Published in final edited form as:

Cell Rep. 2021 August 17; 36(7): 109560. doi:10.1016/j.celrep.2021.109560.

Empirical versus theoretical power and type I error (false-positive) rates estimated from real murine aging research data

Irene Alfaras[1,4,5], **Keisuke Ejima**[2,3,4], **Camila Vieira Ligo Teixeira**[1], **Clara Di Germanio**[1,6], **Sarah J. Mitchell**[1,7], **Samuel Hamilton**[1,8], **Luigi Ferrucci**[1], **Nathan L. Price**[1], **David B. Allison**[2], **Michel Bernier**[1], **Rafael de Cabo**[1,9,*]

[1]Translational Gerontology Branch, National Institute on Aging Intramural Program, National Institutes of Health, Baltimore, MD 21224, USA

[2]Department of Epidemiology and Biostatistics, Indiana University School of Public Health-Bloomington, Bloomington, IN 47405, USA

[3]Graduate School of Medicine, The University of Tokyo, Tokyo 1130033, Japan

[4]Contributed equally to this work

[5]Present address: Aging Institute of UPMC and the University of Pittsburgh, Pittsburgh, PA 15213, USA

[6]Present address: Vitalant Research Institute, San Francisco, CA 94118, USA

[7]Present address: Department of Health Sciences and Technology, ETH Zürich, Zürich 8603, Switzerland

[8]Present address: Department of Biology, Northwestern University, Evanston, IL 60208, USA

[9]Lead contact

SUMMARY

We assess the degree of phenotypic variation in a cohort of 24-month-old male C57BL/6 mice. Because murine studies often use small sample sizes, if the commonly relied upon assumption of a normal distribution of residuals is not met, it may inflate type I error rates. In this study, 3–20 mice are resampled from the empirical distributions of 376 mice to create plasmodes, an approach for computing type I error rates and power for commonly used statistical tests without assuming a normal distribution of residuals. While all of the phenotypic and metabolic variables studied show considerable variability, the number of animals required to achieve adequate power

*Correspondence: decabora@grc.nia.nih.gov.

AUTHOR CONTRIBUTIONS

R.d.C. conceived and designed the study; I.A. and S.J.M. performed most experiments; I.A. conducted correlation and regression analyses of various datasets, while S.H. used R programming for effective data analysis and generation of figures; K.E. performed plasmode-based simulation for statistical analyses; and D.B.A. assisted in data analysis. M.B. wrote the first draft of the manuscript and created the figures. All authors were involved in editing the paper and proofreading the accepted version.

SUPPLEMENTAL INFORMATION

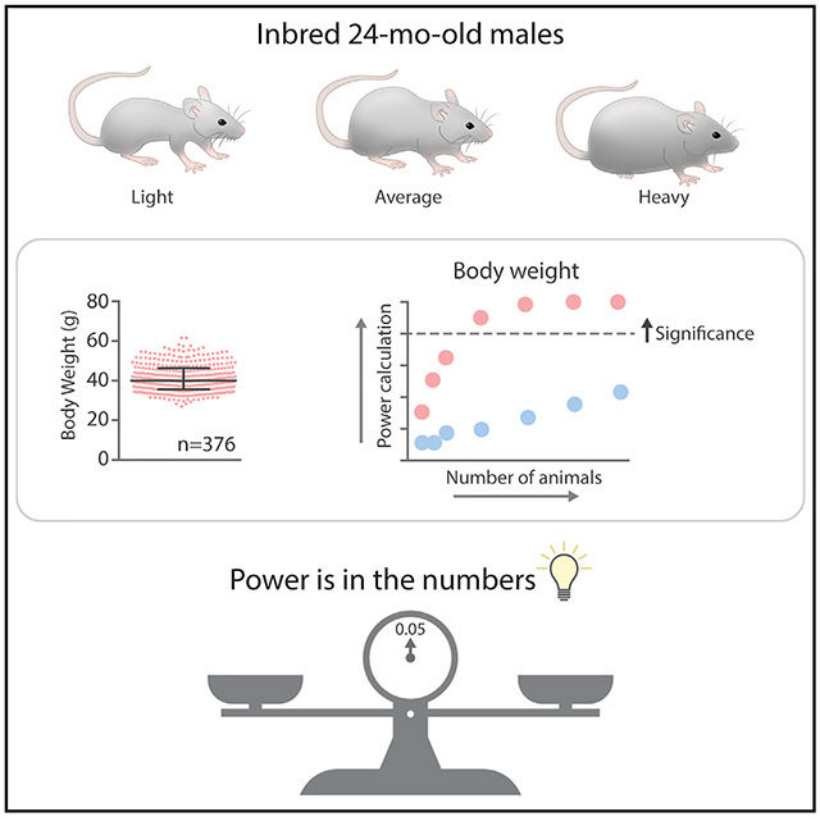
Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2021.109560.

DECLARATION OF INTERESTS

The authors declare no competing interests.

is markedly different depending on the statistical test being performed. Overall, this work provides an analysis with which researchers can make informed decisions about the sample size required to achieve statistical power from specific measurements without a *priori* assumptions of a theoretical distribution.

Graphical Abstract



In brief

Alfaras et al. report that the plasmode approach reveals that differently measured traits have distributions that affect power differently, and that trait type affects the minimal required sample size. Their findings expand the statistical and inferential toolbox of aging research.

INTRODUCTION

Laboratory mice in controlled environments are short-lived with a record lifespan of 4 years (Miller et al., 2002). A wide range of physiological, functional, behavioral, and pathological changes occur with age that determine individual mouse longevity. Mice develop different histopathological characteristics during their lives, such as lymphocyte infiltration, tissue inflammation and necrosis, cancer, and amyloidosis (Mitchell et al., 2016, 2019). To date, there is little to no information on how the aging biomarkers in humans mirror those associated with the normal progression of aging in mice. We surmise that the identification

of murine aging biomarkers may help develop novel approaches and interventions capable of improving health and quality of life that are translatable to humans.

