ANNUAL REVIEWS

*Annual Review of Statistics and Its Application*

# Statistical Methods in Aging Research: Improving Current Practices and Embracing Emerging Approaches

Deependra K. Thapa,[1] Erik S. Parker,[1] Mounika Kandukuri,[2] Xi (Rita) Wang,[2] Thirupathi R. Mokalla,[2] Olivia C. Robertson,[2] Wasiuddin Najam,[1] Andrew E. Teschendorff,[3] Andrew W. Brown,[4,5] John R. Speakman,[6] Yisheng Peng,[7] Bernard S. Gorman,[8] Heping Zhang,[9] Luis-Enrique Becerra-Garcia,[1] Colby J. Vorland,[1] and David B. Allison[1,2]

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

[2]USDA/ARS Children's Nutrition Research Center, Baylor College of Medicine, Houston, Texas, USA; email: david.allison@bcm.edu

[3]CAS Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, Chinese Academy of Sciences, University of Chinese Academy of Sciences, Shanghai, China

[4]Department of Biostatistics, University of Arkansas for Medical Sciences, Little Rock, Arkansas, USA

[5]Arkansas Children's Research Institute, Little Rock, Arkansas, USA

[6]School of Biological Sciences, University of Aberdeen, Aberdeen, United Kingdom

[7]Department of Organizational Sciences and Communication, George Washington University, Washington, DC, USA

[8]Gordon F. Derner School of Psychology, Adelphi University, Garden City, New York, USA

[9]Department of Statistics and Data Science, Yale University, New Haven, Connecticut, USA

Annu. Rev. Stat. Appl. 2026. 13:17.1–17.33

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-042324-060005

## Keywords

aging research, compression of morbidity, data analysis, geroscience, statistical methods, statistical rigor

## Abstract

Aging research relies on varied statistical methods, and applying these methods appropriately is important for scientific rigor. However, proper use of these statistical techniques is a challenge. We discuss two categories of statistical methods in aging research: (*a*) emerging methods requiring further validation, including techniques to examine compression of morbidity, maximum lifespan, immortal time bias, molecular aging clocks, and treatment response heterogeneity, and (*b*) classic and existing methods needing reconsideration and improvement, such as stepwise regression, generalized linear models, methods for accounting for clustering and nesting effects, methods for testing for group differences, methods for mediation and moderation analyses, and nonlinear models. For each method, we review its relevance to aging research, highlight statistical issues, and suggest improvements or alternatives with examples from aging research. We urge researchers to refine traditional approaches and embrace emerging methods tailored to the unique challenges of aging research. This review will help researchers identify and apply sound statistical methods, thereby improving statistical rigor in aging research.

## 1. INTRODUCTION

The foundations of statistical analysis in aging research rest upon well-understood, and widely applied, classic methods such as *t*-tests and their nonparametric alternatives (Lane 2022), chi-squared and Fisher's exact tests for contingency tables (Alexander et al. 2018), multiple regression (Tabachnick & Fidell 2013), and survival analyses such as the Kaplan–Meier estimator, log-rank test, and Cox proportional hazards models (Cox & Oaks 1984, Hosmer et al. 2008). These methods have been well-covered by others. Here we focus on statistical techniques in aging research that are either emerging or in need of reconsideration. First, we discuss emerging statistical techniques specifically relevant to geroscience research that, while promising, require rigorous testing and proper validation. Then, we critically examine some of the widely used methods whose use should be either reconsidered or improved due to their common misapplication. Throughout, we emphasize appropriate use and potential pitfalls that often lead to misuse or misleading conclusions in the field.

## 2. EMERGING STATISTICAL METHODS IN GEROSCIENCE RESEARCH

As aging research seeks to address complex questions—such as whether antiaging interventions can compress morbidity, extend maximum lifespan, or slow biological aging—there is growing reliance on innovative statistical approaches. In this section, we cover several emerging statistical methods that are increasingly shaping the future of geroscience research but that have yet to be broadly adopted and standardized. These include methods for testing compression of morbidity, advancements in analyzing maximum lifespan and estimating species-level lifespan limits, approaches for identifying and correcting immortal time bias, and the development and application of molecular aging clocks. We also explore recent efforts to address treatment response heterogeneity (TRH), an often overlooked but critical source of variation in aging studies. We highlight the promise of these emerging methods while critically discussing their limitations, current gaps in statistical understanding, and the need for robust validation for their appropriate use in geroscience.

## 2.1. Statistical Methods to Test Compression of Morbidity

This section focuses on the statistical methods used to examine whether life-extending interventions prolong health or illness in aging research. We introduce the concept of compression of morbidity, review statistical approaches applied in both human observational studies and model organism experiments, and compare methods for quantifying health and morbid spans relative to lifespan. Recognizing the critical gap in rigorous statistical methods for testing compression of morbidity, we also introduce a novel analytical approach our group is developing that compares rates of health decline and survival decline toward the end of life.

**AUC:** area under the curve

**2.1.1. Introduction to compression of morbidity.** The twentieth century witnessed an unprecedented gain in life expectancy, while research in model organisms identified various life-extending interventions, including caloric restriction, pharmacological interventions, and epigenetic reprogramming. As human longevity rises, a key question emerges: Does extended lifespan imply a longer period of healthy life? Three competing hypotheses exist: "failure of success"—increased survival leads to higher disease prevalence, with a longer lifespan leading to more years of unhealthy life (Gruenberg 1977); "compression of morbidity"—chronic disease onset is delayed toward the end of life, with a longer lifespan leading to more healthy life (Fries 1983); and "dynamic equilibrium"—a balance exists between mortality and morbidity (Manton 1982). Understanding which scenario dominates has important implications for informing health care strategies.

The shorter period of morbidity hypothesized as compression of morbidity is thought to occur either by morbidity rates declining more quickly than mortality rates or through a delay in the age of onset of chronic conditions that is greater than the increase in life expectancy (Fries 1983). Compression of morbidity can be absolute, in which unhealthy life years ("sick span") decrease, or relative, in which longevity increases and the proportion of unhealthy life years relative to the total lifespan decreases. While life-extending interventions could compress morbidity by increasing healthy years, they may also expand years of morbidity by extending the period of illness. **Figure 1** illustrates different scenarios of compression or expansion of morbidity because of life-extending interventions.

**2.1.2. Statistical methods for analyzing compression of morbidity.** The relationship between increased longevity and compression of morbidity in human populations, and the impact of life-extending interventions on morbidity compression in model organisms, remains unclear. A critical gap remains in the availability of quantitative measures for assessing health (and/or sick) span and in standardized quantitative approaches to analyze compression of morbidity (Thapa et al. 2024).

***2.1.2.1. Human observational studies.*** Most human observational studies examining compression of morbidity use the concept of "health expectancy," a summary measure combining morbidity and mortality (Jagger et al. 2014, di Lego 2021). Health expectancy is estimated by dividing the survival curve into slices representing various health states and estimating the specific areas under the curve (AUCs) to quantify how much of total life expectancy is lived in good health. These studies often use one of two approaches to estimate health expectancy. The prevalence-based life table [i.e., Sullivan's Method (Sullivan 1971)] is used with cross-sectional data (e.g., Graham et al. 2004, Manton et al. 2008). This method integrates age-specific mortality rates with data on disease (or disability) prevalence. Alternatively, the incidence-based multistate approach relies on longitudinal data, often using the Markov transition model (e.g., Marioni et al. 2012, Nusselder et al. 2000). The incidence-based multistate method estimates the transition probabilities between all health states (e.g., healthy to disabled, healthy to death), which allows researchers

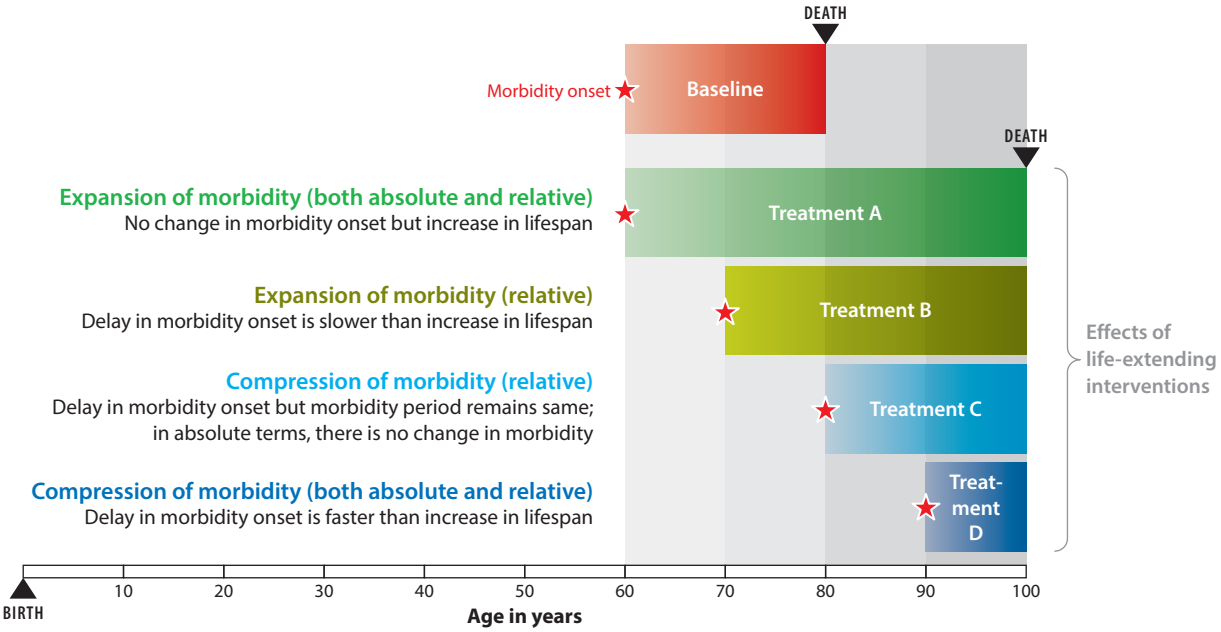


**Figure 1**

Models of compression or expansion of morbidity (models based on Fries 1989, Walter et al. 2016).

to estimate health expectancies for different health states. Under these methods, compression of morbidity is determined to be present if health expectancy increases faster than life expectancy, while morbidity expansion is observed if health expectancy remains stable or increases more slowly than life expectancy.

Other studies compare morbidity across different age groups and over time or examine both morbidity and mortality trends (e.g., Beltrán-Sánchez et al. 2016, Gouveia & Raposo 2019). However, merely observing changes in morbidity or mortality over time may not provide the statistical rigor needed to test the hypothesis of morbidity compression (see the sidebar titled Examining the Impact of Life-Extending Interventions on Compression of Morbidity).

***2.1.2.2. Model organism experimental studies.*** To our knowledge, two studies have used systematic and statistically rigorous methods to assess the effect of life-extending interventions on

## EXAMINING THE IMPACT OF LIFE-EXTENDING INTERVENTIONS ON COMPRESSION OF MORBIDITY

Although numerous studies claim that various antiaging interventions improve the health span of model organisms, a critical gap remains in the availability of quantitative measures for assessing health span and of approaches for statistical validation of compression of morbidity. The lack of a clearly articulated statistical approach with well-defined, specifiable statistical properties has slowed progress in identifying interventions that compress morbidity. We propose a novel statistical approach that compares the rates of health decline and survival decline near the end of life. A smaller difference between the rate of health decline and the rate of survival decline provides evidence for compression of morbidity in the intervention group compared with the control group.

both health span and lifespan, and thus compression of morbidity in model organisms. First, Yang et al. (2025) proposed that interventions leading to a steeper survival curve may compress morbidity. In contrast, interventions that extend longevity while maintaining the original shape of the survival curve should expand morbidity. Second, Lamming (2024) introduced frailty-adjusted mouse years (FAMY) as a summary measure for quantifying health span in mice, which is like the concept of quality-adjusted life years (QALY), a measure of health expectancy in humans. Lamming (2024) calculated FAMY as the AUC of a graph with vitality (the complement of frailty; 1 – frailty index) on the *y*-axis and age of the mouse on the *x*-axis. The estimate of AUC was further divided by 365 to convert the units to years.

**FAMY:** frailty-adjusted mouse years

**QALY:** quality-adjusted life years

**2.1.3. Difference in rate effects—a novel statistical approach for testing compression of morbidity.** Recognizing the critical gap in the availability of statistically valid methods for testing compression of morbidity, our group has developed a statistical method to compare rates of health decline and survival decline toward the end of life (Thapa et al. 2025). This method provides evidence for compression of morbidity if the difference between the rate of decline in health and the rate of decline in survival is lower in an intervention group than in a control group. For our analysis, we utilized publicly available data from Di Francesco et al. (2024), who investigated the effects of intermittent fasting and chronic caloric restrictions on health span and lifespan in 937 genetically diverse outbred mice. Di Francesco et al.'s study measured morbidity by assessing a clinically relevant frailty index (possible score range 0 to 1, with higher values indicating greater frailty) using methods described by Parks et al. (2011). Baseline measurements in the mice were taken at 5 months of age, intervention began at 6 months, and follow-up assessments were conducted approximately every 6 months, for a total of six time points. All animals were monitored until natural death.

For our analysis, we defined vitality as the complement of frailty (i.e., 1 – frailty index) and expressed it at each measurement time point as a proportion of its baseline value. We calculated the average rate of vitality decline by fitting exponential decay models to individual vitality trajectories. The rate of survival decline was estimated from a Cox proportional hazards model. Both vitality and survival decline rates were estimated at 18 months of age (540 days, roughly equivalent to 60 human years) and subsequently at 30-day intervals. Finally, we calculated the difference between the rates of decline in vitality and survival and compared the mean of these difference scores across intervention groups. **Figure 2** presents preliminary results of our analysis. Contrary to our hypothesis, the intervention groups—particularly the 20% and 40% caloric restriction groups—exhibited significantly higher differences between the rates of decline in vitality and survival compared with the control group, suggesting that these life-extending interventions may lead to an expansion of morbidity.

