THE GERONTOLOGICAL SOCIETY OF AMERICA®

Journals of Gerontology: Biological Sciences
cite as: *J Gerontol A Biol Sci Med Sci*, 2022, Vol. 77, No. 11, 2155–2164
https://doi.org/10.1093/gerona/glab382
Advance AccesDecemberon December 24, 2021

OXFORD

Editor's choice

Invited Contribution

From Model Organisms to Humans, the Opportunity for More Rigor in Methodologic and Statistical Analysis, Design, and Interpretation of Aging and Senescence Research

Daniella E. Chusyd, PhD,[1] Steven N. Austad, PhD,[2,3] Andrew W. Brown, PhD,[4] Xiwei Chen, MS,[1] Stephanie L. Dickinson, MS,[1] Keisuke Ejima, PhD,[1] David Fluharty, PhD,[1,5] Lilian Golzarri-Arroyo, MS,[1] Richard Holden, PhD,[6] Yasaman Jamshidi-Naeini, PhD,[1] Doug Landsittel, PhD,[1] Stella Lartey, PhD,[1] Edward Mannix, PhD,[7] Colby J. Vorland, PhD,[4] and David B. Allison, PhD[1,*]

[1]Department of Epidemiology and Biostatistics, Indiana University Bloomington, Bloomington, Indiana, USA. [2]Department of Biology, University of Alabama at Birmingham, Birmingham, Alabama, USA. [3]Nathan Shock Center, University of Alabama at Birmingham, Birmingham, Alabama, USA. [4]Department of Applied Health Science, Indiana University Bloomington, Bloomington, Indiana, USA. [5]Departments of Mathematics and Economics, Ivy Tech Community College, Columbus, Indiana, USA. [6]Department of Health and Wellness Design, Indiana University Bloomington, Bloomington, Indiana, USA. [7]Department of Anatomy, Cell Biology, and Physiology, Indiana University School of Medicine, Indianapolis, Indiana, USA.

*Address correspondence to: David B. Allison, PhD, Department of Epidemiology and Biostatistics, Indiana University Bloomington, 1025 E. 7th St., PH 111, Bloomington, IN 47405, USA. E-mail: allison@iu.edu

Received: August 8, 2021; Editorial Decision Date: December 8, 2021

Decision Editor: Rozalyn M. Anderson, PhD, FGSA

Abstract

This review identifies frequent design and analysis errors in aging and senescence research and discusses best practices in study design, statistical methods, analyses, and interpretation. Recommendations are offered for how to avoid these problems. The following issues are addressed: (a) errors in randomization, (b) errors related to testing within-group instead of between-group differences, (c) failing to account for clustering, (d) failing to consider interference effects, (e) standardizing metrics of effect size, (f) maximum life-span testing, (g) testing for effects beyond the mean, (h) tests for power and sample size, (i) compression of morbidity versus survival curve squaring, and (j) other hot topics, including modeling high-dimensional data and complex relationships and assessing model assumptions and biases. We hope that bringing increased awareness of these topics to the scientific community will emphasize the importance of employing sound statistical practices in all aspects of aging and senescence research.

Keywords: Geroscience, Methodologies, Reproducibility

Profound concerns about the proper use of statistical methods have been the subject of a large scientific literature and many professional discussions. Most notably, in 2016, the American Statistical Association, for the first time in its history, made a formal statement on statistical practice: "The validity of scientific conclusions, including their reproducibility, depends on more than the statistical methods themselves. Appropriately chosen techniques, properly conducted analyses and correct interpretation of statistical results also play a key role in ensuring that conclusions are sound and that uncertainty surrounding them is represented properly." It is noteworthy that the intended audience of this Statement on Statistical Significance and *p* Values "would be researchers, practitioners, and

science writers who are not primarily statisticians" (1). As for all areas of research, the appropriate use and interpretation of statistical methods pertain to the study of aging, senescence, senescent cells, and senolytic agents, and to aging and longevity research in general.

Aging and senescence research requires the utmost efforts to ensure rigor, reproducibility, transparency, and sound inference (2). Yet, the value of published reports is frequently compromised because seemingly adequate methods are underdeveloped (3), available adequate methods are not used (4), or the methods chosen are used improperly (5,6). We highlight these errors here in the hope that they will be avoided in future studies, and we share best practices to assist investigators in choosing the most appropriate methods for their research (Table 1).

Errors

Topic 1: Errors in Randomization

The random assignment of subjects (eg, patients, mice, flies) in aging research bolsters causal inference (7). Randomization, if implemented correctly, is the only known method that allows for the assignment of units of observation (ie, subjects) to treatments to be completely independent of the prerandomization characteristics of those units, both observed and unobserved, that could confound the outcome, providing unbiased estimation of treatment effects (8). More specifically, the difference (or other contrast) between randomization groups is asymptotically (ie, for large sample sizes) equivalent to the expected paired differences between potential outcomes (9), that is, the mean of paired individual differences in outcomes that would have occurred (hypothetically if the outcome of both treatment conditions could have been observed) between the observed and unobserved treatment conditions. Analyzing and reporting randomized experiments according to the intent-to-treat principle, that is, where the treatment status of every subject is based entirely on their randomized assignment (10), therefore, yield unbiased estimates for the causal effect of being randomized to the given treatment.

By extension, both the use of nonrandom methods where it is possible to randomize and the incorrect implementation, analysis, and reporting of randomized designs increase our uncertainty in the knowledge of aging-related research questions. We have identified numerous instances of authors representing nonrandom allocation as random, which has spurred retractions (11). Some of these ways include using control groups that were not randomly allocated, allocating in nonrandom ways to reach a certain sample size, allocating participants in a household or other group setting together and analyzing their data as if they were individually randomized, replacing subjects in ways that are not random, and allocating animals by body weight instead of using an appropriate method to generate random assignments (for specific examples, see (11,12)). In cases where participants drop out of a study, or outcome data are missing, failing to appropriately handle these missing data breaks the random assignment and can introduce bias (13,14). In other instances, authors fail to blind the random allocation of participants, which is key to preventing selection bias and confounding (15–17), and is associated with larger effect estimates than those with adequate concealment (ie, suggestive of bias) (18). Finally, we and others observe that published reports frequently contain insufficient information for readers to understand exactly how subjects were randomized (19–23). Consolidated Standards of Reporting Trials reporting guidelines for parallel-group randomized trials, Animal Research: Reporting of In Vivo Experiments guidelines for animal studies, or related guidelines and extensions assist authors in how to report adequately, which provide readers confidence that experiments were executed rigorously (24). To facilitate their use, journals can require

Table 1. Summary of Common Errors or Challenges and Their Associated Best Practices in Aging Research

	Common Error or Challenge	Best Practices
1	Participants, animals, or organisms are nonrandomly assigned to treatment groups.	Randomize using a random number generator or table with allocation concealment.
2	Conclusions in an RCT are based on within-group differences rather than between-group differences.	Test differences between groups rather than within groups.
3	Clustering in data, such as group-housed animals, is ignored.	Consider correlation among observations in the analysis, especially for cluster-randomized trials.
4	Interference effects, where the treatment of one individual affects another individual, are not considered.	Consider study design should be done carefully to prevent interdependency.
5	Individual studies may report different metrics of effect size.	Standardize effect size metrics in data shared publicly.
6	Comparison of longevity between groups is often limited to overall difference in means.	Consider maximum life-span tests to compare differences at older ages.
7	Standard *t*-tests comparing means may violate assumptions of normality and Type I error rate.	Consider quantile regression and generalized lambda distributions for comparisons beyond the mean, with FWER control.
8	Testing negligible senescence has challenges including limited power. Power calculations are complicated for nonnormally distributed data.	Consider maximum life span and other tests for small differences. Use plasmode and EEE approaches to facilitate power calculations.
9	Compression of morbidity is confused with survival curve squaring.	The 2 concepts should be clearly separated with discussion in the literature.
10	Complicated relationships for aging and senescence are overly simplified in standard comparisons or incorrectly analyzed in complex models. Missing data, outliers, and skewed data are often handled inappropriately leading to biased results.	Analysis for high-dimensional data and machine learning are needed for complicated data. Handle missing data carefully with multiple imputation or linear mixed models. Perform sensitivity analyses without outliers. Transform data to satisfy normality.

