Title: Biomarkers for aging identified in cross-sectional studies tend to be non-causative

Authors: Paul G. Nelson[1,*], Daniel E. L. Promislow[2], Joanna Masel[1]

1) Department of Ecology & Evolutionary Biology, University of Arizona, Tucson, AZ, USA

2) Departments of Pathology and Biology, University of Washington, Seattle, WA, USA

Corresponding author:

Paul G. Nelson

217-714-4232

pgnelson@email.arizona.edu

Running head: Identifying causal biomarkers of aging

Abstract:

Biomarkers are important tools for diagnosis, prognosis, and identification of the causal factors of physiological conditions. Biomarkers are typically identified by correlating biological measurements with the status of a condition in a sample of subjects. Cross-sectional studies sample subjects at a single timepoint, while longitudinal studies follow a cohort through time. Identifying biomarkers of aging is subject to unique challenges. Individuals who age faster have intrinsically higher mortality rates and so are preferentially lost over time, in a phenomenon known as cohort selection. In this paper, we use simulations to show that cohort selection biases cross-sectional analysis away from identifying causal loci of aging, to the point where cross sectional studies are less likely to identify loci that cause aging than if loci had been chosen at random. We go on to show this bias can be corrected by incorporating correlates of mortality identified from longitudinal studies, allowing cross sectional studies to effectively identify the causal factors of aging.

Keywords: Regression, senescence, mortality, epigenetics, longevity, gerontology

31 Body:

32 Despite the universality of aging (1, 2), we do not know which factors are the most important
33 contributors to age-associated decline in physiological function and to increase in mortality rate.
34 Further, we lack consensus regarding what it means to be aged and how to properly measure
35 aging (3, 4). The advent of high throughput "omics" technologies has led to the tantalizing
36 promise that a large enough dataset of informative biomarkers might both elucidate the causal
37 factors behind aging, and provide precise, individualized diagnoses (5). Biomarkers have also
38 been proposed as surrogate endpoints, allowing studies to forgo the slow and expensive process
39 of measuring the symptoms of aging as they develop over time (6). However, the limits of what
40 biomarkers can tell us about aging, and the consequences of the methods by which we discover
41 biomarkers, have yet to be examined theoretically.

42 Biomarkers, broadly speaking, are biological quantities that are convenient to measure, and
43 that give us medically important information (7, 8). Biomarkers can inform three very different
44 categories of medically important information. First, biomarkers can inform diagnosis and
45 associated prognosis. The first molecular biomarker was glucose, used to diagnose diabetes (9).
46 Historically, sweet urine indicated death within a matter of months (10), making glucose in the
47 urine useful for informing prognosis but little else. Today, prognosis without treatment is still
48 useful for patients with untreatable cancer or Alzheimer's disease, who may wish to put their
49 affairs in order. Biomarkers for aging can similarly predict lifespan remaining or future quality
50 of life (e.g. 11).

51 Second, biomarkers can reveal disease mechanisms. Perhaps surprisingly, while glucose is
52 important to the mechanism of diabetes, historically, it did not lead physicians directly to the
53 causal mechanisms of the disease. When medieval physicians noticed that glucose in the urine

54 varied with diet, they mistakenly concluded that diabetes must be a disease of the digestive
55 system (10). Instead, the indirect advantage of the discovery of glucose was that because its
56 presence in the urine aided diagnosis, it facilitated the search for other mechanistic clues,
57 specifically abnormalities in the pancreas noted during autopsy (12). In contrast, after cholesterol
58 was noted as a prognostic biomarker for coronary artery disease, this led more directly to
59 understanding the mechanistic role of plaque build-up (13).

60 Third, a biomarker can provide information that enhances interventions and improves
61 outcomes (14). Such markers are called "predictive" rather than the "prognostic" biomarkers
62 discussed above (15, 16). Insulin treatment, combined with a more accurate blood test (17), made
63 glucose an important biomarker to drive treatment decisions (10). The presence of hormone
64 receptors in breast cancer biopsy tissue is another predictive biomarker used to make treatment
65 decisions. There is hope that the many cancer driver mutations currently being identified might
66 also move from prognostic indicators to drug targets, in the way that oncofusion protein BCR-
67 Abl (18) is targeted by Gleevec.

68 Identifying biomarkers generally begins with noting a correlation between a measurement
69 and patient fate (19-21). Patient fate can be measured longitudinally, or else a known correlate of
70 patient fate can be measured within a contemporaneous cross-section. To measure biomarkers of
71 aging in a longitudinal study, potential markers are measured at the start, and a cohort is then
72 followed through time to determine mortality. The correlation between marker status at the
73 beginning of the study and later mortality rates or other symptoms of aging is then ascertained
74 (22-24). In a cross-sectional study, biomarkers are identified based on their ability to predict
75 either current chronological age (25), or a composite of chronological age and physiological
76 indicators often termed "biological age" (26), even when their intent, as described above, is to

77 predict lifespan remaining (27). New technologies mean that we can now investigate many
78 potential biomarkers simultaneously, which opens the door both to discovering more and/or
79 better markers, and to the risk of spurious results (28).

80 Longitudinal studies are obviously slow and difficult to conduct, but cross-sectional studies
81 may not yield reliable results. Individuals with lower intrinsic mortality rates are likely to survive
82 to older ages than their peers with high mortality rates, and are thereby more likely to be
83 observed at older ages. This bias, known as cohort selection (29), may complicate the search for
84 biomarkers of aging. Here, we simulate core aspects of the search for epigenetic biomarkers in
85 the cross-sectional approaches, specifically those of Horvath (25), Horvath (30), and Levine (26).
86 We include the possibility of cohort selection, in order to determine what cross-sectionally
87 identified biomarkers can, and cannot, tell us about aging.

