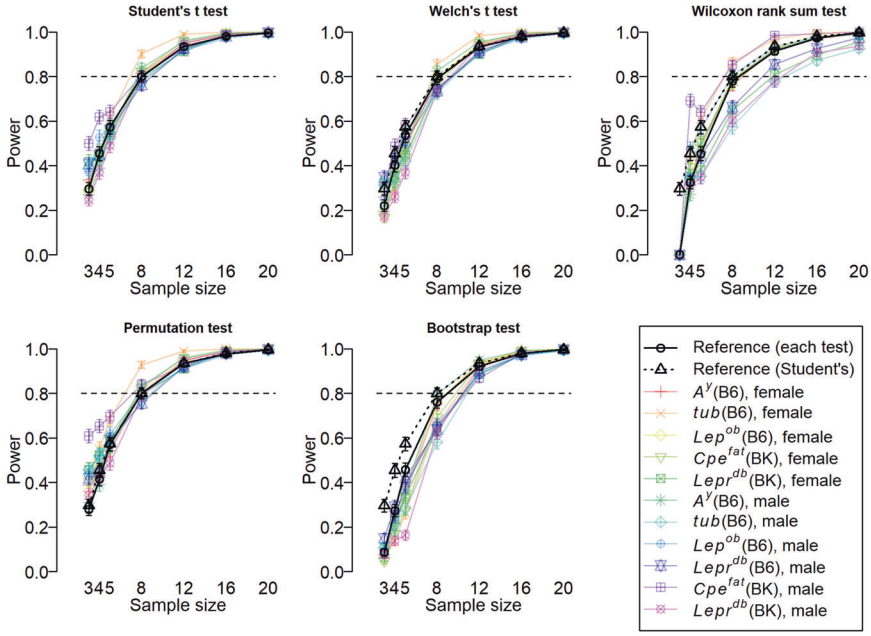

Fig. 5 Estimated power with effect size set to Cohen's *d* of 1.5 for each test. Different colors correspond to different strain × sex combinations. The black thick and dotted lines correspond to the reference values assuming normal distributions for each test and the Student's *t* test, respectively. To obtain reference values for each test, we created 1000 samples randomly sampled from normal distributions $N(0,1)$ and $N(1.5,1)$ and tested with each test and Student's *t* test. Significance level is 0.05.

We observed qualitatively similar results for the other nominal significance levels (Figs. S1–S3); however, we highlight a few differences. As α becomes lower, the ability of the Wilcoxon, permutation, and bootstrap tests to reject the null hypothesis diminishes. When sample sizes become too low, the type I error rates of these tests become 0 because there is no set of results capable of rejecting the null hypothesis at that level.

Overall, the results presented in Figs. 3 and 4 suggest that the bootstrap test controlled the type I error inflation even for small sample sizes. However, the bootstrap test was conservative for small sample sizes, and at some αs and sample sizes, the tests were unable to ever reject the null.

Statistical power and type II error rates

The power (Fig. 5) was computed by adjusting the mean difference between mutant and control mice to the same effect size (Cohen's *d* of 1.5) for all strain × sex combinations. Figure 5 shows the power for each strain × sex combination, and each panel corresponds to a different test. For each test, the power increased as the sample size increased. The power of the reference values approached or reached 80% with a sample size (n) of 8 for any test. Again, the power for the Wilcoxon test was zero for sample size 3 because of the inherent properties of the test. Comparing the reference values of the other four tests, the power of the bootstrap test was the lowest. Departure from the reference was not observed for large sample sizes (≥12) for any of the tests except the Wilcoxon test, for which the power was markedly below the reference for some strain × sex combinations. With small sample size (≤5), departure (both higher and lower) from the reference was observed for all the tests for some strain × sex combinations. Figure S4 shows the powers for each test, and each panel corresponds to a single strain × sex combination (note: the reference line is drawn based on the Student's *t* test). The power of the bootstrap test was consistently the lowest (except for the Wilcoxon test) for any strain × sex combination. This was not surprising because a large type I error rate contributes to a larger power of the test because rejections of the null from type I errors are counted in the calculation of the type II error as well.