## 2.2. Tests for Maximum Lifespan

This section focuses on statistical methods developed to test for evidence of changes in maximum lifespan, rather than only mean or median lifespan. Observation of the full lifespan in animal models enables geroscience researchers to evaluate intervention effects across the entire survival distribution, with particular attention to the upper extreme. We begin by providing background on early efforts to evaluate maximum lifespan, then review recent methodological improvements, and finally describe the Wang–Allison and Gao–Allison tests that have become influential in the field.

**2.2.1. Background and early testing of maximum lifespan.** Unlike human studies and studies of animals in the wild, longevity or geroscience investigations in laboratory animal models typically study the animal's entire lifespan. This offers unique opportunities to explore effects of different

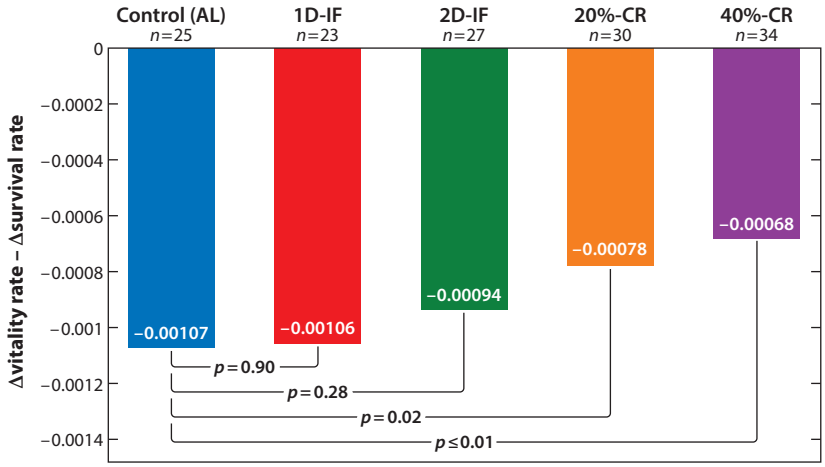


**Figure 2**

Mean scores of the difference between the rates of decline in vitality and survival by diet groups. *p*-Values are obtained from a linear regression. *n* refers to the number of measurement points (ages) at which the vitality and survival were estimated. Contrary to our hypothesis, the intervention groups—particularly the 20% and 40% caloric restriction groups—showed significantly higher mean differences between the rates of decline in vitality and survival, suggesting that life-extending interventions may lead to an expansion of morbidity. Abbreviations: 1D-IF, intermittent fasting one day per week; 2D-IF, intermittent fasting two consecutive days per week; 20%-CR, caloric restriction at 20%; 40%-CR, caloric restriction at 40%; AL, ad libitum.

interventions on lifespan or the survival distribution, including its mean, median, and upper or lower quantiles.

The phrase "maximum lifespan" (which is used differently here than in the later section on maximal lifespan at the species level in Section 2.3) was historically adopted to refer to the lifespans of some proportion of a sample of animals in an experimental study that lived longer than the remaining portion. Usually, this portion was small, such as those animals surviving beyond the 90th sample percentile of the study.

Longitudinal geroscientists opined that interventions that truly slow the rate of aging, as opposed to simply preventing or ameliorating some specific disease or disease process, will increase not only mean or median lifespan but also maximum lifespan. To test this, two or more experimental groups exposed to different treatment conditions would be compared with respect to lifespans beyond this upper percentile. An earlier commonly used statistical approach for this comparison was shown to be invalid (Gao et al. 2008, Wang et al. 2004).

**2.2.2. Recent improvements.** Allison and colleagues recently derived several valid tests for comparing differences in maximum lifespan, one of which, the Wang–Allison test (Wang et al. 2004), has become a standard procedure in geroscience studies in animal models. This test essentially evaluates whether the proportion of animals surviving beyond 90th percentiles (or any other percentile an investigator chooses) differs significantly between intervention groups and is widely used by the intervention testing program of the National Institute on Aging (Nadon et al. 2008). A fundamental problem in longevity research is the practical impossibility of observing maximum lifespan in a finite sample of a population over a fixed observation period. The Wang–Allison test addresses this challenge by comparing the probability of surviving beyond a predefined age (often characterized as old age) in the tail of the survival distribution. This approach enables researchers to examine treatment effects on extreme longevity without relying on the rarely observed

## GAO–ALLISON AND WANG–ALLISON TESTS FOR TESTING MAXIMUM LIFESPAN

Two tests for maximum lifespan (Gao et al. 2008, Wang et al. 2004) have emerged as standards in geroscience research. The newer Gao–Allison test is theoretically more powerful under some circumstances, but the older Wang–Allison test is more widely applicable. Both are easy to perform and allow geroscientists to examine whether aging interventions contribute survival at the right tail of the survival distribution, an essential aspect of identifying true lifespan benefits. The Wang–Allison test, which has become a benchmark in preclinical aging studies, compares the proportion of animals surviving beyond a pooled upper percentile (usually the 90th percentile), thus circumventing the impracticality of observing true maximum lifespan in finite studies. The Gao–Allison test may offer greater statistical power in some settings but requires complete survival data or imputation, making it more complex in application. Together, these tests offer analytical tools to distinguish interventions that extend extreme longevity from those extending average lifespan.

maximum lifespan. The subsequently developed Gao–Allison test (Gao et al. 2008) can be more powerful under some circumstances, but it requires either some form of complex imputation of survival times or waiting until absolutely every animal in the study has died (see the sidebar titled Gao–Allison and Wang–Allison Tests for Testing Maximum Lifespan).

**2.2.3. Details of the Wang–Allison and Gao–Allison tests for maximum lifespan.** The Wang–Allison and Gao–Allison test procedure is simple and involves calculating the 90th percentile survival time by pooling data from all intervention groups. Pooling the data is essential: If one calculates survival time separately in the different intervention groups, the procedure is invalid. After one has calculated the upper percentile of survival time, one simply conducts a test, such as Fisher's exact test, to determine whether the proportion of individuals that survive beyond the pooled upper percentile differs significantly between intervention groups.

**Table 1** (reproduced from Harrison et al. 2014) illustrates survival analysis results using both the log-rank test (for median lifespan) and the Wang–Allison test (for maximum lifespan). The

**Table 1 An illustration of survival analysis using log-rank and the Wang–Allison test**

| Group | Median lifespan | | | Lifespan at 90th percentile | | |
|---|---|---|---|---|---|---|
| | Days | Difference (%) | Log-rank *p*-value[a] | Days | Difference (%) | Wang–Allison *p*-value |
| **Males** | | | | | | |
| Control | 807 | NA | NA | 1,094 | NA | NA |
| ACA | 984 | 21.9 | **<0.001** | 1,215 | 11.1 | **<0.001** |
| EST | 900 | 11.5 | **0.002** | 1,148 | 4.9 | 0.130 |
| MB | 790 | −2.1 | 0.270 | 1,037 | −5.2 | 0.600 |
| **Females** | | | | | | |
| Control | 896 | NA | NA | 1,072 | NA | NA |
| ACA | 939 | 4.8 | **0.010** | 1,167 | 8.9 | **0.001** |
| EST | 893 | −0.3 | 0.800 | 1,068 | −0.4 | 0.900 |
| MB | 902 | 0.7 | 0.170 | 1,138 | 6.2 | **0.004** |

Table reproduced from Harrison et al. (2014). Abbreviations: ACA, acarbose; EST, 17-α-estradiol; MB, methylene blue; NA, not applicable.

[a]Log-rank *p*-values consider all the data, while the Wang–Allison test used Fisher's exact test to compare the proportion of surviving mice in control and treatment groups at the age corresponding to the 90th percentile of the pooled survival distribution—that is, when 90% of all mice had died. *p*-Values $< 0.05$ are presented in bold.

results showed that acarbose significantly increased both median and maximum lifespan in male and female mice compared with controls. 17-α-Estradiol significantly extended median lifespan in male mice but had no effect on maximum lifespan. Conversely, methylene blue improved maximum lifespan in female mice, but not median lifespan.

If a researcher uses multiple percentiles to define "old" and performs multiple tests, they should state clearly how many tests were done, whether the tests were preplanned, and whether any statistically significant results would still be significant if a multiple testing correction (e.g., a Bonferroni correction) were applied.

## 2.3. Estimating Maximal Achievable Lifespan at the Species Level

This section addresses how maximal lifespan is quantified at the species level, treating it as a population-based trait. We first define maximal lifespan and highlight the conceptual and statistical challenges inherent in comparing lifespans across populations and species. Next, we review surrogate measures, such as the 90th quantile of lifespan (Q90) and the average lifespan of the longest-lived 10% of the population (e90), which are less sensitive to sample size and strongly correlated with maximal lifespan. Finally, we discuss the limitations of these metrics for statistical analysis, particularly in interspecific comparisons, and highlight rigorous methods, including the Wang–Allison and Gao–Allison tests, that address some of these challenges. The goal is to provide a clear framework for understanding and comparing species-level longevity while accounting for sampling, environmental, and statistical constraints.

**2.3.1. Definition of maximal lifespan.** Lifespan is an individual trait. A given animal lives a certain number of days, which can be analyzed like any other continuously distributed variable. Maximal lifespan, in contrast, is a population-based trait. One individual in a population lives the longest, and the lifespan of that individual is taken to be the maximal lifespan. Maximal lifespan can only be defined relative to the other members of a sampled population. Maximal lifespan has been used as a metric in studies that compare species (e.g., Austad & Fischer 1991, Holmes & Austad 1995, Hulbert et al. 2007). Several databases are available that have values of maximal lifespan for different species (e.g., De Magalhães & Costa 2009).

**2.3.2. Issues with defining the maximal lifespan of a population.** Maximal lifespan depends critically on the size of the sample taken from a given population (Gavrilov & Gavrilova 1991). This is because adding an extra individual to a sample of $n$ individuals can only maintain or increase the maximal lifespan. This characteristic is not unique to maximal lifespan. It occurs anytime one is looking for the most extreme—maximum or minimum—value of a variable. For example, when measuring the minimal (or maximal) metabolic rate, extending the measurement period can only result in the same or a lower minimum (Hayes et al. 1992) or the same or higher maximum. Comparisons of maximal lifespan among species, or populations, or individuals exposed to different treatments can only be performed when the sample sizes of the different groups are at least approximately equivalent. Analyses of extreme human lifespans suggest that remaining life length after the age of 109 years is approximately exponentially distributed, with any upper limit likely beyond currently observed values, highlighting the difficulties of estimating maximal lifespan from finite samples (Belzile et al. 2022). A classic misuse of maximal lifespan is the observation that maximal lifespan in humans is much greater than in other primates. The value for humans (Jeanne Calment, documented as the longest-lived human, lived to be 122) is based on a sample of literally billions of individuals, while other primates are typically represented by less than 1,000 values.

A second issue is that the maximal lifespan of a given population is only a single value. It has no variance. Hence, comparing the lifespan of different groupings falls outside the scope of classic

frequentist inferential statistical analyses, which depend on comparing the magnitude of differences to the variation of a given variable. Furthermore, relying on single values requires an assumption of the veracity of the value (Austad 2022). The longevity of Jeanne Calment, for example, has been the subject of some dispute (Robine & Allard 1999, Zak 2019). On the other hand, Madame Calment's birth certificate was preserved, and she was listed in 14 census records (Atwal 2024). To overcome the bias caused by sample size, Moorad and colleagues (2012) explored the correlation between maximal reported lifespan and other species-specific traits to identify surrogate measures that capture the essence of extreme longevity in a population sample. They suggested that Q90 and e90 are suitable measures that demonstrate far less sensitivity to sample size and have high correlations to maximal lifespan (0.95 to 0.98). On this basis, they recommended that Q90 and e90 replace maximal lifespan as a measure of comparative longevity, and they encouraged comparative biologists to collect data to better quantify the population-level traits.

**2.3.3. Issues with statistical inference around maximal lifespan.** Despite being better metrics related to maximal lifespan, Q90 and e90 are not without problems that lead to complications in statistical analysis. The Q90 value is still a single value, and the distribution that underlies e90 is severely left-skewed, making it inappropriate to use statistical tests that assume normality in the distributions (see the sidebar titled Metrics for Species-Level Longevity).

Allison and colleagues introduced rigorous statistical approaches that overcome the limitations of Q90 and e90 (the Wang–Allison and Gao–Allison tests), which are covered in depth in Section 2.2.

A final problem, tangential to the statistical issues, particularly for interspecific comparisons of maximal lifespan, is the equivalence of the measures. Most animals in the wild do not die of old age. Hence, lifespans (both average and maximal) are typically short. Consider, for example, the field vole (*Microtus agrestis*), which in the wild typically lives only a few months, but in captivity can live several years (Selman et al. 2008, Weigl 2005). This is principally because animals in captivity are typically not subject to starvation, dehydration, predation, and to some extent infectious diseases. On the face of it, comparing maximal lifespans only of species kept in captivity might seem preferable. But this raises a problem because for most species, we are not aware of the nutritional requirements that optimize lifespan, and hence we may feed them inadequately to generate a meaningful maximal lifespan value. In cats, for example, taurine is an essential amino acid. Feeding cats diets lacking taurine compromises their health (Pion et al. 1987). Moreover, keeping some species in captivity is technically difficult or impossible. Small, mouse-sized bats, for example, may live 30 to 40 years in the wild, but in captivity they cannot be maintained in optimal husbandry conditions for anywhere near such periods.

## METRICS FOR SPECIES-LEVEL LONGEVITY

Using maximal lifespan to compare aspects of aging across species or treatments is fraught with unique conceptual and statistical problems (Moorad et al. 2012). Unlike individual lifespan, maximal lifespan is a population-based trait, defined by the oldest individual in a sample, which is highly sensitive to sample size, lacks variance, and is unsuitable for standard statistical inference. The surrogate metrics such as Q90 and e90 are preferable, as they are less affected by sample size and correlate strongly with maximal lifespan. Rigorous statistical methods for comparing such values are available (Gao et al. 2008, Wang et al. 2004), but obtaining accurate and reliable measures of maximal lifespan remains problematic, especially in interspecies comparisons, where differences between wild and captive environments, nutritional adequacy, and husbandry conditions can confound lifespan estimates.

