

Does AI-predicted biochemical age predict death? A cross-validated re-examination in NHANES 1999–2018 with linked mortality follow-up

John Feng

johngladwit@gmail.com

Savitribai Phule Pune University

Research Article

Keywords: biological age, aging clocks, NHANES, all-cause mortality, prediction-model evaluation, C-index, machine learning

Posted Date: August 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10599297/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Does AI-predicted biochemical age predict death?

A cross-validated re-examination in NHANES 1999–2018
with linked mortality follow-up

John Feng

Independent Researcher, johngladwit@gmail.com

Abstract

Models that predict a person’s chronological age from routine blood chemistry are widely described as measuring “biological age”. The claim only means something if the error such a model makes — predicted minus actual age, usually called age acceleration — carries information about how long the person will live, over and above their actual age. We tested that claim directly. We joined the standard biochemistry and complete blood count panels from ten cycles of the National Health and Nutrition Examination Survey (1999–2018) to the public-use linked mortality file, giving 45,033 adults aged 20–79 with 5,213 deaths over 445,961 person-years. We refit chronological-age predictors from six model families under nested cross-validation, including a sweep of 360 neural networks, extracted age acceleration from each, and tested all of them in Cox models adjusted for chronological age and sex.

Flexible models predict age much better than linear ones: out-of-fold R^2 rose from 0.398 for ridge regression to 0.549 for gradient boosting and 0.557 for a neural network. The usefulness of their residuals moved the other way. Age acceleration from ridge regression raised Harrell’s C-index for all-cause mortality by +0.0124 over age and sex, while age acceleration from the neural network raised it by only +0.0056. Across the six families the correlation between age-prediction R^2 and C-index gain was -0.94 ($p = 0.005$): the more accurately a model hit chronological age, the less its residual was worth. Meanwhile the same 38 analytes fitted against mortality instead of against age gained +0.0367 (95% CI +0.0332, +0.0408) — 3.0 times the best age-trained residual — and retained 85% of that gain even after age acceleration was already in the model. Published PhenoAge acceleration, which uses only nine markers but was built against mortality, gained +0.0182. The limitation is therefore neither the biomarkers nor the model class; it is the training target. Finally, we show that the R^2 values reported for blood-chemistry age predictors are mostly a property of the sample: holding the model and the analytes fixed and varying only the age window moves R^2 from 0.247 to 0.683. Fitting a model to chronological age is a reasonable way to describe how blood chemistry changes with age; it is a poor way to build a mortality biomarker, and reporting R^2 alone hides that.

Keywords: biological age; aging clocks; NHANES; all-cause mortality; prediction-model evaluation; C-index; machine learning.

1 Introduction

The idea behind a biological age is simple and appealing. Two people who have been alive for the same number of years may be in very different physiological condition, and a measurement that captured that difference would be useful: it would let a clinician spot accelerated decline early — the same motivation that drives the frailty phenotype (Fried et al., 2001) and the broader programme

of mapping the biology of aging onto measurable processes (López-Otín et al., 2023; Furman et al., 2019) — and it would let a trial of an anti-aging intervention read out an effect in months rather than in the decades an all-cause mortality endpoint would require (Barzilai et al., 2016; Justice et al., 2018; Moqri et al., 2023).

The usual way to build such a measurement is indirect. One trains a model to predict chronological age from some biological substrate — DNA methylation (Horvath, 2013; Hannum et al., 2013), blood chemistry (Putin et al., 2016; Mamoshina et al., 2018; Kendiukhov, 2020), gene expression (Holzscheck et al., 2021), the plasma proteome (Oh et al., 2023) — and then treats the model’s error as the signal. A person predicted to be older than they are is said to be aging quickly; a person predicted younger is aging slowly. The residual is given a name, age acceleration, and studied as though it were a measured quantity.

This is a strong assumption, and it has an obvious test. If age acceleration tracks the rate of biological decline, it should predict death once chronological age is already accounted for. The field has largely accepted that test. Second-generation clocks such as PhenoAge (Levine et al., 2018; Liu et al., 2018) and GrimAge (Lu et al., 2019) were built against mortality rather than against age, and they outperform age-trained clocks on exactly this criterion (McCrory et al., 2021; Li et al., 2020). Recent consensus statements on validating aging biomarkers put predictive validity for mortality and morbidity at the centre (Moqri et al., 2023, 2024).

Yet much of the work that predicts chronological age from clinical biochemistry never runs the test. Papers in this line report the coefficient of determination R^2 against chronological age and the mean absolute error in years, and treat a higher R^2 as a better clock. A representative and carefully executed example is the study of Kendiukhov (2020), which trained 1,152 neural networks of varying depth and hyper-parameterisation on NHANES blood biochemistry and reported a best R^2 of about 0.78 with a mean absolute error of 6.5 years. That work asked which architectures predict age best and which analytes drive the prediction. It did not ask whether the residual from those networks predicts anything.

This is not a criticism of one paper; it is a description of a genre. And it matters, because the two questions can have opposite answers. A model that predicts chronological age perfectly has, by construction, no residual at all and therefore no biological-age signal whatsoever. Somewhere between a useless model and a perfect one, the residual must stop improving and start getting worse. Pyrkov et al. (2018) made this argument on theoretical and simulation grounds and titled their paper accordingly. What has been missing is a direct empirical test on the kind of data these models are actually trained on, with a mortality endpoint, modern baselines, and an accounting of discrimination rather than only of hazard ratios.

That is what we provide. We use the same public data source — the National Health and Nutrition Examination Survey — joined to the NCHS public-use linked mortality file, which follows participants for death through the end of 2019 (National Center for Health Statistics, 2022). We refit chronological-age models across six families, sweep a large space of network architectures, extract age acceleration from each, and evaluate all of them against all-cause mortality with chronological age and sex already in the model. As reference points we compute PhenoAge from its published coefficients (Levine et al., 2018) and build a mortality-trained score from exactly the same analytes, which tells us how much signal was available to be extracted in the first place.

Contributions.

1. We show that flexible learners predict adult chronological age from a 38-analyte blood panel substantially better than linear ones (out-of-fold R^2 0.377 to 0.557), so the usual concern that neural networks add nothing on tabular data does not apply here.

2. We show that this extra accuracy is not merely wasted but actively counterproductive for the stated purpose. Across model families, the correlation between age-prediction R^2 and the C-index gain of the resulting age acceleration is -0.94 ($p = 0.005$); ridge regression, the least accurate non-degenerate age model we fit, yields the most useful residual.
3. We separate two regimes that are usually conflated. Within a single linear model class, tuned along a ridge penalty path, accuracy and usefulness rise together (correlation 0.89). It is the move to nonlinear model classes, which buys accuracy by capturing age structure unrelated to mortality, that reverses the relationship.
4. We show the ceiling is set by the training target, not the data. The same analytes fitted against mortality gain $+0.0367$ in C-index, against a maximum of $+0.0124$ for any age-trained residual, and retain 85% of that gain after age acceleration is included.
5. We show that the headline R^2 of blood-chemistry age predictors is largely a property of the age window of the sample — 0.247 to 0.683 for one model on one panel — which reconciles our adult-only figure with the higher values reported elsewhere and argues against using R^2 to compare clocks across studies.

2 Background and related work

2.1 Clocks trained on chronological age

The first widely used aging clocks regressed chronological age on DNA methylation at a few hundred CpG sites (Horvath, 2013; Hannum et al., 2013). Their success set the template that later work followed: choose a substrate, fit age, call the residual biological aging. On methylation the template partly works: methylation age acceleration does predict all-cause mortality, in single cohorts (Marioni et al., 2015) and in meta-analysis (Chen et al., 2016), and methylation signatures selected directly against death predict it better still (Zhang et al., 2017). Applied to blood chemistry, the same template produced a series of increasingly elaborate models. Putin et al. (2016) trained an ensemble of deep neural networks on 41 blood markers from about 62,000 people; Mamoshina et al. (2018) extended this across three national populations; Zhavoronkov and Mamoshina (2019) surveyed the resulting family of “deep aging clocks”; Kendiukhov (2020) performed a systematic architecture sweep on NHANES; Gialluisi et al. (2022) built a deep-learning marker on an Italian cohort; Sagers et al. (2020) predicted both chronological and biological age from laboratory data at scale. Reviews of the area (Jylhävä et al., 2017; Ferrucci et al., 2020; Rutledge et al., 2022; Bafei and Shen, 2023) describe the same design repeatedly.

2.2 Clocks trained on an outcome

A parallel line abandoned chronological age as the target. Klemes and Doubal (2006) proposed estimating a latent biological age from the joint behaviour of age-associated markers rather than from a single regression on age. Nakamura and Miyao (2007) developed a related index from age-associated markers, and Cohen et al. (2013) showed that a multivariate distance capturing multi-system physiological dysregulation behaves like an aging measure without reference to a chronological-age regression at all. Levine (2013) showed in NHANES III that a biological age constructed this way predicts mortality better than chronological age. PhenoAge (Levine et al., 2018; Liu et al., 2018) went further: it fits a Gompertz mortality model to nine clinical chemistry markers plus age, then re-expresses the resulting mortality score in units of years. GrimAge (Lu et al., 2019) predicts methylation surrogates of plasma proteins and smoking pack-years and combines them

against time to death, and DunedinPACE estimates the rate of decline from longitudinal physiology rather than a cross-sectional level (Belsky et al., 2022). Head-to-head comparisons consistently favour the outcome-trained measures (McCrary et al., 2021; Li et al., 2020; Belsky et al., 2018). Our mortality-trained score below is built in the same spirit and serves as a ceiling rather than as a proposed new biomarker.

2.3 How age acceleration is defined

The definition of the residual is not a detail. Any regression shrinks predictions towards the sample mean, so the naive difference $\hat{a} - a$ between predicted and actual age is negatively correlated with actual age: old people look young and young people look old, purely as an artefact. The neuroimaging community, which builds “brain age” models on the same template (Cole and Franke, 2017), has studied this carefully and converged on regressing the prediction on chronological age and using that residual (Smith et al., 2019; Le et al., 2018; de Lange and Cole, 2020; Butler et al., 2021). We follow that convention throughout. Some of the disagreement in the blood-chemistry literature about whether age acceleration “works” comes down to which of these two quantities was used.