The results for the other effect sizes are shown in Figs. S5–S8. The power increased as the effect size and sample size increased. For an effect size of 1.0, the departure from the reference value became clearer for large sample sizes for the Wilcoxon test for some strain × sex combinations. For a large effect size ($d = 3.0$) and small sample size, the powers of Welch's *t* test and the bootstrap test were lower than reference values, whereas those for the permutation test were higher than the reference values.

Discussion

We examined type I error rates and power using a plasmode approach, which allowed us to assess both the magnitude and the direction of the bias in these statistical variables in body weight in murine genetic models of obesity. For type I error rates, inflation was observed for Student's *t* test, Welch's *t* test, and the permutation test, especially with small sample sizes (<8) for Case 1. The inflation in the type I error was mitigated as sample size

increased for those three tests; however, for the Student's t test and the permutation test, the inflation remained, though smaller, because equal variance assumptions behind those tests were violated. This result is consistent with previous studies in which normal distributions with the same mean and different variances between groups were used as population distributions, and Student's t test was applied for sample data [33–35]. The Wilcoxon test is not appropriate for small sample sizes when setting a small significance level, as p values theoretically cannot be below the significance levels at small sample sizes (i.e., $n = 3$ or 4). Thus, recommendations to use smaller α (e.g., 0.005) need to account for the requirement for sufficiently large samples. Furthermore, type I error inflation was observed for large sample sizes with the Wilcoxon test when the distributions except mean differed (Case 1). The bootstrap test was consistently conservative, especially for small sample sizes.

These observations for type I error rates are consistent with different significance thresholds. The powers for the bootstrap test were the lowest compared with the other tests, except the Wilcoxon test, for any sample sizes, and even lower compared with the simulated references. As sample size increased, the power consistently increased. Variation in power was observed among strain × sex combinations, especially for small sample sizes. Welch's t test and bootstrap were underpowered, and the permutation test was overpowered, for small sample sizes. Departure from reference values was observed for any sample size for the Wilcoxon test.

We enforced equivalent sample sizes for the group comparisons in this study. However, unequal sample sizes can lead to biased power and type I error rates [33–35]. This could be a problem in murine genetic models, where mutants can be expensive or fewer in number due to breeding outcomes.

A strength of this study is that we used a plasmode simulation approach instead of parametric or contrived distributions. Given that the data generation process and population distributions are unknown in general (i.e., only sample distributions are available), a plasmode approach is useful for computing reliable type I error rates and power without parametric assumptions of the underlying data distributions. Furthermore, we used multiple mutants paired to their respective controls for our simulation and obtained qualitatively similar results, which strengthens the generalizability of our results.

A few limitations of this study should be stated. Although we assumed that the samples represented the populations, we have no data to confirm this. If the sample distributions differ from the population distributions (such as through selection bias or by chance), type I error and power computed on the basis of the samples may differ from what is expected from the population. To avoid and minimize errors due to sampling, minimizing differences between animals (i.e., genetic background and treatments) and increasing sample size are recommended. In the data used for this study, the mice shared a genetic background, were treated in the same environment, and constituted a relatively large sample compared with most preclinical studies (often >20 per sex × strain combination). Another issue is potential limitations in the generalizability of our findings. We focused on weight as an outcome, and type I error rates and power might differ numerically for different outcomes (e.g., fat mass). Because the plasmode approach is adaptable, the same simulation can be implemented on a case-by-case basis. Furthermore, we compared only two groups using tests permitting head-to-head comparisons. Analyses for more groups (e.g., ANOVA) or more complex designs (e.g., repeated measures) are warranted.

Although the plasmode simulation is useful, it may not be easily accessible for many nonstatisticians. We published the code used in the study (https://doi.org/10.5281/zenodo.1488359) to enhance the reproducibility of this study. A potential future direction of this work is to design turnkey plasmode-based simulation utilities.