## ADDRESSING IMMORTAL TIME BIAS IN AGING RESEARCH

Immortal time bias arises when a period during which participants cannot experience the event of interest—such as death—is incorrectly included in survival analysis. This bias typically stems from misaligned eligibility and treatment classification criteria. In aging research, it is particularly relevant because treatments are often only received by individuals who survive long enough to become eligible, leading to selection bias in time-to-event analyses. This problem can be mitigated by target trial emulation, which preserves the causal question's validity by aligning eligibility, treatment classification, and the start of follow-up. In complex scenarios—such as delayed treatments like transplants—sequential trial emulation assigns treatment at multiple time points to avoid bias. Traditional Cox models, even with time-varying extensions, are often inadequate when both immortal time and time-dependent confounding are present. Advances in causal inference, especially g-methods such as marginal structural models, structural nested models, and the parametric g-formula, offer more rigorous tools to estimate causal effects under these conditions. However, these methods rely on strong assumptions, including no unmeasured confounding and correct model specification.

### 2.4. Frozen in Time: Navigating Immortal Time Bias in Aging Research

This section focuses on the concept of immortal time bias, a threat to validity in survival analyses. This bias arises when periods of time during which participants cannot experience the event of interest are mistakenly included in survival analyses, often exaggerating apparent treatment benefits. This issue is particularly relevant in aging research, where treatment initiation typically depends on surviving to a certain age or disease progression. We discuss how the misalignment of eligibility criteria, treatment classification, and the start of the follow-up period produces immortal time and outline how methodological strategies such as target trial emulation and advanced causal inference methods help mitigate its impact.

**2.4.1. Introduction to immortal time bias.** Immortal time refers to a period during which a person, by definition, cannot experience the event of interest (e.g., cannot die when the event of interest is death). When immortal time is included in survival analyses, both the absolute risks and the effect estimates will be biased, and potentially not estimable. The two main reasons immortal time arises are (*a*) incorrect definition of eligibility criteria after the start of follow-up and (*b*) incorrect classification of individuals to treatment strategies based on post-eligibility information (see the sidebar titled Addressing Immortal Time Bias in Aging Research). In essence, immortal time bias obscures the causal question of interest due to misalignment of eligibility criteria, treatment classification, and the start of the follow-up period (Hernán et al. 2025).

**2.4.2. Relevance to aging research.** Many research questions in the aging field involve evaluating treatments (e.g., medications, operations, lifestyle interventions) that are only received if participants survive long enough to be eligible. Accordingly, selection bias may be introduced as frailer individuals may be less likely to survive long enough to receive treatment. Furthermore, many research questions are either unethical [e.g., withholding lifesaving treatment, or randomizing participants to a known toxic intervention (e.g., smoking)] or practically unfeasible (e.g., decades of follow-up time are required) to study mortality or longevity. Consequently, investigators must often resort to observational data and depend on real-world treatment initiation that is influenced by and further complicated by clinical decision-making and patient characteristics (e.g., confounding by indication). These complexities ultimately make observational studies more susceptible to immortal time bias as investigators may inadvertently misalign design elements

(eligibility criteria and treatment classification) away from the causal question they aimed to answer (e.g., Suissa 2007).

**RCT:** randomized controlled trial

**2.4.3. Target trial emulation and immortal time bias.** A target trial emulation framework (a method used in observational studies to mimic the design of a randomized trial) can be used to circumvent immortal time bias. In brief, the framework involves designing a hypothetical randomized controlled trial (RCT) and then emulating that trial using observational data to answer the causal question of interest (Hernán et al. 2022). Target trial emulation circumvents the possibility of immortal time bias by explicitly specifying and synchronizing eligibility and assignment criteria with the beginning of the follow-up period (Hernán et al. 2016).

The first type of immortal time bias occurs when an eligibility criterion is applied after the follow-up period has begun, and participants have already started treatment before the start of the emulated trial. For instance, suppose we are interested in the causal effect of initiating and continuing statin therapy on developing cancer, and we define statin users as those who have used statins continuously for the previous 4 years, compared with nonusers (Khurana et al. 2007). This treatment definition introduces a period of immortal time, during which, by definition, no participants classified as statin users can have developed cancer (i.e., the risk is zero), because if they had, they would not have met eligibility criteria for inclusion in either group. The inclusion of immortal time therefore biases the treatment group by artificially lowering the event rate, which can make statins appear more protective than they actually are (Dickerman et al. 2019). To prevent this immortal time bias, investigators must ensure that all eligibility criteria are determined and applied at the time point when treatment groups are defined and outcome risk begins to accrue (i.e., time zero), just as would be the case in an RCT.

The second type of immortal time occurs when treatment strategies cannot be clearly defined at the start of follow-up (e.g., no heart transplant versus receiving a heart transplant when one becomes available) (Gail 1972, Hernán et al. 2025). As a result, patients are classified into treatment groups based on post-eligibility information that was not initially available (e.g., those who died waiting are classified as no heart transplant). Thus, those in the treatment group must have survived long enough to receive the transplant, thereby inducing a period of immortal time where their risk of dying is 0. To prevent immortal time bias, investigators may reframe their research question in terms of treatment strategies that can be distinguished at the start of follow-up (e.g., immediate treatment initiation versus no treatment initiation), which can be generalized to the emulation of sequential target trials (Gail 1972). Thus, instead of classifying patients by whether they eventually received a transplant (which misclassifies those who died waiting), researchers can emulate sequential trials at regular intervals. At each time point, patients are assigned based on treatment received at that time and followed prospectively (Hernán et al. 2008, Keogh et al. 2023).

**2.4.4. Advances and limitations in statistical methods to address immortal time bias.** Historically, Cox models, including time-varying extensions, have been used to address timing of treatment and treatment changes in time-to-event analyses. However, these models fall short when immortal time bias and time-varying confounding are present. To address these limitations, recent developments in causal inference have led to the use of g-methods, namely, marginal structural models, the parametric g-formula, and structural nested models (Naimi et al. 2017). These methods are specifically designed to estimate causal effects in the presence of time-varying treatments and confounders and are well-suited for addressing immortal time bias (Robins et al. 2000). Marginal structural models, in particular, account for time-varying confounding because they use inverse probability weighting to balance covariates across treatment groups over time (Hernán et al. 2000). The weighted dataset is then analyzed using a Cox model to estimate the effect of

**DNAm:** DNA methylation

treatment on time-to-event outcomes. This allows for unbiased estimation of causal effects even when confounders vary over time and are influenced by earlier treatment. However, all g-methods assume that there is no unmeasured confounding and that models used to estimate treatment and censoring probabilities are correctly specified.

## 2.5. Molecular Aging Clocks: Applications and Challenges

This section focuses on molecular aging clocks and their applications and challenges in aging research. We introduce the concept of molecular clocks, with a particular focus on DNA methylation (DNAm)–based clocks, which are widely used due to their high measurement accuracy. We further review common statistical algorithms including penalized regression and alternative models, and discuss key challenges such as sample size limitations, validation, confounding, and nonlinear age-related changes. We aim to provide a framework for understanding the statistical foundations, applications, and limitations of molecular aging clocks.

**2.5.1. Introduction to molecular aging clocks.** Molecular aging clocks are machine learning–based predictors of chronological or biological age, which have found potential application in diverse areas of science, including forensics, ecological conservation, and healthy aging (Horvath & Raj 2018). In the context of aging research, they serve as promising biomarkers and tools for evaluating the effect of antiaging interventions (Rutledge et al. 2022). Typically, such clocks are derived by applying a machine learning algorithm to large-scale molecular omics datasets (this could include epigenetic, transcriptomic, proteomic, or metabolomic data), which comprise potentially large numbers of molecular features. Molecular clocks are often trained on either chronological age or some quantifiable form of biological age, such as a clinical aging biomarker like C-reactive protein levels in plasma (Levine et al. 2018). Among the different types, DNAm-based clocks are the most popular and promising molecular clocks. DNAm is an epigenetic covalent modification of DNA, which, unlike proteomic or transcriptomic measurements, can be measured with relatively high accuracy, allowing detection of even 1–2% changes in methylation associated with aging (Zhu et al. 2018) or major disease risk factors such as smoking (Joehanes et al. 2016). Although most of the discussion here is focused on DNAm clocks, many of the statistical issues discussed broadly apply to all types of molecular aging clocks.

**2.5.2. Statistical algorithms.** Penalized multivariate linear regression, particularly the elastic net (Friedman et al. 2010), has been by far the most widely applied statistical framework to derive molecular aging clocks. Several factors contribute to its widespread use. First, penalization/regularization is necessary to avoid overfitting when learning predictive models from high-dimensional omics datasets, which can range from hundreds of thousands to over a million features, frequently outnumbering samples by 100- or even 1,000-fold. Second, sparser penalized versions such as lasso regression are less favored because excessive reduction in redundancy can jeopardize validation potential in independent datasets where not all features are measured, or where predictive features drop out because they fail quality-control procedures (Teschendorff & Horvath 2025). Third, although it is now increasingly recognized that many molecular features change nonlinearly with age, linear models have been proven to be valuable approximations (Marioni et al. 2015). In this regard, it is worth pointing out that one of the very first DNAm clocks, the Horvath clock (Horvath 2013), was built by training an elastic net model on a nonlinearly transformed version of chronological age (specifically, a stepwise linear function) to account for observed age-associated log-linear changes. Fourth, the features driving these linear models are, in principle, more readily interpretable. In practice, however, interpretation is more challenging because of the underlying complexity and diversity of age-associated molecular changes (Teschendorff & Horvath 2025).

## MOLECULAR AGING CLOCKS: STATISTICAL ADVANCES AND PERSISTENT CHALLENGES

Although the field of molecular aging clocks is over a decade old, from a statistical perspective, it is still in its infancy. There is much room for improvement in developing statistical methods that can better quantify the uncertainty of estimates, especially given the noisy nature of molecular omics data, hidden latent variables and confounders, nonlinear age-related patterns, and pervasive missing data. Molecular aging clocks, particularly those based on DNAm, have emerged as powerful machine learning methods for estimating biological age, identifying risk factors, and evaluating antiaging interventions. Linear models such as the penalized multivariate linear regression, and particularly elastic net, dominate due to their ability to prevent overfitting in high-dimensional data, maintain predictive power and validation across datasets, approximate complex age-related patterns, and offer simplicity and interpretability. However, alternative methods, such as deep learning and probabilistic models, are increasingly being explored. Major challenges remain, including overfitting, limited generalizability from small sample sizes, inconsistent definitions of biological age acceleration, lack of probabilistic interpretation at the individual level, inadequate validation and benchmarking without uncertainty estimates, confounding from cell-type heterogeneity in bulk tissue data, and inadequate modeling of nonlinear aging trajectories. Addressing these challenges is essential for the robust application of molecular aging clocks in both clinical and research settings.

Depending on the type of molecular clock or predictor being developed, statistical paradigms other than the traditional penalized multivariate regression model have been considered. For instance, adaptive index-like algorithms (Tian & Tibshirani 2010) have been applied to predict cancer risk (Teschendorff et al. 2012), and Markov chain modeling has been used to predict mitotic age (i.e., age-associated cumulative number of stem-cell divisions of a tissue) (Teschendorff 2020). In these applications, the clock can be viewed as a counter that keeps track of age-associated cumulative alterations relative to a defined ground state. More advanced machine learning methods, such as those based on deep learning (LeCun et al. 2015), have also been applied to construct molecular aging clocks. For instance, a deep neural network was used to build a DNAm clock highly predictive of chronological age (de Lima Camillo et al. 2022). However, deep learning clocks have not yet led to demonstrable improvements in predicting chronological or biological age. This may be due in part to the fact that deep learning methods have not yet been applied in a sequential data context (e.g., taking the genomic position of the DNAm measurements into account), where deep learning methods are particularly powerful (see the sidebar titled Molecular Aging Clocks: Statistical Advances and Persistent Challenges).

**2.5.3. Pitfalls and challenges.** Common pitfalls and challenges in the construction and application of molecular aging clocks include issues related to data and modeling (including confounding factors), validation, and benchmarking. First, most molecular aging effects are typically subtle, requiring a large sample size to ensure sufficient power. This is often not done, with some studies building clocks from only a few hundred samples (e.g., Petkovich et al. 2017), when ideally many thousands of samples are needed, thus leading to potential overfitting and limited generalizability. A second pitfall relates to a clock's quantification of biological age acceleration. Although both absolute (estimated age minus chronological age) and relative (residuals from a regression of estimated versus chronological age) measures have been proposed, many studies apply only one of these, often without proper justification. As discussed recently, in most cases, a residual-based definition is preferable; however, there are circumstances where it could lead to false positives or false negatives (Teschendorff & Horvath 2025). A third related pitfall and outstanding challenge is the urgent need to quantify age acceleration probabilistically at the individual sample level. Currently,

**GLM:** generalized linear model

**TRH:** treatment response heterogeneity

molecular aging clock readouts are quantified in terms of numerical age-acceleration values, making it unclear whether an individual's readout falls within a physiologically normal range. This challenge is particularly pertinent for clinical applications of these clocks. While a recent study has developed a Bayesian model for probabilistic quantification of age acceleration (Dabrowski et al. 2024), further research in this direction is needed.

The fourth challenge is how best to quantify uncertainty of clock estimates in the face of missing data, as missingness is a ubiquitous characteristic of molecular omics data, especially when validating a given predictor in independent datasets. The fifth common pitfall and challenge is the comparison and benchmarking of clocks, which is often done without quantification of the magnitude of uncertainty levels associated with performance measures. This is particularly relevant when benchmarking is only conducted in one particular dataset (Teschendorff 2019). The sixth challenge is associated with the existence of nonlinear or nonmonotonic age-related patterns, which, for instance, generalized linear models (GLMs) may not model appropriately. A final pitfall more specific to the construction of DNAm clocks involves ignoring substantial confounders, such as cell-type heterogeneity (Teschendorff & Horvath 2025). In the case of DNAm, the only large-scale data available for clock construction have inevitably been generated in bulk tissues like blood, which comprises a plethora of different cell types, each with its own DNAm profile, and with cell-type proportions changing with age (Jaffe & Irizarry 2014, Luo et al. 2023). As a concrete example, Skinner et al. (2025) recently built a clock for chronic low-grade inflammation ("inflammaging") without adjustment for neutrophil–lymphocyte ratio shifts, resulting in a predictor that captures an acute form of inflammation only, and not inflammaging (Guo & Teschendorff 2025).

