European Actuarial Journal

Biological Age for Prevention in Insurance

--Manuscript Draft--

Manuscript Number:	EUAJ-D-24-00095R2
Full Title:	Biological Age for Prevention in Insurance
Article Type:	Original Research Paper
Keywords:	Biological age; Self-protection; prevention; NHANES; Life table
Corresponding Author:	Oleksandr SOROCHYNSKYI Lyon 1 University: Universite Claude Bernard Lyon 1 FRANCE
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Lyon 1 University: Universite Claude Bernard Lyon 1
Corresponding Author's Secondary Institution:	
First Author:	Oleksandr SOROCHYNSKYI
First Author Secondary Information:	
Order of Authors:	Oleksandr SOROCHYNSKYI
	Frédéric Planchet
	Edouard Debonneuil
	François Robin-Champigneul
Order of Authors Secondary Information:	
Funding Information:	
Abstract:	Biological age (BA) offers a promising approach for encapsulating complex health information into a single interpretable metric. This study evaluates BA methods as tools for prevention in insurance, focusing on their ability to predict mortality and disease incidence. Using National Health and Nutrition Examination Sur vey (NHANES) data, we compare five BA calculation methods—multiple linear regression (MLR), Klemera-Doubal Method (KDM), PhenoAge, calibrated Phe noAge, and Random Forest (RF). We include a practical application of estimating death counts from life tables. Our findings reveal that RF and calibrated PhenoAge consistently outperform other methods in mortality prediction and more accurately estimate observed death counts. While MLR and KDM lag in predictive performance, they demon strate interpretability that may be valuable for some applications. PhenoAge showed the greatest flexibility and adaptability for prevention-focused applica tions, particularly for estimating death counts. However, a key challenge remains in calibrating BA methods to align with absolute mortality risks, as highlighted by their initial biases in estimating death counts. Weargue that BA's primary value lies in its dual role: a reliable risk estimator and an effective communication tool for promoting preventive health behaviors. By addressing calibration issues and tailoring BA methods to specific insurance con texts, this research underscores BA's potential to improve prevention programs, aligning health incentives for both policyholders and insurers.
Response to Reviewers:	Dear Editor, Thank you for the thorough and constructive review of our manuscript. We appreciate the reviewer's recognition that our work "could offer a nuanced understanding grounded in objective methods" and could be "of interest to the readership of the European Actuarial Journal." We have carefully considered all comments and have made substantial revisions to address the concerns raised. Please find below a detailed point-by-point

response to the reviewer's comments.

1. Temporal and Population Context

> The NHANES database is reported to be based on 'survey cycles between 1999 and
> 2018'. Which year is used as the base year for the selected biological markers?

We use data from all NHANES survey cycles between 1999 and 2018. The biological markers are measured across all these cycles, with biological age models trained on the combined data (excluding the test set). Mortality follow-up extends through 2019, resulting in varying follow-up periods for participants across different survey cycles.

> It is reported on page 13 that 'characteristics measured at the time of
> examination remain constant'. What are the potential implications of assuming
> the underlying covariates remain unchanged?

This assumption is a necessary limitation of the data structure. We measure biological markers at a single point in time during the survey examination, then assess their association with mortality outcomes that may occur years later. Essentially, we are asking: "How do markers measured at age 30 predict mortality risk from age 30 onwards (up to 20 years of follow-up in this dataset)?" This differs from a hypothetical repeated-measures design where we would ask: "How do current marker values affect current mortality risk?"—which would require longitudinal covariate measurements at multiple time points.

This limitation does not undermine our objective of assessing whether biological age can predict mortality in general, rather than estimating specific effects of individual markers. We chose not to censor follow-up times to maintain comparability with existing NHANES-based literature, particularly @levine_epigenetic_2018 and @qiu_explainable_2023.

> In Section 3, the calculated BAs are presented and compared across different
> methods. The underlying population/cohort and relevant time are not clear. For
> example, are the results for all individuals or stratified by sex, and are the
> BAs relevant to a single year, e.g. 2019, or averaged over 1999–2018?

Biological age models are fitted separately for each sex using all available NHANES cycles (1999-2018), excluding the test subset. These sex-specific models are then applied to estimate biological ages for the entire test subset, and all metrics presented in Section 3 (now called Section 4 "Evaluation of Biological Age Methods") are calculated on this test set. The models thus capture patterns across the entire time period rather than representing a single year.

> Linked to the previous two comments, the interpretation of the results may
> change. For example, challenges regarding the interpretability of modelling
> outputs are noted on page 15 due to smaller BA numbers compared to calendar
ages
> (CAs). In more recent years, with better health care, diet, and exercise, BAs
> would be expected to be lower—indicating a younger and healthier biological
> state—than the corresponding CAs, compared with earlier years. A comparison of
> trends over time would help put the discussion into context. Has this been
> already considered in the analysis? A discussion on this would be useful.

While examining temporal trends in biological age as a population health indicator is certainly valuable—and historically one of BA's first applications—we deliberately chose not to pursue this analysis for two reasons. First, NHANES purposefully oversamples certain population segments to address specific policy questions, making it non-representative of the general US population without survey-weight adjustments. Since our goal is to demonstrate BA's usefulness as a general methodological approach rather than to estimate health indicators for the US population specifically, we treat NHANES as an abstract population and do not apply survey weights. Second, trends in US population health over this period are complex; while life expectancy increased, health-adjusted indicators have stagnated in the 21st century, making it unclear a priori what trends we should expect in BA. These considerations led us away from this line of inquiry.

We adjusted the phrasing describing the NHANES data to better explain the 'constant covariates' assumption, clarifying that biological markers are measured once during survey examination and used to predict mortality throughout follow-up. We also added a short summary in the Biological Age Comparison section specifying that BA models are fitted on NHANES cycles 1999-2018, stratified by sex, with metrics computed on the test set.

2. Biomarkers Consistency Across Methods

> On page 7, it is stated that 'we calculate BAs on the same four sets of biomarkers for all methods'. Is this different from the biomarkers reported in Tables 1 and 2? Or, perhaps, what is meant are the response variables, such as CAs, all-cause mortality, and so on?

The "four sets of biomarkers" refers to the four covariate sets described in Section 3.2 (Biological Markers). We have clarified this passage to avoid confusion---these are the predictor variables (biological markers/covariates), not the response variables (CA or mortality).

> Related to this, would it not be useful to see the model outputs based on identical biomarkers across the different methods—even in the Supporting Information? Although each method may have an optimal set of biomarkers (as listed in Tables 1 and 2), using the same biomarkers could provide a fairer comparison in terms of the predictive power of these methods.

Indeed, all biological age methods are applied to all four biological marker sets, allowing us to both (i) compare how each method performs across different covariate regimes, and (ii) compare methods head-to-head on the same marker sets. This is why all biological age comparison plots contain 24 data points: 6 methods (5 BA methods + chronological age as reference) × 4 marker sets.