2.4 How predictive value should be measured

A hazard ratio answers a different question from a discrimination statistic. With tens of thousands of participants, a marker can have an unambiguous association with mortality — a hazard ratio well above one with a vanishing p -value — and still change almost nobody’s predicted risk ordering (Pepe et al., 2004; Cook, 2007). Because chronological age is such a powerful mortality predictor on its own, this trap is particularly easy to fall into in aging research. We therefore report Harrell’s C-index (Harrell et al., 1996) and, specifically, the change in C-index when a score is added to a model that already contains age and sex, with bootstrap confidence intervals (Efron and Tibshirani, 1993); related discussions of incremental value appear in Pencina and D’Agostino (2004); Uno et al. (2011); Steyerberg et al. (2010); Vickers and Elkin (2006).

2.5 Simple baselines

Across clinical prediction problems on tabular data, flexible learners often fail to beat well-regularised linear models (Christodoulou et al., 2019; Shwartz-Ziv and Armon, 2022; Grinsztajn et al., 2022), and the same pattern has been reported inside aging research: simple linear composites track elaborate ones closely (Belsky et al., 2018; Hastings et al., 2019; Li et al., 2020), and in single-cell transcriptomics a recent study found that frozen foundation-model representations carried aging signal no better than far simpler baselines once donor structure was respected (Kendiukhov, 2026). Our accuracy results are a partial exception — flexible models really do predict age better here — which makes the mortality results below more, not less, pointed.

3 Data

3.1 Source and assembly

NHANES is a continuous, nationally representative survey of the civilian non-institutionalised United States population, released in two-year cycles (Johnson et al., 2013). We used all ten cycles from 1999–2000 through 2017–2018. From each cycle we took the demographic file, the standard biochemistry profile and the complete blood count with five-part differential, and merged them on the participant sequence number. This gives 23 serum chemistry analytes and 15 haematology

analytes, 38 in total; with a sex indicator the models see 39 inputs. The panel is listed in full in Table 7.

Mortality status came from the NCHS public-use Linked Mortality File, which matches survey participants to the National Death Index with follow-up through 31 December 2019 (National Center for Health Statistics, 2022). We used the linkage-eligibility flag, the vital status flag, person-months of follow-up from the examination date, and the underlying-cause-of-death recode.

3.2 Harmonisation across cycles

Two measurement issues required attention. Serum creatinine was assayed differently in some cycles, and NCHS publishes regression equations to place all cycles on a common standardised scale; we applied the published corrections for 1999–2000 and 2005–2006 (Selvin et al., 2007). C-reactive protein was measured as standard CRP in mg/dL from 1999 to 2010, was not measured in 2011–2014, and returned as high-sensitivity CRP in mg/L in 2015–2018; we converted the latter to mg/dL. Because CRP is missing for two cycles, and because PhenoAge requires it, all PhenoAge comparisons are made on the 35,486 participants with a CRP value (4,633 deaths), and we repeat the head-to-head comparison of the other scores on that same subsample so that nothing turns on the difference in sample.

3.3 Sample definition

We restricted to adults aged 20 years and older, and imposed an upper bound of 79 years. The upper bound is not cosmetic: NHANES top-codes age at 85 in 1999–2006 and at 80 from 2007 onward, so 79 is the highest age at which the chronological-age variable is uncensored in every cycle. Including top-coded participants would place a spike of heterogeneous true ages at a single recorded value and would distort both the age model and the residual. We then required eligibility for mortality linkage, non-missing follow-up time, and complete data on all 38 analytes.

The analytic sample contains 45,033 participants, 48.2% male, with mean age 47.3 years (SD 16.6). Over a median follow-up of 9.6 years (maximum 20.8) there were 5,213 deaths in 445,961 person-years. Table 1 gives the characteristics; the full attrition log is in the released code.

4 Methods

4.1 Preprocessing

Fifteen analytes whose distributions are strongly right-skewed on physiological grounds — the transaminases, alkaline phosphatase, gamma-glutamyl transferase, lactate dehydrogenase, triglycerides, creatinine, bilirubin, glucose, blood urea nitrogen, iron, white cell count, and the three rarer differential percentages — were modelled on the natural-log scale. This choice was fixed in advance from the marginal distributions and uses no outcome information. Features were then standardised inside each training fold.

4.2 Chronological-age models

We fit six model families: ordinary least squares; ridge regression; principal component regression (PCA followed by ridge); k -nearest neighbours; histogram gradient boosting; and a multilayer perceptron. A model that predicts age from sex alone serves as a floor.

Accuracy comes from nested cross-validation: an outer 5-fold split produces the out-of-fold predictions we report and use downstream, and hyper-parameters are chosen by an inner 3-fold

search inside each outer training fold. This matters, because selecting hyper-parameters on the same data used to report performance inflates the estimate, sometimes substantially (Varma and Simon, 2006; Cawley and Talbot, 2010). Every prediction used anywhere in this paper, including in every survival model, is out-of-fold. The multilayer perceptron in the nested-CV comparison uses a fixed 256–128 architecture with a fixed weight decay, specified in advance rather than tuned, because the architecture question is answered separately and at much larger scale by the sweep described next; this keeps the nested loop affordable without giving the network an unfair selection advantage.

4.3 Architecture sweep

To reproduce the design of Kendiukhov (2020) rather than only its conclusion, we defined a full factorial grid over network architecture: twelve hidden-layer configurations from a single layer of 32 units up to four layers of 128 and a 512–256–128 funnel, crossed with three activation functions, four L_2 penalties, three learning rates and three batch sizes, for 1,296 combinations in total. Under our compute budget we trained a uniform random sample of 360 of them on a fixed 70/30 split. The purpose is to establish what the ceiling of this model class is on these data, so that the later mortality results cannot be dismissed as a consequence of an underpowered network.

4.4 Age acceleration

For each model we define age acceleration as the residual from a linear regression of the out-of-fold predicted age on chronological age, following the convention established in the brain-age literature (Smith et al., 2019; de Lange and Cole, 2020). This residual is uncorrelated with chronological age by construction, which is what we want when the outcome model already contains age. Every score is divided by its standard deviation so that hazard ratios are per one standard deviation and are comparable across scores.

4.5 Reference standards

PhenoAge. We computed phenotypic age from the published coefficients of Levine et al. (2018), using albumin (g/L), creatinine ($\mu\text{mol/L}$), glucose (mmol/L), log CRP (mg/dL), lymphocyte percentage, mean cell volume, red cell distribution width, alkaline phosphatase, white cell count and chronological age. Levine et al. define a Gompertz mortality score M and then map it back to year units; substituting the first expression into the second cancels the exponential exactly, giving an equivalent closed form that we use to avoid floating-point underflow in the tails. We verified the coefficients and units against the reference **BioAge** implementation (Kwon and Belsky, 2021), and verified the closed form against the two-step form to twelve decimal places over the observed range. PhenoAge acceleration is the residual of phenotypic age on chronological age, matching how it is used in the source papers.

Mortality-trained score. This is the key control. Using the same outer folds and the same 39 inputs, we fit a ridge-penalised Cox model on each training fold with the standardised analytes *and* chronological age, and scored the held-out fold with the biomarker part of the linear predictor only — that is, the analytes weighted by their age-adjusted log hazard ratios. We then residualise that score on chronological age exactly as for age acceleration, so that every score in the results table is an age-orthogonal quantity measured in standard deviations. This construction is the recipe behind PhenoAge, re-estimated on the present panel. It is not offered as a new biomarker. It answers one question: given these inputs, how much mortality information was there to find?

4.6 Survival analysis

We fit Cox proportional hazards models (Cox, 1972) with time measured in years from the examination date and all-cause death as the event. The reference model contains chronological age and sex. Each score is then added, one at a time, and we report the hazard ratio per standard deviation, Harrell’s C-index of the full model, and the change in C-index relative to the reference. The change in C-index is accompanied by a 95% percentile interval from 250 bootstrap resamples in which both models are refit; because the scores are already out-of-fold, this interval reflects sampling variability in the statistic rather than optimism from model fitting. We also report cumulative-case/dynamic-control areas under the curve at five and ten years.

Sensitivity analyses add race/ethnicity, education and the poverty–income ratio; add survey cycle; and repeat the primary models with NHANES examination weights divided by ten, with variance estimation clustered on stratum-by-primary-sampling-unit, following the survey’s analytic guidance for pooling cycles (Johnson et al., 2013). Cause-specific analyses use the underlying-cause recode, treating other causes as censored.

4.7 Why the targets differ: an objective-mismatch analysis

To see *why* an age-trained residual should underperform, we compared what the two objectives lean on. For each of the 39 inputs we computed its absolute correlation with chronological age and its standardised log hazard ratio for mortality from a Cox model containing that analyte, chronological age and sex. We use marginal rather than multivariable quantities here because several panel members are algebraically derived from others — osmolality from sodium, glucose and urea nitrogen; globulin from total protein and albumin — so multivariable coefficients are inflated and cancelling. We report the multivariable version as a check.

4.8 The accuracy–utility path

Finally we varied a single knob. Along a ridge penalty path of twenty values spanning heavy under-fitting to essentially unpenalised regression, we recomputed out-of-fold predictions, extracted age acceleration at each penalty, and measured its C-index gain. This traces accuracy against usefulness within one fixed model class, which we can then contrast with the same trace across model classes.

4.9 Software

Analyses used scikit-learn (Pedregosa et al., 2011) and lifelines (Davidson-Pilon, 2019). All random seeds are fixed.