## 2.6. Treatment Response Heterogeneity in Aging Research

This section focuses on TRH—the variation in individual responses to the same intervention—which is particularly salient in aging research. We first outline the evidence for TRH in aging populations, then review emerging statistical and computational methods (from frailty models to machine learning and causal inference approaches) designed to capture individual variability, and finally discuss current challenges and opportunities for translating these methods into personalized antiaging strategies.

**2.6.1. Introduction.** TRH refers to the phenomenon in which different individuals experience varying effects when treated with the same (antiaging) interventions (Loop et al. 2012, Zoh et al. 2023). For example, some older adults who follow caloric restriction or take metformin may receive consequent (not just coincident) improvements in metabolic health and longevity, whereas others may not receive any consequent benefits or may even experience adverse effects. Genetic factors, environment, lifestyle choices, and rate of aging define this variation (Beavers et al. 2022, Soukas et al. 2019). Given the highly heterogeneous nature of aging, designing universal treatments is challenging, underscoring the importance of personalized medicine (Ferrucci & Kuchel 2021, Perrie et al. 2012).

**2.6.2. Relevance of treatment response heterogeneity in aging research.** In general, older individuals exhibit more variability in health measures (or health heterogeneity) than their younger counterparts (Nguyen et al. 2021). Erickson et al. (2023) opine on the importance of addressing interindividual variability in exercise responses in older adults. The National Institutes of Health National Institute on Aging workshop report states that the (presumed) variability of exercise intervention effects in aging populations is caused by both intrinsic (e.g., genetics, biological age) and extrinsic (e.g., medications, comorbidities) factors (Erickson et al. 2023).

**Figure 3** (adapted from Erickson et al. 2023) highlights various factors that plausibly induce variability in exercise responses in older adults. These intrinsic (e.g., inflammation, epigenomics,

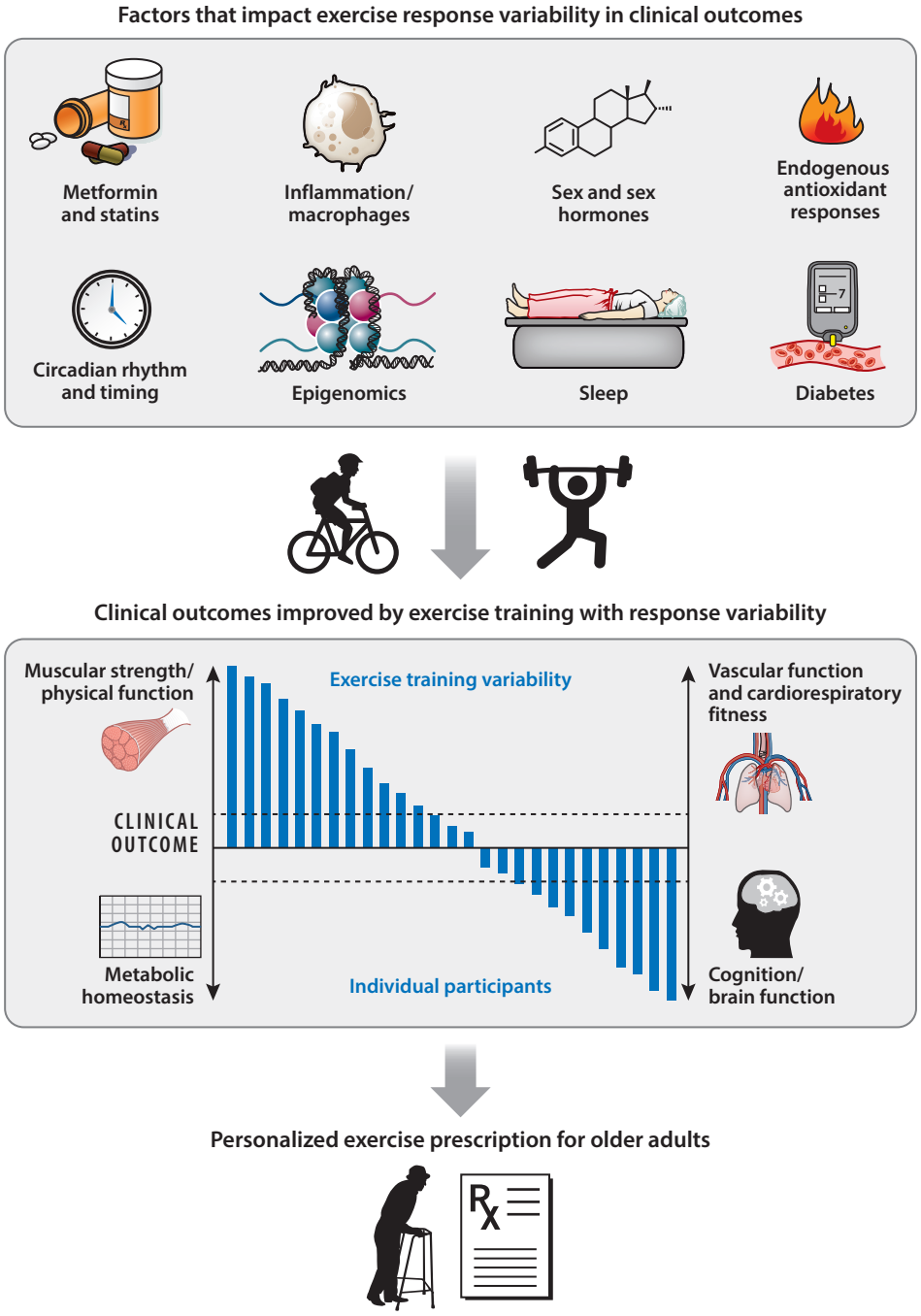


**Figure 3**

National Institute on Aging–identified factors contributing to variation in exercise response among older adults and a theoretical waterfall plot of individual outcomes. Adapted with permission from Erickson et al. (2023).

sex hormones, sleep) and extrinsic (e.g., medications such as metformin and statins, circadian rhythm, comorbidities like diabetes) variables presumptively modify exercise-induced physiological adaptations. The figure also shows the wide range of clinical outcomes speculated to be affected by TRH, including physical strength, cardiorespiratory fitness, metabolic homeostasis, and cognition. This variability suggests the limitations of a one-size-fits-all approach and the postulated value of personalized exercise prescriptions for older adults.

When designing personalized health care and precision medicine to improve older individuals' health, it is important to consider the conceivably large variability in individual responses (Argentieri et al. 2025, Castruita et al. 2022, Ferrucci & Kuchel 2021, Mitnitski et al. 2001). When analyzing aging-related events, traditional statistical methods, which sometimes treat subjects as a uniform population, may produce misleading or inaccurate results (Argentieri et al. 2025, Fittipaldi et al. 2024).

**2.6.3. Novel techniques to address treatment response heterogeneity.** To address TRH, advanced statistical methods that account for individual differences in aging trajectories and treatment responses are essential. Caswell (2014) developed a matrix approach to compute aging-related statistics in heterogeneous frailty models based on the gamma-Gompertz (G-G) model, which assumes mortality increases exponentially with age. "The marginal mortality rate for the G-G model is a sigmoid function of age" (Caswell 2014, p. 556) and is given by

$$\mu^*(t) = \frac{ae^{bt}}{1 + \frac{a\sigma^2}{b}(e^{bt} - 1)},$$

where

$$\mu^*(t) = \text{marginal mortality rate at age } t,$$

$$a = \text{baseline mortality rate,}$$

$$b = \text{rate of increase in mortality with age (Gompertz parameter), and}$$

$$\sigma^2 = \text{variance of initial frailty distribution.}$$

This model is useful for analyzing TRH by capturing individual variations in frailty and their impact on mortality and longevity.

Machine learning models are emerging as a powerful tool to address TRH. These computational approaches enable precision medicine by predicting individual variability and improve personalized care (Wilczok 2025). For example, one study used mixed-effects multilevel regression and machine learning techniques to create a health score and analyze how sociodemographic factors affect health patterns over time (Caballero et al. 2017). A recent review highlights the increasing role of supervised models, such as logistic regression, random forests, and XGBoost, in predicting health risks and guiding early interventions (Das & Dhillon 2023, Speiser et al. 2021).

**2.6.4. Pitfalls.** Despite the oft-stated importance of TRH to aging, unequivocal demonstrations are not always forthcoming (Kelley et al. 2023). Many published studies, interventional or observational, do not account for or discuss individual variability or heterogeneity in aging and instead emphasize average differences between groups. In cross-sectional data, failing to account for individual aging trajectories over time can be misleading as it may obscure true age-related changes and misrepresent the variability in the aging process across individuals (Ferrucci & Kuchel 2021). Furthermore, the kinds of machine learning algorithms discussed above can introduce potential pitfalls when applied to aging research, including hallucination (large language model–induced

## PERSONALIZING ANTIAGING INTERVENTIONS: ADDRESSING TREATMENT RESPONSE HETEROGENEITY

Treatment response heterogeneity (TRH)—the variation in individuals' responses to the same treatment—is a major challenge in aging research. The varying effectiveness of antiaging interventions among individuals due to genetic, environmental, and lifestyle factors greatly complicates the development of universally applicable treatments and has motivated a push toward personalized medicine approaches tailored to individual characteristics. This work is still new, but advanced statistical (e.g., mixed-effects models and the G-G framework) and machine learning (e.g., random forests and XGBoost) methods have been suggested as essential tools to address TRH. However, challenges with issues of model validation and data biases require careful study and management if we wish to use these methods to improve health outcomes in aging populations. Rigorous model validation and integration of high-dimensional data are essential for improving prediction accuracy.

false or misleading information presented as fact) and underrepresented population bias (where certain groups, such as older adults from minority backgrounds or those with unique health conditions, are not well-represented in data), leading to inaccurate predictions and health disparities. Other machine learning challenges include model overfitting; confounding and collinearity in aging datasets; a lack of explainability; high computational cost; ethical considerations, such as lack of transparency; data privacy; and accountability in decision-making. Additionally, rigorous clinical validation is crucial before applying machine learning–based biomarkers and deep aging clocks in real-world settings (Wilczok 2025) (see the sidebar titled Personalizing Antiaging Interventions: Addressing Treatment Response Heterogeneity).

**2.6.5. Refinements and improvements.** To improve the accuracy of aging research, several advanced approaches can be used. One such approach is the use of causal inference and counterfactual models such as causal forests (Jawadekar et al. 2023), Bayesian hierarchical models (Belasso et al. 2023, Meunier-Duperray et al. 2025), and structural equation modeling (SEM), which, in some designs and with some assumptions, can be employed to estimate individual treatment effects in aging interventions, thereby reducing bias, though they need careful validation with high-dimensional data. Alternatively, integrating multimodal data sources such as genomic, epigenetic, proteomic, and imaging biomarkers can provide a comprehensive view of aging trajectories (Lehallier et al. 2019). However, these methods require large, well-curated datasets and advanced computational resources, which can be a limitation for some studies (Fittipaldi et al. 2024). Future research should focus on model validation and generalizability, diverse datasets, and transparency in statistical assumptions to improve accuracy and generalizability (Argentieri et al. 2025, Fittipaldi et al. 2024).

## 3. METHODS WE SHOULD IMPROVE

In this section, we discuss statistical methods and issues that remain frequently misapplied and overlooked in aging research. These issues can compromise the scientific rigor and statistical validity of geroscience research. We focus on commonly used methods and analytical techniques that warrant reconsideration or refinement in aging research, such as stepwise regression, repeated-measures analysis of variance (ANOVA) and GLMs, failure to account for clustering and nesting, errors in testing for group differences, limitations in mediation and moderation analysis, and inappropriate use of nonlinear models. By addressing how these foundational problems arise in the context of aging studies and offering practical alternatives, we aim to support more robust and interpretable analyses across the field.

**SEM:** structural equation modeling

**ANOVA:** analysis of variance

### 3.1. Stepwise Regression: A Method We Should Reconsider

**ANCOVA:** analysis of covariance

Stepwise regression is commonly used in aging studies for variable selection to be included in the final regression model due to its simplicity (Scialfa & Games 1987). However, even though it involves multiple hypothesis tests, as at each step one or more tests are performed to determine which variables should be included, excluded, or retained, the final results of stepwise regression are typically reported as if only one test was performed. The final model may thus inflate type I error (i.e., the error one makes when one says that a null hypothesis is false, when in reality it is true) (Mundry & Nunn 2009). This form of model selection bias often remains uncorrected. Additionally, stepwise regression methods select variables based on predefined statistical criteria, rather than their theoretical or biological relevance, which is particularly problematic in aging research, where multicollinearity is present and results in misleading conclusions (Scialfa & Games 1987, Streiner 1994). Furthermore, stepwise regression ignores model uncertainty and tends to overfit, resulting in models that are nongeneralizable across datasets (Smith 2018). Stepwise regression is widely discouraged for hypothesis testing (Streiner 1994).

To address these issues, researchers have developed various modifications for selecting variables. These include regularization techniques, such as lasso regression, which reduce overfitting and improve model generalization (Tibshirani 2018); stopping rules that define when to stop adding or removing variables (Forsythe et al. 1973); the bootstrapping method, which involves variable selection across repeated resampled datasets (Austin & Tu 2004); data splitting into training and validation sets (Thall et al. 1997); and use of conditional instead of traditional *p*-values (Grechanovsky & Pinsker 1995). More recently, Zhu et al. (2020, p. 33117) proposed a polynomial algorithm that "exploits the idea of sequencing and splicing to reach a stable solution in finite steps when the sparsity level of the model is fixed but unknown."