Note: RCT = randomized controlled trial; FWER = family-wise error rate; EEE = Elston's excellent estimator.

and enforce adherence to reporting guidelines for publication. Each of these issues increases the risk of producing results that are not trustworthy and slows scientific discovery in aging research. In geroscience, when allocation of participants, animals, or well plates to conditions is needed for their evaluation, randomization using a random number generator or table is recommended unless it is determined to be not feasible or ethical. In those cases, it may still be possible to randomize outcome or sample measurements. In all cases, the method of allocation should be clearly communicated so readers can appropriately evaluate the potential for bias.

Topic 2: Making Conclusions Based on Differences in Nominal Significance in Effects (or Associations), Effect Modification, and Regressions

A common statistical procedure is to declare a comparison nominally significant if it has a p value below a threshold like $p < .05$. Unfortunately, numerous studies make conclusions based on tests of differences *within* groups instead of *between* groups. Consider a parallel-arm study in which units (eg, mice) are measured at Time 1 and again at Time 2. In treatment group A, there is a statistically significant change from baseline, while in treatment group B there is not. Some authors use this to conclude that the treatment worked compared to the control. However, this represents 2 *within*-group comparisons, but the purpose of having a control group is to conduct a *between*-group comparison. In the extreme, further imagine that those in treatment group A grew by 5 g with a p value of .049, while those in treatment group B grew by 5 g but with a p value of .051 (Figure 1A). We can conclude that the between-group comparison is not statistically significant (1 − 1 = 0), even without calculating the p value. Several of the present authors call this the "differences in nominal significances" (DINS) error because authors are comparing nominal significance within groups instead of between groups (25). Another perspective on this error is identifying subgroup effects without first testing the interaction. To avoid this error, a significant result should be observed for the group-by-treatment interaction before making any statements about individual subgroups. Although this error has been addressed multiple times (26–29), including a related warning in the American Statistical Association's statement on p values (1), and is now part of formal methodology standards (standard HT-3) produced by the Patient-Centered Outcomes Research Institute (30), it still occurs. We have identified it in multiple papers (summarized here (25,31,32)) and in at least one aging-related example (33).

The same inferential error can occur for various comparisons. Imagine a study stratified by male and female with a treatment group and a control group. If there was a statistically significant difference in treatment versus control in males but not in females, a DINS error may conclude that the treatment "worked" better in males than in females (34). However, no formal between-group test was done (eg, through an interaction effect in a 2-way analysis of variance [ANOVA]). The error can occur in mediation models, in which models with and without the mediator are not formally tested. They can also arise in regression models in which one group's slope is not statistically significantly different from zero while another is, but with no between-group test (eg, Figure 1B). It can occur in randomized, nonrandomized, or observational designs. The use of appropriate tests for making comparisons between groups of interest is essential for appropriate inference.

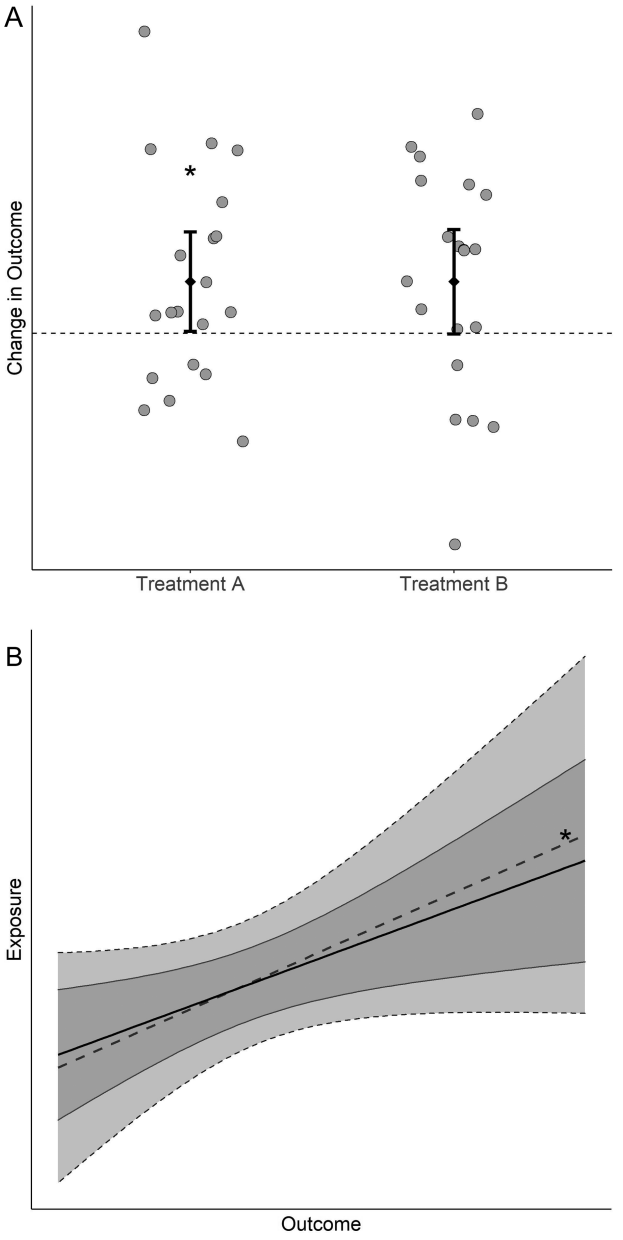

Figure 1. Two examples of the differences in nominal significance (DINS) error. The DINS error occurs when the nominal significance (eg, if a result is called "statistically significant" by meeting a threshold like $p < .05$) of 2 different results is compared, rather than a direct between-group comparison. (**A**) The change in outcome during Treatment A is marked as statistically significantly greater than 0 (demarked with an asterisk), while the change during Treatment B is not statistically significantly different from 0. A DINS error would occur if authors concluded Treatment A worked better than Treatment B, when there is no compelling difference between the 2 groups, if they were compared directly. (**B**) The regression lines for 2 different groups (eg, different sexes) as a function of an exposure are nearly identical, with the dashed line having a slope significantly different from 0 and the solid line having a slope not significantly different from 0. If authors concluded these 2 regression lines were different based on statistical significance (eg, that there was a sex difference in the exposure–outcome relationship), they would be committing a DINS error, as the lines are not statistically different from each other if compared directly.