In this study, we performed an extensive characterization of morphometric, glucoregulatory, and physical performance measures in a genetically homogeneous population of 24-month-old male C57BL/6 mice (n = 376) and assessed the degree of relationship between various outcomes and survival. Strikingly, there was considerable variability in outcome distributions across physiological and metabolic readouts, raising the thorny question of the assumptions made in commonly used statistical tests, such as normality. In the present study, a plasmode-based approach, based on utilizing datasets derived from real data from a cohort study (Gadbury et al., 2008), was used for assessing type I error rates and power without assuming a normal distribution of residuals (Ejima et al., 2020). Specifically, the influence of sample size (n = 3–20 per group), resampled with replacement from the overall population distribution of 376 mice, and choice of commonly used statistical tests on the testing of mean differences for various physiological and metabolic outcomes were investigated.

RESULTS AND DISCUSSION

Phenotypic heterogeneity of a genetically homogeneous inbred mouse strain

The lifespan of laboratory animals is affected by a number of factors such as sex, diet type, husbandry, and environmental conditions. In this study, we assessed the survival trajectory of a cohort of male C57BL/6 (n = 366) mice in our animal facility and recorded a median lifespan of 130.9 weeks (916 days) and maximum lifespan of 157 weeks (1,099 days) (Figure 1A).

Next, the variability in body weight (BW), morphometric measurements, and multiple biological and physical function outcomes was determined in this cohort of 366 animals (Figures 1B-1F). Statistical dispersion of the data was described as interquartile range (IQR), whereby the difference between the 25% and 75% IQR values is known as H-spread.

BW ranged from 26.8 to 61.4 g (median 40 g, H-spread 10.5 g) (Figure 1B; Table S1), despite mice having identical genetic background and being exposed to the same environmental factors, including air quality, water, food, temperature, bedding, and housing. Body fat, fluid content, lean body mass, and lean-to-fat ratio were among the phenotypic measures collected (Figure 1B; Table S1). After normalization as % BW, fat mass and lean-to-fat ratio datasets showed the greatest variability (median 22.85%, H-spread 6.66%; median 2.56, H-spread 0.86, respectively), whereas fluid and lean body mass had a tighter data clustering (median 6.87%, H-spread 0.54%; median 58.44%, H-spread 2.31%, respectively). Body temperature ranged from 33.5°C to 41.9°C (median 35.8°C, H-spread 0.7°C).

Venous blood from 6-h fasted mice was collected for fasting blood glucose (FBG) and insulin level determination, and calculation of homeostasis model assessment of insulin resistance (HOMA-IR), an accepted surrogate for estimating insulin resistance. Blood glucose ranged from 53 to 264 mg/dL (median 148 mg/dL, H-spread 46 mg/dL), whereas

plasma insulin concentrations ranged from 0.17 to 6.85 ng/mL (median 1.33 ng/mL, H-spread 0.8 ng/mL) (Figure 1C; Table S1; n = 355). HOMA2 values ranged from 0.50 to 13.89 (median 4.52, H-spread 2.48). Anxiety-like behaviors have been associated with aging in healthy mice (Morgan et al., 2018); and a connection exists between anxiety level and resting venous blood lactate levels in murine models (Hatchell and Mac-Innes, 1973). Here, blood lactate concentrations ranged from 0.5 to 5.8 mmol/L (median 1.3 mmol/L, H-spread 0.80 mmol/L) (Figure 1C; Table S1). Mouse behavior and motor performance were assessed during the light cycle using a standardized battery of phenotyping tests, including wire hang and cage top, grip strength measurement, and the rotarod test (Figure 1D; Table S1). For instance, the latency to fall off a hanging wire ranged from 0.061 to 1.43 s/g BW (median 0.312 s/g BW, H-spread 0.262 s/g BW), whereas the latency to fall off an accelerating rotarod ranged from 0.657 to 6.109 s/g BW (median 2.28 s/g BW, H-spread 1.777 s/g BW) (Figure 1D; Table S1). As in human muscle aging (Ballak et al., 2014), there were clear deficits in multiple motor performance tasks requiring muscle strength, coordination, and balance in old male mice, consistent with an age-related reduction in overall physical fitness.

Measuring energy expenditure in mice can be accomplished via an indirect respiration calorimetry system, known as Comprehensive Lab Animal Monitoring System (CLAMS) metabolic chambers (Martin-Montalvo et al., 2016). Here, O_2 consumption, CO_2 generation, heat production, and spontaneous locomotor activity were determined in 24-month-old mice (n = 64) during a 48-h period (Figure 1E; Table S1). The volumes of O_2 consumption and CO_2 generation in the dark phase were significantly higher compared to the light phase ($p < 0.0001$ by two-tailed paired t test, Table 1), consistent with the active/feeding period. In the dark phase, the rate of oxygen consumption (VO_2) ranged from 92.2 to 190.0 mL/h (median 129.0 mL/h, H-spread 25.3 mL/h), whereas the rate of CO_2 production (VCO_2) ranged from 81.7 to 176.4 mL/h (median 116.8 mL/h, H-spread 18.6 mL/h). Similarly, the resultant VCO_2/VO_2 ratio, known as respiratory exchange ratio (RER), was significantly higher during the dark cycle ($p < 0.0001$, Table 1), ranging from 0.772 to 0.954 (median 0.906, H-spread 0.043). Heat production was also elevated in the dark period and ranged from 0.463 to 0.815 kcal/h (median 0.639 kcal/h, H-spread 0.090 kcal/h) ($p < 0.0001$; Figure 1E; Table S1). Figure 1F depicts the average amplitude of VO_2, VCO_2, and the associated RER during the two light/dark cycles. Once again, mice showed significantly greater spontaneous locomotor activity during the dark phase than during the light period ($p < 0.001$, Table 1). Total activity at night ranged from 264 to 2,909 counts (median 1,736 counts, H-spread 901 counts), highlighting again a large data dispersion (Figure 1E; Table S1). Taken together, these results indicated clear phenotypic heterogeneity in an isogenic group of male mice of similar age, and constant animal husbandry practices.