Although these modifications may improve data-driven variable selection, challenges related to correct model specification, causal interpretation of predictor effects, and hypothesis testing for causal inference persist. Gerontological researchers should clearly define whether their goal is causal inference or prediction and should select variables based on subject-matter expertise, as "the validity of causal inferences cannot be exclusively data-driven" (Hernán et al. 2019, p. 49). This hypothesis-driven approach is supported by DNAm studies, in which selection of biologically relevant variables in relation to clinical biomarkers (e.g., inflammatory markers) or exposures (e.g., smoking pack years) resulted in more generalizable, interpretable, and accurate biological clocks (Levine et al. 2018, Lu et al. 2022). Additionally, directed acyclic graphs, which are developed based on domain expertise, theoretical frameworks, and assumptions about causal relationships among variables, offer a robust framework for determining which variables to include to more plausibly estimate a causal effect (Tennant et al. 2020).

### 3.2. Beyond Repeated-Measures ANOVA: Modern Techniques for Complex Data in Aging Research

Aging research often seeks to relate changes in an aging-related outcome to an observed or experimental condition. The analysis of complex data structures such as repeated measures, multiple experimental groups, and covariate adjustment demands sophisticated analytical approaches. Analysis of variance/covariance (ANOVA/ANCOVA) modeling approaches are widely used for complex data because they can readily be adjusted for covariates and can handle multiple-group comparisons.

However, ANOVA models are less useful when researchers are collecting data with noncontinuous outcomes (e.g., the number of medications taken daily, word recall, Mini-Mental State Examination scores). Treating such data as continuous violates key model assumptions and can

lead to biased results (Schober & Vetter 2021). GLMs provide a more appropriate alternative by extending the linear model to accommodate various noncontinuous outcome distributions (e.g., Poisson, binomial) and link functions (e.g., log, logit) (Gelman 2005, Maxwell et al. 2018), enabling appropriate modeling of binary data, counts, and frequencies.

**rANOVA:** repeated-measures ANOVA

**LMM:** linear mixed model

Repeated-measures data, which are ubiquitous in geroscience, introduce additional challenges. Traditional statistical methods (e.g., linear regression, ANOVA) assume independence of observations, an assumption that is violated when measurements are taken repeatedly from the same experimental unit. When outcome data are continuous, accounting for repeated measures is often done using the repeated-measures ANOVA (rANOVA), but this approach has important caveats. rANOVA assumes that the variation in the differences between all pairs of related groups in the data is roughly equal (i.e., sphericity), which, when violated, can inflate type I error (Huynh & Feldt 1976). Corrections such as suitable adjustment for degrees of freedom (such as the Greenhouse–Geiser or Huynh–Feldt correction) are suggested when the assumption of sphericity is violated (Haverkamp & Beauducel 2017). Another common issue is that when data are missing, rANOVA employs listwise deletion, which removes all subjects with incomplete data, resulting in a potentially biased nonrandom sample and subsequently biased results and lower statistical power (Ma et al. 2012). A more robust alternative is to use linear mixed models (LMMs), which are a more general case of the rANOVA model family (Gelman 2005). LMMs accommodate correlated data through random effects and handle missing data more flexibly without requiring listwise deletion. Importantly, these models do not share the rANOVA assumption of sphericity (de Melo et al. 2022, Krueger & Tian 2004).

In studies with noncontinuous repeated measures, generalized LMMs are often the best solution. They are much like LMMs as they allow researchers to model more complex repeated-measures data by incorporating random effects in addition to fixed effects. Like GLMs, they also allow for specifying many alternative distributions and link functions, which allows for modeling a wide variety of outcome measures. Marginal models, such as those based on generalized estimating equations, are popular alternatives for analyzing repeated-measures data, which, although less flexible to handling missing data, are robust to misspecification and can be used for nonnormally distributed outcome variables (Hardin & Hilbe 2012).

### 3.3. The Importance of Considering Clustering and Nesting in Geroscience Research

Clustering and nesting arise when study units share environments, treatments, or other contextual factors that induce correlation among observations, thus violating the independence assumption of many traditional statistical methods, such as linear regression and ANOVA, that assume independence among observations. Clustering refers to situations in which individuals are grouped together, inducing statistical dependence among their observations (Jamshidi-Naeini et al. 2022). Nesting occurs when a treatment is applied to all individuals in a cluster, wherein the experimental unit becomes the cluster instead of the individual (Jamshidi-Naeini et al. 2022).

Clustering and nesting are particularly prominent in human and animal aging research (Chusyd et al. 2022). In human research, a cluster-randomized controlled trial may allocate treatments at the facility or clinic level (Brown et al. 2015). In animal studies, rodents are often group-housed, creating correlated observations within each cluster because of the shared housing environment or diet (Chusyd et al. 2022, Klatt et al. 2023, Landes 2024, Luciano & Churchill 2025). Treatments may further be nested within clusters, such as a cage of mice allocated to receive a high-fat diet. Even in studies of model organisms, like worms or cellular systems, multiple measurements

recorded from the same environment (e.g., agar plate, agar medium, bacterial lawn) or individual (e.g., multiple cells, tissues, or organs taken from a single animal) introduce dependencies (Lazic et al. 2018).

Failing to account for clustering and nesting can result in invalid analyses. Ignoring these data dependencies can result in underestimated variance, inflated type I error rates, and overstated statistical significance by increasing false-positive rates. For instance, analyzing cognitive outcomes in 40 nursing home residents clustered within 2 facilities as if there were 40 independent observations overlooks facility-level variance and can lead to overstated statistical significance. If all residents clustered in each nursing home received the same treatment, the residents are also nested within that unit. In this case, the degrees of freedom also change to account for both the individuals and the units of allocation. Simulations based on intraclass correlation from murine studies of lifespan (Luciano & Churchill 2025), and pilot analyses of the National Institute on Aging's Intervention Testing Program data (Parker et al. 2025), showed that failing to account for clustering and nesting effects can lead to underestimated power and potentially biased results. Nesting can have an even more substantial impact, especially when there are very few clusters. At the extreme, one cluster per treatment arm reduces the degrees of freedom to test for treatment effects to zero, resulting in the treatment effect being completely confounded by the cluster-level effects.

To prevent errors related to clustering and nesting, researchers should consider the following. (*a*) Identify the unit of allocation: If interventions are delivered to a group (such as a cage or facility), consider the group as the unit for power calculations and analyses. (*b*) Adjust power calculations: Incorporate the intraclass correlation when estimating the number of clusters (and individuals per cluster) needed to maintain adequate power (Brown et al. 2015). (*c*) Use an appropriate statistical model to account for clustering: Mixed-effects models (multilevel models) and generalized estimating equations account for correlated data. They incorporate random effects or robust standard errors to accommodate clustering, ensuring accurate standard errors and *p*-values. (*d*) Account for the correct degrees of freedom: When using nested designs, the effective number of independent units is the number of clusters, not the number of individuals. Using methods such as Kenward–Roger approximations can help adjust degrees of freedom (Kenward & Roger 2009).

### 3.4. Errors in Testing for Group Differences: The Need for Formal Tests Between Groups

Researchers sometimes draw conclusions about within-group comparisons when between-group comparisons are of primary interest. Consider the following two examples:

- A randomized trial of a form of exercise called Nordic walking versus usual overground walking to test improvement of gait speed in geriatric rehabilitation (Figueiredo et al. 2013)
- An observational study assessing differential age acceleration defined by brain cortical thinning between males and females during the COVID-19 pandemic (Corrigan et al. 2024)

In both examples, two types of comparisons could be reported: within-group and between-group. For within-group comparisons, the first example could test whether there were significant changes in gait speed from baseline to follow-up within each assignment group. The second example could compare whether the cortical thinning deviated significantly from normative values within each sex.

Yet, in both cases, there is a particular interest in between-group comparisons. In the first case, the purpose of the randomized trial was to test whether assignment to Nordic walking is

## DO NOT COMPARE SIGNIFICANCE—TEST THE CONTRAST

When between-group conclusions are of interest, a formal between-group test should be conducted. The CONSORT (Consolidated Standards of Reporting Trials) guidelines for randomized trials suggest that "Confidence intervals should be presented for the contrast between groups. A common error is the presentation of separate confidence intervals for the outcome in each group rather than for the treatment effect" (Moher et al. 2010, p. 44). The between-group contrast, rather than the significance of the two groups separately, should be the basis for conclusions from the study. There are other situations and designs in which the DINS (differences in nominal significance) error can occur (see Allison et al. 2016; Bland & Altman 2011, 2015; Maney & Rich-Edwards 2023). Each situation may be unique in how best to compare between groups. If it is unclear what approach will provide valid comparisons for between-group comparisons of interest, collaborating with a professional biostatistician is recommended.

better than overground walking. In the second example, the emphasis was testing whether cortical thinning differed between sexes.

A common mistake is to emphasize within-group comparisons at the expense of formal between-group comparisons when the between-group comparisons are of interest. In the Nordic walking example, gait speed changed significantly from baseline within the Nordic walking group but not within the overground walking group. Yet, when the two groups were directly compared, there was no significant between-group difference (Allison et al. 2015). Similarly, in the cortical thinning example, the article reported significant thinning in 30 of 68 brain regions within females and only 2 regions within males after false discovery rate correction. Yet, only 1 region was significantly different between males and females after false discovery rate correction (Brown et al. 2025).

Making conclusions based on discordant significance without expressly testing whether the change in one group is different from the change in the other group has been referred to as the differences in nominal significance (DINS) error (George et al. 2016). The DINS error can lead to type I error rates of up to 50% when comparing two groups, and in some circumstances up to 95%; however, the target type I error rate is often only 5% (i.e., $P < 0.05$) (Bland & Altman 2011). Emphasizing within-group comparisons also often undermines inferences of interest. The purpose of a control or comparator group in randomized trials, for instance, is to account for other factors unrelated to the intervention itself (e.g., placebo effects, aging itself, social trends) that may affect the outcome. Comparing only within a treatment group loses the comparator group's information in making inferences about the treatment (see the sidebar titled Do Not Compare Significance—Test the Contrast).

### 3.5. Mediation and Moderation with Structural Equation Modeling

Mediation and moderation analyses are regression-based techniques used to evaluate the effects of third variables on the association between an independent and a dependent variable (Baron & Kenny 1986). Mediation analysis explains how or why the independent variable is related to the dependent variable, while moderation analysis helps to identify when and how the relationships exist. **Figure 4** shows their conceptual differences.

**DINS:** differences in nominal significance

Specifically, mediation analysis quantifies the indirect effects of an independent variable ($X$) on a dependent variable ($Y$) through a mediator ($M$). For example, in older adults, the negative relationship between depression and cognitive function can be mediated by social activities and activities of daily living (Fan et al. 2024). The following is the equation typically used to

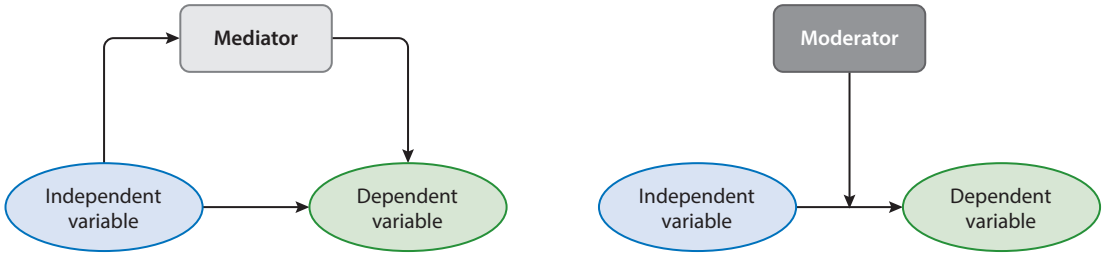


**Figure 4**

Conceptual diagrams of mediation and moderation models (models based on Hayes 2013).

test mediation:

$$\text{Mediator model: } M = \alpha_0 + \alpha X + \varepsilon_M,$$

$$\text{Outcome model: } Y = \beta_0 + c'X + \beta M + \varepsilon_Y,$$

where $\alpha_0$ and $\beta_0$ are intercepts and $\alpha$, $c'$, and $\beta$ are path coefficients. Errors $\varepsilon_M$ and $\varepsilon_Y$ are uncorrelated with $X$ and $M$. The corresponding effects typically reported for mediation are as follows:

$$\text{Indirect (mediated) effect: } \alpha\beta,$$

$$\text{Direct effect: } c',$$

$$\text{Total effect: } c = c' + \alpha\beta.$$

On the other hand, moderation analysis examines interactive effects of the independent variable ($X$) and a moderator ($W$) on the dependent variable ($Y$). The relationship between the independent and the dependent variable varies based on the level of the moderator. For example, sex can moderate the relationship between aging and various types of function (Gur & Gur 2002). The following is the equation typically used for moderation analysis:

$$Y = \beta_0 + \beta_1 X + \beta_2 W + \beta_3 (XW) + \varepsilon,$$

where $\beta_3$ captures how the slope of $X \rightarrow Y$ changes per one-unit change in $W$. Error $\varepsilon$ is uncorrelated with $X$ and $W$. This can plot simple slopes at meaningful values of $W$ (e.g., $\pm 1$ standard deviation) to visualize the interaction.

Traditional mediation and moderation analyses rely on assumptions of independence, linearity, and homogeneity of effects (Baron & Kenny 1986), which are often hard to meet when studying the complex nature of aging processes (e.g., nonlinear changes), thus highlighting the importance of examining such phenomena with advanced statistical techniques (e.g., Cohen 2016, Shen et al. 2024).

SEM, with its extension of generalized linear SEM (Rabe-Hesketh et al. 2004), allows for simultaneously testing multiple paths and for incorporating latent variables, defined as unobserved constructs inferred by observed measurement. Estimating shared variance across observed indicators partitions item-specific error, thereby allowing estimation of the construct free from measurement error (Cai 2012, Christ et al. 2014, Rush et al. 2019). In aging research, mental, cognitive, and behavioral concepts are usually latent variables that are indirectly inferred (Folstein et al. 1975, Prince et al. 2008). **Figure 5** shows a conceptual model of SEM with two latent variables.