88 Methods:

89 We first conduct simulations based on the procedure of Horvath (25), who trained an
90 algorithm to predict chronological age from DNA methylation status at 21,389 sites in a cross-
91 section of 7,844 individuals of different ages. The resulting regression model used 353 of those
92 sites, and was used to assess the relative rate of aging (assessed as the difference between
93 predicted age and chronological age) in a wide range of test datasets, from cancer tissue to
94 patients with progeria.

95 Guided by this procedure, we simulate the methylation status of many loci during aging,
96 followed by cross-sectional approaches to select loci as biomarkers. Specifically, we simulate
97 L=20,000 loci in each individual. Each locus has two states, degraded and non-degraded, and
98 begins the simulation in the undegraded state. We simulate enough individuals to obtain 1000
99 living individuals at each target age of 10, 20, 30,…80 years, with simulations proceeding in

discrete time steps of one year. Each year, each un-degraded locus i has a probability ρ_i of degrading, making the mean age at which the locus degrades $(1-\rho_i)/\rho_i$.

We consider two models for the probability of dying m. First, we make the probability of dying increase exponentially with time, independently of which loci have degraded, according to a Gompertz mortality curve (47):

$$m = \alpha \exp(\gamma t) \tag{1}$$

Using mortality rates from Arias et al. (48), we estimate $\alpha = e^{-10}$ deaths/year and $\gamma = 0.0807$/year to obtain a survival curve that approximates human demography in a developed nation, excluding elevated infant and early adult mortality, as well as any late life mortality deceleration.

In the second model, we make mortality a function of the number of degraded loci, choosing a function that yields a mortality curve that increases approximately exponentially with age. Specifically, we make mortality a power function of the sum of the effects of all degraded loci:

$$m = a(1 - \textstyle\sum_i b_i x_i)^k \tag{2}$$

where b_i is the effect size of locus i (often 0) and x_i is 0 if the locus is in an undegraded state and 1 if the locus is degraded. Log mortality then has a linear relationship with age (i.e. we have a Gompertz mortality curve) in the special case where $\sum_i b_i = 1$, with k determining its slope. With $k < 0$, mortality varies from a minimum of a when no causal sites have degraded to a maximum of infinity (instant death) when all sites have degraded. In our simulations, we set b_i for all causative sites to equal $1/L$ where L is the number of causative loci. We refer to the "effect size" as the percent increase in mortality resulting from a single degraded locus in an otherwise undegraded individual: $100 \times ((1 - b_i)^k - 1)$.

To parameterize Equation 2 to approximate the Gompertz curve parameter values of Equation (1), we assume (just for the purpose of this parameterization) that all loci degrade at the same rate ρ, and thus the expected number of degraded loci as a function of time t is $\mathrm{E}(\sum_i x_i) = L(1-(1-\rho)^t)$. If we ignore variation both in effect sizes of causative loci and in the number of causative loci degraded, and we also ignore cohort selection (i.e. ignoring the fact that individuals with more than average degradations are less likely to be alive to have their mortality assessed), substitution of this expectation into Equation 2 would yield:

$$m = a\exp(kt\ln(1-\rho)) \tag{3}$$

We therefore set $a = \alpha$ and $k = \gamma/\ln(1-\rho)$ in Equation 2. Figure 1 shows that when these simplifying assumptions are relaxed, mortality in our Equation (2) simulations (red and blue for small and large b_i, respectively) still exhibits the exponential relationship with time characteristic of a Gompertz mortality curve (solid black), with a slight deviation toward late life due to cohort selection when loci are of large effect (blue).

Some more recent efforts to identify biomarkers of aging (e.g. Levine (26)) train on a measure of "biological age" that incorporates phenotypic indicators in addition to chronological age. Phenotypic indicators are physiological measurements identified by their ability to predict, within a linear regression, mortality rates measured in a longitudinal study.

To determine how including measures of biological age affects the search for causal biomarkers of aging, we construct a simulated "biological age" phenotype p. We assume that phenotype can be scaled such that the mean phenotype of each age group is equal to the mean mortality rate of that age group (i.e., $\overline{m_t} = \overline{p_t}$), and that once this is done, the variance in phenotype among individuals of age t is equal to the variance in mortality (i.e., $\mathrm{var}_t(m_i) = \mathrm{var}_t(p_i)$). We also assume that within any age cohort, the phenotype correlates with mortality

with the same correlation coefficient $r_{p,m}$. Finally, we assume p is normally distributed, making individual i's simulated phenotype:

$$p_i = \overline{m_t} + r_{p,m}(m_i - \overline{m_t}) + N\left(0, \left(1 - r_{p,m}^2\right) \mathrm{var}_t(m_i)\right) \tag{4}$$

where m_i is the true mortality rate of individual i, and $N(0, \sigma^2)$ is a random number drawn from a normal distribution with mean 0 and variance σ^2. This formulation lets $r_{p,m}$ determine the fraction of variance in p that stems from variation in m, without changing $\mathrm{var}_t(p_i)$.

Next we consider how phenotypes can be used to construct a biological age. Simulated phenotypic values could be used, along with mean mortality $\overline{m_t}$ for each age group, to predict individual mortality rates using a linear model:

$$\widehat{m_i} = \beta_t \overline{m_t} + \beta_p p_i \tag{5}$$

Instead of simulating the fit of this regression model to a data sample, we consider the best possible predictor by obtaining values for β analytically when the number of datapoints n goes to infinity. First, we take the expectation of both sides of Eq. 5, giving $\overline{m_t} = \lim_{n\to\infty} \beta_t \overline{m_t} + \lim_{n\to\infty} \beta_p \overline{m_t}$, hence for the best predictor we will have $\beta_t = 1 - \beta_p$.