Next, we conducted plasmode-based simulations (Ejima et al., 2020) to determine appropriate sample sizes with sufficient power (80% as an example) to detect significant changes in outcome distributions while avoiding type I error rate deviation from a prespecified significance level (0.05 as an example). Three to 20 mice (per group) were resampled with replacement from the empirical distributions of the 366 animals to create control and treatment groups. The datasets thereby created are called plasmodes. Note that control and treatment groups are created by resampling from the same original outcome distributions for type I error simulation, whereas the treatment group was resampled from

shifted distributions (i.e., 10%–50% of the mean was added to the original data) for power simulation. Type I error rates and power were computed using five common tests based on 1,000 plasmodes: Student's t test, Welch's t test, Wilcoxon rank sum test (also known as Mann-Whitney U test), permutation test, and a bootstrap test. The computed powers were further compared with those computed from the conventional approach assuming the normal distribution (Jones et al., 2003).

The type I error rate did not diverge from the nominal significance level (0.05) for any outcome measure or sample size for Student's t test, Welch's t test, and permutation test (Figure 2A; Figure S1; Table S2). As was noted before, type I error deflation was observed for small sample size (e.g., n = 3 or 4) for Wilcoxon and bootstrap tests (Dwivedi et al., 2017). However, the lack of type I error inflation in small sample sizes is likely due in part to the fact that the control and treatment group distributions are assumed to be identical. As expected, the power increased with sample size and the percent increase in treatment group (Figure 2B; Figure S2; Table S3). The Welch's test is known to avoid type I error rate inflation by accounting for different variances between control and treatment groups (Ejima et al., 2020). Thus, the effect of effect size (% increase in treatment group) on the minimal sample size required to achieve sufficient power (80%) of all outcome measures using the Welch's test is illustrated as a heatmap (Figure 2D). The cumulative distributions of empirical data along with the normal distributions with the same means and variances are depicted (Figure 2C; Figure S3). A summary of the p values of the Shapiro-Wilk test (a test for normality) and the skewness and the kurtosis for each variable, both of which quantify the magnitude of violation of the normality assumption (both the skewness and the kurtosis are zero with a normal distribution), is provided (Table S4). The powers of the four tests were close to those of the conventional approach except for the Wilcoxon rank sum test, which is presumably because most of the outcomes followed, or were close to, the normal distribution. When the normality assumption was largely violated such as the lean/fat ratio (skewness = 5.31, kurtosis = 55.82), the power of those four approaches was far from that of the conventional approach assuming the normal distribution (Figure S2; Table S3). The Wilcoxon rank sum test's power was larger than the others when the normality assumption was violated.

The use of ratios for variables that were not normally distributed (e.g., morphometric measurements) could have various implications in subsequent parametric statistical analyses (Allison et al., 1995). These results attest to the inherent biological variation in a homogeneous mouse population even in a well-controlled environment.

Lifespan expectancy is extremely variable among mammalian species, ranging from 2.1 to 211 years (AnAge database at https://genomics.senescence.info/species/). In all cases, organisms undergo complex structural and functional changes during life that affect all levels of their organization. Genetic variants, lifestyle choices, and environmental factors all contribute to differential rates of aging, and the human body is not exempt.

To identify scientific explanations for the differential rates of organismal aging among humans, the Baltimore Longitudinal Study on Aging (BLSA) was initiated more than 6 decades ago to study changes that occur due to normal aging and distinguish them

from those caused by environmental factors, lifestyle choices, and diseases. A number of milestones have been achieved in aging research, particularly when comparing the rate of decline due to age versus disease-related factors, the relationship between health risk factors and aging, and the impact of behavioral trends on health and disease risk (National Institute on Aging et al., 2008). There is great variability in survival in the older population that cannot be solely accounted for by age and sex. Assessment of physical performance measures, such as gait speed, has provided important information about individualized estimates of survival (Studenski et al., 2011).

Aging is associated with distinct phenotypic changes ranging from the classical loss in physical performance to a decline in glucoregulation and energy homeostasis. Metabolic dysregulation leads to increased susceptibility to age-associated chronic diseases. This study shows a large phenotypic heterogeneity in multiple morphometric, physical, and metabolic data collected in male C57BL/6 mice at 24 months of age, which corresponds to ~70-year-old individuals (Mercken et al., 2017). Even though environmental variables were well controlled (e.g., food, temperature, husbandry), there are still large phenotypic variations among littermates, indicating that several factors could contribute to the overall population distribution in this isogenic cohort.