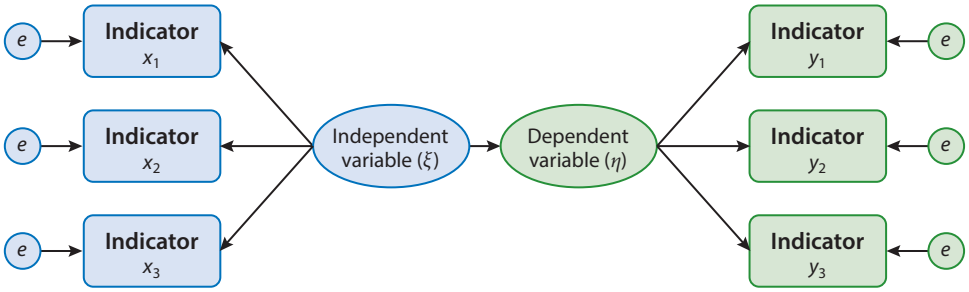


**Figure 5**

Conceptual model of structural equation modeling with two latent variables.

The following is the equation typically used for SEM analysis:

$$\eta = B\eta + \Gamma\xi + \zeta, \text{ with}$$

$$y = \Lambda_y\eta + \varepsilon \text{ (indicators of } \eta),$$

$$x = \Lambda_x\xi + \delta \text{ (indicators of } \xi),$$

$$\eta = \text{ exogenous variables,}$$

$$\xi = \text{ endogenous variables,}$$

$$B = \text{ regression coefficients among endogenous latents,}$$

$$\Gamma = \text{ path coefficients from exogenous to endogenous latents,}$$

$$\zeta = \text{ structural disturbances, and}$$

$$\Lambda = \text{ factor-loading matrices,}$$

where errors $\varepsilon$ and $\delta$ are uncorrelated with factors.

However, SEMs still have other limitations, such as the assumption of multivariate normality and the issue of low statistical power (Tomarken & Waller 2005). Researchers can use robust estimation (Yuan & Bentler 2000), bootstrapping techniques (Bollen & Stine 1992, Cheung & Lau 2008), and Bayesian approaches (Muthén & Asparouhov 2012) to address limitations.

Ultimately, the study of causal relationships requires rigorous designs (e.g., longitudinal studies, experiments, Mendelian randomization) (Carter et al. 2021, Tomarken & Waller 2005). Temporality is crucial but is often overlooked in mediation analysis. In addition, specifying not only the time lags among the variables but also the durations and rates of change is fundamental to a rigorous study design (Wang et al. 2017).

## 3.6. Nonlinear Models in Aging Research

Linear and nonlinear models are used to assess relationships between a variable of interest and explanatory variables. Linear models are less flexible but are often preferred for their interpretability and suitability for hypothesis testing. Nonlinear models, which are not linear in parameters, are applied when linear models are inadequate and not necessarily because of curvature (Jarantow et al. 2023).

A model that is linear in parameters, i.e., in which the exponent of the $\beta$s is 1, is termed a linear model and is represented as

$$f(x) = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n.$$

**DICNAP:** data-driven identification and classification of nonlinear aging patterns

A nonlinear model is a model where the $\beta$s can have exponents other than 1 or can be in nonlinear functional forms (e.g., exponential, logarithmic, sine functions) that are not expressible as a linear combination of $\beta$s (Konishi 2014).

Aging may involve nonlinear biological shifts rather than gradual decline. Assuming linear age–outcome relationships can mislead analyses. Initial categorical modeling and visualization (e.g., log odds plots) can reveal patterns. When nonlinear trends emerge, methods like splines or segmented regression improve accuracy, aiding precision aging strategies (Massa et al. 2025, Shen et al. 2024). The data-driven identification and classification of nonlinear aging patterns (DICNAP) method has found that DNAm changes with age in a nonlinear way. Although DICNAP is not a nonlinear model, it uses data and statistics to spot these patterns, helping us better understand how aging works (Okada et al. 2023). Furthermore, nonlinear models can be used to investigate health trajectories over time. For instance, Chen et al. (2016) found that semiparametric and graphical models were useful in modeling late-life cognitive changes.

Nonlinear models pose challenges in model selection, interpretation, computational complexity, and overfitting, where models capturing noise rather than true patterns can reduce validity (Bilger & Manning 2015, Massa et al. 2025). Also, confounding remains a concern in both linear and nonlinear models, especially when age is modeled differently (continuous versus categorical) (Brenner & Blettner 1997).

Several approaches have been developed to tackle the challenges. DICNAP, for example, applies functional data analysis to identify biomarkers and determine aging patterns (Okada et al. 2023). Other methods, such as spline-based models, piecewise LMMs, and sigmoidal models, provide interpretable nonlinear parameters for deeper insights (Capuano & Wagner 2023, Wagner et al. 2024). While these established and emerging methods enhance the precision and capacity to identify complex aging patterns, they do have some limitations, such as computational complexity and the need for larger datasets. Researchers must balance these factors to effectively apply these advanced models in aging research (Offermann-Van Heek et al. 2019, Thompson 2022). Given that nonlinear relationships are more common than linear ones, these methods offer substantial insights into the aging process and health outcomes; however, their effective implementation requires the precise selection and validation of models to prevent common pitfalls.

## 4. CONCLUSION

Proper use of statistical methods is central to advancing aging research and improving our understanding of longevity, disease progression, and the effect of life-extending interventions on lifespan and health span. Emerging techniques for examining compression of morbidity, estimating maximum lifespan, addressing immortal time bias, deriving molecular aging clocks, and addressing TRH offer notable avenues for future research in aging. While these methods offer powerful tools to advance the field, they also present conceptual and methodological challenges that require further validation and refinement. The complex, longitudinal, and heterogeneous datasets arising from aging studies further demand careful and rigorous application of different statistical methods. Use of commonly applied techniques such as stepwise regression, inappropriate use of ANOVA and other methods under the GLM family, failure to address clustering and nesting, erroneous use of techniques for testing group differences, and use of oversimplified linear models introduce bias and inaccuracies, thus compromising research findings. Appropriate use of advanced statistical

techniques, including mixed methods, nonlinear models, mediation and moderation analyses, and other causal inference techniques, can provide robust and accessible alternatives. **Table 2** summarizes the key issues pertinent to aging research for each of these statistical techniques, along with recommended suggestions, improvements, and alternatives. We urge researchers in the field of aging to discontinue using outdated and invalid methods, refine traditional approaches, and embrace emerging and innovative methods tailored to the unique challenges of aging research. Adopting best practices, integrating advanced statistical methods, and ensuring transparency

**Table 2 A summary of the statistical techniques with major issues pertinent to aging research and key takeaways discussed in this review**

| Topic (statistical technique) | Issue(s) pertinent in aging and geroscience research | Key takeaways |
|---|---|---|
| Compression of morbidity | There exists no standard statistical method to integrate longitudinal health span and lifespan measurements to examine compression of morbidity. | A few newly developed methods, such as the steepness of survival curve, use of FAMY, and the comparison of the rates of declines between survival and vitality curves, should be rigorously tested and validated. |
| Tests for maximum lifespan | Aging researchers might also wish to compare how long the longest-lived animals in different groups survive, in addition to the average or median survival times. | The Wang–Allison and Gao–Allison methods extend comparisons of the lifespans of animals in different groups beyond the mean or median, enabling explicit comparison of the longest-lived animals in different groups. |
| Maximal achievable lifespan | Different species are known to survive different lengths of time, even under ideal conditions. | While statistically rigorous tests for maximum lifespan exist, they are based on quantiles and not true maximal values for each species. Determining and comparing maximal lifespan values is difficult, and more work is needed. |
| Immortal time bias | Older adults often receive treatments based on survival, which introduces bias when time periods before treatment initiation are misclassified as exposure. This can lead to inaccurate estimates of treatment effects, especially in studies where survival and treatment eligibility are intertwined. | Immortal time bias can be mitigated by proper specification of selection and assignment criteria within a target trial emulation framework. |
| Molecular aging clocks | Current molecular aging clocks do not deliver personally quantifiable age-acceleration estimates, preventing successful commercial or clinical application. Confounders like cell-type heterogeneity are not part of a clock's modeling and training, which can lead to irreproducibility or misinterpretation. | The effective development and application of clocks require addressing key challenges, particularly the probabilistic quantification of age acceleration and improved modeling of confounders when training and validating clocks. |
| Treatment response heterogeneity | Most studies do not account for or discuss individual variability in aging research and instead focus on average differences between groups. Failing to account for the individual aging trajectories over time can be misleading as it may obscure true age-related changes and misrepresent the variability in aging process across different groups. | Future research should focus on accounting for individual variability, model validation, and transparency in statistical assumptions to improve accuracy and generalizability in aging research. |

(*Continued*)

**Table 2** (*Continued*)

| Topic (statistical technique) | Issue(s) pertinent in aging and geroscience research | Key takeaways |
|---|---|---|
| Stepwise regression | Stepwise regression, though widely used in aging research, poses significant challenges by inflating type I error rates due to multiple hypothesis tests and selecting variables based on statistical criteria rather than theoretical relevance, leading to misleading conclusions. It ignores model uncertainty, often results in overfitting, and produces nongeneralizable models. | Gerontological research should prioritize hypothesis-driven variable selection informed by domain expertise to enhance model generalizability, interpretability, and validity, rather than relying solely on data-driven methods. |
| ANOVA and GLMs | Aging data are often complex and involve repeated measures, so standard ANOVA-like approaches may not apply. | GLMs are useful for noncontinuous outcomes, and mixed models (linear and generalized linear) are powerful tools for analyzing longitudinal data. |
| Clustering and nesting | Clustering and nesting are common in aging research, as individuals or animals often share environments, treatments, or housing, leading to correlated observations that violate the independence assumption of many statistical tests. Failing to account for clustering can result in underestimated variance, inflated type I error rates, and biased conclusions. | To prevent errors from clustering and nesting in aging research, it is crucial to identify the correct unit of allocation, incorporate intraclass correlation in power analyses, use appropriate statistical models such as mixed-effects models or generalized estimating equations to account for correlated data, and properly adjust for degrees of freedom. |
| Errors in testing for group differences | Estimates are sometimes calculated within groups rather than formally testing between groups. This error is common, occurs across fields and different types of group comparisons, and is sometimes difficult to recognize. | For any conclusions about differences between groups, formal tests should be conducted between groups. |
| Mediation and moderation | Mediation and moderation techniques are essential for understanding complex relationships (e.g., how and when) between outcomes and exposures in aging research. | Researchers should consider moderation and mediation analyses for complex relationships and theory testing. With SEM, it is possible to study complex effects more accurately, particularly in the presence of latent variables. |
| Nonlinear models | Nonlinear models in aging research pose significant issues such as misconceptions and misapplications, overfitting, confounding (age as confounder), model misrepresentation, and validation checks. | Nonlinear models offer substantial insights into the aging process and health outcomes. However, their effective implementation requires the precise selection and validation of models to prevent common pitfalls. |

Abbreviations: ANOVA, analysis of variance; FAMY, frailty-adjusted mouse years; GLM, generalized linear model; SEM, structural equation modeling.

in statistical application and reporting will improve the rigor and reproducibility of aging research.

## DISCLOSURE STATEMENT

Dr. Allison holds equity in one company (Zero Longevity Science), and he and his institutions (Indiana University and the Indiana University Foundation) have received grants, contracts, in-kind donations, and consulting fees from numerous governmental agencies, nonprofit organizations, and for-profit organizations, including litigators and dietary supplement, food,

pharmaceutical, health care, research, medical device, and publishing companies. However, none funded or are directly relevant to this article or the issues addressed herein.

## ACKNOWLEDGMENTS

We thank Professor A. Miguel Hernán for providing the initial draft of the immortal time bias section, Professor John D. Sorkin and Dr. Scott W. Keith for critical review and feedback, and Jennifer Holmes for proofreading the manuscript. This work was supported in part by the National Science Foundation award FAIN 2318478 and by the National Institutes of Health under award numbers R25DK099080, P30AG050886, U24AG056053, R25HL124208, R01GM152543, R25HL124208, UL1TR003107. The assertions expressed are those of the authors and not necessarily those of the National Science Foundation, the National Institutes of Health, or any other organization. Some authors utilized generative artificial intelligence to refine certain sentences for clarity and conciseness in specific sections.