5 Results

5.1 Flexible models predict adult chronological age considerably better than linear ones

Table 2 and Figure 1a give out-of-fold accuracy. Ridge regression reaches $R^2 = 0.398$ with a mean absolute error of 10.48 years, and ordinary least squares is indistinguishable from it. Gradient boosting reaches $R^2 = 0.549$ (MAE 8.83 years) and the multilayer perceptron $R^2 = 0.557$ (MAE 8.69), an improvement of 0.159 in R^2 over ridge — about a forty per cent relative gain. Principal component regression, with far fewer effective parameters, trails at 0.377. The relationship between

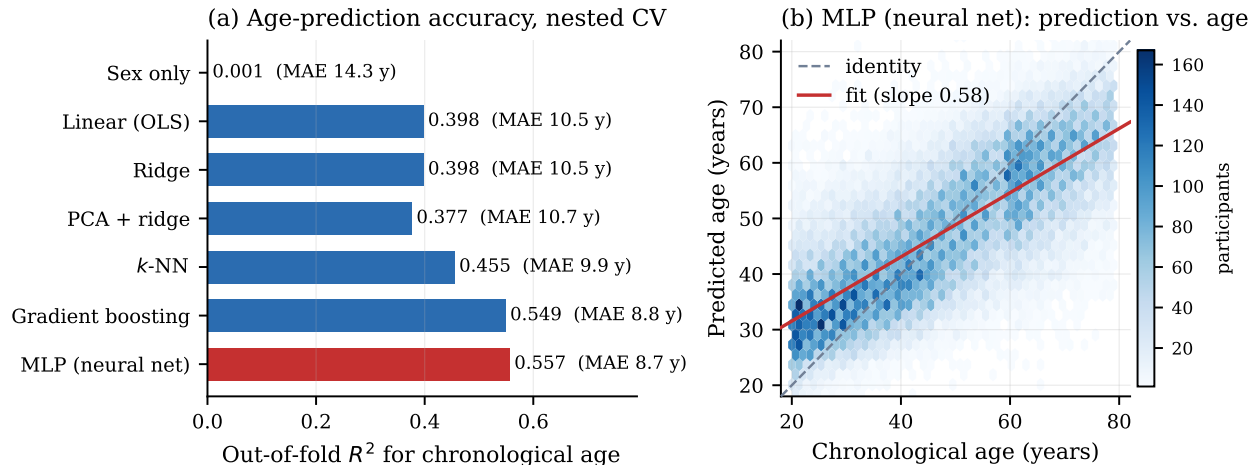


Figure 1: Predicting chronological age from 38 blood analytes and sex, adults aged 20–79, NHANES 1999–2018 ($n = 45,033$). (a) Out-of-fold R^2 under nested cross-validation, with mean absolute error in parentheses. (b) Out-of-fold predictions from the most accurate family against chronological age. The fitted slope is well below one: predictions are shrunk towards the sample mean, which is why age acceleration must be defined as a residual on age.

blood chemistry and adult age is therefore appreciably nonlinear, and the flexible models find that structure. This is worth stating plainly because it is the opposite of the pattern usually reported for tabular clinical prediction (Christodoulou et al., 2019; Grinsztajn et al., 2022), and because it means the mortality results below cannot be attributed to a weak neural network.

Figure 1b shows the characteristic geometry of any such model. The cloud of predictions is tilted relative to the identity line: the fitted slope of predicted on actual age is well below one, so the oldest participants are systematically predicted young and the youngest predicted old. This is ordinary regression shrinkage, not a defect, but it is why age acceleration must be defined as a residual on age rather than as a raw difference (Smith et al., 2019; de Lange and Cole, 2020).

5.2 The architecture sweep confirms the ceiling

Of 360 networks sampled from the 1,296-point grid, 256 reached a positive held-out R^2 ; the remaining 104 configurations failed to converge, almost all of them combinations of a high learning rate with a saturating activation. Among those that learned (Figure 2), the best achieved $R^2 = 0.551$ (MAE 8.82 years) with a 256–128 hidden structure, and the median reached 0.509. Within the top five per cent of the sweep, R^2 varied by only 0.013, and the best network at each depth varied by 0.014 from one to four hidden layers. Depth buys almost nothing; the ceiling of this model class on these data is around $R^2 = 0.55$, and the fixed architecture used in the nested cross-validation sits at that ceiling.

5.3 Age acceleration predicts death, but adds little discrimination

Chronological age and sex alone give a C-index of 0.8117 for all-cause mortality. Table 3 and Figure 3 report what happens when each age-acceleration score is added.

The associations are unambiguous. Age acceleration from ridge regression carries a hazard ratio of 1.431 per standard deviation (95% CI 1.394–1.468), and from the multilayer perceptron 1.265 (1.230–1.300). With more than five thousand deaths, all of these have p -values below 10^{-60} .

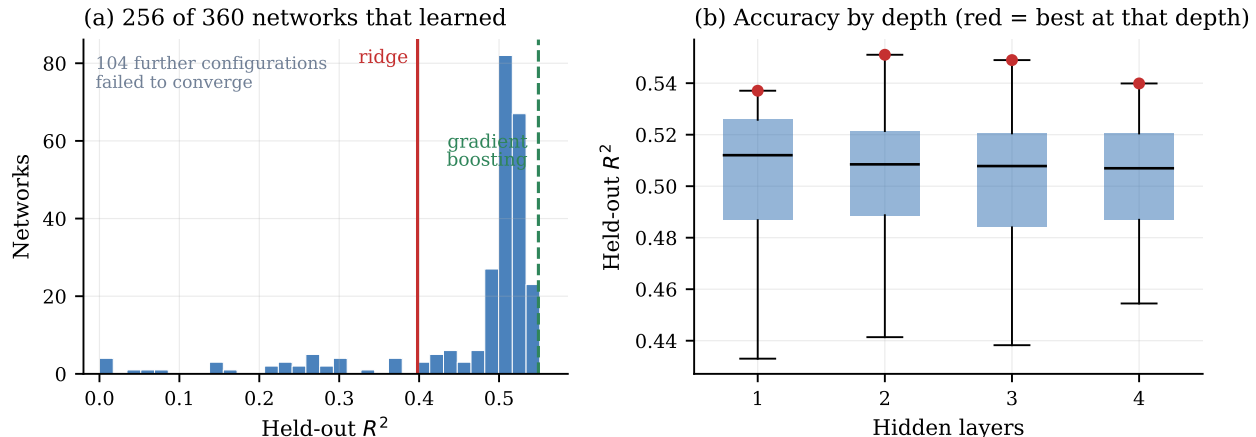


Figure 2: Architecture sweep over a random sample of 360 multilayer perceptrons from a 1,296-point factorial grid, on a fixed 70/30 split. (a) Distribution of held-out R^2 among networks that learned, with ridge regression and gradient boosting marked. (b) Held-out R^2 by number of hidden layers; red points mark the best network at each depth.

The discrimination gains are small. The best of them, ridge regression, moves the C-index by +0.0124 (95% CI +0.0107, +0.0143), from 0.8117 to 0.8242. The neural network moves it by +0.0056 (+0.0046, +0.0071). The ten-year AUC moves from 0.827 to 0.835. These are increments visible in the fourth decimal place of a table and invisible in any decision a clinician or a trialist would take.

5.4 The more accurate the age model, the less useful its residual

The ordering of the six families in Figure 3b is the reverse of their ordering in Figure 1a. Ridge regression and ordinary least squares, the least accurate non-degenerate age models we fit, produce the most informative residuals (+0.0124 and +0.0124). Gradient boosting and the neural network, which predict age far better, produce the least (+0.0063 and +0.0056). Across the six model families the Pearson correlation between out-of-fold R^2 and C-index gain is -0.94 ($p = 0.005$; Spearman -0.77). Going from ridge regression to the neural network improves age prediction by forty per cent relative and cuts the usefulness of the residual by more than half.

Figure 4a shows why this is not a paradox. Within the linear model class, traced along a ridge penalty path, accuracy and usefulness rise *together*: as the penalty falls and out-of-fold R^2 climbs from 0.017 to 0.398, the C-index gain rises to a maximum of +0.0133 around $R^2 = 0.39$ (correlation 0.89 along the path). A model too weak to capture the age-related drift in blood chemistry leaves that drift in the residual, where it is redundant with chronological age. So some accuracy is necessary.

The reversal appears only when one leaves that model class. The nonlinear families sit far to the right of the ridge path and far below its trend line. The extra accuracy they buy comes from resolving age-specific structure — curvature, interactions, threshold effects — that carries little prognostic weight, and in fitting it they absorb into the prediction variation that a linear model would have left in the residual. Accuracy on chronological age is thus not a proxy for the quality of a biological-age measure in either direction: it helps until the model can represent the age trend, and hurts thereafter.

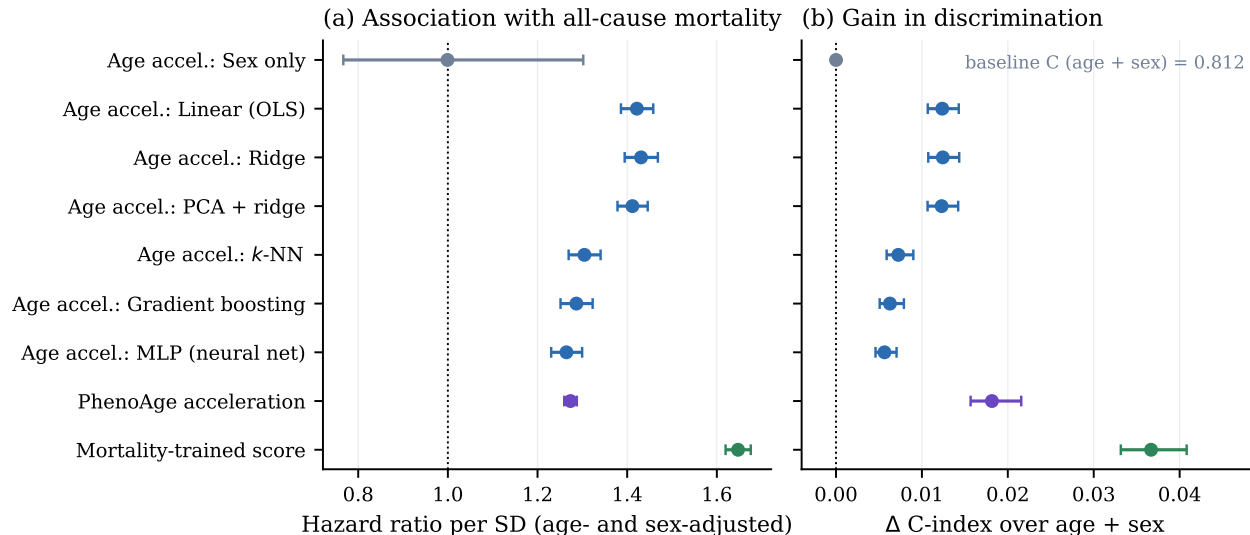


Figure 3: All-cause mortality. Each score is added to a Cox model already containing chronological age and sex. (a) Hazard ratio per standard deviation. (b) Change in Harrell’s C-index, with 95% bootstrap percentile intervals. Blue: age acceleration from chronological-age models. Purple: PhenoAge acceleration. Green: the score trained directly on mortality from the same 38 analytes.

