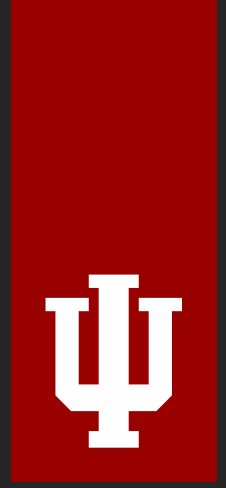


SCHOOL OF PUBLIC HEALTH BLOOMINGTON

Rigor and reproducibility in caloric restriction experiments

David B. Allison, Ph.D.

INDIANA UNIVERSITY BLOOMINGTON

Rigor

Scientific rigor is the **"strict application of the scientific method to ensure robust and unbiased experimental design, methodology, analysis, interpretation and reporting of results"**.

“Rigor does not guarantee that a study can be replicated [or reproduced and verified] but conducting a study with rigor - with a well-thought-out plan and strict adherence to methodological best practices - makes it more likely.”

National Institutes of Health. 2019. Rigor and Reproducibility in Grant Applications. https://grants.nih.gov/policy/reproducibility/guidance.htm

Reproducibility and Replicability

A National Academies of Sciences report utilizes the following working definition for reproducibility and replicability:

"Obtaining consistent results using the same input data; computational steps, methods, and code; and conditions of analysis. This definition is synonymous with 'computational reproducibility'".

"Disqualifying reproducibility criteria include nonpublic data and code, inadequate record keeping, nontransparent reporting, obsolescence of the digital artifacts, flawed attempts to reproduce others' research, and barriers in the culture of research."

""Replicability" refers to instances in which **a researcher collects new data** to arrive at **the same scientific findings as a previous study**."

National Academies of Sciences, Engineering, and Medicine. (2019). doi.org/10.17226/25303.
Yu et al. (2024) eLife. doi.org/10.7554/eLife.82268.

Verifiability

Verifiability subsumes, but goes beyond, reproducibility.

That is, a study is said to be verifiable, and to have been verified, when: (a) **the study is reproducible**, and the results have been reproduced; and (b) a determination is made that **the methods used to generate the results are valid methods** and that **the interpretations validly and logically follow from the obtained results**.

Stability

Stability reflects **an aspect similar to sensitivity analysis but focuses more on the methodological approach**. It examines whether a reasonable minor alteration in the analytic approach yields a markedly different answer.

For example, this includes assessing how outlier removal affects conclusions; the effects of using transformations or non-parametric statistics if model residuals do not seem normal; using different adjustment variables, and so on.

Transparency

Complete disclosure of the methods used, the procedures for selecting those methods, and the results obtained, including the underlying data, to **reveal the 'whole truth'.**

Transparency encompasses **sufficient clarity** that enables a competent investigator to **both reproduce and replicate the study.**

Siddique … & Allison. Hidden: A Baker's Dozen Ways In Which The Scientific Community Falls Short Of And Can Better Fulfill Einstein's Principle of The Whole Truth (manuscript in preparation)

An example of improving rigor: Developing a valid statistical test for “maximum lifespan”

- Comparing “average” (mean or median) lifespan vs. "maximum" lifespan
- Conditional t-test, which is often used to compare means is invalid in terms of Type I error rate.
- We have proposed more rigorous tests based on quantile regression to compare maximum lifespan:
  - **Wang-Allison test:** compares the proportion of animals reaching the threshold (e.g., 90th percentile of life span) between treatment groups.
  - **Gao-Allison maximum lifespan test:** Compares the treatment groups on the likelihood to live past the old age threshold and the magnitude of how long the individuals lived past the threshold.

Example of poor rigor

Using a single example as "proof of principle" without considering the stochastic component of statistical analysis.

- High-profile journals' need to include real data examples in manuscripts in general, particularly those with desirable findings, is not always necessary (scientifically unsound) and may have adverse implications.
- Researchers might selectively publish examples that show their method in the most favorable context, ignoring less favorable findings.
- **Simulation**: Haseman-Elston (**HE**) vs. weighted Haseman-Elston (**wHE**) tests.
  - Both tests are valid under H_0 but wHE is more powerful than HE under H_1 (*[**wHE** = 68% power] > [**HE** = 53% power] at 5% level of significance*).
  - **HE** test produced a lower p-value: more significant result than **wHE** test for at least 29% of the generated data sets.