Second, we minimize the sum of squares $\sum_i^n (\widehat{m_i} - m_i)^2$. From Equation 5,

$$\sum_i^n (\widehat{m_i} - m_i)^2 = \sum_i^n \left(\beta_t \overline{m_t} + \beta_p p_i - m_i\right)^2. \tag{6.1}$$

Using $\beta_t = 1 - \beta_p$ yields:

$$\sum_i^n (\widehat{m_i} - m_i)^2 = \sum_i^n \left(\left(1 - \beta_p\right)\overline{m_t} + \beta_p p_i - m_i\right)^2. \tag{6.2}$$

Substituting in Equation 4 yields:

$$\sum_i^n (\widehat{m_i} - m_i)^2 = \sum_i^n \left(\overline{m_t} + \beta_p \left(r_{p,m}(m_i - \overline{m_t}) + N\left(0, \left(1 - r_{p,m}^2\right) \mathrm{var}_t(m_i)\right)\right) - m_i\right)^2. \tag{6.3}$$

We next take the limit as the number of data points goes to infinity. Using

$$\lim_{n\to\infty}(1/n)\sum_i^n N\left(0,\left(1-r_{p,m}^2\right)\mathrm{var}_t(m_i)\right)=0 \tag{7.1}$$

$$\lim_{n\to\infty}(1/n)\sum_i^n(m_i-\overline{m_t})^2=\mathrm{var}_t(m_i) \tag{7.2}$$

$$\lim_{n\to\infty}(1/n)\sum_i^n\left(N\left(0,\left(1-r_{p,m}^2\right)\mathrm{var}_t(m_i)\right)\right)^2=\left(1-r_{p,m}^2\right)\mathrm{var}_t(m_i), \tag{7.3}$$

we obtain, after some rearrangement,

$$\lim_{n\to\infty}(1/n)\sum_i^n(m_i-\widehat{m_i})^2=\left(1+\beta_p^2-2\beta_p r_{p,m}\right)\mathrm{var}_t(m_i). \tag{8}$$

This reaches a minimum when the derivative by β_p is zero, which occurs when $\beta_p = r_{p,m}$.

To obtain an estimated biological age $\widehat{t_i}$ from the combination of an observed phenotype p_i and a chronological age *t*, using this idealized linear model, we back-transform the estimated mortality from Equation 5 using the Gompertz mortality function (Equation 1):

$$\widehat{t_i}=\frac{1}{\gamma}(\ln\widehat{m_i}-\ln\alpha) \tag{9}$$

For each age cohort, we simulate each individual until they either die or reach the target age; only the latter are included in the training dataset. The process is repeated until we obtain 1000 individuals for each of the 8 ages of interest. Each locus is coded as 1 if degraded at the age of sampling and 0 otherwise. We correlate loci status either with chronological age, as in Horvath (25), or with biological age, as in Levine et al. (26), using the glmnet package in R (49). Regression coefficients between locus status and age are generated with a ridge lasso elastic net with the elastic net mixing parameter α=0.5 as in Horvath (25). During our analysis we noticed subtle biases in the regression coefficients generated by glmnet as a function of the order in which the loci appear in the training data set. To prevent such biases from affecting our results, we randomized the order of loci in the data set.

188 Results:

189 Here we examine the effect of two locus attributes – degradation rate, and the effect size
190 of degradation on mortality – on the regression coefficient assigned by glmnet as a predictor of
191 age. Loci with expected ages of degradation less than 40 make poor biomarkers of age, and the
192 most commonly selected biomarkers have expected ages of degradation between 60 and 100
193 (Figure 2A). Similarly, the most informative loci (i.e. those assigned the largest weights by
194 glmnet) are expected to degrade between ages 70 through 90 (Figure 2B).

195 Loci that have no causal effect on mortality are more likely to be chosen as informative
196 markers (Figure 2C) than their age-causing counterparts. This counter-intuitive result arises
197 because cohort selection makes biomarkers of aging unsuitable for identifying mechanisms of
198 aging. In the rare cases when age-causing loci of large effect are chosen, they have slightly
199 smaller regression coefficients than neutral loci (Figure 2D). Together, these results show that
200 loci that cause aging are worse predictors of chronological age than neutral loci.

201 To illustrate how cohort selection makes age-causing loci worse predictors of age, we
202 track the mean number of degraded loci as a function of cohort age. Let n_i be the number of
203 individuals with i degraded loci, $m_i(t)$ the probability of dying from age t to age t+1 of
204 individuals with l degraded loci, and $p_{binom}(l-i, L-l, \rho_i)$ the probability from a binomial
205 distribution, of $l-i$ loci degrading out of $L-l$ previously non-degraded loci, when the
206 probability of a locus degrading per year is p_i. The expected number of individuals of age t years
207 with l degraded loci is then:

208 $$n_l(t) = \sum_i^l n_i(t-1)\big(1-m_i(t-1)\big)p_{binom}(l-i, L-l, \rho_i) \quad (4)$$

209 To explore the dynamics of neutral loci, we make mortality a function of time as given by
210 Equation 1 (Figure 3, dashed black); for causative loci, mortality is given by Equation 2 (Figure

211 3, solid grey). Early in life, the frequency of degraded age-causing loci is dominated by the rate
212 of degradation and is thus indistinguishable from neutral loci. Later in life, the additional
213 mortality incurred by age-causing loci results in a slight decline in frequency relative to non-
214 causal loci. Even though cohort selection is subtle (as shown in Figure 3), it is enough to make
215 loci that cause aging significantly worse predictors of chronological age than neutral loci.

216 Training on chronological age biases regression analysis away from identifying causative
217 loci of aging, but more recent work instead trains biomarkers on "biological age," a combination
218 of chronological age and phenotypic measurements that are known to correlate with mortality
219 (26). Such phenotypic measurements are identified by regressing phenotype and mortality over
220 the course of a longitudinal study (e.g. 23).