Topic 3: Errors Involving Clustering

The incorrect identification of the experimental unit in the design and analysis of aging studies is common (35,36). The experimental unit is the smallest unit independently allocated to a particular treatment (37). Experimental design of aging research may involve hierarchical data structures and, thus, underlying correlations among

observations (eg, participants in hospitals or care facilities, cells/tissues/organs from the same animal, animals housed within the same cage, or a single cell line being independently sampled at different time points) (36,38). When groups of subjects are assigned as units to treatment conditions (eg, by a physician clinic), or when individuals are assigned to treatment conditions, but the treatment is applied at the aggregate level (eg, a nursing home social program), subjects in the same group are not independent, have common influences, and influence each other on many variables (36,38,39). In this case, treating the individual subject as the experimental unit for analysis ignores variability due to clustering and nesting. This leads to an incorrect estimation of the standard error of treatment differences, which is a type I error that is often inflated, and p values that are too small (ie, concluding a treatment effect when there is no evidence for one) (37,38). The degree of independent information depends on the number and size of clusters (as captured by their coefficient of variation) and the intracluster correlation (ICC), where fewer clusters, with smaller cluster sizes and higher ICCs, lead to less independent information. A number of formulas have been derived to quantify the inflation factor in the context of cluster-randomized designs (40). The most severe type of error results in designs that cannot be rescued by reanalysis. To effectively determine treatment effects in scenarios with group housing, experimental units (eg, cages, clinics) are needed within each treatment group, taking the correlated observations into consideration when performing power calculations and statistical analyses (36,37). More efforts are needed to prevent and correct errors related to clustering and nesting in aging research to maximize the return on investment.

Best Practices

Topic 4: Accounting for Interference Effects

To make causal inferences about the effects of interventions, investigators typically use design and analysis procedures predicated on the concept that an individual's outcome is independent of treatment assignment and the outcomes of other individuals in the study (41). However, this independence may not always hold owing to interference effects. Interference occurs when the treatment of one individual affects the outcome of another (42). This is often referred to as "contamination" in human aging and cluster-randomized studies (43). When interference is present, interpreting causal inference is more complex, particularly when interference is in the same direction as the direct effects. In this case, direct effects alone do not capture the full impact of the intervention (43). The role of interference on causal inference has begun to be acknowledged (43–46). However, the discussion of methods to address and estimate interference effects has primarily occurred in the vaccine literature (47–50), with few articles in other areas (51,52). The field of senescence and aging research has largely overlooked interference, even though there is ample evidence of it occurring. For example, housing conditions can affect aging-related outcomes, as shown in animal models (53,54), as can the interplay between an individual's position in a social hierarchy and energy availability when food is supplied to the group (55). Interference can also occur when individuals in an education intervention share their materials with control group participants (56). Interestingly, the ability of senescent cells to influence neighboring cells through secreted soluble factors could also contribute to interference effects (57). In some instances, methods for determining mechanistic interaction, direct and indirect effects (eg, cluster-randomized trials, sensitivity analysis), and principal stratification can be applied to address interference (42). However, interference can be quite complicated, so there is no one strategy to address it. Rather, different strategies to prevent or avoid such interdependency will depend on the setting. This further highlights the need to develop additional methodologies. Thus, it is important investigators are aware of the inherent assumptions in their study design and analyses, and the implications of these, for the interpretations of their results.

Topic 5: Reporting Comprehensible, Sensible, and Consistent Effect Size Indicators

One challenge in interpreting individual studies is that different investigators use different metrics of effect size, and not all metrics used are easily interpreted. For instance, in one study of the impact of rapamycin (initiated at 20 months of age) on male mouse longevity, effect size was reported as increased mean longevity of both 9% and 28%, depending on whether the effect was calculated on total longevity or longevity from the initiation of treatment, respectively (58).

To facilitate clearer interpretation and comparison across studies, it would be useful if all investigators reported the same, and sound, metrics of effect size. The adoption of a common set of metrics would improve clarity and provide common grounds for correct comprehension, interpretation, and further research. In addition, it is advisable that results be reported in both relative and absolute terms because the implications of these differ. For example, in survival analysis, hazard ratios (a relative term), which inform on how large (or small) the hazard is in one group using another group as a reference, are commonly reported. However, combining the hazard ratio with the median survival time (an absolute term) of each group would allow readers to roughly assess the magnitude of the ratio on the absolute scale. The ideal situation, which allows the best of both freedom of reporting for the investigator and comparison across studies, is that the raw data along with the analytical codes be publicly deposited or published with corresponding articles whenever possible. This would allow everyone to have access to the raw data.

Because not all investigators are likely to make all data and codes publicly available and not all people who wish to make comparisons across studies will necessarily have equal ability to download and analyze the raw data, it will be valuable to devise a list of standardized metrics. Ideal effect size metrics should inform both the magnitude of the effect of interest and the precision of the estimated effect (59). Furthermore, the pros and cons of each metrics need to be clearly understood. To promote more consistent reporting and interpretation, we recommend the adoption of a standard whereby a common suite of effect size indicators is always reported at a minimum, with the option to supplement that reporting with other effect size indicators of interest. In Table 2, we list a variety of potential effect sizes, a reference, and a mention of advantages and disadvantages.

Topic 6: Maximum Life-Span Testing

While the vast majority of survival analyses will compare the "average" (mean or median) longevity between treatments (eg, Cox models or log-rank tests), it is often overlooked and valuable to also consider differences in the so-called maximum life-span test (74,75) among the animals that lived the longest. Here, the right tail of the survival distribution determines how a treatment affects animals later in life, even when the median life spans were not different. In these methods, a threshold (τ) is set for "old" age, such as the

Table 2. Advantages and Disadvantages of Commonly Used Effect Sizes

Effect size indicator	Advantages	Disadvantages
Cohen's *d*, Hedge's *g*, and Glass's delta (60–62)	• Allows the comparisons across different continuous measures. • Allows the quantification of effect size using a unit-free measure.	• Standardizing effect size measure leads to a loss of units of the measure. • Standardized effect size measures are determined not only by the size of the effect, but also by the amount of variance in the data. • With this and other metrics involving variance in the denominator, miscalculations are common (63).
Cohen's *f* (64–66)	• Enables the evaluation of effect size. • Makes it easy to compare results across studies and forms the basis for meta-analysis. • Useful and appropriate for measuring local effect size in hierarchical and repeated-measures data. • Cohen's *f* used in repeated-measures studies often yields stronger effect sizes compared to an equivalent independent design.	• Estimated effect size can be reduced by variables with low reliability/high variance. • It could be unstable across studies with different designs. • Effect size can be distorted by the sampling procedure, and this can consequentially affect the generalizability of the effect.
Eta squared (62,67)	• Easy to interpret. • Easy to evaluate.	• The addition of more variables to the model leads to a decrease in the proportion explained by any one of the variables.
Common language effect size indicator (68,69)	• Effect can be generalized to a variety of research designs. • Although it assumes normal distributions, it is robust to the violation of its assumptions. • It can easily be translated into Cohen's *d*. • Quickly and easily calculated and associated with ease of interpretation. • Reporting using CL allows all including nonstatisticians to intuitively evaluate the effect.	• The disadvantages listed for Cohen's *d* apply here.
Hazard ratio reduction (70–72)	• The tests behind the hazard ratios can account for censoring problems that are common in longevity studies. • Takes the entire survival curve into account. • Almost always available for studies including meta-analysis of effects. • Practically reasonable, and interpretation is comprehensible.	• Interpretation of the metric is challenging. • Need to meet the proportional hazards assumption. • Only applicable to survival data.
Life expectancy difference and life expectancy ratio (73)	• The difference is presented in the same unit as the outcome (ie, week or year), thus easy to interpret. • Applicable even when the proportional hazard assumption is violated.	• Dependent on units. • Only applicable to survival data.

Note: CL = common language.