## LITERATURE CITED

Alexander LK, Lopes B, Ricchetti-Masterson K, Yeatts KB. 2018. *Common statistical tests and applications in epidemiological literature*. Course Notes, UNC Gillins School of Global Public Health. **https://sph.unc.edu/wp-content/uploads/sites/112/2015/07/nciph_ERIC2-rev.pdf**

Allison DB, Brown AW, George BJ, Kaiser KA. 2016. Reproducibility: a tragedy of errors. *Nature* 530(7588):27–29

Allison DB, Williams MS, Hand GA, Jakicic JM, Fontaine KR. 2015. Conclusion of "Nordic walking for geriatric rehabilitation: a randomized pilot trial" is based on faulty statistical analysis and is inaccurate. *Disabil. Rehabil.* 37(18):1692–93

Argentieri MA, Amin N, Nevado-Holgado AJ, Sproviero W, Collister JA, et al. 2025. Integrating the environmental and genetic architectures of aging and mortality. *Nat. Med.* 31:1016–25

Atwal S. 2024. Oldest person ever: 122-year-old Jeanne Calment's extraordinary life. *Guinness World Records News Blog*, Feb. 21. **https://www.guinnessworldrecords.com/news/2024/2/oldest-person-ever-122-year-old-jeanne-calments-extraordinary-life-765016**

Austad SN. 2022. *Methuselah's Zoo: What Nature Can Teach Us About Living Longer, Healthier Lives*. MIT Press

Austad SN, Fischer KE. 1991. Mammalian aging, metabolism, and ecology: evidence from the bats and marsupials. *J. Gerontol.* 46(2):B47–53

Austin PC, Tu JV. 2004. Bootstrap methods for developing predictive models. *Am. Stat.* 58(2):131–37

Baron RM, Kenny DA. 1986. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. *J. Pers. Soc. Psychol.* 51(6):1173–82

Beavers DP, Hsieh KL, Kitzman DW, Kritchevsky SB, Messier SP, et al. 2022. Estimating heterogeneity of physical function treatment response to caloric restriction among older adults with obesity. *PLOS ONE* 17(5):e0267779

Belasso CJ, Cai Z, Bezgin G, Pascoal T, Stevenson J, et al. 2023. Bayesian workflow for the investigation of hierarchical classification models from tau-PET and structural MRI data across the Alzheimer's disease spectrum. *Front. Aging Neurosci.* 15:1225816

Beltrán-Sánchez H, Jiménez MP, Subramanian S. 2016. Assessing morbidity compression in two cohorts from the Health and Retirement Study. *J. Epidemiol. Commun. Health* 70(10):1011–16

Belzile LR, Davison AC, Gampe J, Rootzén H, Zholud D. 2022. Is there a cap on longevity? A statistical review. *Annu. Rev. Stat. Appl.* 9:21–45

Bilger M, Manning WG. 2015. Measuring overfitting in nonlinear models: a new method and an application to health expenditures. *Health Econ.* 24(1):75–85

Bland JM, Altman DG. 2011. Comparisons against baseline within randomised groups are often used and can be highly misleading. *Trials* 12(1):264

Bland JM, Altman DG. 2015. 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.* 102(5):991–94

Bollen KA, Stine RA. 1992. Bootstrapping goodness-of-fit measures in structural equation models. *Sociol. Methods Res.* 21(2):205–29

Brenner H, Blettner M. 1997. Controlling for continuous confounders in epidemiologic research. *Epidemiology* 8(4):429–34

Brown AW, Chung S, Koscik T, Vorland CJ, Maney DL. 2025. No compelling evidence of sex differences in brain maturation during COVID-19 lockdowns when the sexes are compared statistically. *PNAS* 122(14):e2421462122

Brown AW, Li P, Bohan Brown MM, Kaiser KA, Keith SW, et al. 2015. Best (but oft-forgotten) practices: designing, analyzing, and reporting cluster randomized controlled trials. *Am. J. Clin. Nutr.* 102(2):241–48

Caballero FF, Soulis G, Engchuan W, Sánchez-Niubó A, Arndt H, et al. 2017. Advanced analytical methodologies for measuring healthy ageing and its determinants, using factor analysis and machine learning techniques: the ATHLOS project. *Sci. Rep.* 7(1):43955

Cai L. 2012. Latent variable modeling. *Shanghai Arch. Psychiatry* 24(2):118–20

Capuano AW, Wagner M. 2023. nlive: An R package to facilitate the application of the sigmoidal and random changepoint mixed models. *BMC Med. Res. Methodol.* 23(1):257

Carter AR, Sanderson E, Hammerton G, Richmond RC, Davey Smith G, et al. 2021. Mendelian randomisation for mediation analysis: current methods and challenges for implementation. *Eur. J. Epidemiol.* 36(5):465–78

Castruita PA, Piña-Escudero SD, Rentería ME, Yokoyama JS. 2022. Genetic, social, and lifestyle drivers of healthy aging and longevity. *Curr. Genet. Med. Rep.* 10(3):25–34

Caswell H. 2014. A matrix approach to the statistics of longevity in heterogeneous frailty models. *Demogr. Res.* 31:553–92

Chen H, Zhao B, Cao G, Proges EC, O'Shea A, et al. 2016. Statistical approaches for the study of cognitive and brain aging. *Front. Aging Neurosci.* 8:176

Cheung GW, Lau RS. 2008. Testing mediation and suppression effects of latent variables: bootstrapping with structural equation models. *Organ. Res. Methods* 11(2):296–325

Christ SL, Lee DJ, Lam BL, Zheng DD. 2014. Structural equation modeling: a framework for ocular and other medical sciences research. *Ophthalmic Epidemiol.* 21(1):1–13

Chusyd DE, Austad SN, Dickinson SL, Ejima K, Gadbury GL, et al. 2022. Randomization, design and analysis for interdependency in aging research: no person or mouse is an island. *Nat. Aging* 2(12):1101–11

Cohen AA. 2016. Complex systems dynamics in aging: new evidence, continuing questions. *Biogerontology* 17(1):205–20

Corrigan NM, Rokem A, Kuhl PK. 2024. COVID-19 lockdown effects on adolescent brain structure suggest accelerated maturation that is more pronounced in females than in males. *PNAS* 121(38):e2403200121

Cox DR, Oaks D. 1984. *Analysis of Survival Data*. Chapman and Hall/CRC

Dabrowski JK, Yang EJ, Crofts SJC, Hillary RF, Simpson DJ, et al. 2024. Probabilistic inference of epigenetic age acceleration from cellular dynamics. *Nat. Aging* 4(10):1493–507

Das A, Dhillon P. 2023. Application of machine learning in measurement of ageing and geriatric diseases: a systematic review. *BMC Geriatr.* 23(1):841

de Lima Camillo LP, Lapierre LR, Singh R. 2022. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *npj Aging* 8(1):4

De Magalhães JP, Costa J. 2009. A database of vertebrate longevity records and their relation to other life-history traits. *J. Evol. Biol.* 22(8):1770–74

de Melo MB, Daldegan-Bueno D, Menezes Oliveira MG, de Souza AL. 2022. Beyond ANOVA and MANOVA for repeated measures: advantages of generalized estimated equations and generalized linear mixed models and its use in neuroscience research. *Eur. J. Neurosci.* 56(12):6089–98

Di Francesco A, Deighan AG, Litichevskiy L, Chen Z, Luciano A, et al. 2024. Dietary restriction impacts health and lifespan of genetically diverse mice. *Nature* 634(8034):684–92

di Lego V. 2021. Health expectancy indicators: What do they measure? *Cad. Saúde Colet.* 29:115–29

Dickerman BA, García-Albéniz X, Logan RW, Denaxas S, Hernán MA. 2019. Avoidable flaws in observational analyses: an application to statins and cancer. *Nat. Med* 25(10):1601–6

Erickson ML, Allen JM, Beavers DP, Collins LM, Davidson KW, et al. 2023. Understanding heterogeneity of responses to, and optimizing clinical efficacy of, exercise training in older adults: NIH NIA workshop summary. *Geroscience* 45(1):569–89

Fan P, Li H, Xu H, Rong C. 2024. A chain mediation model reveals the association between depression and cognitive function in the elderly. *Sci. Rep.* 14(1):31375

Ferrucci L, Kuchel GA. 2021. Heterogeneity of aging: individual risk factors, mechanisms, patient priorities, and outcomes. *J. Am. Geriatr. Soc.* 69(3):610–12

Figueiredo S, Finch L, Mai J, Ahmed S, Huang A, Mayo NE. 2013. Nordic walking for geriatric rehabilitation: a randomized pilot trial. *Disabil. Rehabil.* 35(12):968–75

Fittipaldi S, Legaz A, Maito M, Hernandez H, Altschuler F, et al. 2024. Heterogeneous factors influence social cognition across diverse settings in brain health and age-related diseases. *Nat. Ment. Health* 2(1):63–75

Folstein MF, Folstein SE, McHugh PR. 1975. "Mini-Mental State": a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12(3):189–98

Forsythe AB, Engelman L, Jennrich R, May PRA. 1973. A stopping rule for variable selection in multiple regression. *J. Am. Stat. Assoc.* 68(341):75–77

Friedman JH, Hastie T, Tibshirani R. 2010. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* 33(1):1–22

Fries JF. 1983. The compression of morbidity. *Milbank Q.* 83(4):801–23

Fries JF. 1989. The compression of morbidity: near or far? *Milbank Q.* 67(2):208–32

Gail MH. 1972. Does cardiac transplantation prolong life? A reassessment. *Ann. Intern. Med.* 76(5):815–17

Gao G, Wan W, Zhang S, Redden DT, Allison DB. 2008. Testing for differences in distribution tails to test for differences in 'maximum' lifespan. *BMC Med. Res. Methodol.* 8(1):49

Gavrilov LA, Gavrilova NS. 1991. *The Biology of Life Span: A Quantitative Approach*. Harwood Academic

Gelman A. 2005. Analysis of variance—why it is more important than ever. *Ann. Stat.* 33(1):1–53

George BJ, Beasley TM, Brown AW, Dawson J, Dimova R, et al. 2016. Common scientific and statistical errors in obesity research. *Obesity* 24(4):781–90

Gouveia M, Raposo P. 2019. Aging and the compression of disability in Portugal. *Popul. Dev. Rev.* 45(2):401–18

Graham P, Blakely T, Davis P, Sporle A, Pearce N. 2004. Compression, expansion, or dynamic equilibrium? The evolution of health expectancy in New Zealand. *J. Epidemiol. Commun. Health* 58(8):659–66

Grechanovsky E, Pinsker I. 1995. Conditional *p*-values for the F-statistic in a forward selection procedure. *Comput. Stat. Data Anal.* 20(3):239–63

Gruenberg EM. 1977. The failures of success. *Milbank Q.* 55(1):3–24

Guo X, Teschendorff AE. 2025. Epigenetic clocks and inflammaging: pitfalls caused by ignoring cell-type heterogeneity. *Geroscience* 47(3):2707–19

Gur RE, Gur RC. 2002. Gender differences in aging: cognition, emotions, and neuroimaging studies. *Dialogues Clin. Neurosci.* 4(2):197–210

Hardin JW, Hilbe JM. 2012. *Generalized Estimating Equations*. Chapman and Hall/CRC

Harrison DE, Strong R, Allison DB, Ames BN, Astle CM, et al. 2014. Acarbose, 17-α-estradiol, and nordihydroguaiaretic acid extend mouse lifespan preferentially in males. *Aging Cell* 13(2):273–82

Haverkamp N, Beauducel A. 2017. Violation of the sphericity assumption and its effect on type-1 error rates in repeated measures ANOVA and multi-level linear models (MLM). *Front. Psychol.* 8:1841

Hayes AF. 2013. *Introduction to Mediation, Moderation, and Conditional Process Analysis: Methodology in the Social Sciences*. Guilford Press

Hayes JP, Speakman JR, Racey PA. 1992. Sampling bias in respirometry. *Physiol. Zool.* 65(3):604–19

Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, et al. 2008. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology* 19(6):766–79

Hernán MA, Brumback B, Robins JM. 2000. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology* 11(5):561–70

Hernán MA, Hsu J, Healy B. 2019. A second chance to get causal inference right: a classification of data science tasks. *Chance* 32(1):42–49

Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. 2016. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J. Clin. Epidemiol.* 79:70–75

Hernán MA, Sterne JAC, Higgins JPT, Shrier I, Hernández-Díaz S. 2025. A structural description of biases that generate immortal time. *Epidemiology* 36(1):107–14

Hernán MA, Wang W, Leaf DE. 2022. Target trial emulation: a framework for causal inference from observational data. *JAMA* 328(24):2446–47

Holmes DJ, Austad SN. 1995. The evolution of avian senescence patterns: implications for understanding primary aging processes. *Am. Zool.* 35(4):307–17

Horvath S. 2013. DNA methylation age of human tissues and cell types. *Genome Biol.* 14(10):3156

Horvath S, Raj K. 2018. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* 19(6):371–84

Hosmer DW, Lemeshow S, May S. 2008. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. John Wiley & Sons

Hulbert A, Pamplona R, Buffenstein R, Buttemer W. 2007. Life and death: metabolic rate, membrane composition, and life span of animals. *Physiol. Rev.* 87(4):1175–213

Huynh H, Feldt LS. 1976. Estimation of the box correction for degrees of freedom from sample data in randomized block and split-plot designs. *J. Educ. Stat.* 1(1):69–82

Jaffe AE, Irizarry RA. 2014. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome Biol.* 15(2):R31

Jagger C, Van Oyen H, Robine J. 2014. *Health Expectancy Calculation by the Sullivan Method: A Practical Guide*. Institute of Aging, Newcastle University

Jamshidi-Naeini Y, Brown AW, Mehta T, Glueck DH, Golzarri-Arroyo L, et al. 2022. A practical decision tree to support editorial adjudication of submitted parallel cluster randomized controlled trials. *Obesity* 30(3):565–70

Jarantow SW, Pisors ED, Chiu ML. 2023. Introduction to the use of linear and nonlinear regression analysis in quantitative biological assays. *Curr. Protoc.* 3(6):e801

Jawadekar N, Kezios K, Odden MC, Stingone JA, Calonico S, et al. 2023. Practical guide to honest causal forests for identifying heterogeneous treatment effects. *Am. J. Epidemiol.* 192(7):1155–65

Joehanes R, Just AC, Marioni RE, Pilling LC, Reynolds LM, et al. 2016. Epigenetic signatures of cigarette smoking. *Circ. Cardiovasc. Genet.* 9(5):436–47

Kelley GA, Kelley KS, Stauffer BL. 2023. Effects of resistance training on body weight and body composition in older adults: an inter-individual response difference meta-analysis of randomized controlled trials. *Sci. Prog.* 106(2). **https://doi.org/10.1177/00368504231179062**

Keogh RH, Gran JM, Seaman SR, Davies G, Vansteelandt S. 2023. Causal inference in survival analysis using longitudinal observational data: sequential trials and marginal structural models. *Stat. Med.* 42(13):2191–225

Kenward MG, Roger JH. 2009. An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Comput. Stat. Data Anal.* 53(7):2583–95

Khurana V, Bejjanki HR, Caldito G, Owens MW. 2007. Statins reduce the risk of lung cancer in humans: a large case-control study of US veterans. *Chest* 131(5):1282–88

Klatt KC, Bass K, Speakman JR, Hall KD. 2023. Chowing down: diet considerations in rodent models of metabolic disease. *Life Metab.* 2(3):load013