We revised the passage introducing the biological age comparison to clarify that "biomarkers" refers to the four covariate sets described in Section 3.2, not to response variables or method-specific optimal marker selections.

3. Data Imputation and Completeness

> Could you please comment on the completeness of the data and the percentage of missing values in the underlying cohort used for the main results?

NHANES data exhibits a specific missingness pattern often referred to as "block-missingness": the set of biological markers measured varies across survey cycles. Some markers are only available starting from certain cycles, others are discontinued, and some are measured intermittently (measured, then absent for several cycles, then measured again). Within-cycle missingness for measured variables is relatively small, typically due to survey design decisions (e.g., grip strength only measured for those aged 40+).

To provide concrete numbers: the overall missingness rate for variables used in the models is 34%. Variables appear in an average of 7.8 NHANES cycles (out of 10). The within-cycle missingness rate, considering only cycles where the variable is measured, is 19%.

> Have you checked the distribution of related cohort(s) by age and year before and after imputation? This is important to ensure that model bias is avoided.

Yes, we followed the diagnostic procedures recommended by @van_buuren_flexible_2018, including checks for convergence of MICE chains and comparison of means and standard deviations of variables by year before and after imputation. Some modeling choices were informed by these diagnostics, notably the chain length (30 iterations) and the exclusion of certain variables that behaved erratically or caused instability in other variables.

However, we note two important points: First, changes in distribution are not necessarily problematic---when missingness is not random, the distribution of missing items may legitimately differ from observed items. Second, it may not be appropriate to speak of "bias" in the traditional sense, as our goal is not to create an imputed dataset representative of the United States population, but rather to generate plausible datasets that allow us to compare BA methods under realistic conditions.

We added concrete missingness statistics to the Multiple Imputation section, clarifying the block-missingness pattern in NHANES data (34% overall missingness, 19% within-cycle). We also documented the diagnostic procedures followed to validate the imputation process, including convergence checks and distributional comparisons across years.

4. Readability and Flow

Introduction

> The introduction is detailed, but too long and somewhat repetitive.
> - I suggest focusing on a more concise introduction, with greater emphasis on the contribution of this study, where the narrative could be centred on BA and its value as a preventative measure.
> - A brief summary of the following sections could also be useful.
> - Section 1.1 could be shortened and combined with Section 1.3. Some of the arguments in Section 1.1 can be moved to Discussion, e.g. Section 5.4, to tighten the concluding remarks.

We restructured the introduction into a single, flowing narrative without subsections, reducing its length by approximately 30% while emphasizing this study's contribution more prominently. We streamlined the prevention program discussion while retaining essential motivating examples.

> The NHANES dataset is first mentioned in Section 1.3. It is not clear which population this dataset covers or what NHANES stands for.

We now introduce NHANES earlier in the revised introduction, providing its full name (National Health and Nutrition Examination Survey) and specifying that it is a large-scale U.S. health examination survey.

> Incidence rates and BA are compared in terms of interpretation, but an example on mortality is provided. Why not use an example more aligned with the argument?

Mortality is the primary outcome we consider in this study. The example uses mortality rates to illustrate that when absolute risk remains low (as it does for mortality throughout most of the lifespan), expressing differences as biological age better communicates the relative magnitude of risk change than probability differences alone.

Data and Methods Organization

> The NHANES dataset is introduced under Section 2, which primarily covers
> methods. I suggest introducing it in a separate section, e.g. Section 3 titled
> 'Data'. Furthermore, a data imputation method, MICE, is implemented. This is
> important and should be explained more clearly in the main text. Currently,
> there are references to 'imputation' but no clear description of how missing
> data are handled. I note a separate section in the Supplementary Information.
> Perhaps the method and its purpose could be briefly explained in the main text,
> with full details left for the Supplementary Information.

We have restructured the manuscript to place NHANES data in a separate "Data" section (now Section 3), distinct from the methods sections. In this section, we now provide key missingness statistics (as detailed in our response to your earlier comment) and briefly explain that we use multiple imputation to address this missingness, with the motivation that the relatively high missingness rate requires propagating uncertainty rather than using simpler imputation approaches. Full methodological details of the MICE implementation remain in the Supplementary Materials.

Results Section

> Section 3 reports the numerical results. A broader title can be more
> appropriate than 'Biological Age Comparison'.

We have renamed the sections to better reflect their scope. Section 4 is now titled "Evaluation of Biological Age Methods" (previously "Biological Age Comparison") and Section 5 remains "Using BA to Estimate Death Counts." This structure maintains the distinction between methodological evaluation and practical application.

Methods Overview

> There are four methods listed in the abstract but five methods explained in
> Section 2. It would be useful to provide a table that clearly maps the different
> methods to their names and acronyms.

We have added a table (Table 1) that clearly maps all five BA calculation methods to their names, acronyms, primary outcomes, and source references. Moreover, the 'Methods' section has been renamed 'Biological Ages' to reflect its new focus.

Tables and Supporting Information

> Table 1 could be moved to the Supporting Information. It would also be helpful
> to have a look-up table in the main text summarising which table in the
> Supporting Information corresponds to each method.

We have moved the detailed marker lists to the Supplementary Materials. In the main text, the table providing a summary of the four marker sets with their names, number of variables, selection criteria now also includes explicit references to the corresponding detailed tables in the Supplementary Materials.

Terminology and Statistical Measures

> There are comments and captions based on R^2, such as 'Association with the
> number of diseases (R^2)'. The wording here sounds imprecise. Would it be
> possible to use more accurate terminology to put this into context? For example,
> is this about how much variability in the response variable is explained by BA?

We have revised figure captions to use more precise terminology. Captions now explicitly state that R^2 represents the variance in the outcome variable (chronological age, disease count, etc.) explained by biological age.

Self-rated Health and Other Variables

> Self-rated health is mentioned twice in the main text, including Section 2,
> but the related outcomes do not contribute to the narrative. A brief mention in
> the Discussion would be sufficient. I also note that the corresponding R² values
> are below 5%. It is therefore difficult to view this as a 'modest' contribution,
> as suggested on page 23. A similar comment applies to behavioural and social
> factors.

We have removed behavioral and social factors from the manuscript entirely. Self-rated health is retained in the Supplementary Materials as it is commonly examined in biological age studies and provides a useful benchmark for comparison with prior work. However, we now provide only a brief discussion of these weak associations (R² < 0.05) in the Discussion section, focusing the narrative on mortality and disease associations where biological ages show substantive improvements over chronological age.