90th percentile of life span, across animals in all groups combined. Where the Wang-Allison test (75) compares the proportion of animals reaching the threshold between treatment groups, its successor, the Gao-Allison maximum life-span test (74) compares the groups on both the likelihood to live past the old age threshold and the magnitude of how long the individuals lived past the threshold. The Gao-Allison test is performed such that a new variable Z is calculated where Z_i = survival time (Y_i) for animal i if survival (Y_i) is greater than the threshold (τ), and $Z_i = 0$ otherwise, where τ is commonly the 90th percentile of survival across the 2 groups. An exact Wilcoxon–Mann–Whitney test is then used to compare the distributions of Z between the 2 groups. The Kruskal–Wallis test could replace the exact Wilcoxon–Mann–Whitney test for multiple group comparisons. The maximum life-span tests in the studies of Wang et al. (75) and Gao et al. (74) provide tools for comparing treatment effects on how animals age later in life.

Topic 7: Testing for Effects Beyond Mean but Controlling the Family-Wise Error Rate

Conventional statistical methods often test for group differences in a single parameter of a distribution, usually the conditional mean under specific distributional assumptions (such as a normal distribution for the *t*-test). However, distributional assumptions of conventional statistical methods may be violated in some situations, especially with a small sample size that fails to satisfy the central limit theorem. Parameters other than the mean may be of interest in geroscience. For example, the distribution of longevity is skewed to the right (ie, early death events are often not observed). Therefore, measures of central tendency such as the mean may change only a little when the distribution changes at the tails. Because of this, some studies report and test the difference in percentiles of life span using Fisher's exact test (76). If the independent variables are continuous variables, quantile regression is increasingly used (77,78). Several of the present authors also invented a statistical approach to test the difference in maximum life span (74). In such approaches, one must specify the percentile (such as the 95th percentile) to test. If multiple percentiles are tested, some would opine that there is a need to control the family-wise error rate (FWER). Such FWER control may result in reduced statistical power and a failure to detect effects when they exist. One potential approach to lessening this concern is the use of a flexible distribution, the generalized lambda distribution, to test the difference beyond central tendency (58). The generalized lambda distribution is characterized by 4 parameters, which relate to (a) median, (b) interquartile range, (c) asymmetry, and (d)

steepness. Thus, it has 4 basic shapes: (a) unimodal (with or without symmetry), (b) U-shape, (c) monotone (decline or increase), and (d) S-shape, including the normal distribution as a special case of a unimodal distribution (79). By fitting the model to the data and subsequently performing a likelihood ratio test, it is possible to test the difference among those 4 parameters. Therefore, with this test, one arguably does not have to specify percentiles to be tested or control the FWER (80).

Topic 8: Power and Sample Size

Tests and power for negligible senescence

Senescence is often measured by monitoring decline in global functional capacity (81,82). Previous studies have tested negligible senescence using the time required for mortality rate to double, initial mortality rate, and survival methods (82,83), including Bayesian survival trajectory analyses (84) and the Gompertz model of survival curves (85). However, limitations and challenges have been noted regarding the power of these tests to detect minimal effects, and alternatives for negligible senescence are lacking (86). Researchers have suggested statistical tests with power to detect minimal effects in other areas of gerontological studies, such as the Wang-Allison (75) and Gao-Allison (74) tests for maximum life span, and Hall et al. (87,88) to detect a small departure from a monotonic shape. The suggested methods can also be extended for use in negligible senescence testing. Finally, researchers have suggested that success could be achieved in testing for the presence of negligible senescence by following sequential steps along the pathway to discovery and by using specific and validated methods (82). These sequential steps are necessary and set the exact parameters and boundaries in the process of testing. The steps include clearly defining negligible senescence; identifying biomarkers that satisfy the requirements implied by the definition; developing appropriate techniques and tools to measure senescence; and understanding and accurately interpreting the results, manipulation of data, and development of therapeutics. Once a concrete definition of negligible senescence is determined, a primary emphasis should be placed on then identifying proper analyses and validation criteria (86). Thus, extending existing and developing new statistical models to test and improve the power of the effect of negligible senescence has become vital.

Power and sample size—plasmode approaches

Sample size and power calculation is now essential both when writing grant applications and when designing (and registering) experiments. The conventional approach used for sample size and power calculation makes several strong assumptions about the distributions of the outcome, such as equal variance and normal distribution, and depends on strictly unadjusted results. However, typical longevity studies using animal models have data that may violate assumptions of common statistical tests. We have proposed a data-driven "plasmode simulation approach" (89). In a plasmode simulation, multiple data sets are created by resampling from the original (empirical) data set allowing for replacement. Statistical tests are performed on the plasmodes, and the results (eg, p values) are summarized to compute power. The strength of a plasmode simulation approach is that it preserves the data structure of the original data set, which may not necessarily follow a normal distribution. The approach is flexible enough to be extended to survival analysis, quantile regression, and other types of tests used in geroscience. Familiarizing geroscience researchers with the plasmode approach and developing and providing a plasmode-based power (and sample size) calculator that is accessible to the scientific community are suggested.

Other issues and techniques

Power and sample size calculations are vital to ensure that an appropriate number of animals are included to provide adequate power for detecting statistically significant and biologically meaningful effects. While simple calculations can be done in freely available software, assuming normally distributed survival times with no censored data and no clustering of individuals, these conditions are not always met. More complex cases require careful thought with more sophisticated methods. Methods for Cox proportional hazards models and exponential models are available in commercial software such as PASS from NCSS (90). The more general Weibull distribution may also be considered, which accommodates increasing, decreasing, or flat hazards across the life span rather than restricting to the constant hazards required in Cox models. Heo et al. (91) discuss power and sample size calculations assuming the Weibull distribution. Where calculations are not available in closed-form formulas or software, Tiwari et al. (92) provide an option called the Elston's excellent estimator in which the power, alpha level, and sample size reported in previous literature can be used to calculate approximate power for a new study with different sample sizes (when other parameters are expected to remain the same). It is important to perform power calculations that most closely align with the ultimate statistical models and assumptions to be performed in data analysis and to give careful consideration to the consequences on estimates of power and sample size when simplified methods are utilized. When available techniques do not sufficiently match the ultimate statistical models and assumptions, simulation studies may be required to characterize the statistical properties of those models.

Topic 9: The Compression of Morbidity Versus Survival Curve Squaring

A common point of confusion involves discussion of the compression of morbidity and its inadvertent conflation with so-called survival curve squaring (93,94). These are 2 different concepts. Yet, because they both involve examination of curves in which the abscissa (X-axis) is age and the ordinate (Y-axis) is some function that slowly declines with age, they seem to be confused and conflated. In a survival curve, the ordinate is the proportion of a population that has or is predicted to survive up to that age (ie, point on the abscissa). Many survival curves are characterized by slow and steady declines after the population has reached maturity. In contrast, one could imagine that roughly all members of an animal population survive until approximately the same age and then all die shortly thereafter. This would lead to a curve that more closely approximates a step function and visually appears rectangular or "squared." Hence, the achievement of such a curve is sometimes referred to as "curve squaring." Such curve squaring might be taken as a sign of egalitarianism or a "disparity-free" environment in that all members of the population achieve roughly the same life span. However, the phrases *curve squaring* or *squaring the survival curve* also seem to be used to imply that one has achieved a situation in which members of a population do not decline in their functional ability, health, or freedom from morbidity and disability until just before death. This concept is sometimes referred to as *compression of morbidity* (95), and yet it is not implied by a squaring of the population survival curve. In contrast, one could imagine an individual curve in which the ordinate was health, functional ability, or freedom of disability,

and the abscissa was again age. Such a curve would represent *an individual's* degree of decline in health or functional ability with age. A compression of morbidity would imply a squaring of these curves in which the ultimate loss of health or functionality coincided with the point of death. After all, when it comes to people aging, compression of morbidity is a highly desired outcome, that is, most people want to live independently as long as possible before death. Although it is easy to see how these 2 ideas are conflated, the compression of morbidity in individuals and the squaring of survival curves in populations are indeed distinct. Therefore, we advocate both a clear discussion of this in the literature to separate the ideas (others have addressed confusion and misconception in this area (93,96)) and the development of methods in which the degree of compression of morbidity for individuals can be analyzed as a function of antisenescence interventions.