221 To determine if training on longitudinally validated correlates of aging can help identify
222 causative loci of aging in a cross-sectional study, we modified our analysis to incorporate
223 chronological age, a known correlate of individual mortality, and simulated noise. We construct
224 a simulated biological age-revealing phenotype such that the correlation coefficient $r_{p,m}$ between
225 an appropriately transformed version of that phenotype and mortality has the same value within
226 each age cohort. Biological age is then calculated as an optimal function of phenotype and
227 chronological age (see Methods). When $r_{p,m}^2 = 1$, biological age perfectly reflects an
228 individual's true mortality; when $r_{p,m}^2 = 0$, it gives no information beyond that already given by
229 chronological age. Figure 4 shows that when phenotype provides little additional information
230 ($r_{p,m}^2 < 0.04$), non-causative loci (blue circles) are preferentially chosen as biomarkers to predict
231 biological age. However, when phenotype is a reliable indicator of mortality ($r_{p,m}^2 > 0.7$) most,
232 or even all, causative loci (red diamonds) are chosen as biomarkers of age.

233 To evaluate the quality of existing age-revealing phenotypes in the context of the search
234 for causative biomarkers of aging, we used NHANES III linked laboratory and mortality data
235 from the CDC (31, 32). We regressed (using the lm function in R) the nine phenotypic indicators
236 used in Levine et al. (26) (white blood cell count, serum alkaline phosphatase, red cell width
237 distribution, mean blood cell volume, lymphocytes percent, ln(serum C-reactive protein), serum
238 creatinine, serum albumin, and serum glucose) on lifespan remaining, including only the 6898
239 individuals who had died over the course of the study. Chronological age explained r^2 =0.184 of
240 variation in lifespan remaining, while the combination of chronological age plus phenotype
241 explained r^2 =0.28, indicating that the amount of unique information provided by phenotype is r^2
242 = 0.043 with respect to lifespan remaining, on the cusp of where biological age is useful for
243 identifying causative loci of aging in our idealized model with respect to mortality rate.
244 However, there is also considerable stochastic variation in lifespan remaining even when there is
245 no variance in underlying mortality rate. It therefore seems clear that the phenotypic age markers
246 used by Levine et al. (26) are above the threshold quality needed to identify causal loci of aging
247 more often than chance.
248 Next, we examine whether biomarkers of aging may be useful prognostic indicators of
249 variation in the overall rate of degradation between individuals. To test this, we generated
250 populations where the rate of aging varies among individuals, and populations where individuals
251 vary in initial status, but then continue to age at the same rate. We trained the model to predict
252 chronological age from loci status, as above. Comparing the difference between predicted age
253 and chronological age to the known underlying rate of aging, shows that biomarkers, even
254 neutral biomarkers, can provide useful information regarding the rate of aging (Figure 5A), but
255 not regarding initial mortality rate (Figure 5B). Note that we examined the special case in which

256 variation in either the slope or intercept of the aging curve arises at birth. These results should
257 still hold when it is later events that alter longevity, e.g. following an intervention. Biomarkers of
258 aging could therefore inform the efficacy of interventions that affect the slope of the Gompertz
259 curve, but not its intercept.
260
261 Discussion:
262 Like other biomarkers discussed in the Introduction, biomarkers of aging can have three uses:
263 to estimate lifespan remaining (prognosis), to identify the mechanisms that cause aging (33), and,
264 most ambitiously and usually late in their development, to distinguish classes of individuals for
265 whom different treatments are appropriate. In cross-sectional studies, epigenetic biomarkers of
266 aging have been selected for their ability to predict age at the time of sampling (21). However,
267 the individuals in the training dataset are not random samples from their age cohort. Specifically,
268 some individuals will not be sampled because they died before reaching the age of interest, and
269 these individuals will on average be faster-aging. The resulting bias is known as cohort selection.
270 Here, we have shown that cohort selection can make a cross-sectional study design spectacularly
271 ineffective at identifying causal factors of aging.
272 While we have focused on cohort selection associated with causal locus-induced mortality, a
273 similar sampling bias can arise from study exclusion criteria. Sicker individuals might be less
274 likely to enroll in such studies, and may be excluded, e.g. if they have defined diseases whose
275 prevalence increases with age, such as cardiovascular disease or diabetes (34). In a cross-
276 sectional study, any alleles or epigenetic events that cause a predisposition toward such diseases
277 will thus be underrepresented in older individuals.

278 Even if cross-sectional analyses trained on chronological age cannot identify loci that
279 contribute to accelerating mortality, such analyses are not without value. Much effort has been
280 put into quantifying natural variation in the slope and intercept of the aging curve (35) as well as
281 the impact of interventions (36, 37). Our simulations show that aging biomarkers can provide
282 prognostic (and hence potentially diagnostic) information regarding variation in the overall rate
283 of degradation among individuals (24), but not in the basal mortality rate. They may therefore be
284 useful in identifying factors affecting the overall rate of aging, as suggested by (25).
285 Biomarker selection enriches not just for non-causal loci, but also for loci with an expected
286 age of degradation toward the end of human lifespan. To obtain this result, we considered the
287 degradation rate of a locus to be independent of its causal effect. Whether selection against age-
288 related mortality, which is necessarily subtle (38), can lead to lower degradation rates in age-
289 causing loci than in the rest of the genome is an important theoretical and empirical question.
290 Degradation rates are a function of evolvable factors such as sequence context, chromatin
291 structure, or enzyme recruitment (39-41). The evolution of site-specific error rates has been
292 observed for species with larger population sizes than humans (42-44). Comparing methylation
293 changes associated with the expression of genes known *a priori* to play a role in aging to those in
294 putatively neutral sites may help illuminate whether loci that cause aging are under tighter
295 regulation than the genome as a whole.
296 While we have focused on molecular biomarkers of aging, our results apply to non-molecular
297 markers as well. Indeed, the most commonly used non-molecular markers of aging, such as grip
298 strength or skin elasticity (45, 46), are non-causal.
299 Our results show that such physiological measurements, if sufficiently strongly correlated
300 with mortality, can be used to identify causal loci of aging. However, a longitudinal study is