The importance of achieving sufficient power and controlling type I error rate is paramount for animal researchers who are considering choosing an empirical dataset that is as small as possible while avoiding false-positive rate inflation. Our recent study has illustrated the use of plasmode-based simulation to compute both the magnitude and the direction of the bias in type I error rates and power in BW in a murine obesity model (Ejima et al., 2020). Ejima et al. (2020) observed type I error inflation when the treatment and control group distributions are not the same (especially the variances). As we do not have treatment groups in the current study, we assumed that the control and treatment group distributions are identical. Therefore, we did not observe type I error inflation even with small sample size. However, it is widely reported that the variances are not the same between the groups. In clinical trials assigning a specific diet to a treatment group, for example, people/animals in a control group keep the same diet as before, whereas those in a treatment group vary in magnitude of compliance to the assigned protocol. Thus, the variance of BW tends to be larger in the treatment group than in the control group (Kaiser and Gadbury, 2013). In theoretical studies, small (less than five) and/or unequal sample sizes are known to lead to reduced power and inflated type I error rates (Zimmerman, 1987, 1988); however, as sample size increases, power consistently increases. Different results may be achieved if the population distributions are different (in variance, for example). We demonstrated that the type I error inflates with small sample size when the two groups' variances are different using our original data (BW) (Figure 3). In this simulation, the population distribution of the treatment group was created maintaining the same mean but different variance. Denoting the BW of animal i in the control and treatment group as x_i (original data) and y_i, respectively, y_i is represented by x_i as follows: $y_i = x_i + \sqrt{k}(x_i - \mu_x)$, where μ_x is the mean BW of the control group and k is the ratio of the variances (i.e., the variance in the treatment group being k times larger than that in the control group). The results illustrate type I error inflation with small sample size for Student's t test, Welch's t test, and the permutation test, whereas

deflation in type I error rate was observed for the Wilcoxon test and the bootstrap test with small sample sizes (Figure 3). However, because the Wilcoxon test is the test for difference in distributions, results should be interpreted with caution considering the purpose of test (i.e., are we testing the difference in mean or distribution?). Although the bootstrap test provides conservative p values at the 0.05 alpha level with small sample size, the power is lower than the other tests.

As anticipated, power differed numerically for different outcomes, tests, and percent increase in outcomes, indicating the need to optimize the sample size for each measured readout (Figure 2D). Our results demonstrate that the distribution of outcomes, whether physiological or metabolic, may not necessarily be normal, and thus the plasmode approach can be useful for sample size calculation. Although not directly tested in this study, one should not expect equal variance in animals of varying ages given the likelihood that distributions will differ among young, middle-aged, and old populations for any particular outcome (Petr et al., 2021).

Limitations of the study

This study was carried out solely in males of an inbred strain of mice, genetically homogeneous and subject to strain-specific pathologies. To gain a better understanding of the stochastic nature of the aging phenotype, further studies will require examination of heterogeneous strains of mice of both sexes at various ages. Although this work highlights the importance of proper sample size estimation in order to ensure adequate power to detect significant group differences or treatment effects in physiological and metabolic outcomes while controlling the type I error rate, we can use these estimates to inform adequately about the complexity of the aging phenotype. Because the results presented in this study are specific to a mean shift, without change in distribution shape, we are planning future investigations aimed at examining how anti-aging interventions affect distributions in a large cohort of animals and how such changes affect inference when inappropriate tests are used.

STAR★METHODS

RESOURCE AVAILABILITY

Lead contact—Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Rafael de Cabo (decabora@mail.nih.gov).

Materials availability—This study did not generate new unique reagents.

Data and code availability—This study has generated datasets or code. All simulations were performed using the statistical computing software R 4.0.1 (R Development Core Team). The data and codes used in this study will be available online (https://doi.org/10.5281/zenodo.4574094).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Twenty-four-month-old male C57BL/6J mice were purchased from the Jackson Laboratories (Bar Harbor, ME) and 24-mo-old male C57BL/6JN mice were obtained from the NIA Aged Rodent Colony located at Charles River Laboratories (Frederick, MD). Mice (n = 366) were maintained on standard diet (Teklad Global 18% Protein Extruded Rodent Diet, #2018SX, Envigo, Frederick, MD). Animal rooms were maintained at 22.2 ± 1°C and 30%–70% humidity. The lights were turned off at 6:00 PM and back on at 6:00 AM each day. Mice were group-housed up to four per cage with *ad libitum* access to food and water. All animals were provided shepherd shacks and nestlets for enrichment. Animal procedures, housing and diets were in accordance with the guidelines issued by the Intramural Research Program of the National Institutes of Health protocol numbers 444TGB2019 and 458TGB2018. All tests in animals were performed by non-blinded investigators.

After completion of the baseline measurements, a subgroup of mice (n = 112) was switched to AIN-93G diet (Dyets, Inc., Bethlehem, PA) for survival studies. Survival curves were plotted using the Kaplan-Meier method, which includes all available animals at each time point. The criteria for euthanasia were based on an independent assessment by a veterinarian, according to AAALAC guidelines and only cases, where the condition of the animal was considered incompatible with continued survival, are represented in the curves.

METHOD DETAILS

Core temperature—Body temperature was measured using implantable temperature transponder system (BMDS IPTT-300; Seaford, DE, USA).

Body composition—Lean, fat, and fluid mass measurements were obtained from unanesthetized mice by nuclear magnetic resonance (NMR) using the Minispec LF90 (Bruker Optics, Billerica, MA, USA).

Physical performance tests—Grip strength measurements and latency to fall off a metal wire, cage top, and accelerating rotarod were determined. A detailed explanation of all physical performance tests performed is described in Alfaras et al. (2017) and Bellantuono et al. (2020).

***In vivo* metabolism—**Mouse metabolic rate was assessed with an open circuit indirect calorimeter (Oxymax) with Columbus Instruments Comprehensive Lab Animal Monitoring System (CLAMS; Columbus Instruments International, Columbus, OH) as previously described (Martin-Montalvo et al., 2013). Various features of mouse locomotor and behavioral exploratory activity were also measured by dual axis detection using infrared photocell technology.

Blood and serum markers—Glucose concentrations in blood were measured from the submandibular vein in 6-h fasted mice with the Blood Glucose Monitoring System Breeze 2 (Bayer, Mishawaka, IN). Lactate concentrations in blood were measured with Lactate plus Meter (Nova Biomedical Corporation, Waltham, MA). Coagulated blood was centrifuged at 12,000 x *g*, 4°C for 10 min. Serum was aliquoted and kept frozen at −80°C. Insulin levels

were determined according to the manufacturer's protocol (Crystal Chem, Inc., Downers Grove, IL). Homeostasis model assessment-insulin resistance (HOMA-IR) was calculated to assess changes in insulin resistance using the HOMA2 Calculator software available from the Oxford Centre for Diabetes, Endocrinology and Metabolism Diabetes Trials Unit website (http://www.dtu.ox.ac.uk/homacalculator/index.php).