"These data suggest that had Wang and Elston been required to vet their new method by showing that it produced more significant results than the older method, with realistic data sets, there would have been roughly a 30% chance of finding that the newer method produced less significant results, perhaps leading a naive reader to conclude that the new method was not really more powerful." (Williams K.Y. et al (2011))

GeroScience
https://doi.org/10.1007/s11357-024-01161-9

ORIGINAL ARTICLE

The Gehan test identifies life-extending compounds overlooked by the log-rank test in the NIA Interventions Testing Program: Metformin, Enalapril, caffeic acid phenethyl ester, green tea extract, and 17-dimethylaminoethylamino-17-demethoxygeldanamycin hydrochloride

Nisi Jiang · Jonathan Gelfond · Qianqian Liu · Randy Strong · James F. Nelson

Received: 17 February 2024 / Accepted: 10 April 2024

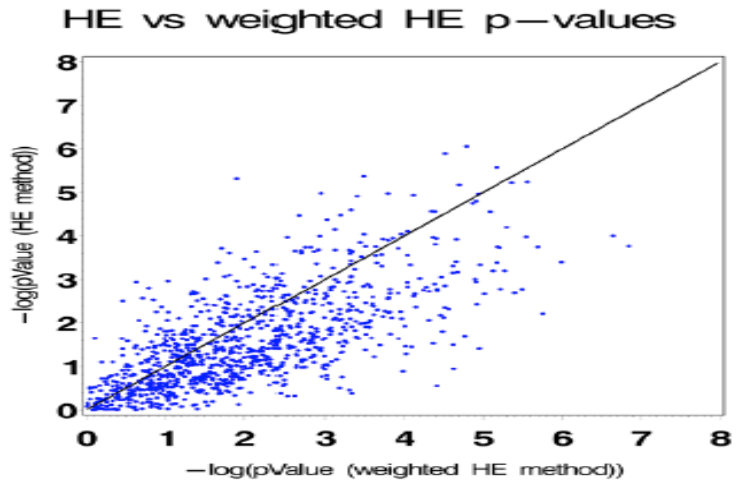


Figure: Results (p-values) from analysis of 1,000 simulated data sets analyzed with the original Haseman-Elston (HE) method and a newer weighted Haseman-Elston method.

- Jiang, N., Gelfond, J., Liu, Q., Strong, R., & Nelson, J. F. (2024). The Gehan test identifies life-extending compounds overlooked by the log-rank test in the NIA Interventions Testing Program: Metformin, Enalapril, caffeic acid phenethyl ester, green tea extract, and 17-dimethylaminoethylamino-17-demethoxygeldanamycin hydrochloride. *bioRxiv*, 2024-02.
- Williams, K. Y., Yoo, Y. J., Patki, A., & Allison, D. B. (2011). Real data examples in statistical methods papers: Tremendously valuable, and also tremendously misvalued. *Statistics and its interface*, *4*(3), 267.

Failures and Successes in Research Rigor and Transparency

Methodological Errors:

Jamshidi-Naeini *et al. BMC Psychiatry* (2023) 23:683
https://doi.org/10.1186/s12888-023-05180-2

BMC Psychiatry

MATTERS ARISING Open Access

Corrected analysis of "the effects of bright light treatment on affective symptoms in people with dementia: a 24-week cluster randomized controlled trial" that accounts for clustering and nesting verifies conclusions

Yasaman Jamshidi-Naeini[1], Lilian Golzarri-Arroyo[1], Abu Bakkar Siddique[2], Colby J. Vorland[1], Miriam Jocelyn Rodriguez[3], Richard J. Holden[3,4,5] and David B. Allison[1*]

J Nutr Health Aging. 2019;23(4):397

LETTER TO THE EDITOR

I. Nykänen, *Institute of Public Health and Clinical Nutrition, University of Eastern Finland, P.O. Box 1627, FI-70211 Kuopio, Finland Phone: +358 40 355 2991, Fax: +358 17 162 131, E-mail: Irma.Nykanen@uef.fi*

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"

Excellence in Transparency and Data Sharing:

Journal of Alzheimer's Disease Reports 8 (2024) 677–679
DOI 10.3233/ADR-240006
IOS Press