301 required first to identify reliable physiological markers of mortality rates, whether molecular or
302 non-molecular. What we have shown is that it is possible for a cross-sectional study to "piggy
303 back" off markers identified from a longitudinal study to identify other, causative, loci of aging.
304 To better identify causal factors of aging we need a better understanding of what it means
305 to be "aged". Regression analyses using subjects' chronological ages implicitly assume that
306 "aging" is the process of having survived for a certain duration of time. Alternatively, "age" can
307 be interpreted as morbidity, with associated mortality rate following a Gompertz curve.
308 Ultimately, we are interested in the causal factors that prevent survival or decrease quality of life
309 beyond a certain age. Our results confirm the critical importance of long-term longitudinal
310 studies, whether directly conducted or indirectly incorporated, in finding both correlates and
311 causes of aging.
312
313
314
315 Author Contributions:
316 Paul Nelson designed the methods and wrote the manuscript. Daniel Promislow raised the initial
317 questions regarding biomarkers, and contributed to the writing of the manuscript. Joanna Masel
318 contributed to the methods and the writing of the manuscript.
319
320 This work was supported by the National Institute of Health (grant number K12GM000708);
321 and the John Templeton Foundation (grant number 60814).
322
323 Conflicts of Interest: The authors declare no conflicts of interest.

References:

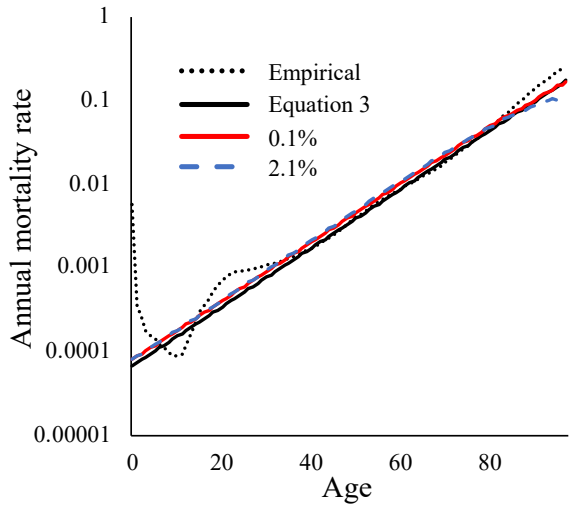
1. Medawar PB. An Unsolved Problem of Biology... An Inaugural Lecture Delivered at University College, London, 6 December, 1951. London: H.K. Lewis & Co; 1952.
2. Nelson P, Masel J. Intercellular competition and the inevitability of multicellular aging. Proc Natl Acad Sci USA. 2017;**114**:12982-12987.
3. Moorad JA, Promislow DE, Flesness N, Miller RA. A comparative assessment of univariate longevity measures using zoological animal records. Aging Cell. 2012;**11**:940-948.
4. Margolick JB, Ferrucci L. Accelerating aging research: how can we measure the rate of biologic aging? Experimental Gerontology. 2015;**64**:78-80.
5. Wu IC, Lin C-C, Hsiung CA. Emerging roles of frailty and inflammaging in risk assessment of age-related chronic diseases in older adults: the intersection between aging biology and personalized medicine. BioMedicine. 2015;**5**:1.
6. Group BDW, Atkinson Jr AJ, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ*, et al.* Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clinical Pharmacology & Therapeutics. 2001;**69**:89-95.
7. Strimbu K, Tavel JA. What are biomarkers? Curr Opin HIV AIDS. 2010;**5**:463-466.
8. Lara J, Cooper R, Nissan J, Ginty AT, Khaw K-T, Deary IJ*, et al.* A proposed panel of biomarkers of healthy ageing. BMC Med. 2015;**13**:222.
9. Zarshenas MM, Khademian S, Moein M. Diabetes and related remedies in medieval Persian medicine. Indian J Endocrinol Metab. 2014;**18**:142-149.
10. King KM, Rubin G. A history of diabetes: from antiquity to discovering insulin. British Journal of Nursing. 2003;**12**:1091-1095.
11. Zhao Y, Li S, Liu H. Estimating the survival advantage based on telomere length and serum biomarkers of aging. J Transl Med. 2017;**15**:166.
12. Cawley T. A singular case of diabetes, consisting entirely in the quality of the urine; with an inquiry into the different theories of that disease. The London Medical Journal. 1788;**9**:286.
13. Levine GN, Keaney Jr JF, Vita JA. Cholesterol reduction in cardiovascular disease—clinical benefits and possible mechanisms. New England Journal of Medicine. 1995;**332**:512-521.
14. Mandrekar SJ, Sargent DJ. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. J Clin Oncol. 2009;**27**:4027.
15. Søreide K. Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. J Clin Pathol. 2009;**62**:1-5.
16. Ballman KV. Biomarker: Predictive or Prognostic? J Clin Oncol. 2015;**33**:3968-3971.
17. Conn JW. Interpretation of the glucose tolerance test. The necessity of a standard preparatory diet. Am J Med Sci. 1940;**199**:555-564.
18. Druker BJ. STI571 (Gleevec™) as a paradigm for cancer therapy. Trends Mol Med. 2002;**8**:S14-S18.
19. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. Pivotal Evaluation of the Accuracy of a Biomarker Used for Classification or Prediction: Standards for Study Design. Am J Med Sci. 2008;**100**:1432-1438.
20. Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS*, et al.* Metagenomic biomarker discovery and explanation. Genome Biology. 2011;**12**:R60.