Plasmode simulation—We assume in the plasmode simulation that the empirical data from each outcome represent a whole population. A plasmode was composed of $2n$ (n = {3,4,5,8,12,16,20}) samples (a half is for the control group and the other half is for the treatment group) resampled from the outcome data of the cohort of 24-mo-old male C57BL/6J mice with allowing replacement, and 1,000 plasmodes were created, as described (Ejima et al., 2020). The five different statistical tests (Student's t test, Welch's t test, Wilcoxon rank sum test [aka, Mann-Whitney U test], permutation test, and bootstrap test), all of which are commonly used to test mean difference between two groups, were implemented for each of the plasmodes, and the p values obtained were summarized to compute type I error rates or power: type I error and power were defined as the proportion of the plasmodes with p value below the significance level (0.05): $\hat{\alpha} = \sum_{i=1}^{1000} I\left(p_i < 0.05\right) / 1000$, where I is a indicator function and p_i is the p value of the ith plasmode. The 95% CI is computed using the normal approximation: $\hat{\alpha} \pm 1.96\sqrt{\hat{\alpha}(1-\hat{\alpha}) / 1000}$. Note that the Wilcoxon rank sum test is a statistical test for distributional difference per se, whereas the other four tests are for mean difference. However, the Wilcoxon rank sum test is frequently and mistakenly used to test mean difference, and thus was included in the analyses. Note that the power of the Wilcoxon rank sum test is zero for sample size = 3, when alpha < 0.05, by theory (Janusonis, 2009).

QUANTIFICATION AND STATISTICAL ANALYSIS

Unless otherwise indicated, the data are depicted as five-number summary (minimum, 25% interquartile range (IQR), median, 75% IQR, and maximum) in addition to the statistical dispersion of the data known as H-spread (Tukey, 1977), which is a measure of the difference between the 75% and 25% IQRs (Figure 1—source data 1). Outliers were not omitted. The normality hypothesis was rejected for body composition data (NMR-generated morphometric analysis) and physical performance/motor coordination results (D'Agostino & Pearson normality test). In contrast, the normality hypothesis was not rejected for metabolic outputs and locomotor activity between light and dark cycles ($p > 0.05$). Student's two-tailed paired t test was applied for comparing metabolic output variables and locomotor activity data between light and dark phases. $P\ value \leq 0.05$ was considered statistically significant.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported, in part, by the Intramural Research Program of the National Institute on Aging, NIH, and by NIH grants 3P30DK056336, R25DK099080, and R25HL124208 (to D.B.A.) and Japan Society for Promotion

of Science (JSPS) KAKENHI grant 18K18146 (to K.E.). The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. The authors thank Nancy Zhang, Devin Wahl, and Vincent Gutierrez from the Translational Gerontology Branch for technical assistance, and the members of the Comparative Medicine Section of the National Institute on Aging.

REFERENCES

Alfaras I, Mitchell SJ, Mora H, Lugo DR, Warren A, Navas-Enamorado I, Hoffmann V, Hine C, Mitchell JR, Le Couteur DG, et al. (2017). Health benefits of late-onset metformin treatment every other week in mice. NPJ Aging Mech. Dis 3, 16. [PubMed: 29167747]

Allison DB, Paultre F, Goran MI, Poehlman ET, and Heymsfield SB (1995). Statistical considerations regarding the use of ratios to adjust data. Int. J. Obes. Relat. Metab. Disord 19, 644–652. [PubMed: 8574275]

Ballak SB, Degens H, de Haan A, and Jaspers RT (2014). Aging related changes in determinants of muscle force generating capacity: A comparison of muscle aging in men and male rodents. Ageing Res. Rev 14, 43–55. [PubMed: 24495393]

Bellantuono I, de Cabo R, Ehninger D, Di Germanio C, Lawrie A, Miller J, Mitchell SJ, Navas-Enamorado I, Potter PK, Tchkonia T, et al. (2020). A toolbox for the longitudinal assessment of healthspan in aging mice. Nat. Protoc 15, 540–574. [PubMed: 31915391]

Dwivedi AK, Mallawaarachchi I, and Alvarado LA(2017).Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. Stat. Med36, 2187–2205. [PubMed: 28276584]

Ejima K, Brown AW, Smith DL Jr., Beyaztas U, and Allison DB (2020). Murine genetic models of obesity: Type I error rates and the power of commonly used analyses as assessed by plasmode-based simulation. Int. J. Obes 44, 1440–1449.

Gadbury GL, Xiang Q, Yang L, Barnes S, Page GP, and Allison DB (2008). Evaluating statistical methods using plasmode data sets in the age of massive public databases: An illustration using false discovery rates. PLoS Genet. 4, e1000098. [PubMed: 18566659]

Hatchell PL, and MacInnes JW (1973). A quantitative analysis of the genetics of resting blood lactic acid levels in mice. Genetics 75, 191–198. [PubMed: 4762874]

Janusonis S (2009). Comparing two small samples with an unstable, treatment-independent baseline. J. Neurosci. Methods 179, 173–178. [PubMed: 19428524]

Jones SR, Carley S, and Harrison M (2003). An introduction to power and sample size estimation. Emerg. Med. J 20, 453–458. [PubMed: 12954688]

Kaiser KA, and Gadbury GL (2013). Estimating the range of obesity treatment response variability in humans: Methods and illustrations. Hum. Hered 75, 127–135. [PubMed: 24081228]