5.5 The ceiling is set by the training target, not by the data

The same analytes, fitted against mortality rather than against age, behave completely differently. The mortality-trained score — an out-of-fold ridge-penalised Cox score built from the identical 39 inputs and residualised on age in exactly the same way — carries a hazard ratio of 1.647 per standard deviation (1.620–1.676) and raises the C-index by +0.0367 (95% CI +0.0332, +0.0408), from 0.8117 to 0.8484. That is 3.0 times the gain from the best age-trained residual and 6.5 times the gain from the neural network’s, using the same measurements on the same people. Its ten-year AUC is 0.867 against 0.827 for age and sex alone. PhenoAge acceleration, also outcome-trained but using only nine markers with coefficients estimated in a different NHANES sample two decades ago, sits in between at 1.274 and +0.0182.

Table 4 makes the point in the sharpest available way. If age acceleration had captured the mortality-relevant part of the blood panel, adding the mortality-trained score to a model that already contains age acceleration should add little. It adds +0.0310 (95% CI +0.0277, +0.0348) on top of the neural network’s age acceleration and +0.0304 on top of gradient boosting’s — 85% of its standalone gain of +0.0367. The age-trained residual has not extracted the mortality signal; it has mostly discarded it. Consistently, the correlation between the neural network’s age acceleration and the mortality-trained score is only 0.33 (11% of variance), against 0.70 (48%) for PhenoAge acceleration. Adding the mortality-trained score on top of PhenoAge acceleration still gains +0.0145, indicating that even the published nine-marker formula leaves recoverable signal in the wider panel.

Figure 5 shows the same fact without any modelling. Among participants aged 55–74 at examination — a narrow band in which chronological age can explain little of the contrast — survival curves separate modestly by quartile of age acceleration and sharply by quartile of the mortality-trained score.

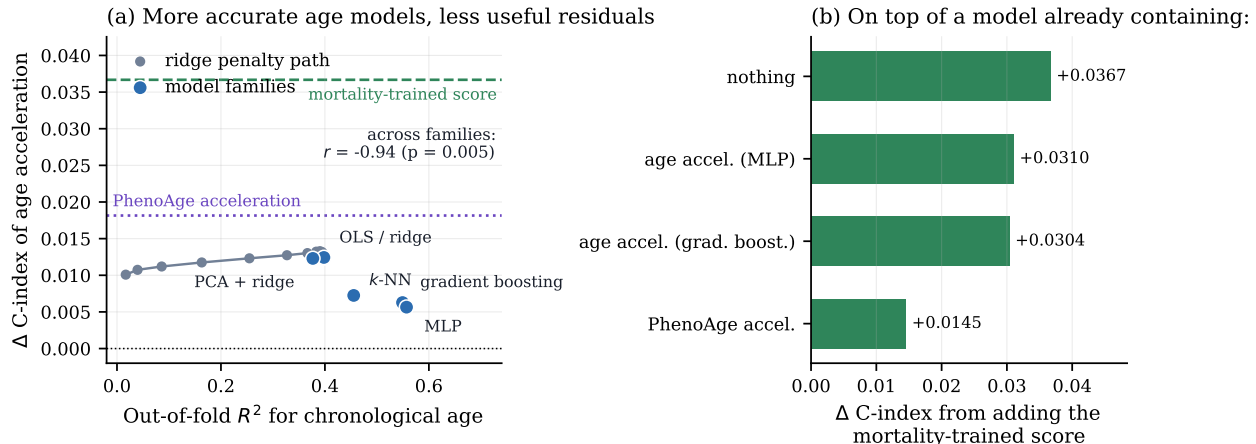


Figure 4: Accuracy against usefulness. (a) Grey points trace a ridge penalty path within one model class, where accuracy and usefulness rise together; blue points are the six model families, where they do not. Horizontal lines mark PhenoAge acceleration and the mortality-trained score built from the same analytes. (b) The C-index gain from adding the mortality-trained score to a model that already contains each of the listed scores.

5.6 Why: the two objectives lean on different analytes

Figure 6 plots, for each input, how strongly it correlates with chronological age against how strongly it predicts mortality once age and sex are controlled. The rank correlation between the two magnitudes is $+0.02$ ($p = 0.91$); using multivariable coefficients instead gives $+0.09$ ($p = 0.61$). Knowing how well a blood marker tracks age tells one essentially nothing about how well it predicts death.

The individual rankings are instructive. Blood urea nitrogen is the single most age-correlated analyte in the panel (rank 1 of 39) and ranks 30 for mortality. Albumin is the strongest mortality marker (rank 1) and only 15 for age; red cell distribution width ranks 2 for mortality and 12 for age. Both are long-established mortality markers (Goldwasser and Feldman, 1997; Patel et al., 2010; Perlstein et al., 2009) and both are components of PhenoAge. A model fit to chronological age is not a noisy version of a mortality model; it weights the evidence by a criterion that is close to orthogonal to the one that matters.

5.7 Sensitivity and cause-specific analyses

Table 5 repeats the primary comparisons under three alternative adjustment sets and under survey weighting. Adding race/ethnicity, education and the poverty-income ratio, or adding survey cycle, leaves the ordering and the magnitudes essentially unchanged. Survey-weighted estimates with cluster-robust variance tell the same story; we report unweighted models as primary because the question concerns the internal validity of a biomarker rather than a population parameter. Restricting every comparison to the CRP-available subsample, so that PhenoAge and the other scores are evaluated on identical participants, also leaves the conclusion intact: there the mortality-trained score gains $+0.0328$ against $+0.0051$ for the neural network’s age acceleration.

Table 6 splits deaths by underlying cause. For cardiovascular death (1,512 deaths) the mortality-trained score gains $+0.0336$ against $+0.0070$ for age acceleration; for cancer death (1,298 deaths) the corresponding figures are $+0.0212$ and $+0.0011$. Age acceleration is particularly weak for cancer mortality, which is consistent with the panel’s cancer signal residing in markers that are not strongly age-graded. In the heterogeneous residual category of other causes, which includes external causes

Participants aged 55–74 at examination, by quartile of each score

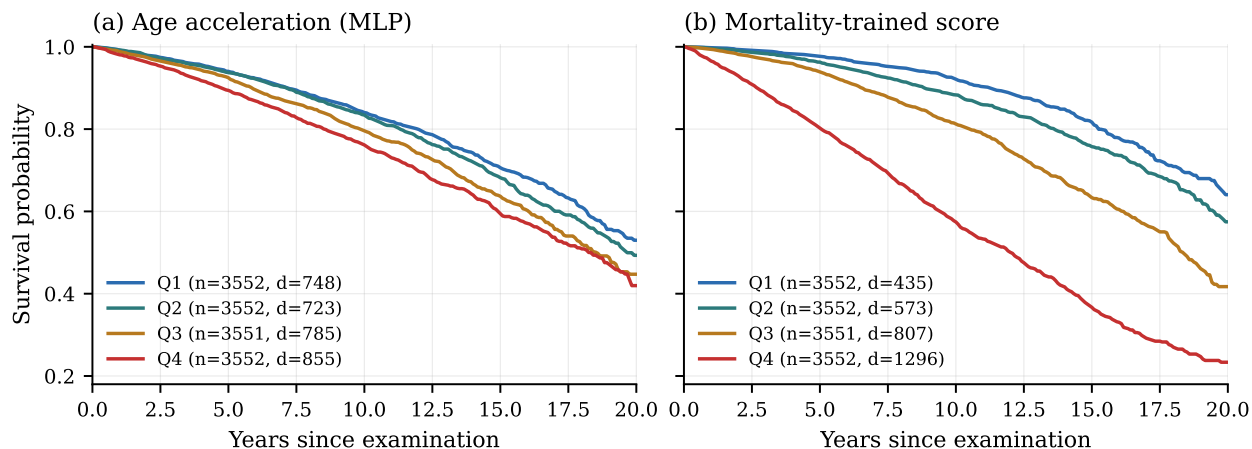


Figure 5: Kaplan–Meier survival by quartile of each score, restricted to participants aged 55–74 at examination so that chronological age is nearly constant across the groups being compared.

of death, PhenoAge acceleration is weakly inverse (hazard ratio 0.939, $p = 0.014$); we do not read anything into this beyond the limitations of a mixed category evaluated under competing risks.

5.8 Reconciling the R^2 gap: the age window, not the model

Our best adult out-of-fold R^2 of 0.557 is well below the roughly 0.78 reported by Kendiukhov (2020) on the same survey. Several things differ between the two analyses — the analyte panel, the evaluation protocol and the sample — and we can only test the ones we control. Having swept model class thoroughly and found a ceiling near 0.557 on adults, we asked how much of the gap the sample alone can account for. Figure 7 refits the same models over 8 different age windows.

The effect is large and monotone. Extending the lower bound down to age 12 — that is, including adolescents, in whom alkaline phosphatase, phosphorus and haemoglobin change steeply with growth and puberty — raises gradient-boosting R^2 from 0.544 to 0.683 and lowers the mean absolute error from 8.90 to 8.61 years. Narrowing the window to ages 20–45 pushes R^2 down to 0.247, with a mean absolute error of 5.47 years — a far “better” MAE attached to a far worse model. Across all windows, R^2 tracks the standard deviation of age in the sample with a correlation of 1.00.

Extending the window therefore closes most, though not all, of the gap: with adolescents included the same panel and the same learner reach 0.683, against 0.557 on adults alone and 0.78 in the earlier report. Whatever accounts for the remainder — a wider analyte set, a different split, a different age ceiling — the sample composition is plainly the dominant term.