Commentary

Promoting Trustworthiness of Science: Reproducing and Verifying Agarwal et al.'s (2022) Findings Through Collaborative Endeavors

Yasaman Jamshidi-Naeini[a], Nicolas Escobar Velasquez[a], Lilian Golzarri-Arroyo[a], Sumayyah Ali[a], Luke R. Howard[b], Stephanie Dickinson[a] and David B. Allison[a,*]
[a]*Department of Epidemiology and Biostatistics, Indiana University School of Public Health, Bloomington, IN, USA*
[b]*Department of Applied Health Science, Indiana University School of Public Health, Bloomington, IN, USA*

› J Alzheimers Dis. 2022;88(2):653-661. doi: 10.3233/JAD-215600.

Pelargonidin and Berry Intake Association with Alzheimer's Disease Neuropathology: A Community-Based Study

Puja Agarwal [1 2 3], Thomas M Holland [2 4], Bryan D James [1 2], Laurel J Cherian [5], Neelum T Aggarwal [1 5], Sue E Leurgans [1 5], David A Bennett [1 5], Julie A Schneider [1 5 6]

Affiliations + expand

PMID: 35694918 PMCID: PMC10903634 DOI: 10.3233/JAD-215600

Exploring the Effects of Clustering and Cage Mate on Survival Time in the Interventions Testing Program (ITP)

GeroScience (2024) 46:795–816
https://doi.org/10.1007/s11357-023-01011-0

ORIGINAL ARTICLE

Astaxanthin and meclizine extend lifespan in UM-HET3 male mice; fisetin, SG1002 (hydrogen sulfide donor), dimethyl fumarate, mycophenolic acid, and 4-phenylbutyrate do not significantly affect lifespan in either sex at the doses and schedules used

David E. Harrison · Randy Strong · Peter Reifsnyder · Nadia Rosenthal · Ron Korstanje · Elizabeth Fernandez · Kevin Flurkey · Brett C. Ginsburg · Meredith D. Murrell · Martin A. Javors · Marisa Lopez-Cruzan · James F. Nelson · Bradley J. Willcox · Richard Allsopp · David M. Watumull · David G. Watumull · Gino Cortopassi · James L. Kirkland · Tamar Tchkonia · Young Geun Choi · Matthew J. Yousefzadeh · Paul D. Robbins · James R. Mitchell · Murat Acar · Ethan A. Sarnoski · Michael R. Bene · Adam Salmon · Navasuja Kumar · Richard A. Miller

Received: 28 August 2023 / Accepted: 7 November 2023 / Published online: 2 December 2023

Effect of Clustering on Results

- We highlight that cages and clustering by cage were not accounted for in the original analysis, potentially leading to biased results.

Impact of Number of Cage Mates on Survival Time

- We used frailty Cox models with a time-varying covariate for the number of cage mates alive at any point.

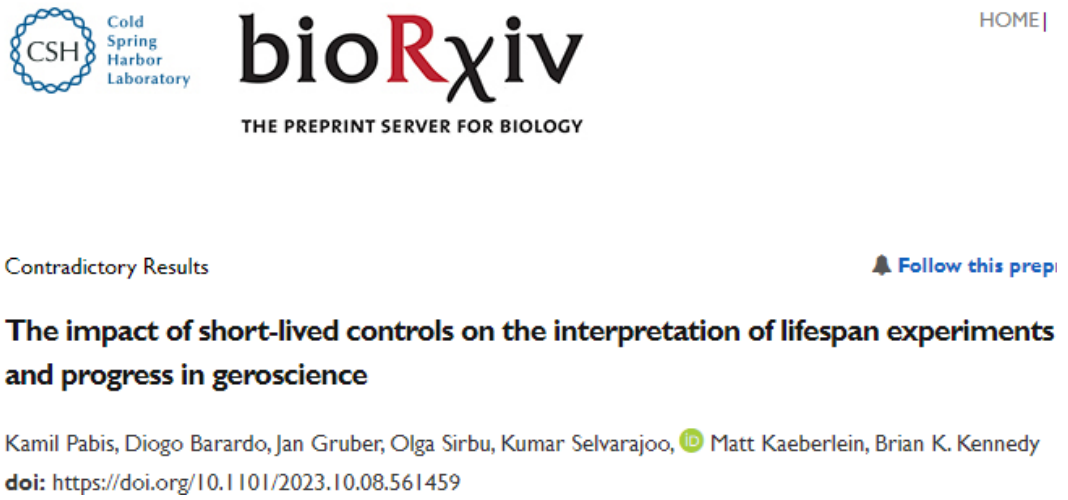
- The number of cage mates alive was a highly significant predictor of survival time ($p < 0.001$).