Martin-Montalvo A, Mercken EM, Mitchell SJ, Palacios HH, Mote PL, Scheibye-Knudsen M, Gomes AP, Ward TM, Minor RK, Blouin MJ, et al. (2013). Metformin improves healthspan and lifespan in mice. Nat. Commun 4, 2192. [PubMed: 23900241]

Martin-Montalvo A, Sun Y, Diaz-Ruiz A, Ali A, Gutierrez V, Palacios HH, Curtis J, Siendones E, Ariza J, Abulwerdi GA, et al. (2016). Cytochrome b_5 reductase and the control of lipid metabolism and healthspan. NPJ Aging Mech. Dis 2, 16006. [PubMed: 28721264]

Mercken EM, Capri M, Carboneau BA, Conte M, Heidler J, Santoro A, Martin-Montalvo A, Gonzalez-Freire M, Khraiwesh H, González-Reyes JA, et al. (2017). Conserved and species-specific molecular denominators in mammalian skeletal muscle aging. NPJ Aging Mech. Dis 3, 8. [PubMed: 28649426]

Miller RA, Harper JM, Dysko RC, Durkee SJ, and Austad SN (2002). Longer life spans and delayed maturation in wild-derived mice. Exp. Biol. Med. (Maywood) 227, 500–508. [PubMed: 12094015]

Mitchell SJ, Madrigal-Matute J, Scheibye-Knudsen M, Fang E, Aon M, González-Reyes JA, Cortassa S, Kaushik S, Gonzalez-Freire M, Patel B, et al. (2016). Effects of Sex, Strain, and Energy Intake on Hallmarks of Aging in Mice. Cell Metab. 23, 1093–1112. [PubMed: 27304509]

Mitchell SJ, Bernier M, Mattison JA, Aon MA, Kaiser TA, Anson RM, Ikeno Y, Anderson RM, Ingram DK, and de Cabo R (2019). Daily Fasting Improves Health and Survival in Male Mice Independent of Diet Composition and Calories. Cell Metab. 29, 221–228.e3. [PubMed: 30197301]

Morgan JA, Singhal G, Corrigan F, Jaehne EJ, Jawahar MC, and Baune BT (2018). The effects of aerobic exercise on depression-like, anxiety-like, and cognition-like behaviours over the healthy adult lifespan of C57BL/6 mice. Behav. Brain Res 337, 193–203. [PubMed: 28912012]

National Institute on Aging; National Institutes of Health; U.S. Department of Health & Human Services (2008). Healthy Aging: Lessons from the Baltimore Longitudinal Study of Aging (National Institute on Aging, National Institutes of Health). https://www.giorgiannirehab.com/docs/healthy_aging_lessons_from_the_baltimore_longitudinal_study_of_aging.pdf.

Petr MA, Alfaras I, Krawcyzk M, Bair WN, Mitchell SJ, Morrell CH, Studenski SA, Price NL, Fishbein KW, Spencer RG, et al. (2021). A cross-sectional study of functional and metabolic changes during aging through the lifespan in male mice. eLife 10, e62952. [PubMed: 33876723]

Studenski S, Perera S, Patel K, Rosano C, Faulkner K, Inzitari M, Brach J, Chandler J, Cawthon P, Connor EB, et al. (2011). Gait speed and survival in older adults. JAMA 305, 50–58. [PubMed: 21205966]

Tukey JW (1977). Exploratory Data Analysis (Addison-Wesley).

Zimmerman DW (1987). Comparative power of Student t test and Mann-Whitney U test for unequal sample sizes and variances. J. Exp. Educ 55, 171–174.

Zimmerman DW (1988). Invalidation of parametric and nonparametric statistical tests by concurrent violation of two assumptions. J. Exp. Educ 67, 55–68.

Highlights

- Extensive characterization of aging phenotypes in inbred 24-month-old male mice
- Association of aging phenotypes and survival is assessed
- High heterogeneity on outcome distributions across variables
- Plasmode-based approach is used for assessing type I error rates and power