This is not a claim that any published number is wrong. It is a statement about what the number means. An R^2 for age prediction is a joint property of the model and the age composition of the sample, so it cannot be used to compare clocks across studies, and a higher value does not indicate a better biological-age measure. Mean absolute error in years has the opposite pathology, falling as the sample narrows. Neither should be reported without the age range and its standard deviation.

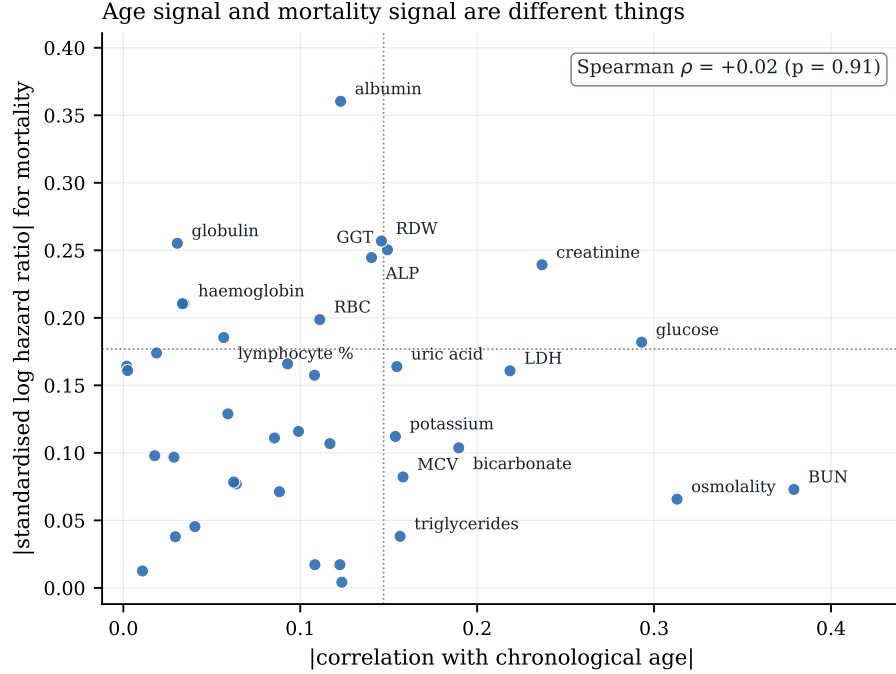


Figure 6: Age signal and mortality signal are different things. Horizontal axis: absolute correlation of each analyte with chronological age. Vertical axis: absolute standardised log hazard ratio for all-cause mortality, adjusted for age and sex. Dotted lines mark the upper tertiles; analytes in either upper region are labelled.

6 Discussion

6.1 What we found

Fitting blood chemistry to chronological age is a well-posed regression problem, and on NHANES adults flexible learners solve it appreciably better than linear ones. The residual from that regression is strongly associated with death. But it improves discrimination over chronological age and sex by at most $+0.0124$ in C-index, and the models that predict age best give the smallest gains of all. The same measurements fitted against mortality improve discrimination by $+0.0367$ and keep 85% of that value after the age residual is included. The information was in the blood panel; the chronological-age objective threw most of it away.

The mechanism is not subtle once stated. A regression on age distributes weight according to how strongly each analyte tracks age. Mortality risk is distributed differently, and in this panel almost independently: the rank correlation between the two is $+0.02$. Markers such as albumin and red cell distribution width are only weakly age-associated but strongly prognostic; the age model down-weights them, and its residual therefore contains little of what they carry. Nothing about a more flexible function class fixes this. Worse, flexibility makes it slightly harder, because the additional age-specific structure a nonlinear model captures is prognostically inert, and capturing it removes from the residual variation that would otherwise have remained there.

6.2 Relation to earlier work

Pyrkov et al. (2018) argued on theoretical grounds that pushing a deep network towards perfect chronological-age prediction would destroy the biological signal in its residual. Our contribution is a direct test of the consequence on public data with a hard endpoint: we hold the inputs and

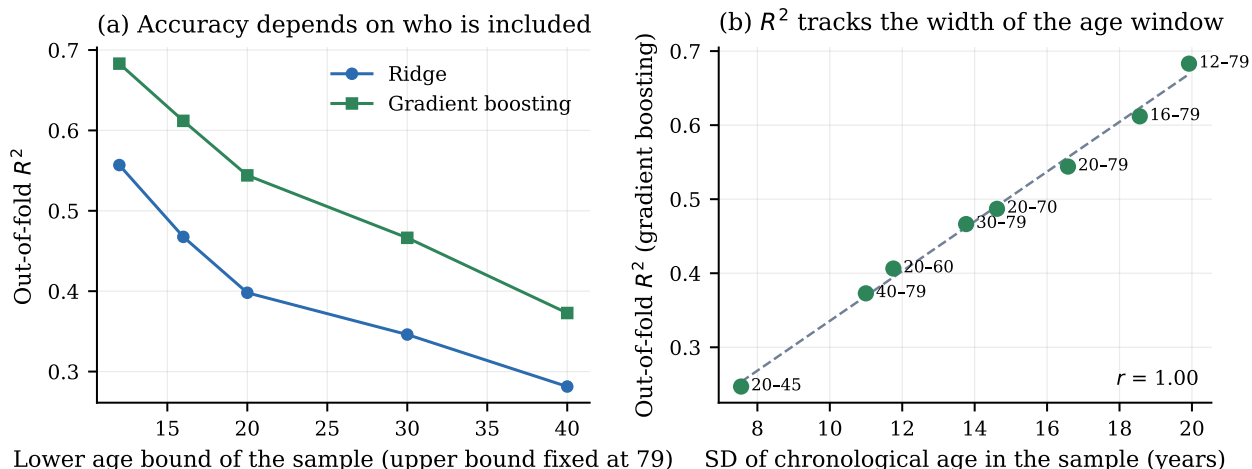


Figure 7: Reported accuracy is mostly a property of the sample. (a) Out-of-fold R^2 as the lower age bound is moved, with the upper bound fixed at 79. (b) R^2 against the standard deviation of chronological age across all windows tested; labels give the window.

the sample fixed, vary only the model class and the training target, and measure the difference in discrimination. We also separate two regimes that are easy to conflate. Within a model class, tuned by regularisation, accuracy and usefulness move together, which is why studies that tune a single model can reasonably conclude that better fit helps. Across model classes, they move apart. The practical implication is that the common strategy of swapping in a more powerful learner and reporting the improved R^2 is, for this purpose, moving in the wrong direction.

The finding also sits naturally alongside comparative studies of biological-age measures. Belsky et al. (2018) found that eleven quantifications of biological aging in the same cohort agreed poorly with one another; Li et al. (2020) found that outcome-trained measures dominated in mortality prediction; Hastings et al. (2019) found substantial disagreement among measures computed in NHANES itself. Our result explains part of that pattern: measures built against different targets are not noisy estimates of one latent quantity, they are estimates of different quantities.

More broadly, this is an instance of a recurring lesson about evaluation. Foundation-model representations of single-cell data appeared to carry aging and inflammation signal until the evaluation was made donor-aware and compared against honest baselines, at which point most of the apparent advantage disappeared (Kendiukhov, 2026). The pattern is the same here: an impressive number that answers a question nobody needs answered, standing in for the harder number that would settle the matter.

6.3 What this implies for practice

Three recommendations follow directly.

First, if the intended use of a clock is prognostic, train it against the prognosis. PhenoAge and GrimAge do this, and it is why they perform as they do. Building an age-prediction model and hoping its residual is prognostic is a detour that costs most of the signal — here, roughly two thirds of it.

Second, report incremental discrimination, not only hazard ratios. Every score we tested had an overwhelmingly significant hazard ratio. They differed by a factor of six in what they were actually worth. In a dataset with five thousand deaths a p -value says almost nothing about clinical value (Cook, 2007).

Third, stop treating R^2 against chronological age as a figure of merit. It is comparable across studies only when the age distributions match, which they rarely do; within a study it does not track the usefulness of the residual; and across model classes it tracks it negatively. If an age-prediction R^2 must be reported, it should be accompanied by the age range and its standard deviation, and it should not be presented as evidence that one clock is better than another.

6.4 What this does not show

It does not show that chronological-age models are useless. They are legitimate descriptions of how blood chemistry varies with age, and their coefficients are interpretable summaries of that variation — which is what the 2020 study set out to obtain, and obtained. Nor does it show that biological age is an empty concept: the mortality-trained score here does carry information beyond chronological age, and second-generation clocks built on richer substrates do better still (Lu et al., 2019; McCrory et al., 2021). Outcome-trained scores on other substrates behave the same way: a metabolomic score selected against death outperforms conventional risk factors for short-term mortality (Deelen et al., 2019), and blood-marker dynamics followed longitudinally reveal a loss of resilience that a single cross-sectional fit cannot see (Pyrkov et al., 2021).

It also does not show that mortality is the only endpoint worth targeting. Morbidity, functional decline and healthspan are arguably more relevant to intervention trials (Justice et al., 2018; Moqri et al., 2024), and a measure optimised for death may not be optimal for them; outcome-trained aging measures have also been shown to predict acute clinical vulnerability (Kuo et al., 2021). Our claim is narrower and, we think, harder to argue with: whatever the endpoint, the model should be trained against it, and the evaluation should be reported in units that reflect decisions rather than in units that reflect sample size.