5. Technical Corrections and Proofreading

> The manuscript requires proofreading and copy-editing.
> - Acronyms must be defined before use and then used consistently thereafter,
> without repeated re-definitions.
> - References do not include publication years. Please double-check the entries.
> Notation is not always coherent or intuitive. For example, t is often used for
> time, not age. It would be clearer to use more intuitive notation, such as a for
> age. Then, in $F_j(t)$ on page 8, t is referred to as 'a certain age'. Is this CA
> or one of the candidate BAs? Also, s_j is defined as the residual standard error
> for the jth marker on page 8, but then s_CA is referred to as 'an estimate from a
> regression on CA'. This is confusing and should be clarified.
> - The Supporting Information appears to continue the numbering of the main text,
> which is not standard practice.

Thank you for this feedback. We went over the manuscript making numerous adjustments. Notably, we slightly reformulated equations as well as their descriptions, and removed redefinitions of acronyms. These points as well as other typographical errors have been addressed.

We believe these revisions have substantially improved the manuscript and address all the reviewer's concerns. We hope the revised manuscript now meets the standards for publication in the European Actuarial Journal and look forward to your decision.

Biological Age for Prevention in Insurance

Oleksandr Sorochynskyi [1,2], Frédéric Planchet[1,2],
Édouard Debonneuil[3], François Robin-Champigneul[4]

[1]University of Lyon, Université Claude Bernard Lyon 1, ISFA,
Laboratoire SAF EA2429, , 69366 Lyon France.
[2]Prim'Act actuarial consulting firm, 42 avenue de la Grande Armée,
75017 Paris, France.
[3]ActuRx actuarial consulting firm, France.
[4]Independent researcher.

Contributing authors: oleksandr.sorochynskyi@primact.fr;

Abstract

Biological age (BA) offers a promising approach for encapsulating complex health information into a single interpretable metric. This study evaluates BA methods as tools for prevention in insurance, focusing on their ability to predict mortality and disease incidence. Using National Health and Nutrition Examination Survey (NHANES) data, we compare five BA calculation methods—multiple linear regression (MLR), Klemera-Doubal Method (KDM), PhenoAge, calibrated PhenoAge, and Random Forest (RF). We include a practical application of estimating death counts from life tables.

Our findings reveal that RF and calibrated PhenoAge consistently outperform other methods in mortality prediction and more accurately estimate observed death counts. While MLR and KDM lag in predictive performance, they demonstrate interpretability that may be valuable for some applications. PhenoAge showed the greatest flexibility and adaptability for prevention-focused applications, particularly for estimating death counts. However, a key challenge remains in calibrating BA methods to align with absolute mortality risks, as highlighted by their initial biases in estimating death counts.

We argue that BA's primary value lies in its dual role: a reliable risk estimator and an effective communication tool for promoting preventive health behaviors. By addressing calibration issues and tailoring BA methods to specific insurance contexts, this research underscores BA's potential to improve prevention programs, aligning health incentives for both policyholders and insurers.

Keywords: Biological age, Self-protection, Prevention, NHANES, Life table

1 Introduction

Biological age (BA) offers a promising approach for encapsulating complex health information into a single interpretable metric. While chronological age (CA) serves as a proxy for biological aging, factors such as genetics, behavior, and environment cause individuals to age at different rates. BA methods attempt to capture this variation by combining biological markers through statistical models to produce an age measure that better reflects underlying health status and outcomes such as mortality risk.

Beyond general descriptions, BA lacks a precise, universally accepted definition. Instead, following Klemera and Doubal [1], we characterize BA as a combination of biological markers and a statistical method to synthesize them into an age-interpretable quantity. This characterization leads to numerous possible BAs, and the aging process itself has many causes and consequences, making it unlikely that a single BA could capture all aspects equally well. The exercise is thus best viewed pragmatically: finding combinations of markers and methods that provide useful predictions for specific quantities of interest [2].

Historical approaches to BA construction have evolved considerably. Early methods used multiple linear regression (MLR) or principal component analysis to combine biological markers [3, 4]. More principled constructions emerged, notably the Klemera-Doubal Method (KDM) [1]. The field transformed with omics-based markers: Horvath [5] introduced DNA methylation-based BA, while Levine et al. [6] developed PhenoAge by explicitly optimizing for mortality prediction, giving rise to "second-generation epigenetic clocks." More recently, machine learning approaches have been applied to improve model quality [7].

Most BA methods involve regressing an outcome on biomarkers, differing primarily in: (1) the outcome used (e.g., chronological age or mortality), (2) the biomarkers included (e.g., clinical measures or epigenetic data), and (3) the statistical method applied. Considerable work has catalogued various BAs and their properties [7–9], with studies comparing their associations with mortality and health outcomes [10, 11].

This study complements prior work by providing a quantitative comparison of BA calculation methods themselves, rather than pre-defined BAs. Unlike Hastings et al. [10], we evaluate how methods perform when applied to different biomarker sets, using four distinct sets to examine consistency and robustness. We assess methods across multiple outcomes including mortality prediction, association with CA, and chronic disease burden. Crucially, we include a practical application: estimating death counts from life tables, which reveals calibration challenges that must be addressed for actuarial applications.

Our focus on insurance and prevention contexts is motivated by BA's dual role. In insurance, BA can serve as a risk metric that, when properly calibrated, may replace CA in premium calculations while incorporating richer biological and lifestyle data. For prevention programs, BA's interpretability makes it an effective communication tool. The key assumption is that BA is easier to interpret than direct risk measures. For risks strongly associated with age and with relatively low incidence, presenting risk differences as biological vs. chronological age can be more striking than probability differences. For example, telling a policyholder their health profile corresponds to "biological age of 35" rather than "biological age 30" may resonate more strongly than explaining a shift from 0.25% to 0.30% annual risk, encouraging prevention when interventions are most effective.

BA's relevance to prevention extends beyond communication. Many insurers already operate prevention programs that reward healthy behaviors. Examples include Generali Vitality and UnitedHealthcare's UHC Rewards (employer-sponsored programs offering monetary rewards for physical activity and preventive care), and tools like the QalyDays calculator [12], which uses BA-like metrics to communicate long-term care insurance needs. Some programs reward specific activities directly (e.g., daily step counts), while others construct health scores—the "Vitality Age" being one example of a BA used in practice. These programs align with the economic framework of Ehrlich and Becker [13], which showed that insurance markets can increase demand for self-protection (risk-reducing behaviors) when premiums reflect such efforts. BA provides a natural mechanism for this alignment: factors influencing BA—lifestyle and behavior—are often directly tied to risk levels, allowing prevention programs to encourage healthier behaviors while aligning financial incentives with reduced risk.

We use data from the National Health and Nutrition Examination Survey (NHANES), a large-scale U.S. health examination survey conducted between 1999 and 2018 [14]. NHANES encompasses demographic, questionnaire, examination, and laboratory data, including diverse health indicators. This breadth allows flexibility in developing BAs for specific contexts. However, NHANES presents challenges such as varying biological marker sets across survey cycles. We address these through multiple imputation, ensuring data continuity and robust statistical comparisons—an approach not previously applied to NHANES for BA comparison purposes. While NHANES is valuable for initial implementation, longer-term applications in insurance will require more frequent and individualized data collection.