Topic 10: Other Topics

Modeling high-dimensional data and complex relationships

Methods for aging and senescence research will likely continue to depend on increasingly complex measurement approaches (eg, for dietary intake or physical activity), which then necessitate statistical models that effectively incorporate high-dimensional data and model the complex relationships between predictors and outcome. Specification of the associated statistical models represents a critical aspect of formulating clinical prediction models (97), for example, for identifying high-risk subjects or subjects most likely to respond to treatment. While the best general approach to model specification for a given problem depends largely on the underlying scientific construct and is difficult to specify in practice, relevant factors to consider include the number and degree of correlation among predictors, the sample size available to estimate model coefficients (including main effects, nonlinear relationships, and interactions), and the need to balance precision (ie, minimizing random variability of model predictions) with bias (ie, minimizing systematic differences between predictions and observed outcomes) and to balance model complexity with interpretability. If a large number of predictor variables (or risk factors) are highly correlated and tend to represent a smaller number of lower-dimensional characteristics, dimension reduction methods, such as principal component analysis (PCA) (98), may be preferable to repeated testing of individual variables or groups of variables. Methods such as PCA avoid errors due to multiple testing and may retain a high percentage of variability in the data through a much smaller number of variables in the model. PCA may be used as an initial step, possibly after multiple testing, to reduce the number of variables while still retaining most of the variability in predictor variables, especially when the number of variables is larger than the number of subjects.

Another challenge in the model specification is optimally accounting for the complexity of the relationship between predictors, for example, subject characteristics, and outcome. Standard regression methods assume linearity on some scale and require explicit specification of any interaction or nonlinear effects (99). Machine learning, or modern regression methods, provides an alternative set of methods. Machine learning tends to automate both the variable selection and the complexity of the predictor–outcome relationship by using either a deterministic algorithm (eg, tree models that recursively partition the data into increasingly homogenous subsets of data (100)) or variations of more standard statistical modeling approaches combined with data reduction techniques (eg, LASSO regression, which uses shrinkage methods and automated variable selection (101,102)). While the methods of machine learning approaches vary substantially, they essentially provide different ways to automate the process of specifying the variables in the model and the functional form of the relationship between predictors and outcome. Different machine learning methods yield more (or less) complex relationships, which then produces more (or less) precise (or alternatively less (or more) biased) predictions. While machine learning methods are often described relative to observational data, they are becoming increasingly useful for randomized trials to identify the high-risk group or those most likely to respond to treatment.

Missing data and outliers

Robust and rigorous data analysis requires careful assessment of model assumptions and diagnostics. Errors and incorrect conclusions will result from failure to assess and satisfy model assumptions such as normally distributed residuals with constant variance (89). While ANOVA is shown to be robust to modest violations of normality with large sample sizes, and consequences of unequal variance are mitigated with balanced sample sizes, the risks to the type I error rate must be evaluated carefully and reported transparently. Transformations such as logs are commonly helpful for skewed data. Outliers and missing data can also cause bias and compromise the reliability of results (103). While analysts are tempted to throw out values deemed to be outliers to ensure well-behaved data distributions, results could be drastically affected and potentially biased by removing valid data, particularly in experiments with small sample sizes. Sensitivity analyses may be used to perform analyses with and without outliers in full transparency. Concerns about missing data have been reported extensively and should be handled thoughtfully to ensure unbiased results (104). Two suggested options for handling missing data to provide unbiased results are multiple imputation and the use of mixed models for longitudinal data (105,106).

Multiple testing

Although not unique to aging, the use of the same data to calculate multiple tests increases the likelihood of a significant finding by chance when no true difference exists (ie, Type I errors). Specifically, as some of the present authors previously pointed out, testing 59 comparisons on any set of independent calculations results in a 95% chance of finding at least one significant difference at the $p < .05$ level when no differences exist (107). When discovery work, such as evaluating large data sets, testing each gut microbe independently, sifting through specific genes, or associating all possible foods or nutrients against aging outcomes, is being undertaken, it is almost certain that such false discoveries will be made. Controlling for multiple comparisons, such as through adjusting the FWER (eg, Bonferroni or Tukey) or controlling the false discovery rate (108), can help the aging community from being overly confident in apparent differences in the data that are, in fact, just there by chance.

Conclusion

The goal of this report is to provide investigators valuable information to aid in avoiding errors while implementing best practices during the design and execution of aging and senescence research. In many cases, immense value is gained by including a professional statistician early. The statistician should be involved in virtually all aspects of the project—from design to measurement to data analysis, interpretation, and presentation, and including selection

and evaluation for randomization, data recording practices, and data cleaning and checking. We realize this does not always occur. Even in our own research, we are often brought in long after many of these steps are done. Therefore, we further suggest continued awareness through the promotion and production of educational materials and programs. Some existing sources include the American Statistical Association, which curates an online list of biostatistics degree programs and offers several professional development events (109,110); certain universities, which offer online courses (eg, University of California San Diego, Drexel University, and Washington University); and the Society for Epidemiologic Research, which holds preconference workshops in connection with its annual meeting (111). We encourage readers to not only carefully consider the American Statistical Association statement (1) but also to review the articles in a special issue of *The American Statistician* devoted to exploring the statement's implications (112). Investing in methodologic training programs, and developing new methodologic research, all of which emphasize the implementation of rigor, reproducibility, and transparency, may also help move the discipline forward. Collectively, we hope this perspective brings awareness and value to investigators as it relates to the appropriate use of statistical and methodologic practices while conducting geroscience research.

Funding

This work was supported in part by the National Institute on Aging (grant number P30 AG050886 to S.N.A., D.E.C., and A.W.B.; U24 AG056053 to A.W.B. and D.E.C.); the National Institute of Diabetes and Digestive and Kidney Diseases (R25 DK099080 to A.W.B.); the National Heart, Lung, and Blood Institute (R25 HL124208 to A.W.B.); and Gordon and Betty Moore Foundation (to C.J.V.).

Conflict of Interest

None declared.

Acknowledgments

We graciously thank Jennifer Holmes from Medical Editing Services for providing language editing. The idea was conceived by D.B.A., and all coauthors contributed to the writing and editing of the manuscript.