7 Limitations

Blood was measured once, at the examination, so we cannot separate a stable individual trait from transient illness at the time of the visit; repeated measures would sharpen every estimate here (Higgins-Chen et al., 2022; Li et al., 2020). Follow-up is right-censored at the end of 2019 and is much shorter for the later cycles, so recent cycles contribute few deaths; we adjust for cycle in a sensitivity analysis and the conclusions do not move. Complete-case restriction on all 38 analytes removed about seven per cent of otherwise eligible participants and may induce mild selection. The upper age bound of 79, forced on us by top-coding, excludes the oldest adults, in whom the relationships may differ. The architecture sweep covers a uniform random 360 of 1,296 configurations rather than the full grid, so the reported best is a lower bound on the class ceiling; the flatness of accuracy across depth and across the top of the sweep makes a materially higher ceiling unlikely. The mortality-trained score is a linear penalised Cox model; a more flexible survival learner might raise the ceiling further, which would strengthen rather than weaken our conclusion. PhenoAge is applied with its original published coefficients to a partly different population, which is the standard use but means its performance is not an upper bound for what its nine markers could achieve if refit. Finally, all analyses are observational and associational; none of these scores is shown to respond to intervention, which is the property a surrogate endpoint ultimately needs (Moqri et al., 2024).

8 Conclusion

We asked whether AI-predicted biochemical age predicts death. On 45,033 NHANES adults with 5,213 deaths, the answer is: barely, and much less than the same data can deliver. Age acceleration from chronological-age models adds at most +0.0124 to the C-index over chronological age and sex, and the more accurately a model predicts age the smaller that gain becomes — the correlation between R^2 and C-index gain across model families is -0.94 . The same analytes fitted against mortality add +0.0367 and keep 85% of that after age acceleration is accounted for. The bottleneck is the training target. A model that hits chronological age accurately is, for the purpose of predicting death, systematically the wrong model, and the metric conventionally used to judge such models — R^2 against chronological age — is largely a description of the sample rather than of the model.

Data and code availability

All data used here are public. NHANES survey and laboratory files are distributed by the National Center for Health Statistics at <https://wwwn.cdc.gov/nchs/nhanes/>, and the public-use Linked Mortality Files at https://ftp.cdc.gov/pub/HEALTH_STATISTICS/NCHS/datalinkage/linked_mortality/. The codebase that downloads the source files, builds the analytic dataset, fits every model and reproduces every figure and table in this paper will be released upon acceptance.

Declarations

Funding. No funding was received for this work.

Competing interests. The author declares no competing interests.

Ethics approval. This analysis uses de-identified public-use data and did not require additional ethical review. The NHANES protocols were approved by the NCHS Research Ethics Review Board and all participants gave written informed consent.

Table 1: Characteristics of the analytic sample. Cells are mean (SD) unless stated otherwise. Deaths are from the NCHS public-use linked mortality file, with follow-up through 31 December 2019.

	All	Alive at censoring	Died
Participants, n	45,033	39,820	5,213
Age at examination, years	47.3 (16.6)	45.1 (15.8)	63.9 (12.5)
Male, %	48.2	46.9	58.3
Race/ethnicity			
Mexican American, %	18.7	19.0	16.0
Other Hispanic, %	8.6	9.1	4.6
Non-Hispanic White, %	42.4	41.0	52.8
Non-Hispanic Black, %	20.8	20.6	22.9
Other / multiracial, %	9.5	10.3	3.7
Median follow-up, years	9.6	9.8	7.5
Albumin, g/dL	4.23 (0.37)	4.25 (0.36)	4.15 (0.38)
Red cell distribution width, %	13.14 (1.33)	13.11 (1.28)	13.44 (1.62)
White blood cells, $10^3/\mu\text{L}$	7.29 (2.97)	7.27 (2.93)	7.45 (3.28)
Creatinine, mg/dL	0.89 (0.46)	0.86 (0.35)	1.10 (0.90)
Glucose, mg/dL	101.1 (38.4)	99.3 (35.4)	115.1 (54.6)

Table 2: Chronological-age prediction from 38 blood analytes and sex, adults aged 20–79, NHANES 1999–2018 ($n = 45,033$). All figures are computed on out-of-fold predictions from 5-fold nested cross-validation; hyper-parameters were selected in an inner 3-fold loop within each training fold.

Model family	R^2	MAE (y)	RMSE (y)	Pearson r
Sex only (reference)	0.001	14.28	16.57	0.032
Linear (OLS)	0.398	10.48	12.86	0.631
Ridge	0.398	10.48	12.86	0.631
PCA + ridge	0.377	10.68	13.09	0.614
k -nearest neighbours	0.455	9.93	12.23	0.690
Gradient boosting	0.549	8.83	11.13	0.741
MLP (neural network)	0.557	8.69	11.03	0.747

Table 3: All-cause mortality. Each row adds one standardised score to a Cox model already containing chronological age and sex; the C-index of that reference model is 0.8117. Hazard ratios are per one standard deviation of the score. ΔC is the change in Harrell’s C-index relative to the reference model, with a 95% percentile interval from 250 bootstrap resamples. All scores are out-of-fold; every p -value for a score other than the sex-only reference is below 10^{-60} . PhenoAge acceleration is evaluated on the subsample with a C-reactive protein measurement ($n = 35,486$; 4,633 deaths).

Score added to age + sex	Deaths	HR per SD (95% CI)	C-index	ΔC	95% CI
Age acceleration, by model family					
Sex only (reference)	5,213	0.999 (0.77–1.30)	0.8117	+0.0000	(-0.0001, +0.0002)
Linear (OLS)	5,213	1.422 (1.39–1.46)	0.8241	+0.0124	(+0.0107, +0.0143)
Ridge	5,213	1.431 (1.39–1.47)	0.8242	+0.0124	(+0.0107, +0.0143)
PCA + ridge	5,213	1.412 (1.38–1.45)	0.8240	+0.0123	(+0.0107, +0.0142)
k -NN	5,213	1.305 (1.27–1.34)	0.8190	+0.0072	(+0.0059, +0.0090)
Gradient boosting	5,213	1.287 (1.25–1.32)	0.8180	+0.0063	(+0.0051, +0.0079)
MLP (neural net)	5,213	1.265 (1.23–1.30)	0.8174	+0.0056	(+0.0046, +0.0071)
PhenoAge acceleration	4,633	1.274 (1.26–1.29)	0.8377	+0.0182	(+0.0157, +0.0216)
Mortality-trained score, same analytes	5,213	1.647 (1.62–1.68)	0.8484	+0.0367	(+0.0332, +0.0408)

Table 4: The mortality-trained score keeps adding information after age acceleration has been included, which shows that age acceleration does not capture the mortality signal already present in the same 38 analytes. Confidence intervals are 250-replicate bootstrap percentile intervals.

Reference model	C before	C after	ΔC (95% CI)	HR per SD
age + sex	0.8117	0.8484	+0.0367 (+0.0332, +0.0408)	1.647
age + sex + age acceleration (MLP)	0.8174	0.8484	+0.0310 (+0.0277, +0.0348)	1.633
age + sex + age acceleration (gradient boosting)	0.8180	0.8484	+0.0304 (+0.0268, +0.0339)	1.632
age + sex + PhenoAge acceleration	0.8377	0.8522	+0.0145 (+0.0117, +0.0185)	1.707

Table 5: Sensitivity analyses. Hazard ratios are per standard deviation of the score. The survey-weighted models use MEC examination weights divided by ten (the number of pooled two-year cycles) with variance estimation clustered on stratum-by-primary-sampling-unit; a C-index is not reported for weighted fits.

Score	HR per SD	p	ΔC
<i>Adjusted for age+sex</i>			
Age acceleration (MLP)	1.265	$< 10^{-16}$	+0.0056
Age acceleration (gradient boosting)	1.287	$< 10^{-16}$	+0.0063
Mortality-trained score	1.647	$< 10^{-16}$	+0.0367
<i>Adjusted for age+sex+race+educ+PIR</i>			
Age acceleration (MLP)	1.260	$< 10^{-16}$	+0.0050
Age acceleration (gradient boosting)	1.279	$< 10^{-16}$	+0.0056
Mortality-trained score	1.602	$< 10^{-16}$	+0.0290
<i>Adjusted for age+sex+cycle</i>			
Age acceleration (MLP)	1.268	$< 10^{-16}$	+0.0058
Age acceleration (gradient boosting)	1.292	$< 10^{-16}$	+0.0065
Mortality-trained score	1.646	$< 10^{-16}$	+0.0367
<i>Adjusted for age+sex (survey weighted)</i>			
Age acceleration (MLP)	1.254	$< 10^{-16}$	—
Age acceleration (gradient boosting)	1.283	$< 10^{-16}$	—
Mortality-trained score	1.862	$< 10^{-16}$	—

Table 6: Cause-specific mortality, from the underlying-cause-of-death recode in the linked mortality file. Models are adjusted for chronological age and sex; deaths from other causes are treated as censored, so these are cause-specific hazards and are subject to competing-risk depletion of the risk set. PhenoAge acceleration is evaluated on the smaller CRP-available subsample, hence its lower death counts. “All other causes” is a heterogeneous residual category that includes external causes, and estimates for it should be read with care.

Score	Deaths	HR per SD (95% CI)	p	ΔC
<i>cardiovascular (UCOD 1,5)</i>				
Age acceleration (MLP)	1,512	1.327 (1.261–1.397)	$< 10^{-16}$	+0.0070
Mortality-trained score	1,512	1.643 (1.593–1.695)	$< 10^{-16}$	+0.0336
PhenoAge acceleration	1,342	1.276 (1.250–1.301)	$< 10^{-16}$	+0.0168
<i>cancer (UCOD 2)</i>				
Age acceleration (MLP)	1,298	1.124 (1.065–1.186)	$< 10^{-4}$	+0.0011
Mortality-trained score	1,298	1.507 (1.445–1.571)	$< 10^{-16}$	+0.0212
PhenoAge acceleration	1,153	1.220 (1.184–1.257)	$< 10^{-16}$	+0.0096
<i>all other causes</i>				
Age acceleration (MLP)	2,403	1.310 (1.258–1.363)	$< 10^{-16}$	+0.0086
Mortality-trained score	2,403	1.706 (1.667–1.746)	$< 10^{-16}$	+0.0479
PhenoAge acceleration	2,138	0.939 (0.892–0.987)	0.014	-0.0034

A The analyte panel

Table 7: The 38-analyte panel. Means and standard deviations are on the original measurement scale. “ r with age” is the marginal Pearson correlation with chronological age. β_{age} is the standardised weight in a linear model of chronological age on all analytes and log HR the standardised coefficient from a Cox model for all-cause mortality containing all analytes and chronological age; both are multivariable quantities and are inflated and mutually cancelling for the algebraically derived members of the panel (osmolality from sodium, glucose and urea nitrogen; globulin from total protein and albumin), which is why the analysis in Section 5.6 uses marginal associations instead. Analytes marked $\dagger$ were modelled on the natural-log scale.