In the meantime, the simplest advice is to increase sample size, because differences in power and type I error control converge. However, given budgetary constraints, difficulties in making some genetic mutants, and the spirit of the Reduction part of the 3Rs [1], choosing a sufficiently powered sample that is as small as possible is desirable. We therefore make the following recommendations in planning a study:

(1) Consult a statistician for advice given expectations for distributions of outcomes, such as whether the outcomes are expected to be normally distributed, and whether equal variance is a reasonable assumption. Herein, we show that in most cases assuming equal variance in genetic mutants compared against wildtype animals is unreasonable (Fig. 2).
(2) Given the likelihood that distributions will differ between strains, Wilcoxon and related tests should not be used unless there is an interest in differences in distributions as opposed to differences in means or central tendency per se.
(3) Determine the statistical test taking the consequences of type I or type II errors rates into account.

Acknowledgements This study was supported in part by NIH grants 3P30DK056336 (DBA), R25DK099080 (DBA), R25HL124208 (DBA) and Japan Society for Promotion of Science (JSPS) KAKENHI grant 18K18146 (KE). The data analyses and simulation were performed using a supercomputer, Karst, which was supported in part by Lilly Endowment, Inc., through its support for the Indiana University

Pervasive Technology Institute, and in part by the Indiana METACyt Initiative. The Indiana METACyt Initiative at IU was also supported in part by Lilly Endowment, Inc. The opinions expressed are those of the authors and do not necessarily represent those of the NIH or any other organization. All the code which was used in this study will be available through the following webpage: https://doi.org/10.5281/zenodo.1488359. Supplementary information is available at the International Journal of Obesity's website.

Author contributions DBA designed the research. DLSJ and AWB gathered the data. KE and AWB performed statistical analysis. DBA, DLSJ, and UB assisted in data analysis. All authors were involved in writing or editing the paper and had final approval of the submitted and published versions