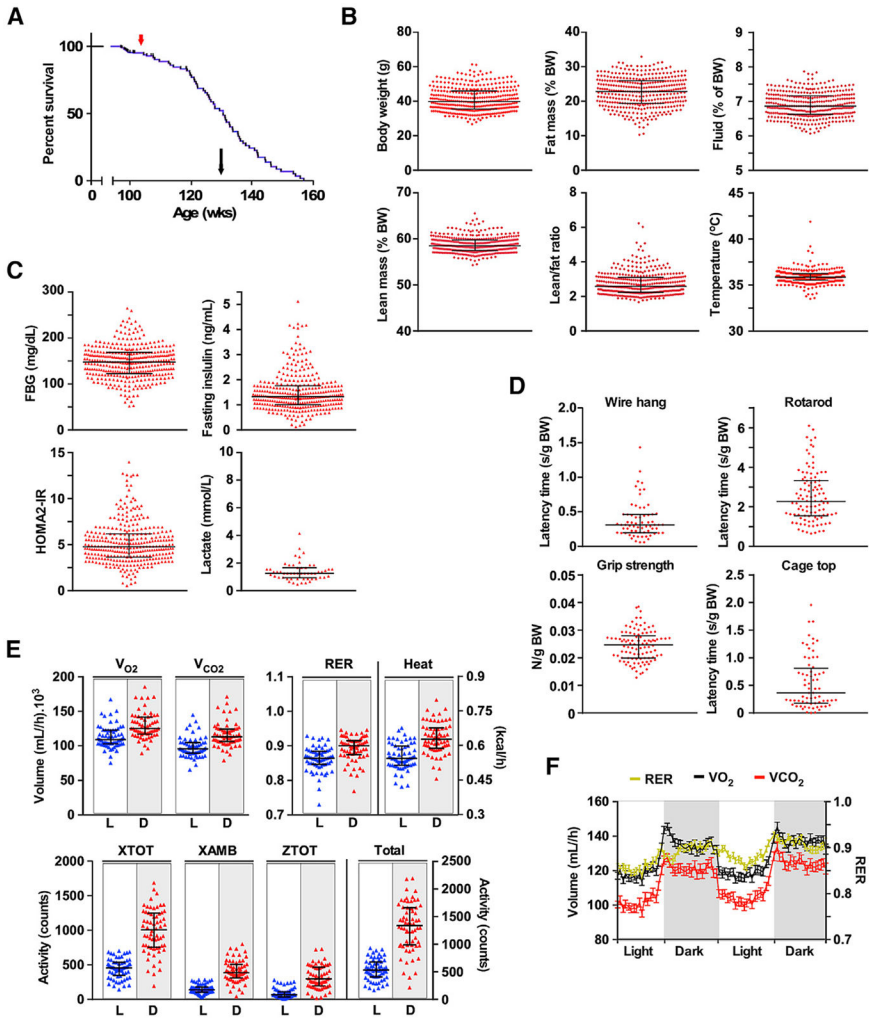


Figure 1. Contribution of phenotypic heterogeneity to longevity of 24-month-old male C57BL/6 mice

(A) Kaplan-Meier survival curve for male C57BL/6 mice fed a standard *ad libitum* diet. Red arrow depicts the age at which baseline measurements were collected (e.g., 104 weeks or 2 years of age), and black arrow shows the median survival. n = 366.

(B) Morphometric analysis and body temperature. Percentages of fat mass, fluid, and lean body mass were determined by nuclear magnetic resonance (NMR) and normalized to body weight (BW). n = 366 mice.

(C) Circulating levels of glucose, insulin, and lactate in animals fasted for 6 h, and calculation of the homeostatic model assessment of insulin resistance (HOMA2-IR). n = 355 mice (n = 57 for lactate).

(D) Physical performance as assessed by wire hang (n = 66), cage top (n = 68), grip strength (n = 101), and rotarod (n = 104) tests. The values were normalized to body weight.

(E) Mice were placed into metabolic cages for the measure of the rates of oxygen consumption (VO_2) and CO_2 production (VCO_2), respiratory exchange ratio (RER), energy expenditure as heat, and voluntary locomotor activity during two light (L) and dark (D) cycles. n = 64 mice.

The data in (B)–(E) represent median with interquartile range (IQR).

(F) Trajectories of VO_2, VCO_2, and RER during 48 h (two light/dark cycles). Each point represents mean ± SEM. n = 64.

See also Table S1.

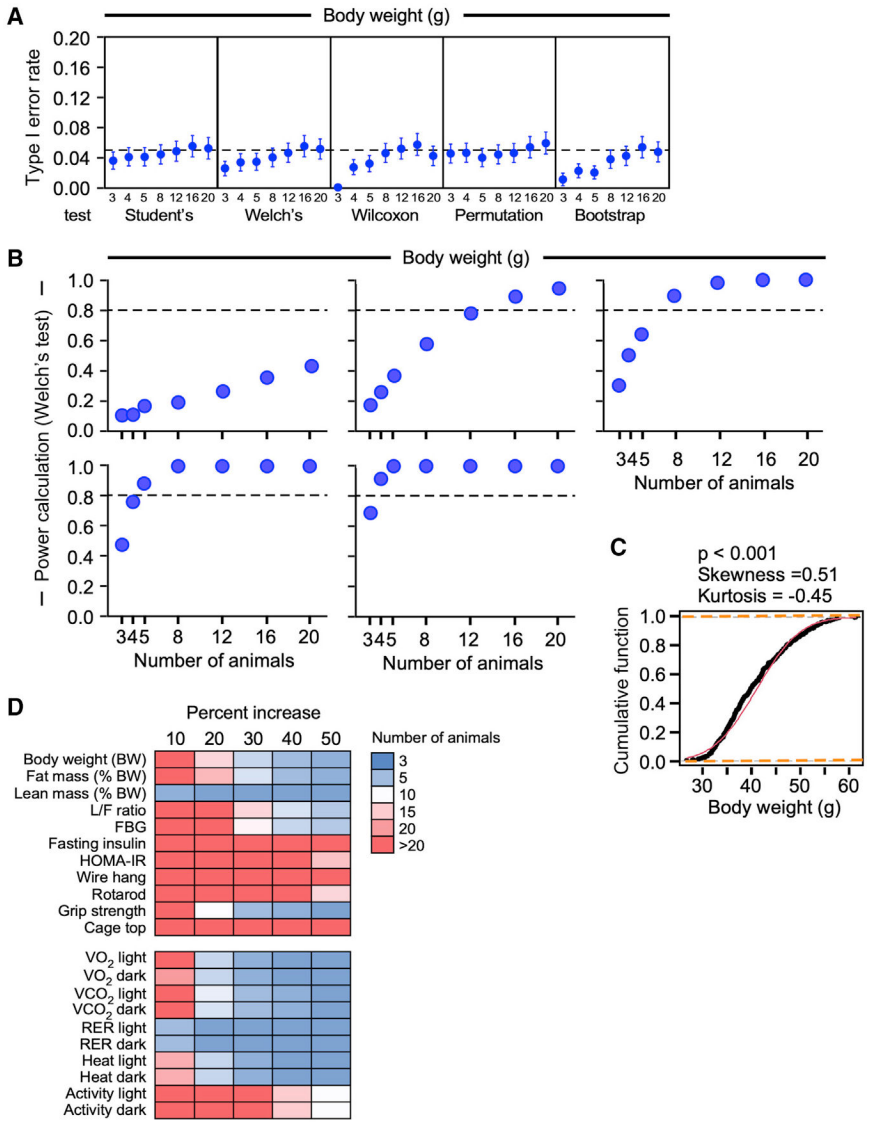


Figure 2. Computed type I error rates and power from plasmode-based simulation

(A) Computed type I error rates for body weight from the plasmode-based simulation and the five different statistical tests with nominal significance level set at 0.05. Filled blue circles are type I error rates and bars are the 95% confidence interval (CI), respectively. A horizontal dotted line corresponds to the significance level.