References

1. Wasserstein RL, Lazar NA. The ASA statement on p-values: context, process, and purpose. *Am Stat*. 2016;70:129–133. doi:10.1080/00031305.2016.1154108
2. Wang C, Keith SW, Fontaine KR, Allison DB. Statistical issues for longevity studies in animal models. In: Conn MP, ed. *Handbook of Models for Human Aging*. Elsevier Academic Press; 2006:153–164. doi:10.1016/B978-0-12-369391-4.X5000-0
3. Belsky DW, Caspi A, Houts R, et al. Quantification of biological aging in young adults. *Proc Natl Acad Sci U S A*. 2015;112(30):E4104–E4110. doi:10.1073/pnas.1506264112
4. Ghisletta P, Aichele S. Quantitative methods in psychological aging research: a mini-review. *Gerontology*. 2017;63(6):529–537. doi:10.1159/000477582
5. Bland JM. Evidence for an 'anti-ageing' product may not be so clear as it appears. *Br J Dermatol*. 2009;161(5):1207–1208; author reply 1208. doi:10.1111/j.1365-2133.2009.09433.x
6. Santen Hv. Nature article is wrong about 115 year limit on human lifespan 2016. 2018. https://indico.cern.ch/event/742158/contributions/3469506/attachments/1870863/3078910/contra1.pdf
7. Imbens GW, Rubin DB. Rubin causal model. In: Durlauf SN, Blume LE, eds. *Microeconometrics. The New Palgrave Economics Collection*. Palgrave Macmillan and Springer; 2010:229–241. doi:10.1057/9780230280816_28
8. Allison DB, Owora AH, Dawson JA, et al. Randomisation can do many things—but it can't "fail". In press. 2021.
9. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. *J Am Stat Assoc*. 2005;100(469):322–331. doi:10.1198/016214504000001880
10. Gupta SK. Intention-to-treat concept: a review. *Perspect Clin Res*. 2011;2(3):109–112. doi:10.4103/2229-3485.83221
11. Vorland CJ, Brown AW, Dawson JA, et al. Errors in the implementation, analysis, and reporting of randomization within obesity and nutrition research: a guide to their avoidance. *Int J Obes (Lond)*. 2021;45(11):2335–2346. doi:10.1038/s41366-021-00909-z
12. Golzarri-Arroyo L, Dickinson SL, Allison DB. Replacement of dropouts may bias results: Comment on "The effect of green tea ointment on episiotomy pain and wound healing in primiparous women: a randomized, double-blind, placebo-controlled clinical trial". *Phytother Res*. 2019;33(8):1955–1956. doi:10.1002/ptr.6394
13. Peos J, Brown AW, Vorland CJ, Allison DB, Sainsbury A. Contrary to the conclusions stated in the paper, only dry fat-free mass was different between groups upon reanalysis. Comment on: "Intermittent energy restriction attenuates the loss of fat-free mass in resistance trained individuals. A randomized controlled trial". *J Funct Morphol Kinesiol*. 2020;5(4):85. doi:10.3390/jfmk5040085
14. Vorland CJ, Mestre LM, Mendis SS, Brown AW. Within-group comparisons led to unsubstantiated conclusions in "Low-phytate wholegrain bread instead of high-phytate wholegrain bread in a total diet context did not improve iron status of healthy Swedish females: a 12-week, randomized, parallel-design intervention study". *Eur J Nutr*. 2020;59(6):2813–2814. doi:10.1007/s00394-020-02287-0
15. Kahan BC, Rehal S, Cro S. Risk of selection bias in randomised trials. *Trials*. 2015;16:405. doi:10.1186/s13063-015-0920-x
16. McKenzie JE. Randomisation is more than a coin toss: the role of allocation concealment. *BJOG*. 2019;126(10):1288. doi:10.1111/1471-0528.15559
17. Chalmers I. Why transition from alternation to randomisation in clinical trials was made. *BMJ*. 1999;319(7221):1372. doi:10.1136/bmj.319.7221.1372
18. Savović J, Jones HE, Altman DG, et al. Influence of reported study design characteristics on intervention effect estimates from randomized, controlled trials. *Ann Intern Med*. 2012;157(6):429–438. doi:10.7326/0003-4819-157-6-201209180-00537
19. Dechartres A, Trinquart L, Atal I, et al. Evolution of poor reporting and inadequate methods over time in 20 920 randomised controlled trials included in Cochrane reviews: research on research study. *BMJ*. 2017;357:j2490. doi:10.1136/bmj.j2490
20. Kilkenny C, Parsons N, Kadyszewski E, et al. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One*. 2009;4(11):e7824. doi:10.1371/journal.pone.0007824
21. Kahathuduwa CN, Allison DB. Letter to the editor: Insufficient reporting of randomization procedures and unexplained unequal allocation: a commentary on "Dairy-based and energy-enriched berry-based snacks improve or maintain nutritional and functional status in older people in home care". *J Nutr Health Aging*. 2019;23(4):396. doi:10.1007/s12603-019-1183-0
22. Vorland CJ, Brown AW, Dickinson SL, Gelman A, Allison DB. The implementation of randomization requires corrected analyses. Comment on "Comprehensive nutritional and dietary intervention for autism spectrum disorder—a randomized, controlled 12-month trial, nutrients 2018, 10, 369". *Nutrients*. 2019;11(5):1126. doi:10.3390/nu11051126
23. Jayawardene WP, Brown AW, Dawson JA, Kahathuduwa CN, McComb B, Allison DB. Conditioning on "study" is essential for valid inference when combining individual data from multiple randomized controlled trials: a comment on Reesor et al's School-based weight management program curbs summer weight gain among low-income Hispanic middle school students. J Sch Health. 2019;89(1):59–67. *J Sch Health*. 2019;89(7):515–518. doi:10.1111/josh.12777