368 21. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of
369 ageing. Nat Rev Genet. 2018;**19**:371-384.
370 22. Belsky DW, Moffitt TE, Cohen AA, Corcoran DL, Levine ME, Prinz JA*, et al.* Eleven
371 Telomere, Epigenetic Clock, and Biomarker-Composite Quantifications of Biological Aging: Do
372 They Measure the Same Thing? Am J Epidemiol. 2018;**187**:1220-1230.
373 23. Ferguson FG, Wikby A, Maxson P, Olsson J, Johansson B. Immune parameters in a
374 longitudinal study of a very old population of Swedish people: a comparison between survivors
375 and nonsurvivors. The Journals of Gerontology Series A: Biological Sciences and Medical
376 Sciences. 1995;**50**:B378-B382.
377 24. Belsky DW, Caspi A, Houts R, Cohen HJ, Corcoran DL, Danese A*, et al.* Quantification
378 of biological aging in young adults. Proc Natl Acad Sci USA. 2015;**112**:E4104-E4110.
379 25. Horvath S. DNA methylation age of human tissues and cell types. Genome Biology.
380 2013;**14**:3156.
381 26. Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S*, et al.* An epigenetic
382 biomarker of aging for lifespan and healthspan. Aging. 2018;**10**:573.
383 27. Sprott RL. Biomarkers of aging and disease: introduction and definitions. Experimental
384 Gerontology. 2010;**45**:2-4.
385 28. Rifai N, Gillette MA, Carr SA. Protein biomarker discovery and validation: the long and
386 uncertain path to clinical utility. Nat Biotechnol. 2006;**24**:971.
387 29. Beard RE. Note on some mathematical mortality models. Ciba Foundation Symposium-
388 The Lifespan of Animals (Colloquia on Ageing): Wiley Online Library; 1959:302-311.
389 30. Horvath S, Oshima J, Martin GM, Lu AT, Quach A, Cohen H*, et al.* Epigenetic clock for
390 skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and *ex vivo* studies.
391 Aging. 2018;**10**:1758-1775.
392 31. Centers for Disease Control and Prevention. National Health and Nutrition Examination
393 Laboratory Protocol National Center for Health Statistics; 2019.
394 32. Centers for Disease Control and Prevention. NHANES III Data Linked to NDI Mortality
395 Files. National Center for Health Statistics; 2019.
396 33. Cohen AA, Morissette-Thomas V, Ferrucci L, Fried LP. Deep biomarkers of aging are
397 population-dependent. Aging. 2016;**8**:2253-2255.
398 34. Golomb BA, Chan VT, Evans MA, Koperski S, White HL, Criqui MH. The older the
399 better: are elderly study participants more non-representative? A cross-sectional analysis of
400 clinical trial and observational study samples. BMJ Open. 2012;**2**:e000833.
401 35. Bronikowski AM, Altmann J, Brockman DK, Cords M, Fedigan LM, Pusey A*, et al.*
402 Aging in the Natural World: Comparative Data Reveal Similar Mortality Patterns Across
403 Primates. Science. 2011;**331**:1325-1328.
404 36. Pletcher SD, Khazaeli AA, Curtsinger JW. Why do life spans differ? Partitioning mean
405 longevity differences in terms of age-specific mortality parameters. The Journals of Gerontology
406 Series A: Biological Sciences and Medical Sciences. 2000;**55**:B381-B389.
407 37. Simons MJ, Koch W, Verhulst S. Dietary restriction of rodents decreases aging rate
408 without affecting initial mortality rate–a meta-analysis. Aging cell. 2013;**12**:410-414.
409 38. Charlesworth B. Fisher, Medawar, Hamilton and the Evolution of Aging. Genetics.
410 2000;**156**:927-931.
411 39. Riggs AD, Xiong Z. Methylation and epigenetic fidelity. Proc Natl Acad Sci USA.
412 2004;**101**:4-5.

413 40. Sormani G, Haerter JO, Lövkvist C, Sneppen K. Stabilization of epigenetic states of CpG
414 islands by local cooperation. Molecular BioSystems. 2016;**12**:2142-2146.
415 41. Margueron R, Reinberg D. Chromatin structure and the inheritance of epigenetic
416 information. Nat Rev Genet. 2010;**11**:285.
417 42. Vvedenskaya IO, Vahedian-Movahed H, Bird JG, Knoblauch JG, Goldman SR, Zhang Y,
418 *et al.* Interactions between RNA polymerase and the "core recognition element" counteract
419 pausing. Science. 2014;**344**:1285-1289.
420 43. Zhang D, Chen D, Cao L, Li G, Cheng H. The Effect of Codon Mismatch on the Protein
421 Translation System. PLoS One. 2016;**11**:e0148302.
422 44. Meer KM, Nelson PG, Xiong K, Masel J. Transcriptional Error Rates Vary by Gene
423 Expression Level in *E. coli* but not *S. cerevisiae*. bioRxiv. 2019:554329.
424 45. Kanaki T, Makrantonaki E, Zouboulis CC. Biomarkers of skin aging. Rev Endocr Metab
425 Disord. 2016;**17**:433-442.
426 46. Oksuzyan A, Demakakos P, Shkolnikova M, Thinggaard M, Vaupel JW, Christensen K,
427 *et al.* Handgrip strength and its prognostic value for mortality in Moscow, Denmark, and
428 England. PloS One. 2017;**12**:e0182684.
429 47. Gompertz B. XXIV. On the nature of the function expressive of the law of human
430 mortality, and on a new mode of determining the value of life contingencies. In a letter to Francis
431 Baily, Esq. FRS &c. Philos Trans R Soc Lond, Ser B: Biol Sci. 1825;**115**:513-583.
432 48. Arias E, Heron M, Xu J. United States Life Tables, 2014. National Vital Statistics Report.
433 2017;**66**:1-64.
434 49. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models
435 via Coordinate Descent. Journal of Statistical Software. 2009;**33**:1-22.
436