Analyte	Unit	Mean	SD	r with age	β_{age}	log HR
Albumin	g/dL	4.23	0.37	-0.123	-2.176	-0.172
Alkaline phosphatase $\dagger$	IU/L	71.77	26.48	+0.140	+0.814	+0.084
Aspartate aminotransferase $\dagger$	IU/L	25.36	19.09	+0.064	+2.001	+0.132
Alanine aminotransferase $\dagger$	IU/L	25.59	24.80	-0.018	-3.649	-0.289
Blood urea nitrogen $\dagger$	mg/dL	13.24	5.62	+0.379	+6.732	-0.045
Bicarbonate	mmol/L	24.76	2.37	+0.190	+1.953	-0.041
Total calcium	mg/dL	9.41	0.38	+0.029	+0.921	+0.010
Total cholesterol	mg/dL	195.78	41.80	+0.117	+1.436	-0.051
Chloride	mmol/L	103.34	2.99	-0.108	-0.746	-0.121
Creatinine $\dagger$	mg/dL	0.89	0.46	+0.237	+0.880	+0.066
Globulin	g/dL	2.98	0.46	+0.031	-0.600	+0.102
Glucose $\dagger$	mg/dL	101.12	38.44	+0.293	+5.679	+0.071
Gamma-glutamyl transferase $\dagger$	IU/L	30.28	44.93	+0.149	+1.521	+0.174
Iron $\dagger$	$\mu\text{g/dL}$	85.35	36.11	+0.002	-0.604	-0.034
Potassium	mmol/L	3.99	0.34	+0.154	+0.846	+0.085
Lactate dehydrogenase $\dagger$	IU/L	134.16	33.30	+0.219	+1.239	+0.103
Sodium	mmol/L	139.21	2.41	+0.099	+6.645	-0.011
Osmolality	mmol/kg	278.21	5.33	+0.313	-6.817	+0.078
Phosphorus	mg/dL	3.71	0.57	-0.085	-1.553	+0.043
Total bilirubin $\dagger$	mg/dL	0.66	0.31	+0.029	+1.457	-0.002
Total protein	g/dL	7.22	0.49	-0.062	-0.150	-0.031
Triglycerides $\dagger$	mg/dL	150.28	114.70	+0.156	+1.933	+0.030
Uric acid	mg/dL	5.39	1.45	+0.155	+0.162	+0.090
White blood cell count $\dagger$	$10^3/\mu\text{L}$	7.29	2.97	-0.093	-0.780	+0.071
Lymphocytes	%	30.58	8.56	-0.057	-1.347	-0.112
Monocytes $\dagger$	%	7.86	2.27	+0.122	+0.804	+0.017
Segmented neutrophils	%	58.08	9.58	+0.002	-0.587	+0.008
Eosinophils $\dagger$	%	2.82	2.13	+0.088	+0.532	-0.020
Basophils $\dagger$	%	0.71	0.47	+0.040	+0.142	-0.012
Red blood cell count	$10^6/\mu\text{L}$	4.69	0.51	-0.111	-0.219	-0.063
Haemoglobin	g/dL	14.14	1.55	-0.034	-6.969	-0.002
Haematocrit	%	41.74	4.37	-0.019	+5.993	+0.016
Mean cell volume	fL	89.23	5.79	+0.158	+1.428	+0.145
Mean cell haemoglobin	pg	30.23	2.36	+0.108	+2.327	+0.032
Mean cell haemoglobin conc.	g/dL	33.86	0.99	-0.059	+0.079	-0.071
Red cell distribution width	%	13.14	1.33	+0.146	+2.189	+0.156
Platelet count	$10^3/\mu\text{L}$	253.85	66.91	-0.123	-1.181	-0.079
Mean platelet volume	fL	8.21	0.92	-0.011	-0.785	-0.034

References

- Solim Essomandan Clémence Bafei and Chong Shen. Biomarkers selection and mathematical modeling in biological age estimation. *npj Aging*, 9:13, 2023. doi: 10.1038/s41514-023-00110-8.
- Nir Barzilai, Jill P. Crandall, Stephen B. Kritchevsky, and Mark A. Espeland. Metformin as a tool to target aging. *Cell Metabolism*, 23(6):1060–1065, 2016. doi: 10.1016/j.cmet.2016.05.011.
- Daniel W. Belsky, Terrie E. Moffitt, Alan A. Cohen, David L. Corcoran, Morgan E. Levine, et al. Eleven telomere, epigenetic clock, and biomarker-composite quantifications of biological aging: do they measure the same thing? *American Journal of Epidemiology*, 187(6):1220–1230, 2018. doi: 10.1093/aje/kwx346.
- Daniel W Belsky, Avshalom Caspi, David L Corcoran, Karen Sugden, Richie Poulton, et al. DunedinPACE, a DNA methylation biomarker of the pace of aging. *eLife*, 11:e73420, 2022. doi: 10.7554/eLife.73420.
- Ellyn R. Butler, Andrew Chen, Rabie Ramadan, Trang T. Le, Kosha Ruparel, et al. Pitfalls in brain age analyses. *Human Brain Mapping*, 42(13):4092–4101, 2021. doi: 10.1002/hbm.25533.
- Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010.
- Brian H. Chen, Riccardo E. Marioni, Elena Colicino, Marjolein J. Peters, Cavin K. Ward-Caviness, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging*, 8(9):1844–1865, 2016. doi: 10.18632/aging.101020.
- Evangelia Christodoulou, Jie Ma, Gary S. Collins, Ewout W. Steyerberg, et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22, 2019. doi: 10.1016/j.jclinepi.2019.02.004.
- Alan A. Cohen, Emmanuel Milot, Jian Yong, Christopher L. Seplaki, et al. A novel statistical approach shows evidence for multi-system physiological dysregulation during aging. *Mechanisms of Ageing and Development*, 134(3–4):110–117, 2013. doi: 10.1016/j.mad.2013.01.004.
- James H. Cole and Katja Franke. Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends in Neurosciences*, 40(12):681–690, 2017. doi: 10.1016/j.tins.2017.10.001.
- Nancy R. Cook. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115(7):928–935, 2007. doi: 10.1161/CIRCULATIONAHA.106.672402.
- David R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2):187–202, 1972. doi: 10.1111/j.2517-6161.1972.tb00899.x.
- Cameron Davidson-Pilon. lifelines: survival analysis in Python. *Journal of Open Source Software*, 4(40):1317, 2019. doi: 10.21105/joss.01317.
- Ann-Marie G. de Lange and James H. Cole. Commentary: correction procedures in brain-age prediction. *NeuroImage: Clinical*, 26:102229, 2020. doi: 10.1016/j.nicl.2020.102229.
- Joris Deelen, Johannes Kettunen, Krista Fischer, der Spek Ashley van, Stella Trompet, et al. A metabolic profile of all-cause mortality risk identified in an observational study of 44,168 individuals. *Nature Communications*, 10:3346, 2019. doi: 10.1038/s41467-019-11311-9.
- Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, New York, 1993.
- Luigi Ferrucci, Marta Gonzalez-Freire, Elisa Fabbri, Eleanor Simonsick, et al. Measuring biological aging in humans: a quest. *Aging Cell*, 19(2):e13080, 2020. doi: 10.1111/accel.13080.

- L. P. Fried, C. M. Tangen, J. Walston, A. B. Newman, C. Hirsch, et al. Frailty in older adults: evidence for a phenotype. *The Journals of Gerontology: Series A*, 56(3):M146–M157, 2001. doi: 10.1093/gerona/56.3.M146.
- David Furman, Judith Campisi, Eric Verdin, Pedro Carrera-Bastos, Sasha Targ, et al. Chronic inflammation in the etiology of disease across the life span. *Nature Medicine*, 25(12):1822–1832, 2019. doi: 10.1038/s41591-019-0675-0.
- Alessandro Gialluisi, Castelnuovo Augusto Di, Simona Costanzo, Marialaura Bonaccio, et al. Exploring domains, clinical implications and environmental associations of a deep learning marker of biological ageing. *European Journal of Epidemiology*, 37(1):35–48, 2022. doi: 10.1007/s10654-021-00797-7.
- Philip Goldwasser and Joseph Feldman. Association of serum albumin and mortality risk. *Journal of Clinical Epidemiology*, 50(6):693–703, 1997. doi: 10.1016/S0895-4356(97)00015-2.
- Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Advances in Neural Information Processing Systems*, volume 35, pages 507–520, 2022.
- Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, 49(2):359–367, 2013. doi: 10.1016/j.molcel.2012.10.016.
- Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4):361–387, 1996. doi: 10.1002/(SICI)1097-0258(19960229)15:4<361::AID-SIM168>3.0.CO;2-4.
- Waylon J. Hastings, Idan Shalev, and Daniel W. Belsky. Comparability of biological aging measures in the National Health and Nutrition Examination Survey, 1999–2002. *Psychoneuroendocrinology*, 106:171–178, 2019. doi: 10.1016/j.psyneuen.2019.03.012.
- Albert T. Higgins-Chen, Kyra L. Thrush, Yunzhang Wang, Christopher J. Minter, et al. A computational solution for bolstering reliability of epigenetic clocks: implications for clinical trials and longitudinal tracking. *Nature Aging*, 2(7):644–661, 2022. doi: 10.1038/s43587-022-00248-2.
- Nicholas Holzschek, Cassandra Falckenhayn, Jörn Söhle, Boris Kristof, et al. Modeling transcriptomic age using knowledge-primed artificial neural networks. *npj Aging and Mechanisms of Disease*, 7:15, 2021. doi: 10.1038/s41514-021-00068-5.
- Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):R115, 2013. doi: 10.1186/gb-2013-14-10-r115.
- Clifford L. Johnson, Ryne Paulose-Ram, Cynthia L. Ogden, et al. National Health and Nutrition Examination Survey: analytic guidelines, 1999–2010. *Vital and Health Statistics, Series 2*, 2(161):1–24, 2013.
- Jamie N. Justice, Luigi Ferrucci, Anne B. Newman, Vanita R. Aroda, et al. A framework for selection of blood-based biomarkers for geroscience-guided clinical trials: report from the TAME Biomarkers Workgroup. *GeroScience*, 40(5–6):419–436, 2018. doi: 10.1007/s11357-018-0042-y.
- Juulia Jylhävä, Nancy L. Pedersen, and Sara Hägg. Biological age predictors. *EBioMedicine*, 21:29–36, 2017. doi: 10.1016/j.ebiom.2017.03.046.
- Ihor Kendiukhov. AI-based investigation of molecular biomarkers of longevity. *Biogerontology*, 21(6):731–744, 2020. doi: 10.1007/s10522-020-09890-y.
- Ihor Kendiukhov. Inflammation-linked aging signals in frozen single-cell foundation models: donor-aware detection and robustness testing. *Biogerontology*, 27(4):132, 2026.