https://doi.org/10.1007/s11357-023-01011-0

Rigor in Lifespan Extension Intervention Studies

"Although differences in genetic background, age of treatment onset, husbandry, and dosing between the original study and the ITP cohorts may explain the failure to replicate, another potential factor is methodological rigor."

Questions are being raised about studies with control groups having shorter than average lifespans, whether these can exaggerate the perceived benefits of interventions and lead to irreplicable results.


Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

HOME|

Contradictory Results

Follow this prepr

The impact of short-lived controls on the interpretation of lifespan experiments and progress in geroscience

Kamil Pabis, Diogo Barardo, Jan Gruber, Olga Sirbu, Kumar Selvarajoo, Matt Kaeberlein, Brian K. Kennedy

doi: https://doi.org/10.1101/2023.10.08.561459

https://doi.org/10.1101/2023.10.08.561459

Data and Code Sharing Practices in Research Funded by Nathan Shock Centers of Excellence in the Basic Biology of Aging

Classified research that generated data into categories

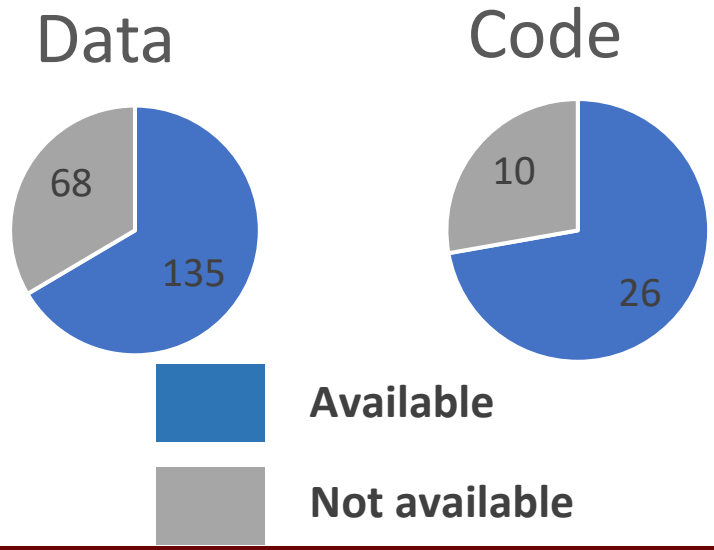
- Data are available in a repository
- Data are available in a supplementary/supporting file(s) (e.g., a csv file)
- Data are included in the paper
- Data are available upon request
- There is an explicit statement that data will not be made available
- There is no statement about data in the paper
- Other

(Similar classification system for code)