437

438 **Figures:**



439

440 **Figure 1: The Gompertz mortality curve (solid black) approximates the empirical estimates**

441 **of human mortality (dotted black), excluding increased infant and early adult mortality**

442 **(48). To include cohort selection, Equation 2 was simulated for a starting population of**

443 **1000 individuals using 20,000 loci of which 6,251 loci have an effect size of 0.1% (red) or**

444 **301 loci have an effect size of 2.1% (blue).**

445

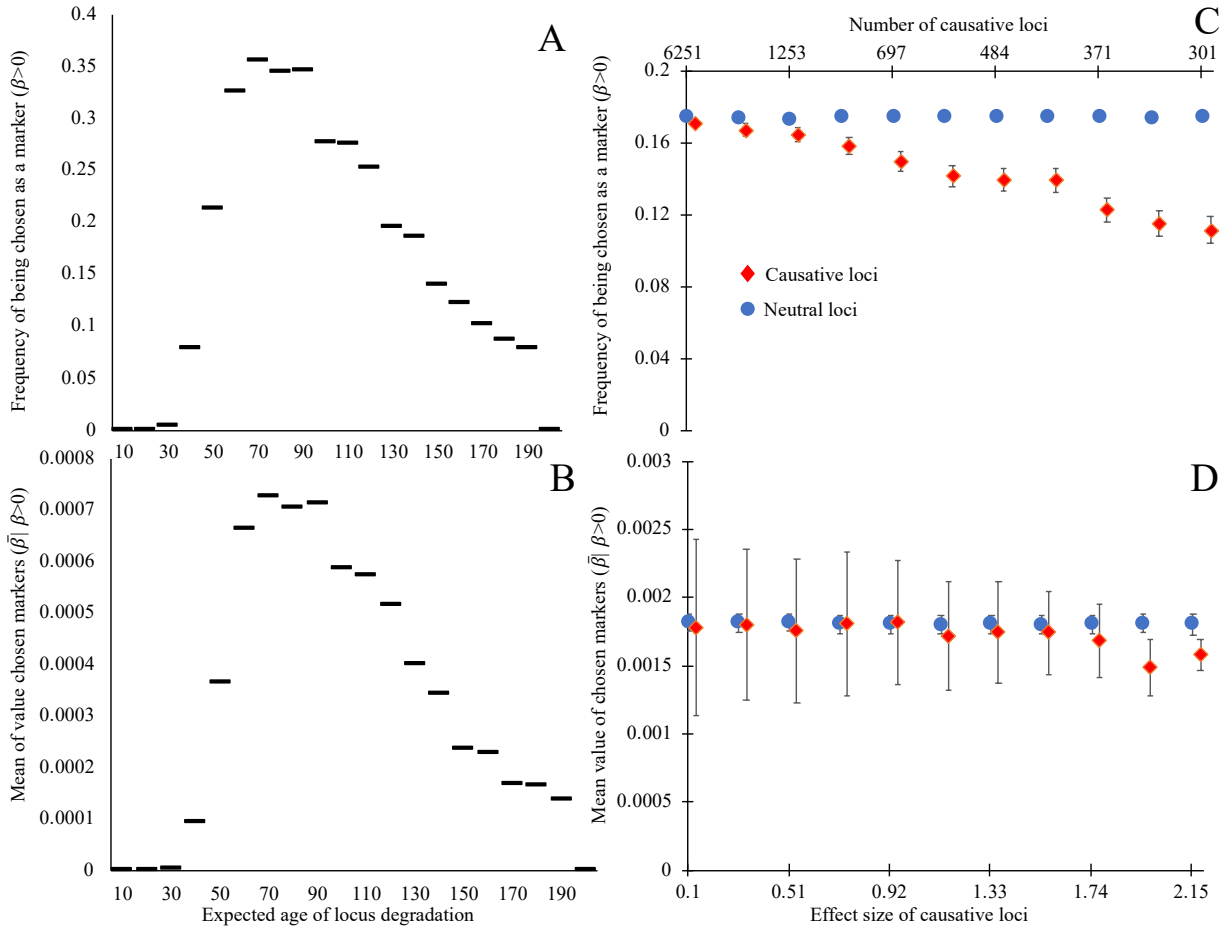


Figure 2: Biomarkers selected on their ability to predict chronological age tend not to be causative for aging, and degrade, on average, late in life. Loci whose expected time of degradation is late in life (age 50+) are chosen as biomarkers of aging (A) more often and have larger regression coefficients when chosen as biomarkers (B), than loci that degrade earlier in life. Causative loci (red diamonds) are less likely to be selected as biomarkers (C) than neutral loci in the same genome (blue circles). Of the loci selected as biomarkers (D), the magnitude of regression coefficients of causative loci is slightly smaller, on average, than those of neutral loci, at least when causative loci are of fairly large effect (>2%). Panels A and B show the outcome of a single simulation in which each of the 20,000 loci

457 **have expected ages of degradation 1+199(*i*/20,000) for integers *i* = 0 through 20,000. All loci**
458 **are neutral and mortality is determined by Equation (1). Horizontal lines indicate the y-**
459 **axis value corresponding to the 200 loci in each bin. A degradation time well above human**
460 **lifespans indicates a locus with a low probability of degrading prior to death. Each point in**
461 **panels C and D shows the mean outcomes of six independent replicate simulations in which**
462 **each of the 20,000 loci are either neutral or have a single causal effect size (non-round-**
463 **number effect sizes are an artifact of making choices using an alternative effect size**
464 **metric). Mortality is determined by Equation (2), with the numbers of causative loci (top x-**
465 **axis C,D) inversely proportional to the effect size of a causal locus (bottom x-axis C,D). In**
466 **C and D, all loci have an expected age of degradation of 75 years. Error bars show**
467 **standard errors; markers for causative loci have been offset slightly to the right.**
468
469