Compliance with ethical standards

Conflict of interest UB has no conflicts of interest. In the last 12 months, DBA has received personal payments or promises for same from for-profit organizations including: Biofortis; Gelesis; Fish & Richardson, P.C.; IKEA; Law Offices of Ronald Marron; Sage Publishing; Tomasik, Kotin & Kasserman LLC; Medpace; Nestle; WW (formerly Weight Watchers International, LLC) and was an unpaid member of the International Life Sciences Institute North America Board of Trustees. In the last 12 months, AWB has received personal payments or paid travel from: American Society for Nutrition, Indiana University, Kentuckiana Health Collaborative, Rippe Lifestyle Institute, Inc. Indiana University has received grants from the following entities to support some of the authors' research or educational activities: NIH; Alliance for Potato Research and Education; American Federation for Aging Research; Dairy Management Inc; Herbalife; Laura and John Arnold Foundation; Oxford University Press; Sloan Foundation; University of Alabama at Birmingham. In the last 12 months, DLSJ has received personal payments or paid travel from: University of Alabama at Birmingham. University of Alabama at Birmingham has received grants from the following entities to support some of the authors' research or educational activities: NIH; Alliance for Potato Research and Education. In the last 12 months, KE has received personal payments or paid travel from: The University of Tokyo. The University of Tokyo has received grants from the following entities to support some of the authors' research or educational activities: Japan Society for the Promotion of Science.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. National Centre for the Replacement, Refinement and Reduction of Animals in Research (NC3Rs). https://www.nc3rs.org.uk/. Accessed 12 Feb 2019.
2. Allison DB, Brown AW, George BJ, Kaiser KA. Reproducibility: a tragedy of errors. Nature. 2016;530:27–9.
3. Brown AW, Kaiser KA, Allison DB. Issues with data and analyses: errors, underlying themes, and potential solutions. Proc Natl Acad Sci. 2018;115:2563–70.
4. National Academies of Sciences, Engineering, and Medicine. Reproducibility issues in research with animals and animal models: workshop in brief. Washington, DC: The National Academies Press; 2015. p. 8.
5. Begley CG, Ioannidis JP. Reproducibility in science: improving the standard for basic and preclinical research. Circ Res. 2015;116:116–26.
6. Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature. 2012;483:531.
7. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol. 2010;8:e1000412.
8. ARRIVE guidelines. https://www.nc3rs.org.uk/arrive-guidelines. Accessed 12 Feb 2019.
9. Smith AJ, Clutton RE, Lilley E, Hansen KEA, Brattelid T. PREPARE: guidelines for planning animal research and testing. Lab Anim. 2018;52:135–41.
10. Student. The probable error of a mean. Biometrika. 1908;6:1–25.
11. Welch BL. The generalisation of student's problems when several different population variances are involved. Biometrika. 1947;34:28–35.
12. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat. 1947;18:50–60.
13. Wilcoxon F. Individual comparisons by ranking methods. Biom Bull. 1945;1:80–3.
14. Pitman EJG. Significance tests which may be applied to samples from any populations. J R Stat Soc. 1937;4:119–30.
15. Hall P, Wilson SR. Two guidelines for bootstrap hypothesis testing. Biometrics. 1991;47:757–62.
16. GEP Box, Andersen SL. Permutation theory in the derivation of robust criteria and the study of departures from assumption. J R Stat Soc Ser B. 1955;17:1–34.
17. Hayes AF. Permutation test is not distribution-free: Testing H_0: $\rho = 0$. Psychol Methods. 1996;1:184–98.
18. Gibbons JD, Chakraborti S. Comparisons of the Mann-Whitney, Student's t, and Alternate t tests for means of normal distributions. The J Exp Educ. 1991;59:258–67.
19. Zimmerman DW, Zumbo BD. Parametric alternatives to the Student t test under violation of normality and homogeneity of variance. Percept Motor Skills. 1992;74:835–44.
20. Zimmerman DW. Statistical significance levels of nonparametric tests biased by heterogeneous variances of treatment groups. J Gen Psychol. 2000;127:354–64.
21. Rogan JC, Keselman HJ. Is the ANOVA F-test robust to variance heterogeneity when sample sizes are equal? An investigation via a coefficient of variation. Am Educ Res J. 1977;14:493–8.
22. Box GEP. Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. Ann Math Stat. 1954;25:290–302.
23. Cattell RB, Jaspers J. A general plasmode (No. 30-10-5-2) for factor analytic exercises and research. Multivar Behav Res Monogr. 1967;67-3:211.
24. Mehta T, Tanik M, Allison DB. Towards sound epistemological foundations of statistical methods for high-dimensional biology. Nat Genet. 2004;36:943.
25. Gadbury GL, Xiang Q, Yang L, Barnes S, Page GP, Allison DB. Evaluating statistical methods using plasmode data sets in the age of massive public databases: an illustration using false discovery rates. PLoS Genet. 2008;4:e1000098.
26. Newton MA, Noueiry A, Sarkar D, Ahlquist P. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics. 2004;5:155–76.
27. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003;31:e15–e.
28. Bouchard G, Johnson D, Carver T, Paigen B, Carey MC. Cholesterol gallstone formation in overweight mice establishes that

obesity per se is not linked directly to cholelithiasis risk. J Lipid Res. 2002;43:1105–13.
29. The Jackson Laboratory. Mouse Phenotype Database. The Jackson Laboratory; 2018. https://phenome.jax.org/projects/Paigen3. Accecced 31 May 2018.
30. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, et al. Redefine statistical significance. Nat Hum Behav. 2018;2:6–10.
31. Goodman SN. How sure are you of your result? Put a number on it. Nature. 2018;564:7.
32. Sawilowsky SS. New effect size rules of thumb. J Mod Appl Stat Methods. 2009;8:26.
33. Zimmerman DW. Comparative power of Student t test and Mann-Whitney U test for unequal sample sizes and variances. J Exp Educ. 1987;55:171–4.
34. Zimmerman DW. A note on homogeneity of variance of scores and ranks. J Exp Educ. 1996;64:351–62.
35. Zimmerman DW. Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. J Exp Educ. 1998;67:55–68.