Konishi S. 2014. *Introduction to Multivariate Analysis: Linear and Nonlinear Modeling*. Chapman and Hall/CRC

Krueger C, Tian L. 2004. A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points. *Biol. Res. Nurs.* 6(2):151–57

Lamming DW. 2024. Quantification of healthspan in aging mice: introducing FAMY and GRAIL. *Geroscience* 46(5):4203–15

Landes RD. 2024. How cage effects can hurt statistical analyses of completely randomized designs. *Lab. Anim.* 58(5):476–80

Lane D. 2022. *Introductory Statistics*. Rice University and LibreTexts

Lazic SE, Clarke-Williams CJ, Munafò MR. 2018. What exactly is '*N*' in cell culture and animal experiments? *PLOS Biol.* 16(4):e2005282

LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. *Nature* 521(7553):436–44

Lehallier B, Gate D, Schaum N, Nanasi T, Lee SE, et al. 2019. Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med.* 25(12):1843–50

Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, et al. 2018. An epigenetic biomarker of aging for lifespan and healthspan. *Aging* 10(4):573–91

Loop MS, Frazier-Wood AC, Thomas AS, Dhurandhar EJ, Shikany JM, et al. 2012. Submitted for your consideration: potential advantages of a novel clinical trial design and initial patient reaction. *Front. Genet.* 3:145

Lu AT, Binder AM, Zhang J, Yan Q, Reiner AP, et al. 2022. DNA methylation GrimAge version 2. *Aging* 14(23):9484–549

Luciano A, Churchill GA. 2025. The impact of co-housing on murine aging studies. *Geroscience* 47(3):3095–110

Luo Q, Dwaraka VB, Chen Q, Tong H, Zhu T, et al. 2023. A meta-analysis of immune-cell fractions at high resolution reveals novel associations with common phenotypes and health outcomes. *Genome Med.* 15(1):59

Ma Y, Mazumdar M, Memtsoudis SG. 2012. Beyond repeated-measures analysis of variance: advanced statistical methods for the analysis of longitudinal data in anesthesia research. *Reg. Anesth. Pain Med.* 37(1):99

Maney DL, Rich-Edwards JW. 2023. Sex-inclusive biomedicine: Are new policies increasing rigor and reproducibility? *Womens Health Issues* 33(5):461–64

Manton KG. 1982. Changing concepts of morbidity and mortality in the elderly population. *Milbank Q.* 60(2):183–244

Manton KG, Gu X, Lowrimore GR. 2008. Cohort changes in active life expectancy in the US elderly population: experience from the 1982–2004 National Long-Term Care Survey. *J. Gerontol. B Psychol. Sci. Soc. Sci.* 63(5):S269–81

Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, et al. 2015. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* 16(1):25

Marioni RE, Valenzuela MJ, van den Hout A, Brayne C, Matthews FE. 2012. Active cognitive lifestyle is associated with positive cognitive health transitions and compression of morbidity from age sixty-five. *PLOS ONE* 7(12):e50940

Massa F, Scavino M, Muniz-Terrera G. 2025. A Bayesian non-linear mixed-effects model for accurate detection of the onset of cognitive decline in longitudinal aging studies. Preprint, arXiv:2502.08418 [stat.ME]

Maxwell O, Mayowa BA, Chinedu IU, Peace AE. 2018. Modelling count data; a generalized linear model framework. *Am. J. Math Stat.* 8(6):179–83

Meunier-Duperray L, Mazancieux A, Souchay C, Fleming SM, Bastin C, et al. 2025. Does age affect metacognition? A cross-domain investigation using a hierarchical Bayesian framework. *Cognition* 258:106089

Mitnitski AB, Mogilner AJ, Rockwood K. 2001. Accumulation of deficits as a proxy measure of aging. *Sci. World J.* 1:323–36

Moher D, Hopewell S, Schulz KF, Montori V, Gøtzsche PC, et al. 2010. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 340:c869

Moorad JA, Promislow DE, Flesness N, Miller RA. 2012. A comparative assessment of univariate longevity measures using zoological animal records. *Aging Cell* 11(6):940–48

Mundry R, Nunn CL. 2009. Stepwise model fitting and statistical inference: turning noise into signal pollution. *Am. Nat.* 173(1):119–23

Muthén B, Asparouhov T. 2012. Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychol. Methods* 17(3):313

Nadon NL, Strong R, Miller RA, Nelson J, Javors M, et al. 2008. Design of aging intervention studies: the NIA Interventions Testing Program. *AGE* 30(4):187–99

Naimi AI, Cole SR, Kennedy EH. 2017. An introduction to g methods. *Int. J. Epidemiol.* 46(2):756–62

Nguyen QD, Moodie EM, Forget MF, Desmarais P, Keezer MR, Wolfson C. 2021. Health heterogeneity in older adults: exploration in the Canadian Longitudinal Study on Aging. *J. Am. Geriatr. Soc.* 69(3):678–87

Nusselder WJ, Looman CW, Marang-van de Mheen PJ, van de Mheen H, Mackenbach JP. 2000. Smoking and the compression of morbidity. *J. Epidemiol. Commun. Health* 54(8):566–74

Offermann-van Heek J, Gohr S, Himmel S, Ziefle M. 2019. Influence of age on trade-offs between benefits and barriers of AAL technology usage. In *Human Aspects of IT for the Aged Population: Design for the Elderly and Technology Acceptance*, ed. J Zhou, G Salvendy. Springer

Okada D, Cheng JH, Zheng C, Kumaki T, Yamada R. 2023. Data-driven identification and classification of nonlinear aging patterns reveals the landscape of associations between DNA methylation and aging. *Hum. Genom.* 17:8

Parker ES, Golzarri-Arroyo L, Dickinson S, Henschel B, Becerra-Garcia L-E, et al. 2025. Improving statistical rigor in animal aging research by addressing clustering and nesting effects: illustration with the National Institute on Aging's Intervention Testing Program data. Preprint, bioRxiv. **https://doi.org/10.1101/2025.03.14.642436**

Parks RJ, Fares E, MacDonald JK, Ernst MC, Sinal CJ, et al. 2011. A procedure for creating a frailty index based on deficit accumulation in aging mice. *J. Gerontol. A Biol. Sci. Med. Sci.* 67A(3):217–27

Perrie Y, Badhan RKS, Kirby DJ, Lowry D, Mohammed AR, Ouyang D. 2012. The impact of ageing on the barriers to drug delivery. *J. Control. Release* 161(2):389–98

Petkovich DA, Podolskiy DI, Lobanov AV, Lee S-G, Miller RA, Gladyshev VN. 2017. Using DNA methylation profiling to evaluate biological age and longevity interventions. *Cell Metab.* 25(4):954–60.e6

Pion PD, Kittleson MD, Rogers QR, Morris J. 1987. Myocardial failure in cats associated with low plasma taurine: a reversible cardiomyopathy. *Science* 237(4816):764–68

Prince SA, Adamo KB, Hamel ME, Hardt J, Gorber SC, Tremblay M. 2008. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *Int. J. Behav. Nutr. Phys. Act.* 5(1):56

Rabe-Hesketh S, Skrondal A, Pickles A. 2004. Generalized multilevel structural equation modeling. *Psychometrika* 69(2):167–90

Robine J-M, Allard M. 1999. Jeanne Calment: validation of the duration of her life. In *Validation of Exceptional Longevity*, ed. B Jeune, J Vaupel. Odense University Press

Robins JM, Hernán MÁ, Brumback B. 2000. Marginal structural models and causal inference in epidemiology. *Epidemiology* 11(5):550–60

Rush J, Rast P, Almeida DM, Hofer SM. 2019. Modeling long-term changes in daily within-person associations: an application of multilevel SEM. *Psychol. Aging* 34(2):163–76

Rutledge J, Oh H, Wyss-Coray T. 2022. Measuring biological age using omics data. *Nat. Rev. Genet.* 23(12):715–27

Schober P, Vetter TR. 2021. Count data in medical research: Poisson regression and negative binomial regression. *Anesth. Analg.* 132(5):1378–79

Scialfa CT, Games PA. 1987. Problems with step-wise regression in research on aging and recommended alternatives. *J. Gerontol.* 42(6):579–83

Selman C, McLaren JS, Collins AR, Duthie GG, Speakman JR. 2008. The impact of experimentally elevated energy expenditure on oxidative stress and lifespan in the short-tailed field vole *Microtus agrestis*. *Proc. R. Soc. B* 275(1645):1907–16

Shen X, Wang C, Zhou X, Zhou W, Hornburg D, et al. 2024. Nonlinear dynamics of multi-omics profiles during human aging. *Nat. Aging* 4(11):1619–34

Skinner CM, Conboy MJ, Conboy IM. 2025. DNA methylation clocks struggle to distinguish inflammaging from healthy aging, but feature rectification improves coherence and enhances detection of inflammaging. *Geroscience* 47(3):3043–60

Smith G. 2018. Step away from stepwise. *J. Big Data* 5(1):32

Soukas AA, Hao H, Wu L. 2019. Metformin as anti-aging therapy: Is it for everyone? *Trends Endocrinol. Metab.* 30(10):745–55

Speiser JL, Callahan KE, Houston DK, Fanning J, Gill TM, et al. 2021. Machine learning in aging: an example of developing prediction models for serious fall injury in older adults. *J. Gerontol. A Biol. Sci. Med. Sci.* 76(4):647–54

Streiner DL. 1994. Regression in the service of the superego: the do's and don'ts of stepwise multiple regression. *Can. J. Psychiatry* 39(4):191–96

Suissa S. 2007. Immortal time bias in observational studies of drug effects. *Pharmacoepidemiol. Drug Saf.* 16(3):241–49

Sullivan DF. 1971. A single index of mortality and morbidity. *HSMHA Health Rep.* 86(4):347–54

Tabachnick BG, Fidell LS. 2013. *Using Multivariate Statistics*. Pearson Education

Tennant PWG, Murray EJ, Arnold KF, Berrie L, Fox MP, et al. 2020. Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations. *Int. J. Epidemiol.* 50(2):620–32

Teschendorff AE. 2019. Avoiding common pitfalls in machine learning omic data science. *Nat. Mater.* 18(5):422–27

Teschendorff AE. 2020. A comparison of epigenetic mitotic-like clocks for cancer risk prediction. *Genome Med.* 12(1):56

Teschendorff AE, Horvath S. 2025. Epigenetic ageing clocks: statistical methods and emerging computational challenges. *Nat. Rev. Genet.* 26(5):350–68

Teschendorff AE, Jones A, Fiegl H, Sargent A, Zhuang JJ, et al. 2012. Epigenetic variability in cells of normal cytology is associated with the risk of future morphological transformation. *Genome Med.* 4(3):24

Thall PF, Russell KE, Simon RM. 1997. Variable selection in regression via repeated data splitting. *J. Comput. Graph. Stat.* 6(4):416–34

Thapa DK, Najam W, Kpormegbey D, Robertson O, Mokalla T, et al. 2024. *Compression of morbidity dynamics: interrogating evidence, measurement challenges, and research horizons.* Paper presented at UAB Aging Research Symposium 2024, University of Alabama, Birmingham, AL

Thapa DK, Najam W, Parker ES, Wang X(R), Smith DL, et al. 2025. Life-extending interventions do not necessarily result in compression of morbidity: a case example offering a robust statistical approach. *Geroscience*. In press

Thompson ME. 2022. Evolutionary approaches in aging research. *Cold Spring Harb. Perspect. Med.* 12(11):a041195

Tian L, Tibshirani R. 2010. Adaptive index models for marker-based risk stratification. *Biostatistics* 12(1):68–86

Tibshirani R. 2018. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 58(1):267–88

Tomarken AJ, Waller NG. 2005. Structural equation modeling: strengths, limitations, and misconceptions. *Annu. Rev. Clin. Psychol.* 1:31–65

Wagner M, Hedeker DR, Wang T, Muniz-Terrera G, Capuano AW. 2024. A comparison of mixed-models for the analysis of non-linear longitudinal data: application to late-life cognitive trajectories. Preprint, arXiv:2402.07806 [stat.AP]

Walter S, Beltrán-Sánchez H, Regidor E, Gomez-Martin C, del-Barrio JL, et al. 2016. No evidence of morbidity compression in Spain: a time series study based on national hospitalization records. *Int. J. Public Health* 61(7):729–38

Wang C, Li Q, Redden DT, Weindruch R, Allison DB. 2004. Statistical methods for testing effects on "maximum lifespan." *Mech. Ageing Dev.* 125(9):629–32

Wang M, Beal DJ, Chan D, Newman DA, Vancouver JB, Vandenberg RJ. 2017. Longitudinal research: a panel discussion on conceptual issues, research design, and statistical techniques. *Work Aging Retire* 3(1):1–24

Weigl R. 2005. *Longevity of Mammals in Captivity: From the Living Collections of the World*. E. Schweizerbart

Wilczok D. 2025. Deep learning and generative artificial intelligence in aging research and healthy longevity medicine. *Aging* 17(1):251–75

Yang Y, Mayo A, Levy T, Raz N, Shenhar B, et al. 2025. Compression of morbidity by interventions that steepen the survival curve. *Nat. Commun.* 16(1):3340

Yuan K-H, Bentler PM. 2000. Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociol. Methodol.* 30(1):165–200

Zak N. 2019. Evidence that Jeanne Calment died in 1934—not 1997. *Rejuvenation Res.* 22(1):3–12

Zhu J, Wen C, Zhu J, Zhang H, Wang X. 2020. A polynomial algorithm for best-subset selection problem. *PNAS* 117(52):33117–23

Zhu T, Zheng SC, Paul DS, Horvath S, Teschendorff AE. 2018. Cell and tissue type independent age-associated DNA methylation changes are not rare but common. *Aging* 10(11):3541–57

Zoh RS, Esteves BH, Yu X, Fairchild AJ, Vazquez AI, et al. 2023. Design, analysis, and interpretation of treatment response heterogeneity in personalized nutrition and obesity treatment research. *Obes. Rev.* 24(12):e13635