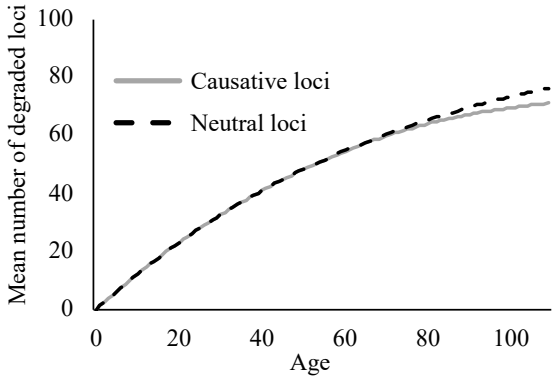


Figure 3: The mean number of degraded loci per individual when loci cause aging (grey), or are neutral (dashed black) out of 301 total loci. We use the largest effect size shown in Figure 2, where the degradation of one causative locus results in a 2.3% increase in mortality rates in an otherwise non-degraded individual, and all loci have an expected age of degradation of 75 years.

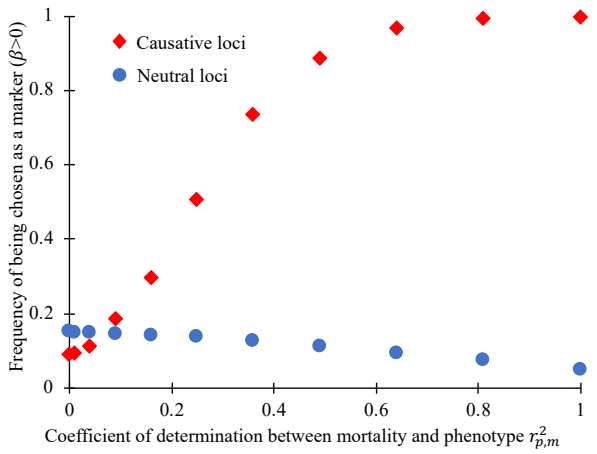


Figure 4: When phenotypes can accurately distinguish individual mortality within a cohort of the same age, regression on biological age (here an optimal function of both chronological age and phenotype) can preferentially choose loci that cause aging (red diamonds) over neutral loci (blue circles) as biomarkers of aging. As in Figure 4 and the rightmost markers in Figure 3, here 301 loci of large effect are causative of aging.

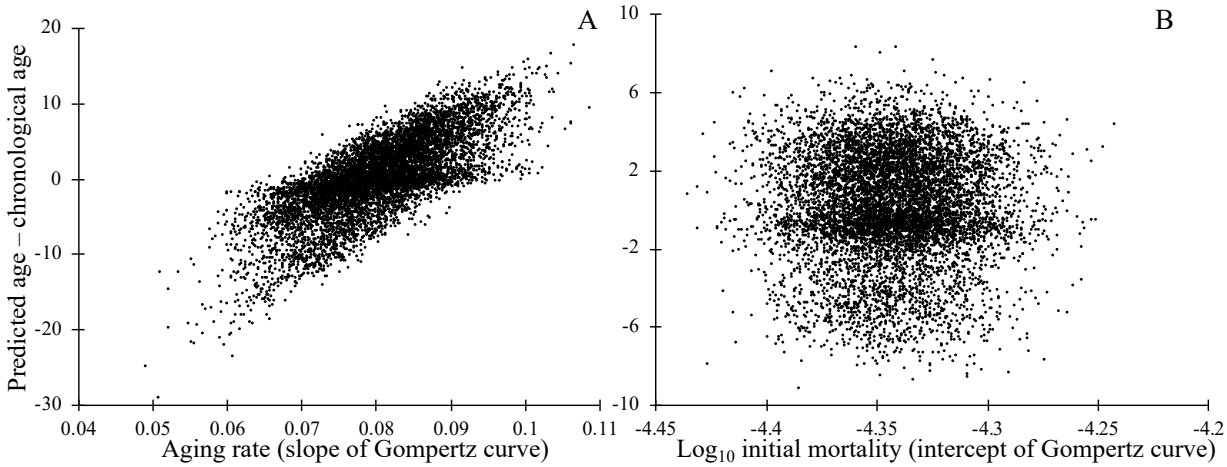


Figure 5: The deviation between an individual's chronological age and predicted age informs the rate of aging, as measured by the slope of the Gompertz mortality curve (A), but not the intercept of the Gompertz mortality curve (B). Mortality depends on causal biomarker loci according to Equation 2. Markers were chosen by regression on chronological age in a training dataset, then applied to a separate testing dataset of independent but identically constructed simulations. In A, each individual's rate of aging was drawn at birth from a normal distribution with mean 0.0807 and standard deviation 0.00807. The rate of aging γ_i of individual i determines the probability of degradation ρ_i of that individuals' loci, where $\rho_i = 1 - \exp(\gamma_i \ln(1-\rho)/0.0807)$ where ρ=1/76, which gives an estimated age of degradation of 75 years as in Figure 3; all loci are non-degraded at birth. In B, a number of loci drawn from a Poisson distribution with mean 10,000 were degraded at birth, making the mean mortality rate at birth of individuals in the top quartile roughly 17% higher than those of the bottom quartile. To keep the mean mortality rate at birth $m_0=e^{-10}$ as above, we make $a=6.65\times10^{-7}$; the rate of degradation of loci is

constant (ρ=1/76) across all individuals. In both A and B, 301 out of 20,000 loci are causative of aging.