(B) Computed power for body weight using the Welch's test when outcome of the treatment group was increased by 10%–50%. A horizontal dotted line denotes an 80% cutoff and corresponds to a significance level of 0.05.

(C) Cumulative distributions of empirical data (black dots and lines) and the normal distributions (red lines) of body weight with the same means and variances. p values from Shapiro-Wilk test (test for normality), skewness, and excess kurtosis are listed.

(D) Heatmap depicting the sample size required to attain 80% power among the indicated outcome measures using the Welch's test. Similar analyses were carried out for all outcome measures and are illustrated in Figure 2 and Figures S1-S3.

See also Figures S1-S3 and Tables S2-S4.

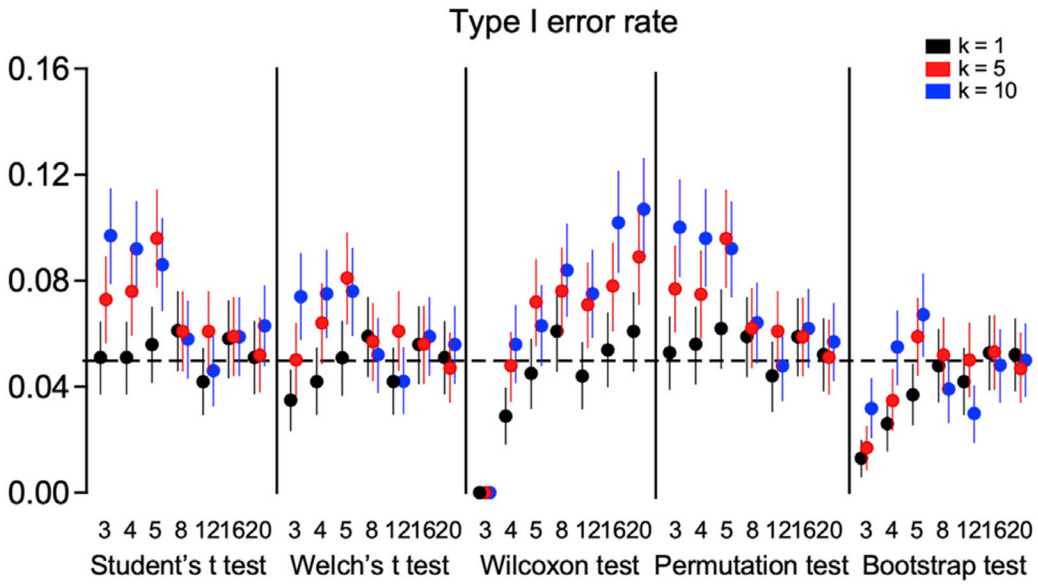


Figure 3. Illustration of a plasmode-based simulation aimed at computing type I error rate

In this simulation, our original data (body weight) was used to create a treatment group by maintaining the same mean of the population distribution but with the variance in the treatment group (*k*) set at 1-, 5-, or 10-fold larger than that in the control group.

Table 1.

Analysis of metabolic and behavioral readouts during continuous 48-h measurements (two light and two dark sessions) in the CLAMS monitoring system

	Light sessions (mean ± SD)	Dark sessions (mean ± SD)	t (df)	p value	n
VO_2 (mL/h)	116.4 ± 16.66	133.3 ± 19.27	7.21 (126)	<0.0001	64
VCO_2 (mL/h)	101.1 ± 14.65	120.0 ± 17.87	7.456 (119.3)	<0.0001	64
RER	0.8677 ± 0.0045	0.8998 ± 0.0046	5.014 (126)	<0.0001	64
Heat (kcal/h)	0.5611 ± 0.0072	0.6454 ± 0.0086	7.545 (126)	<0.0001	64
Total activity (counts)	684.2 ± 31.17	1715 ± 73.2	12.96 (85.12)	<0.0001	64

Paired t test with two-tailed p value. The normality hypothesis was not rejected for metabolic outputs and locomotor activity between light and dark cycles.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
AIN-93G diet	Dyets, Inc.	Cat #110700
Teklad Global 18% Protein Extruded Rodent Diet	Envigo	Cat #2018SX
Critical commercial assays		
Mouse insulin ELISA kit	Crystal Chem, Inc.	Cat# 90080; RRID:AB_2783626
Deposited Data		
Data sets and original codes		https://doi.org/10.5281/zenodo.4574094
Experimental models: Organisms/strains		
Male C57BL/6J mice	The Jackson Laboratory	JAX 000664
Male C57BL/6JN mice	NIA Aged Rodent Colony	N/A
Software and algorithms		
Prism 6.0	GraphPad	https://www.graphpad.com:443/scientific-software/prism; RRID:SCR_015807
Microsoft Excel 2019	Microsoft Corp.	https://www.microsoft.com/en-gb/; RRID:SCR_016137
Canvas Draw 6 for macOS	Canvas GFX	RRID:SCR_014288
R programming language v.4.0.1	R Development Core Team	RRID:SCR_001905
Other		
Rotarod	Med Associates, Inc.	Cat#ENV-574M
Minispec LF90	Bruker Optics	https://www.bruker.com/en/products-and-solutions/magnetic-resonance.html
Oxymax Open Circuit Indirect Calorimeters	Columbus Instruments	https://www.colinst.com/docs/OxymaxBrochure.pdf
Breeze2 Glucometer	Bayer	http://personalcare.manualsonline.com/; Bayer HealthCare Blood Glucose Meter