24. Enhancing the QUAlity and Transparency Of health Research. https://www.equator-network.org/
25. Allison DB, Brown AW, George BJ, Kaiser KA. Reproducibility: a tragedy of errors. *Nature*. 2016;530(7588):27–29. doi:10.1038/530027a
26. Bland JM, Altman DG. Best (but oft forgotten) practices: testing for treatment effects in randomized trials by separate analyses of changes from baseline in each group is a misleading approach. *Am J Clin Nutr*. 2015;102(5):991–994. doi:10.3945/ajcn.115.119768
27. Bland JM, Altman DG. Comparisons against baseline within randomised groups are often used and can be highly misleading. *Trials*. 2011;12:264. doi:10.1186/1745-6215-12-264
28. Bland JM, Altman DG. Comparisons within randomised groups can be very misleading. *BMJ*. 2011;342:d561. doi:10.1136/bmj.d561
29. Gelman A, Stern H. The difference between "significant" and "not significant" is not itself statistically significant. *Am Stat*. 2006;60(4):328–31. doi :10.1198/000313006X152649
30. Hickam D, Totten A, Berg A, Rader K, Goodman S, Newhouse R. *The PCORI Methodology Report*. Patient-Centered Outcomes Research Institute; 2013. https://www.pcori.org/assets/2013/11/PCORI-Board-Meeting-Methodology-Report-for-Acceptance-1118131.pdf
31. Brown AW, Kaiser KA, Allison DB. Issues with data and analyses: errors, underlying themes, and potential solutions. *Proc Natl Acad Sci U S A*. 2018;115(11):2563–2570. doi:10.1073/pnas.1708279115
32. Allison DB, Bassaganya-Riera J, Burlingame B, et al. Goals in nutrition science 2015–2020. *Front Nutr*. 2015;2:26. doi:10.3389/fnut.2015.00026
33. Allison DB, Williams MS, Hand GA, Jakicic JM, Fontaine KR. Conclusion of "Nordic walking for geriatric rehabilitation: a randomized pilot trial" is based on faulty statistical analysis and is inaccurate. *Disabil Rehabil*. 2015;37(18):1692–1693. doi:10.3109/09638288.2014.1002580
34. Sainani K. Misleading comparisons: the fallacy of comparing statistical significance. *PM R*. 2010;2(6):559–562. doi:10.1016/j.pmrj.2010.04.016
35. Huang W, Percie du Sert N, Vollert J, Rice ASC. General principles of preclinical study design. In: Bespalov A, Michel MC, Steckler T, eds. *Good Research Practice in Non-clinical Pharmacology and Biomedicine*. Springer International Publishing; 2020:55–69. doi:10.1007/978-3-030-33656-1
36. Lazic SE, Clarke-Williams CJ, Munafò MR. What exactly is 'N' in cell culture and animal experiments? *PLoS Biol*. 2018;16(4):e2005282. doi:10.1371/journal.pbio.2005282
37. Bello NM, Kramer M, Tempelman RJ, et al. Short communication: on recognizing the proper experimental unit in animal studies in the dairy sciences. *J Dairy Sci*. 2016;99(11):8871–8879. doi:10.3168/jds.2016-11516
38. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health*. 2004;94(3):423–432. doi:10.2105/ajph.94.3.423
39. Murray DM, Taljaard M, Turner EL, George SM. Essential ingredients and innovations in the design and analysis of group-randomized trials. *Annu Rev Public Health*. 2020;41:1–19. doi:10.1146/annurev-publhealth-040119-094027
40. Eldridge SM, Ashby D, Kerry S. Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *Int J Epidemiol*. 2006;35(5):1292–1300. doi:10.1093/ije/dyl129
41. Rubin DB. Randomization analysis of experimental data: the Fisher randomization test comment. *J Am Stat Assoc*. 1980;75(371):591–593. doi:10.2307/2287653
42. VanderWeele T. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University Press; 2015.
43. Benjamin-Chung J, Arnold BF, Berger D, et al. Spillover effects in epidemiology: parameters, study designs and methodological considerations. *Int J Epidemiol*. 2018;47(1):332–347. doi:10.1093/ije/dyx201
44. Sobel ME. What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *J Am Stat Assoc*. 2006;101(476):1398–1407. doi:10.1198/016214506000000636
45. Manski CF. Identification of treatment response with social interactions. *Econom J*. 2013;16(1):S1–S23. doi:10.1111/j.1368-423X.2012.00368.x
46. Rosenbaum PR. Interference between units in randomized experiments. *J Am Stat Assoc*. 2007;102(477):191–200. doi:10.1198/016214506000001112
47. Halloran ME, Struchiner CJ. Causal inference in infectious diseases. *Epidemiology*. 1995;6(2):142–151. doi:10.1097/00001648-199503000-00010
48. Halloran ME, Struchiner CJ. Study designs for dependent happenings. *Epidemiology*. 1991;2(5):331–338. doi:10.1097/00001648-199109000-00004
49. Vanderweele TJ, Tchetgen Tchetgen EJ. Effect partitioning under interference in two-stage randomized vaccine trials. *Stat Probab Lett*. 2011;81(7):861–869. doi:10.1016/j.spl.2011.02.019
50. Vanderweele TJ, Tchetgen Tchetgen EJ, Halloran ME. Components of the indirect effect in vaccine trials: identification of contagion and infectiousness effects. *Epidemiology*. 2012;23(5):751–761. doi:10.1097/EDE.0b013e31825fb7a0
51. Angelucci M, Di Maro V. Programme evaluation and spillover effects. *J Develop Effectiveness*. 2016;8(1):22–43. doi:10.1080/19439342.2015.1033441
52. Bowers J, Fredrickson MM, Panagopoulos C. Reasoning about interference between units: a general framework. *Polit Anal*. 2013;21(1):97–124. doi:10.1093/pan/mps038
53. Ban S, Tenma H, Mori T, Nishimura K. Effects of physical interference on life history shifts in *Daphnia pulex*. *J Exp Biol*. 2009;212(19):3174–3183. doi:10.1242/jeb.031518
54. Stefana MI, Driscoll PC, Obata F, et al. Developmental diet regulates *Drosophila* lifespan via lipid autotoxins. *Nat Commun*. 2017;8(1):1384. doi:10.1038/s41467-017-01740-9
55. Arslan-Ergul A, Erbaba B, Karoglu ET, Halim DO, Adams MM. Short-term dietary restriction in old zebrafish changes cell senescence mechanisms. *Neuroscience*. 2016;334:64–75. doi:10.1016/j.neuroscience.2016.07.033
56. Estruch R, Ros E, Salas-Salvadó J, et al. Retraction and republication: primary prevention of cardiovascular disease with a Mediterranean diet. N Engl J Med. 2013;368:1279–90. *N Engl J Med*. 2018;378(25):2441–2442. doi:10.1056/NEJMc1806491
57. Childs BG, Baker DJ, Kirkland JL, Campisi J, van Deursen JM. Senescence and apoptosis: dueling or complementary cell fates? *EMBO Rep*. 2014;15(11):1139–1153. doi:10.15252/embr.201439245
58. Harrison DE, Strong R, Sharp ZD, et al. Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature*. 2009;460(7253):392–395. doi:10.1038/nature08221
59. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev Camb Philos Soc*. 2007;82(4):591–605. doi:10.1111/j.1469-185X.2007.00027.x
60. Hunter SB, Miles JNV, Paddock SM, D'Amico EJ. Evaluating treatment efficacy. In: Miller PM, ed. *Interventions for Addiction*. Academic Press; 2013:589–597.
61. Hedges LV, Olkin I. *Statistical Methods for Meta-analysis*. Academic Press; 2014. doi:10.1016/C2009-0-03396-0
62. Cohen J. Things I have learned (so far). *Am Psychol Assoc*. 1990;8:3–18. doi:10.1037/0003-066X.45.12.1304
63. George BJ, Beasley TM, Brown AW, et al. Common scientific and statistical errors in obesity research. *Obesity (Silver Spring)*. 2016;24(4):781–790. doi:10.1002/oby.21449
64. Baguley T. Standardized or simple effect size: what should be reported? *Br J Psychol*. 2009;100(Pt 3):603–617. doi:10.1348/000712608X377117
65. Selya AS, Rose JS, Dierker LC, Hedeker D, Mermelstein RJ. A practical guide to calculating Cohen's f(2), a measure of local effect size, from PROC MIXED. *Front Psychol*. 2012;3:111. doi:10.3389/fpsyg.2012.00111
66. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. 2nd ed. L. Erlbaum Associates; 1988:567. doi:10.4324/9780203771587
67. Richardson JTE. Measures of effect size. *BRMIC*. 1996;28(1):12–22. doi:10.3758/BF03203631
68. McGraw KO, Wong SP. A common language effect size statistic. *Psychol Bull*. 1992;111(2):361–365. doi:10.1037/0033-2909.111.2.361
69. Björgvinsson T, Kerr P. Use of a common language effect size statistic. *Am J Psychiatry*. 1995;152(1):151. doi:10.1176/ajp.152.1.151a