The remainder of this article is structured as follows. Section 2 describes BA calculation methods. Section 3 describes the data used in this article. Section 4 compares BA methods quantitatively across mortality prediction, CA association, and chronic disease burden. Section 5 examines the practical application of estimating death counts from life tables, highlighting calibration requirements. Section 6 discusses implications for prevention-focused applications, interpretability considerations, and future directions. These findings demonstrate BA's potential to serve as both a reliable risk estimator and an effective tool for promoting preventive health behaviors in insurance contexts.

2 Biological ages

In this article, we compare five biological age methods using a battery of criteria that evaluate the ages' associations with various outcomes, notably mortality. In order to gain insight into the inner workings of these BA methods across various regimes, we calculate each BA on four sets of biomarkers. This section describes these methods for calculating BA. Table 1 summarizes the five methods considered.

Each biological age is presented as described in the literature, along with brief commentary on potential variants that may be of interest in actuarial applications.

Among these is the question of how to estimate the uncertainty of BA itself—a prerequisite, in our view, for its practical use in individual risk assessment. While we outline possible approaches, a full treatment of this question, including the derivation of confidence intervals and posterior distributions, lies beyond the scope of this article.

Table 1 Biological age calculation methods evaluated in this study.

Method	Acronym	Primary outcome	Source
Multiple Linear Regression	MLR	Chronological age	[3]
Klemera-Doubal Method	KDM	Chronological age	[1]
Phenotypic Age	PhenoAge	Mortality	[6]
Calibrated Phenotypic Age	Calibrated PhenoAge	Mortality	This study
Random Forest	RF	Mortality	A variant of [15]

Throughout this section, we use the following notation:

- X is the matrix of available markers, with X_j representing the individual columns of this matrix,
- J is the total number of available markers,
- CA is the chronological age and BA is the biological age.

All BAs here are fit separately for each sex, following standard practice in prior works, although this fact is not reflected in notation.

2.1 Multiple Linear Regression

One of the earliest proposed methods for calculating BA was introduced in Hollingsworth et al. [3]. BA is calculated as the predicted value from a multiple linear regression (MLR) model with CA as the outcome and the observed markers as covariates. The model can be described by the following equation:

$$CA = \sum_{j=1}^{J} \beta_j X_j + \epsilon. \tag{1}$$

Where β_j are the model coefficients, and ϵ is the error term, assumed to be normally distributed with constant variance. Under this model, MLR BA is defined as the age predicted by this model,

$$BA := \sum_{j=1}^{J} \hat{\beta}_j X_j. \tag{2}$$

With $\hat{\beta}$ the least-squares estimate. By construction, this model ensures that the difference between CA and BA is zero on average.

This model is known to have some drawbacks. The choice of biomarkers is difficult, as the obvious criterion of being correlated with chronological age does not necessarily result in a useful BA. Consider a set of biological markers that is able to perfectly predict chronological age. In this case, BA is equal to CA and brings no additional information. This is known as the “biomarker paradox” [16]. Regression to the mean may also result in bias in observations far from the global age average [17].

This model can be trivially improved by drawing on the extensive literature on regression methods. Generalized linear models (GLMs) or generalized additive models (GAMs) can serve as drop-in replacements for the linear model used here, offering greater flexibility in capturing nonlinear relationships between age and biomarkers. A key advantage of this class of models is that the estimation uncertainty is readily available in the form of standard errors for the predicted values, allowing for direct quantification of the uncertainty in the estimated biological age.

2.2 Klemera and Doubal Method

Klemera and Doubal [1] introduced a method for calculating BA, now known as KDM. KDM is motivated by the need for a more principled construction of BA, particularly to avoid the “biomarker paradox.” The core idea behind KDM is to view BA as a latent variable that determines the values of observed biological markers. The problem is thus reversed. This method constructs a model for markers (and not CA, as in MLR), and then defines BA as the age that most plausibly generates the observed markers under this model. Formally, KDM-BA is the age t that explains the observed marker values under the model, i.e., the value that minimizes the distance between expected and observed biomarkers:

$$BA := \text{argmin}_t Q(t|X) = \text{argmin}_t \sum_{j=1}^{J} \alpha_j \left(F_j(t) - X_j\right)^2 . \tag{3}$$

Where:

- $Q(t|X)$ is a measure of how plausible the biological age t is given the markers X,
- α_j is the weight given to the j-th marker,
- $F_j(t) = E(X_j|t)$ is the conditional expectation of the j-th marker given an age.

Assuming that $F_j(t)$ is linear, i.e., $F_j(t) = k_j t + q_j$, and the root mean squared error after regressing X_j on CA is s_j, the BA expression has an explicit solution:

$$BA := \frac{\sum_{j=1}^{J}(X_j - q_j)\frac{k_j}{s_j^2} + \frac{\text{CA}}{s_{\text{BA}}^2}}{\sum_{j=1}^{J}\left(\frac{k_j}{s_j}\right)^2 + \frac{1}{s_{\text{BA}}^2}}. \tag{4}$$

Where s_{BA} is the root mean squared error of regressing CA on the hypothetical latent BA, in practice it provides a weight for CA. This is the formula most commonly associated with KDM.

KDM is noteworthy in that it treats CA in a manner similar to other biological markers. It is also noteworthy that if the above expression is expanded, it becomes obvious that KDM-BA is a linear combination of biological markers and CA, just as

MLR is, albeit with different weights and with CA included:

$$BA = \underbrace{\frac{-\sum_{j=1}^{J} q_j \frac{k_j}{s_j^2}}{\sum_{j=1}^{J} \left(\frac{k_j}{s_j}\right)^2 + \frac{1}{s_{\text{BA}}^2}}}_{\text{Constant}} + \underbrace{\sum_{j=1}^{J} \left(\frac{k_j/s_j^2}{\sum_{i=1}^{J} \left(\frac{k_i}{s_i}\right)^2 + \frac{1}{s_{\text{BA}}^2}} \right) X_j}_{\text{Linear combination of markers}} + \underbrace{\frac{1/s_{\text{BA}}^2}{\sum_{j=1}^{J} \left(\frac{k_j}{s_j}\right)^2 + \frac{1}{s_{\text{BA}}^2}} \text{CA}}_{\text{CA contribution}} \,. \tag{5}$$

As KDM assumes that markers are uncorrelated, all marker sets are first centered, reduced, and then transformed to the principal component basis. This eliminates all correlations from the marker set and insures that the linear independence assumption holds.

Potential improvements to KDM include relaxing the linearity assumption for $F_j(t)$ and allowing s_{BA} to vary with age, as suggested by Klemera and Doubal [1] and implemented by Levine [18]. In this study, we use KDM as presented above with linear $F_j(t)$ and constant s_{BA}.