1. Are there data and code availability statements?

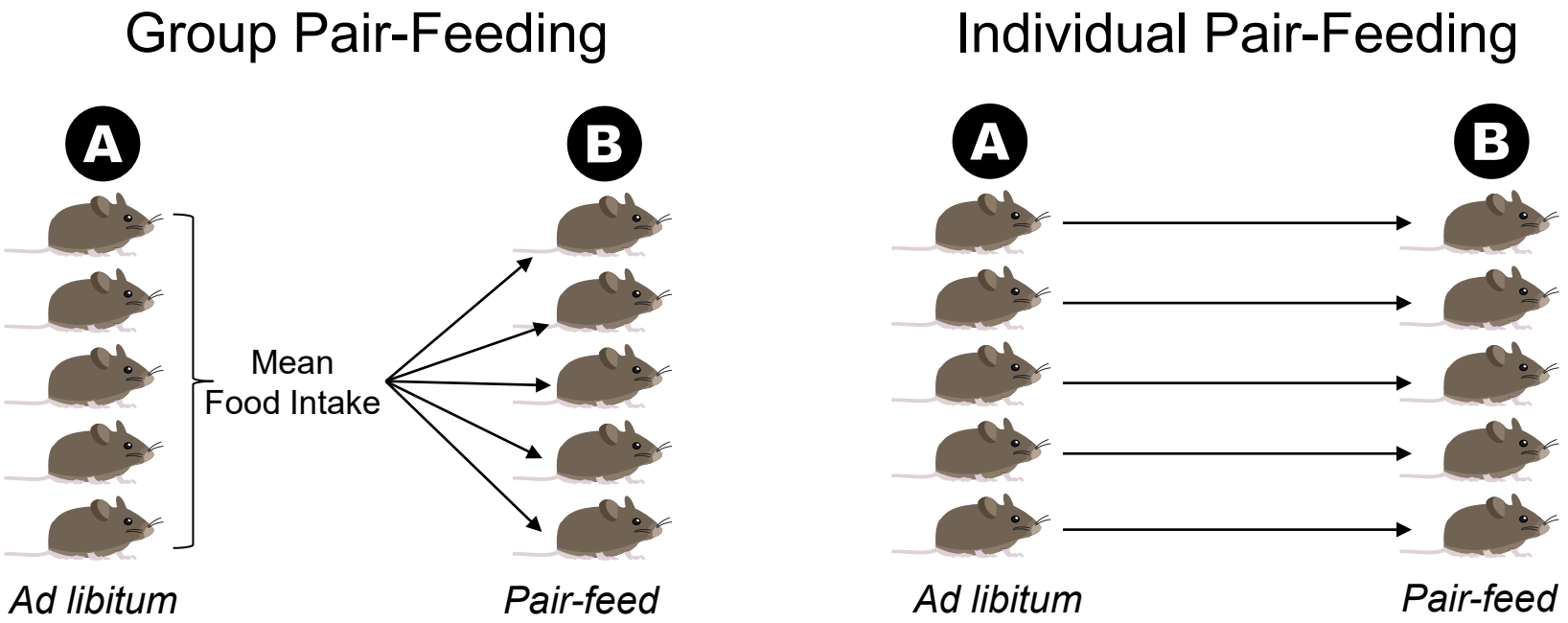
	Data	Code
No statement	45%	90%
In repository	30%	6%
In supplemental files / in paper	35%	0.8%
On request	13%	1.3%
Not available	0.3%	0.5%
Other	1.8%	1.3%

2. Are data and code *actually* available?



Vorland, Kramer, Hultquist, Escobar Velasquez, Allison, Brown. In Preparation.