70. Saad ED, Zalcberg JR, Péron J, Coart E, Burzykowski T, Buyse M. Understanding and communicating measures of treatment effect on survival: can we do better? *J Natl Cancer Inst*. 2018;110(3):232–240. doi:10.1093/jnci/djx179
71. Sashegyi A, Ferry D. On the interpretation of the hazard ratio and communication of survival benefit. *Oncologist*. 2017;22(4):484–486. doi:10.1634/theoncologist.2016-0198
72. Stensrud MJ, Hernán MA. Why test for proportional hazards? *JAMA*. 2020;323(14):1401–1402. doi:10.1001/jama.2020.1267
73. Dehbi HM, Royston P, Hackshaw A. Life expectancy difference and life expectancy ratio: two measures of treatment effects in randomised trials with non-proportional hazards. *BMJ*. 2017;357:j2250. doi:10.1136/bmj.j2250
74. Gao G, Wan W, Zhang S, Redden DT, Allison DB. Testing for differences in distribution tails to test for differences in 'maximum' lifespan. *BMC Med Res Methodol*. 2008;8:49. doi:10.1186/1471-2288-8-49
75. Wang C, Li Q, Redden DT, Weindruch R, Allison DB. Statistical methods for testing effects on "maximum lifespan". *Mech Ageing Dev*. 2004;125(9):629–632. doi:10.1016/j.mad.2004.07.003
76. Ramsey JJ, Tran D, Giorgio M, et al. The influence of Shc proteins on life span in mice. *J Gerontol A Biol Sci Med Sci*. 2014;69(10):1177–1185. doi:10.1093/gerona/glt198
77. Beyerlein A. Quantile regression—opportunities and challenges from a user's perspective. *Am J Epidemiol*. 2014;180(3):330–331. doi:10.1093/aje/kwu178
78. Redden DT, Fernández JR, Allison DB. A simple significance test for quantile regression. *Stat Med*. 2004;23(16):2587–2597. doi:10.1002/sim.1839
79. Chalabi Y, Scott DJ, Wuertz D. Flexible distribution modeling with the generalized lambda distribution. *MPRA*. 2012;43333. Available at: https://mpra.ub.uni-muenchen.de/43333/
80. Ejima K, Pavela G, Li P, Allison DB. Generalized lambda distribution for flexibly testing differences beyond the mean in the distribution of a dependent variable such as body mass index. *Int J Obes (Lond)*. 2018;42(4):930–933. doi:10.1038/ijo.2017.262
81. Barzilai N. *Age Later: Health Span, Life Span, and the New Science of Longevity*. 1st ed. St. Martin's Press; 2020.
82. Heward CB. Negligible senescence: how will we know it when we see it? *Rejuvenation Res*. 2006;9(2):362–366. doi:10.1089/rej.2006.9.362
83. Finch CE. Variations in senescence and longevity include the possibility of negligible senescence. *J Gerontol A Biol Sci Med Sci*. 1998;53(4):B235–B239. doi:10.1093/gerona/53a.4.b235
84. Cayuela H, Olgun K, Angelini C, et al. Slow life-history strategies are associated with negligible actuarial senescence in western Palaearctic salamanders. *Proc Biol Sci*. 2019;286(1909):20191498. doi:10.1098/rspb.2019.1498
85. Finch CE. Update on slow aging and negligible senescence—a mini-review. *Gerontology*. 2009;55(3):307–313. doi:10.1159/000215589
86. Palliyaguru DL, Vieira Ligo Teixeira C, Duregon E, et al. Study of longitudinal aging in mice: presentation of experimental techniques. *J Gerontol A Biol Sci Med Sci*. 2021;76(4):552–560. doi:10.1093/gerona/glaa285
87. Hall P, Van Keilegom I. Testing for monotone increasing hazard rate. *Ann Stat*. 2005;33(3):1109–1137. doi:10.1214/009053605000000039
88. Hall P, Heckman NE. Testing for monotonicity of a regression mean by calibrating for linear functions. *Ann Stat*. 2000;28(1):20–39, 20. doi:10.1214/aos/1016120363
89. Ejima K, Brown AW, Smith DL Jr, Beyaztas U, Allison DB. 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 (Lond)*. 2020;44(6):1440–1449. doi:10.1038/s41366-020-0554-2
90. *Software PPAaSS*. NCSS, LLC.; 2017. ncss.com/software/pass
91. Heo M, Faith MS, Allison DB. Power and sample size for survival analysis under the Weibull distribution when the whole lifespan is of interest. *Mech Ageing Dev*. 1998;102(1):45–53. doi:10.1016/s0047-6374(98)00010-4
92. Tiwari HK, Birkner T, Moondan A, et al. Accurate and flexible power calculations on the spot: applications to genomic research. *Stat Interface*. 2011;4(3):353–358. doi:10.4310/sii.2011.v4.n3.a9
93. Manton KG, Tolley HD. Rectangularization of the survival curve: implications of an ill-posed question. *J Aging Health*. 1991;3(2):172–193. doi:10.1177/089826439100300204
94. Le Couteur DG, Simpson SJ, de Cabo R. Are glycans the Holy Grail for biomarkers of aging? *J Gerontol A Biol Sci Med Sci*. 2014;69(7):777–778. doi:10.1093/gerona/glt202
95. Fries JF. The compression of morbidity. 1983. *Milbank Q*. 2005;83(4):801–823. doi:10.1111/j.1468-0009.2005.00401.x
96. Manton KG, Stallard E, Tolley HD. Limits to human life expectancy: evidence, prospects, and implications. *Popul Dev Rev*. 1991;17(4):603–637. doi:10.2307/1973599
97. Steyerberg EW. *Clinical Prediction Models*. 2nd ed. Springer; 2019. doi:10.1007/978--3-030-16399-0
98. Anderson TW. *An Introduction to Multivariate Statistical Analyses*. 1st ed. Wiley; 1962.
99. Kleinbaum DG, Kupper LL, Muller KE. *Applied Regression Analysis and Other Multivariable Methods*. 1st ed. PWS-Kent Publishing Company; 1988.
100. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. 1st ed. Routledge; 2017. doi:10.1201/9781315139470
101. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Ser B Stat Methodol*. 1996;58(1):267–288. doi:10.1111/j.2517-6161.1996.tb02080.x
102. James G, Witten D, Hastie T, Tibshirani R. *An Introduction to Statistical Learning*. 1st ed. Springer; 2013. doi:10.1007/978-1-4614-7138-7
103. Kwak SK, Kim JH. Statistical data preparation: management of missing values and outliers. *Korean J Anesthesiol*. 2017;70(4):407–411. doi:10.4097/kjae.2017.70.4.407
104. Rubin DB. Inference and missing data. *Biometrika*. 1976;63(3):581–592. doi:10.1093/biomet/63.3.581
105. Austin PC, White IR, Lee DS, van Buuren S. Missing data in clinical research: a tutorial on multiple imputation. *Can J Cardiol*. 2021;37(9):1322–1331. doi:10.1016/j.cjca.2020.11.010
106. Chakraborty H, Gu H. *A Mixed Model Approach for Intent-to-Treat Analysis in Longitudinal Clinical Trials with Missing Values [Internet]*. RTI Press; 2009. https://www.ncbi.nlm.nih.gov/books/NBK538904/. doi:10.3768/rtipress.2009.mr.0009.0903
107. Brown AW, Ioannidis JP, Cope MB, Bier DM, Allison DB. Unscientific beliefs about scientific topics in nutrition. *Adv Nutr*. 2014;5(5):563–565. doi:10.3945/an.114.006577
108. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol*. 1995;57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x
109. Professional Development. American Statistical Association website. https://www.amstat.org/ASA/Your-Career/Professional-Development.aspx?hkey=17b3580d-fd8f-4ca2-a9be-d9c3f136e9da. Published 2020. Accessed June 4, 2021.
110. *The Growing Value of Statistics Education & Experience. Biostatistics and Statistics Programs website*. American Statistical Association. https://sites.google.com/view/biostats-stats-programs. Published 2021. Accessed June 4, 2021.
111. 2021 Workshops. Society for Epidemiologic Research Website. https://epiresearch.org/annual-meeting/2021-meeting/workshop/. Published 2021. Accessed June 4, 2021.
112. Wasserstein R, Schirm A, Lazar N. Statistical inference in the 21st century: a world beyond p< 0.05 [Special issue]. *Am Stat*. 2019;73:1–19. doi:10.1080/00031305.2019.1583913.