Estimating the uncertainty associated with KDM-BA presents additional challenges due to the PCA transformation typically applied beforehand to decorrelate biomarkers. The simplest approach is to treat the PCA rotation matrix as fixed, allowing the regression models for each principal component to be handled independently. However, this ignores the uncertainty introduced by the PCA step itself. A more robust alternative is to apply a bootstrap procedure, re-estimating both the PCA and the regression models on resampled datasets. This captures the full variability of the estimated BA and can be applied to any BA method. Its main limitation is increased computational cost, especially compared to methods with analytically tractable posteriors.

2.3 PhenoAge

Levine et al. [6] introduced two novel clocks aimed at better association with mortality and morbidity: an epigenetic clock based on 513 DNA methylation markers, called "DNAm PhenoAge", and a clock based on 9 blood markers and CA, called "Phenotypic Age", which also serves as the outcome for fitting DNAm PhenoAge. Phenotypic Age is constructed by matching 10-year survival probability as estimated by a multivariate Gompertz proportional hazards model to the same probability in a univariate reference model. This BA is thus notable for choosing mortality as the criterion for its definition. In this study, we focus on the Phenotypic Age method (not the full DNAm PhenoAge) and apply it to various marker sets, which we shall simply refer to as "PhenoAge".

To formally define PhenoAge, consider the Gompertz proportional hazards model, which describes the hazard rate as a function of time-on-study, t. In this model, the hazard rate is given by:

$$h(t) = \alpha \exp\left(t/\beta + \gamma_0 \text{CA} + X\gamma\right) . \tag{6}$$

Where CA is the chronological age at $t = 0$, α is the shape parameter, β is the scale parameter, γ_0 is the coefficient associated with CA, and γ is a vector of coefficients for the remaining covariates. For the reference population, where only age is considered (i.e., no additional covariates), the Gompertz model takes the univariate form:

$$h'(t) = \alpha' \exp\left(t/\beta' + \gamma_0' \mathrm{CA}\right). \tag{7}$$

Here, α', β', and γ_0' represent the same parameters as their non-primed counterparts.

PhenoAge defines biological age as the age x under the reference survival distribution at which the 10-year survival probability matches that predicted by the full model. Let:

- S_{ref} be the function mapping age to 10-year survival probability under the reference model,
- and p be the 10-year survival probability predicted by the full model for a given individual.

Then PhenoAge is implicitly defined as the solution to:

$$S_{\mathrm{ref}}(\mathrm{BA}) = p, \quad \text{or equivalently,} \quad \mathrm{BA} = S_{\mathrm{ref}}^{-1}(p). \tag{8}$$

Expanding this expression yields an explicit formula for PhenoAge as a linear combination of CA and biomarkers:

$$BA := \frac{\log\left(\frac{\alpha\beta\left(e^{10/\beta}-1\right)}{\alpha'\beta'\left(e^{10/\beta'}-1\right)}\right)}{\gamma_0'} + \frac{\gamma_0}{\gamma_0'}\mathrm{CA} + \frac{1}{\gamma_0'}\sum_{j=1}^{J}\gamma_j x_j. \tag{9}$$

All parameters are estimated from the training data. In Levine et al. [6] the reference model is fit on the same dataset as the full model but without the covariates X. The covariates were selected by using a penalized Cox proportional hazards model, only keeping markers with non-zero coefficients at a penalty level chosen by cross-validation.

This formula changes little for a number of variants of PhenoAge. For one, changing the matched survival time from 10 years to any other time t only changes the constant term. If instead of matching survival probabilities, one wishes to match hazard functions or indeed the cumulative hazard functions, then it is sufficient to take the limit of $t \to 0$, which will again only impacts the constant term. Finally, the two final terms of PhenoAge equation are just the linear predictor of a proportion hazard model. This suggests an arguably simpler algorithm of fitting a Cox proportional hazards model, and then fitting a linear regression to CA with the linear predictor from the Cox model.

Estimating the uncertainty of PhenoAge presents additional challenges. Since both Gompertz models are fitted on the same dataset—one using only CA and the other including additional covariates—their parameter estimates are statistically dependent. As a result, their posterior distributions cannot be assumed to be independent, making direct sampling from the joint posterior nontrivial. The most practical solution is to apply a bootstrap procedure: repeatedly resampling the data, refitting both models, and recalculating PhenoAge for each resample. This approach provides an empirical posterior distribution that captures the full estimation uncertainty, though at the cost of additional computational effort.

2.4 Calibrated PhenoAge

The only novel BA considered in this paper is a variant of PhenoAge. This version replaces the reference distribution used in the original PhenoAge with a mortality table. The motivation is simple: if the resulting BA is then used to query the same mortality table, the survival probability predicted by the underlying Gompertz model is approximately recovered. We refer to this variant as Calibrated PhenoAge, as it is explicitly aligned with a given mortality table.

To formally define Calibrated PhenoAge, we reuse the notation from the previous section and additionally define S_{tab} as the map from age to 10-year survival probability under the considered mortality table. The calibrated version of PhenoAge then substitutes S_{tab} in place of the original S_{ref}, yielding:

$$\text{PhenoAge}_{\text{calib}} := S_{\text{tab}}^{-1}(p), \tag{10}$$

where p is the 10-year survival probability predicted by the full Gompertz model underlying the original PhenoAge.

This recalibration has a useful property: when the calibrated PhenoAge is plugged back into the same mortality table, the original survival probability is approximately recovered:

$$S_{\text{tab}}(\text{PhenoAge}_{\text{calib}}) = S_{\text{tab}}(S_{\text{tab}}^{-1}(p)) \approx p. \tag{11}$$

In practice minor deviations occur due to :

1. Age flooring : as is customary for chronological age, calibrated BAs are floored to the nearest integer before use,
2. Time horizon mismatch : the time horizon used to compute the BA may not match the horizon at which the survival probability is estimated.

2.5 Random Forest BA

The last BA method considered in this article is based on ENABL Age from Qiu et al. [15]. Qiu et al. [15] use Gradient Boosting Machines (GBM) to predict a mortality score, in this case a hazard ratio. This score is then transformed to a quantity interpretable as age by fitting a curve to CA, in essence, regressing CA on the mortality score. In this article we adapt a very similar approach, but use Random Forest as the underlying algorithm instead. Qiu et al. [15] defined the mortality score as the hazard ratio predicted by GBM. We, instead, use the sum of the cumulative hazard function evaluated at each distinct exit time in following Ishwaran et al. [19]. Once a valid mortality score is obtained, it is transformed to a scale interpretable as age by applying an exponential curve:

$$\text{mortality score} = \exp(a \cdot \text{CA} + b) + \min(\text{mortality score}) - c, \tag{12}$$

with real parameters a, b and a positive parameter c. The final BA is thus given by:

$$BA := \frac{\log(\text{mortality score} - \min(\text{mortality score}) + \hat{c}) - \hat{b}}{\hat{a}}. \tag{13}$$

Unlike other methods, this BA cannot be expressed as a linear combination of biological markers and CA. In fact, random forests is a non-parametric model does not have a simple closed form expression.