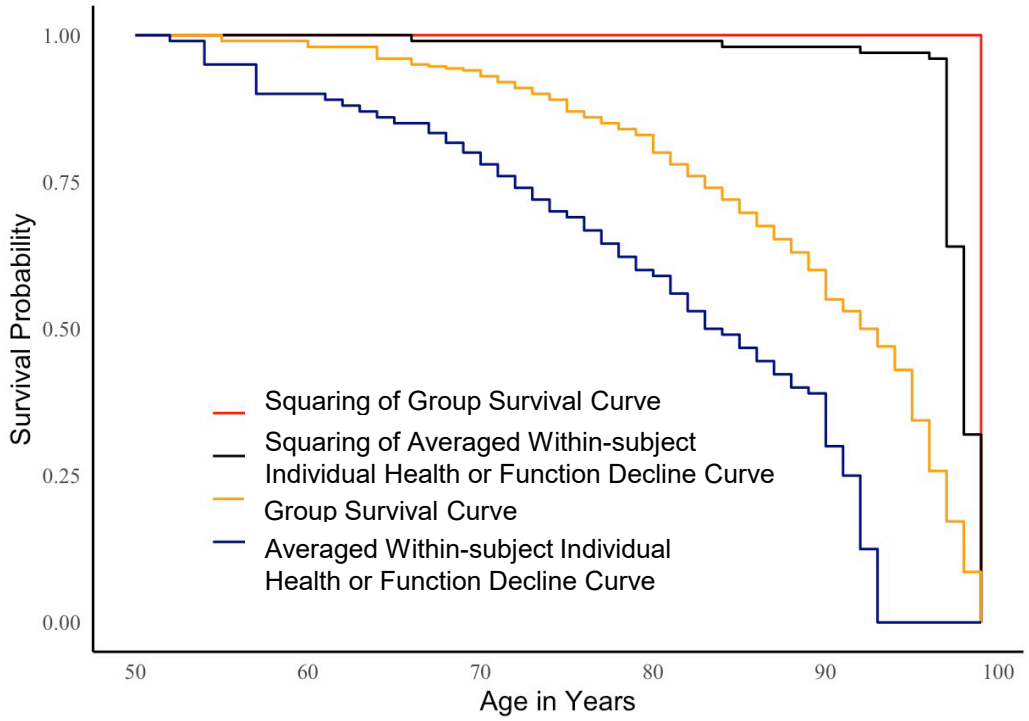
Pair-Feeding



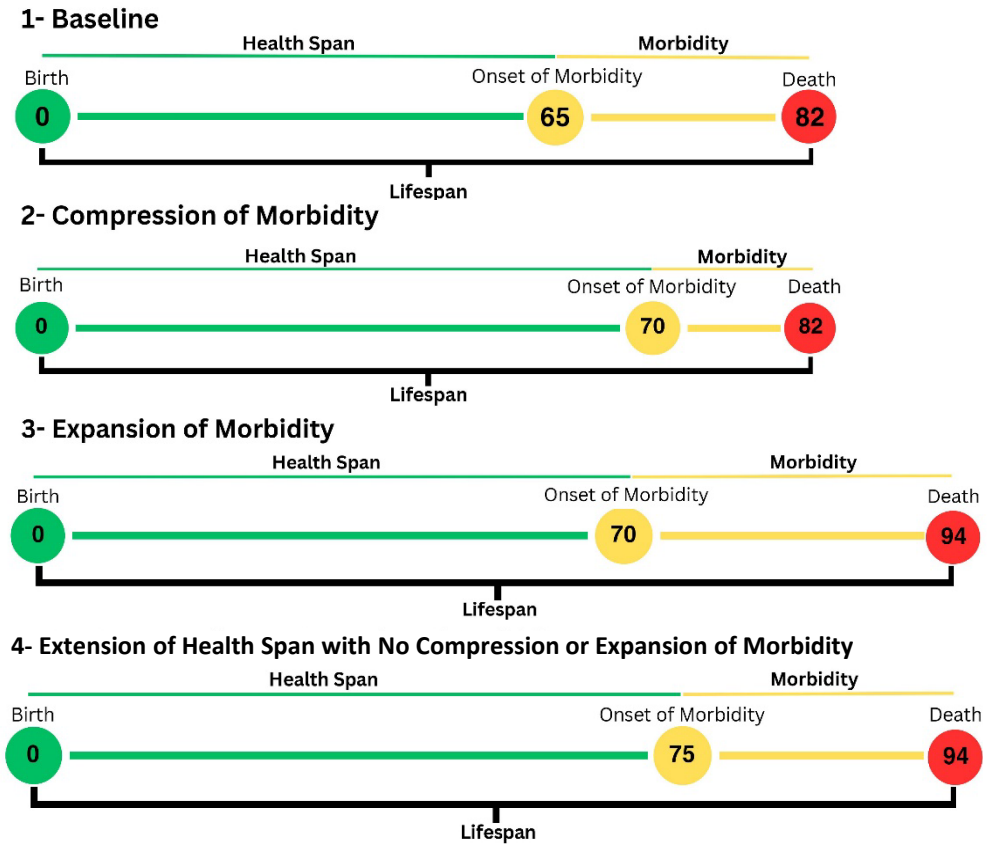
- Pair-feeding violates the independence assumptions, that the independent variable and the covariate must be independent from each other.

Compression of Morbidity

Curve Squaring vs Compression of Morbidity



Healthspan vs Compression of Morbidity



Additional Gaps & Topics We are Thinking About

1. Reliable biomarkers that predict lifespan and healthspan are needed.

2. Standardization of frailty measures is a big need.

3. Best methodologic practices for looking at mediation effects of variables that can only be measured with terminal measurements.

4. How to identify (separately estimate) how effects of interventions on longevity are affected by age at which the intervention is first received, duration for which the intervention is received, from age at which the effect on mortality rate is studied.

https://www.cartoonstock.com/cartoon?searchID=CS136373

Opportunity for More Rigor in Aging and Senescence Research

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.

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

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,*]

Editor's choice

“…let us take this path through the woods…”

~ Jean-Jacques Rousseau

謝謝您

EXTRA SLIDES