- Petr Klemra and Stanislav Doubal. A new approach to the concept and computation of biological age. *Mechanisms of Ageing and Development*, 127(3):240–248, 2006. doi: 10.1016/j.mad.2005.10.004.
- Chia-Ling Kuo, Luke C Pilling, Janice L Atkins, Jane A H Masoli, João Delgado, et al. Biological aging predicts vulnerability to COVID-19 severity in UK Biobank participants. *The Journals of Gerontology: Series A*, 76(8):e133–e141, 2021. doi: 10.1093/gerona/glab060.
- Dayoon Kwon and Daniel W. Belsky. A toolkit for quantification of biological age from blood chemistry and organ function test data: BioAge. *GeroScience*, 43(6):2795–2808, 2021. doi: 10.1007/s11357-021-00480-5.
- Trang T. Le, Rayus T. Kuplicki, Brett A. McKinney, Hung-Wen Yeh, et al. A nonlinear simulation framework supports adjusting for age when analyzing BrainAGE. *Frontiers in Aging Neuroscience*, 10:317, 2018. doi: 10.3389/fnagi.2018.00317.
- Morgan E. Levine. Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age? *The Journals of Gerontology: Series A*, 68(6):667–674, 2013. doi: 10.1093/gerona/gls233.
- Morgan E. Levine, Ake T. Lu, Austin Quach, Brian H. Chen, Themistocles L. Assimes, et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging*, 10(4):573–591, 2018. doi: 10.18632/aging.101414.
- Xia Li, Alexander Ploner, Yunzhang Wang, Patrik KE Magnusson, Chandra Reynolds, et al. Longitudinal trajectories, correlations and mortality associations of nine biological ages across 20-years follow-up. *eLife*, 9:e51507, 2020. doi: 10.7554/eLife.51507.
- Zuyun Liu, Pei-Lun Kuo, Steve Horvath, Eileen Crimmins, Luigi Ferrucci, et al. A new aging measure captures morbidity and mortality risk across diverse subpopulations from NHANES IV: a cohort study. *PLOS Medicine*, 15(12):e1002718, 2018. doi: 10.1371/journal.pmed.1002718.
- Ake T. Lu, Austin Quach, James G. Wilson, Alex P. Reiner, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging*, 11(2):303–327, 2019. doi: 10.18632/aging.101684.
- Carlos López-Otín, Maria A. Blasco, Linda Partridge, Manuel Serrano, et al. Hallmarks of aging: an expanding universe. *Cell*, 186(2):243–278, 2023. doi: 10.1016/j.cell.2022.11.001.
- Polina Mamoshina, Kirill Kochetov, Eugene Lane, Franco Cortese, et al. Population specific biomarkers of human aging: a big data study using South Korean, Canadian, and Eastern European patient populations. *The Journals of Gerontology: Series A*, 73(11):1482–1490, 2018. doi: 10.1093/gerona/gly005.
- Riccardo E Marioni, Sonia Shah, Allan F McRae, Brian H Chen, Elena Colicino, et al. DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biology*, 16:25, 2015. doi: 10.1186/s13059-015-0584-6.
- Cathal McCrory, Giovanni Fiorito, Belinda Hernandez, Silvia Polidoro, et al. GrimAge outperforms other epigenetic clocks in the prediction of age-related clinical phenotypes and all-cause mortality. *The Journals of Gerontology: Series A*, 76(5):741–749, 2021. doi: 10.1093/gerona/glaa286.
- Mahdi Moqri, Chiara Herzog, Jesse R. Poganik, Jamie Justice, Daniel W. Belsky, et al. Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell*, 186(18):3758–3775, 2023. doi: 10.1016/j.cell.2023.08.003.
- Mahdi Moqri, Chiara Herzog, Jesse R. Poganik, Kejun Ying, Jamie N. Justice, et al. Validation of biomarkers of aging. *Nature Medicine*, 30(2):360–372, 2024. doi: 10.1038/s41591-023-02784-9.
- Eiji Nakamura and Kiyoshi Miyao. A method for identifying biomarkers of aging and constructing an index of biological age in humans. *The Journals of Gerontology: Series A*, 62(10):1096–1105, 2007. doi: 10.1093/gerona/62.10.1096.

- National Center for Health Statistics. The linkage of National Center for Health Statistics survey data to the National Death Index — 2019 linked mortality file: methodology overview and analytic considerations. Technical report, National Center for Health Statistics, Hyattsville, MD, 2022.
- Hamilton Se-Hwee Oh, Jarod Rutledge, Daniel Nachun, Róbert Pálovics, Olamide Abiose, et al. Organ aging signatures in the plasma proteome track health and disease. *Nature*, 624(7990): 164–172, 2023. doi: 10.1038/s41586-023-06802-1.
- K. V. Patel, R. D. Semba, L. Ferrucci, A. B. Newman, L. P. Fried, et al. Red cell distribution width and mortality in older adults: a meta-analysis. *The Journals of Gerontology: Series A*, 65(3): 258–265, 2010. doi: 10.1093/gerona/glp163.
- Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- Michael J. Pencina and Ralph B. D’Agostino. Overall C as a measure of discrimination in survival analysis: model specific population value and confidence interval estimation. *Statistics in Medicine*, 23(13):2109–2123, 2004. doi: 10.1002/sim.1802.
- M. S. Pepe et al. Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *American Journal of Epidemiology*, 159(9):882–890, 2004. doi: 10.1093/aje/kwh101.
- Todd S. Perlstein, Jennifer Weuve, Marc A. Pfeffer, and Joshua A. Beckman. Red blood cell distribution width and mortality risk in a community-based prospective cohort. *Archives of Internal Medicine*, 169(6):588–594, 2009. doi: 10.1001/archinternmed.2009.55.
- Evgeny Putin, Polina Mamoshina, Alexander Aliper, Mikhail Korzinkin, Alexey Moskalev, et al. Deep biomarkers of human aging: application of deep neural networks to biomarker development. *Aging*, 8(5):1021–1033, 2016. doi: 10.18632/aging.100968.
- Timothy V. Pyrkov, Konstantin Slipensky, Mikhail Barg, Alexey Kondrashin, et al. Extracting biological age from biomedical data via deep learning: too much of a good thing? *Scientific Reports*, 8:5210, 2018. doi: 10.1038/s41598-018-23534-9.
- Timothy V. Pyrkov, Konstantin Avchaciov, Andrei E. Tarkhov, Leonid I. Menshikov, Andrei V. Gudkov, et al. Longitudinal analysis of blood markers reveals progressive loss of resilience and predicts human lifespan limit. *Nature Communications*, 12:2765, 2021. doi: 10.1038/s41467-021-23014-1.
- Jarod Rutledge, Hamilton Oh, and Tony Wyss-Coray. Measuring biological age using omics data. *Nature Reviews Genetics*, 23(12):715–727, 2022. doi: 10.1038/s41576-022-00511-7.
- Luke Sagers, Luke Melas-Kyriazi, Chirag J. Patel, and Arjun K. Manrai. Prediction of chronological and biological age from laboratory data. *Aging*, 12(9):7626–7638, 2020. doi: 10.18632/aging.102900.
- Elizabeth Selvin, Jane Manzi, Lesley A. Stevens, Lente Frederick Van, et al. Calibration of serum creatinine in the National Health and Nutrition Examination Surveys (NHANES) 1988–1994, 1999–2004. *American Journal of Kidney Diseases*, 50(6):918–926, 2007. doi: 10.1053/j.ajkd.2007.08.020.
- Ravid Shwartz-Ziv and Amitai Armon. Tabular data: deep learning is not all you need. *Information Fusion*, 81:84–90, 2022. doi: 10.1016/j.inffus.2021.11.011.
- Stephen M. Smith, Diego Vidaurre, Fidel Alfaro-Almagro, Thomas E. Nichols, et al. Estimation of brain age delta from brain imaging. *NeuroImage*, 200:528–539, 2019. doi: 10.1016/j.neuroimage.2019.06.017.
- Ewout W. Steyerberg, Andrew J. Vickers, Nancy R. Cook, Thomas Gerds, et al. Assessing the

- performance of prediction models: a framework for traditional and novel measures. *Epidemiology*, 21(1):128–138, 2010. doi: 10.1097/EDE.0b013e3181c30fb2.
- Hajime Uno, Tianxi Cai, Michael J. Pencina, Ralph B. D’Agostino, and L. J. Wei. On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in Medicine*, 30(10):1105–1117, 2011. doi: 10.1002/sim.4154.
- Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7:91, 2006. doi: 10.1186/1471-2105-7-91.
- Andrew J. Vickers and Elena B. Elkin. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making*, 26(6):565–574, 2006. doi: 10.1177/0272989X06295361.
- Yan Zhang, Rory Wilson, Jonathan Heiss, Lutz P. Breitling, et al. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nature Communications*, 8:14617, 2017. doi: 10.1038/ncomms14617.
- Alex Zhavoronkov and Polina Mamoshina. Deep aging clocks: the emergence of AI-based biomarkers of aging and longevity. *Trends in Pharmacological Sciences*, 40(8):546–549, 2019. doi: 10.1016/j.tips.2019.05.004.