Note that the final transformation makes no explicit reference to the way mortality score is calculated. Indeed, this transformation can be applied to any risk score whatever. This makes this approach easy to adapt to results of regressions on the outcome of interest.

Estimating uncertainty in this approach is non-trivial. In principle, one could sample from individual trees in the Random Forest to approximate the variability of the mortality score. This can be combined with repeated estimation of the transformation curve to partially account for uncertainty in the final BA. While more efficient than full resampling, this approach still underestimates sampling variability. As with other complex BA models, a bootstrap procedure—repeating the entire pipeline, including Random Forest fitting and curve transformation, across resampled datasets—remains the most robust solution. Although computationally intensive, it flexibly captures all relevant sources of uncertainty.

3 Data

3.1 NHANES data

This article uses National Health and Nutrition Examination Survey (NHANES) data together with linked mortality data as of 2019 [14, 20]. NHANES is a large-scale health examination survey conducted in the United States. NHANES encompasses a wide range of demographic, questionnaire, examination, and laboratory data, including dietary habits, laboratory blood work, cardiovascular stress tests, dental health, grip strength, and many other health indicators. This extensive range of indicators, coupled with its public availability, makes NHANES an attractive choice for studies on BA.

This article uses survey cycles between 1999 and 2018. NHANES data exhibits block-missingness: the set of biological markers measured varies across survey cycles. The overall missingness rate for variables used for biological calculations is 34%, with variables appearing in an average of 7.8 out of 10 NHANES cycles. The within-cycle missingness rate, considering only cycles where a variable is measured, is 19%. These missing values are treated using multiple imputation which we describe in detail in Supplementary Materials, Section A.5. Multiple imputation was chosen over simpler imputation methods because the missingness rate is somewhat high, and we felt it more appropriate to propagate the uncertainty due to missing values.

In this article, we include only individuals aged 20 to 79. This age range is motivated by data availability constraints. First, while linked mortality data is available for individuals aged 18 and older, many NHANES data files only include information on adults, which is typically defined as 20 years and older. Second, those aged 80 and above are top-coded as 80 in some cycles (with some cycles beginning top-coding at 85) to preserve anonymity.

For mortality-based models (i.e., Cox and Gompertz), we assume that characteristics measured at the time of examination remain constant throughout the follow-up period. This assumption is necessary to exploit the available data and is required to keep results comparable to prior works, such as Levine et al. [6] and Qiu et al. [15].

3.2 Biological markers

The key choice for any BA method is the set of biological markers used to construct it. To enable a systematic comparison of how different BA methods behave under varying input conditions, we separate variable selection from BA construction. This modular approach allows us to assess each model's robustness across different marker sets and the impact of selection criteria, while also including the "natural" pairings where methods are matched with their preferred selection approaches.

We define four marker sets chosen from over 100 biological markers available in NHANES, including blood markers, physical examinations, and engineered features derived from available data. Socio-economic and dietary variables are excluded to maintain comparability with existing literature.

The first two sets are based on mortality-focused selection using LASSO Cox models. The first set ("Pheno") contains 9 blood biomarkers selected in Levine et al. [6] using penalized Cox regression—the natural selection method for PhenoAge. The second set ("LASSO Cox") applies the same methodology to our full NHANES dataset, yielding 15 markers with only 5 overlapping with the first set.

The third and fourth sets use penalized regression with chronological age as the outcome—the natural approach for MLR-based BA. The third set ("LASSO CA restricted") contains approximately the same number of variables as set 2, while the fourth ("LASSO CA full") uses the standard "1se" rule, resulting in by far the largest marker set of 79 markers.

As data are multiply imputed, the penalized regression models were first fit each imputation replication independently, then penalty paths were aggregated, and the optimal penalization level was chosen from the aggregated path. Then, variables that were chosen in more than half of imputation replications were chosen. This corresponds to the "majority" strategy proposed in Brand [21].

Table 2 summarizes the marker sets. The marker sets are listed in full in Supplementary Materials, Section A.4.

Table 2 The four selected marker sets, with corresponding names, number of biomarkers and the selection criterion.

Set number	Set name	Number of variables	Selection criterion	Table
1	Pheno	9	mortality	Table 3
2	LASSO Cox	15	mortality	Table 4
3	LASSO CA (restricted)	17	CA	Table 5
4	LASSO CA (full)	79	CA	Table 6

3.3 Train and test sets

Twenty percent of the data are reserved for testing, specifically for comparing the performance of various BA methods. These test data are not used either for imputations or for fitting the BA models themselves or for choosing the marker sets. This ensures that presented metrics estimate the performance of these BAs on a hypothetical new data drawn from the same population.

4 Evaluation of Biological Age Methods

This section provides a quantitative comparison of biological age methods across three primary criteria: association with chronological age, mortality prediction, and association with chronic diseases. These criteria reflect different applications of biological age in insurance contexts—mortality prediction for life insurance, chronic disease association for long-term care, and age association for interpretability.

The association is measured as the proportion of the variance explained, that is the R^2. Biological age models are fitted separately for each sex and multiple imputation replication, with the test set excluded. Data used includes NHANES cycles 1999-2018. All metrics are computed on the test set with 95% confidence intervals. R^2 values use Fisher transformation with Rubin's rules for multiple imputation. Concordance standard errors use infinitesimal jackknife estimation.

Our analysis reveals a fundamental trade-off in biological age methods: those optimized for mortality prediction (RF, PhenoAge) achieve higher concordance with survival outcomes but show weaker correlation with chronological age, while methods closely aligned with CA (KDM, MLR) maintain interpretability but offer limited predictive advantage over chronological age itself. Importantly, correlation with CA should not be viewed as a metric to maximize—perfect correlation yields chronological age itself, providing no new information. Rather, it serves as a soft constraint: biological age methods should maximize other criteria of interest while remaining sufficiently close to CA to maintain interpretability.

It is important to highlight a practical consideration: mortality-based methods require longitudinal data with follow-up, representing a higher data standard than cross-sectional approaches. This makes CA-based methods particularly valuable when mortality data is unavailable, and their performance is admirable considering they are not trained on mortality outcomes.

Mortality-optimized methods (RF and PhenoAge) achieve concordance scores significantly above chronological age (0.83–0.87 vs 0.81) but show deviations from CA—RF notably underestimates biological age below 35 and shows a plateau effect in younger adults. In contrast, age-based methods (KDM especially) maintain close correspondence with chronological age ($R^2 > 0.8$) but provide minimal mortality prediction advantage. Calibrated PhenoAge generally performs slightly worse than the original PhenoAge criteria, suggesting that calibration improves mortality prediction at the cost of all other metrics.

All biological age methods show weak association with chronic disease burden ($R^2 \approx 0.3$), suggesting that neither age or mortality serve as good criteria for morbidity estimation.

4.1 Association with Chronological Age

The relationship between biological age and chronological age reflects each method's interpretability and face validity. Figure 1 shows how average biological age varies with chronological age across different biomarker sets, while Figure 2 quantifies these relationships using R^2.

Most methods achieve R^2 values above 0.6, indicating reasonable correspondence with chronological age. However, substantial differences emerge:

KDM shows the strongest age alignment ($R^2 > 0.82$ across all biomarker sets, reaching 0.99 with the full LASSO CA set). This exceptionally high correlation stems from KDM's direct inclusion of CA, making it nearly indistinguishable from CA itself.

MLR demonstrates the most variable performance, with R^2 ranging from 0.21 (PhenoAge biomarkers) to 0.83 (full LASSO CA). Since MLR cannot directly use chronological age, its performance depends entirely on the available biomarkers' ability to capture age-related variation.

PhenoAge maintains strong age correlation (R^2 0.73–0.90) despite its mortality optimization, though correlation decreases with larger biomarker sets as it incorporates more mortality-specific rather than age-specific information.

RF shows unique age patterns, displaying a systematic underestimation of biological age, especially before 50. This pattern, while potentially reflecting true biological aging dynamics, poses interpretability challenges as it assigns biological ages below chronological age to much of the population.

4.2 Mortality Prediction

Figure 3 compares concordance with all-cause mortality using time-on-study timescale. The hierarchy clearly reflects each method's optimization target.

RF achieves the highest concordance (e.g., 0.87 with full LASSO Cox *biomarkers). The similar performance across these biomarker sets likely reflects the model's flexibility in capturing mortality-relevant information.

PhenoAge performs second-best (0.84–0.87), significantly outperforming chronological age. Its survival-model foundation enables it to capture mortality-relevant biological variation while maintaining reasonable age correspondence. Being less flexible than RF, it fits mortality less well but also deviates less from CA.

Age-based methods (KDM and MLR) show concordance scores close to chronological age (0.81), with small improvement despite their additional complexity. This suggests that strong age alignment may constrain the ability to capture mortality-relevant biological variation. KDM on the LASSO Cox is notable for performing particularly well, comparably with PhenoAge.

4.3 Chronic Disease Association

Figure 4 shows that all biological age methods have modest association with chronic disease burden ($R^2 \approx 0.3$), with only limited variation across methods and biomarker sets.

PhenoAge performs well across most settings, with KDM holding its own, particularly in the LASSO Cox and restricted LASSO CA biomarker sets. Random Forest

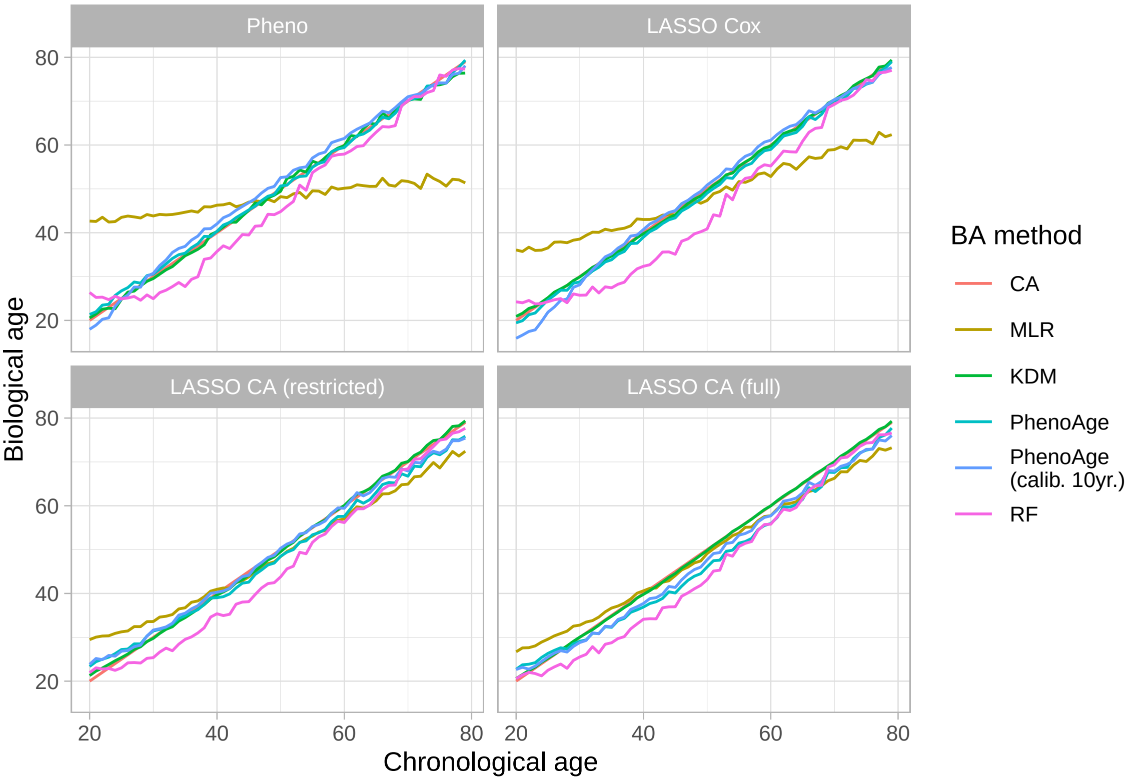


Fig. 1 Average biological age by chronological age, for each marker set. The chronological age (CA) line corresponds to the diagonal line.

underperforms in this task, highlighting that high predictive performance for mortality does not necessarily translate into strong association with disease burden.

A disease-specific breakdown is provided in the Supplementary Materials. Most disease-level associations are modest, ranging from near zero (e.g., for asthma or obesity) to about 0.25 (e.g., for hypertension). While CA slightly outperforms most BAs for cancer, hyperlipidemia, and arthritis, BAs—especially PhenoAge and RF—perform better for diabetes and kidney disease.

These results highlight that while some BA methods offer modest improvements over CA for disease burden, none capture morbidity especially well. Designing biological ages directly optimized for disease risk may be a more promising direction.

4.4 Implications for Method Selection

The choice among biological age methods depends critically on the intended application and available data.

For mortality prediction applications, RF and PhenoAge offer clear advantages, with RF providing the highest concordance. However, RF's unusual age patterns may limit interpretability in some contexts.

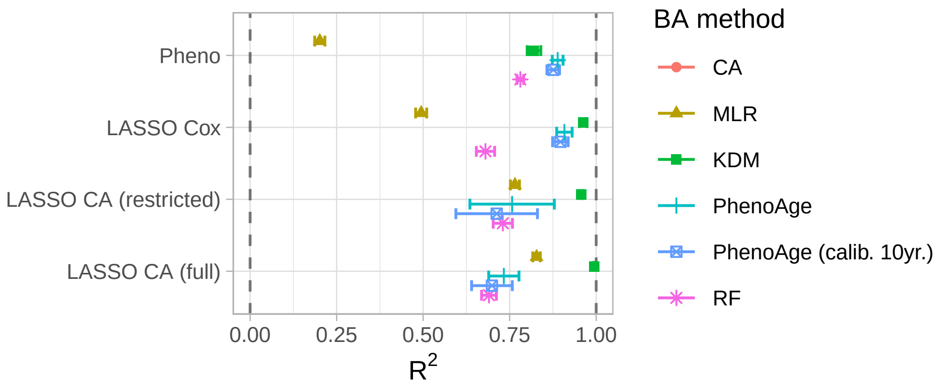


Fig. 2 Variance in Chronological Age Explained by Biological Age (R^2).

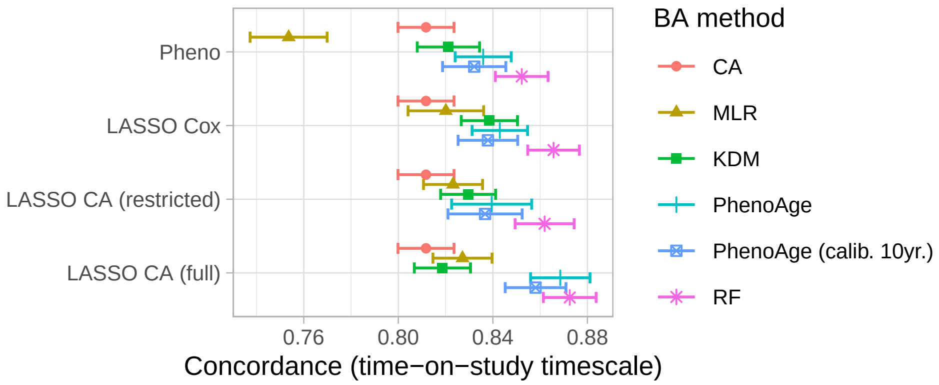


Fig. 3 Concordance with respect to all-cause mortality.

For applications with smaller datasets or when longitudinal follow-up is unavailable, KDM maintains a stronger age correspondence and has lesser data requirements. KDM performs surprisingly well considering it is not trained on mortality data, making it particularly valuable when mortality data is unavailable. However, KDM's extreme age correlation ($R^2 \approx 0.99$) in the full LASSO CA biomarker set highlights the importance of variable selection.

MLR presents a more complex choice, as it requires careful balance between providing sufficient variables for good performance while avoiding overfitting. The dramatic variation in performance across biomarker sets makes it more difficult to implement reliably.

For chronic disease applications, all methods show similar performance, suggesting that biological age based on either age or mortality may not be the optimal approach for chronic disease risk assessment.

These findings highlight that mortality-based training provides more robust performance criteria, but requires longitudinal data with follow-up—a much higher standard than cross-sectional approaches. CA-based methods remain valuable alternatives, especially given their lower data requirements.

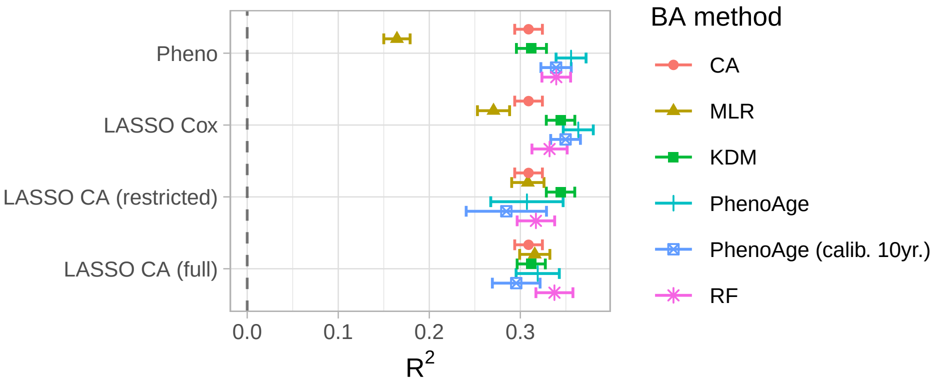


Fig. 4 Variance in Disease Count Explained by Biological Age (R^2).

5 Using BA to Estimate Death Counts

In this section, we explore the use of BA to estimate the number of deaths. The hope is that BAs, combined with usual mortality-table-based methods, allow for a way to integrate various covariates into the estimates. To reproduce a usual approach, we use an external life table—the 2010 period table for the U.S. Social Security population [22]. This life table serves as a reasonable estimation of the mortality for the NHANES population, with 2010 chosen as a midpoint of the observation period.

Therefore, there are two sources of bias: the first arises from discrepancies between the life table and NHANES mortality data, and the second from substituting CA with other BAs. The first source of bias can be observed as the difference between the number of deaths estimated using CA and the observed number of deaths. The second source of bias is reflected in the difference between the number of deaths estimated using CA and the number estimated using other BAs.

We assume age is measured as civil age (the integer part of exact age since birth). Assuming a uniform distribution within each age, individuals are, on average, half a year older than their civil age. To account for this, we adjust death probabilities from the life table as follows:

$$q'_x = 0.5q_x + 0.5q_{x+1} \tag{14}$$

Additionally, when calculating multi-year survival probabilities (e.g., 5-year survival), we use the formula:

$$_nq_x = 1 - \prod_{k=x}^{x+n-1} (1 - q_k) \tag{15}$$

For a population of size N with ages $\{x_i\}, i \in [1, N]$, we estimate the number of deaths as $\sum_{i=1}^{N} q_{x_i}$, with an approximate variance of $\sum_{i=1}^{N} q_{x_i}(1 - q_{x_i})$. Here, $\{x_i\}$ represents either BA or CA, depending on the context.

For this analysis, we disregard the distinction between training and test datasets due to the small number of one-year deaths.

To complement this analysis, we also compute **Brier Skill Scores (BSS)**, which reflect the squared prediction error relative to CA:

$$\text{BSS} = 1 - \frac{\sum_{i=1}^{N}(d_i - q_{x_i}^{\text{BA}})^2}{\sum_{i=1}^{N}(d_i - q_{x_i}^{\text{CA}})^2}, \tag{16}$$

where $d_i \in \{0, 1\}$ indicates whether individual i died within the follow-up window, and $q_{x_i}^{\text{BA}}$ and $q_{x_i}^{\text{CA}}$ are the death probabilities from the life table applied to BA or CA. This metric penalizes both bias and poor individual-level discrimination, unlike concordance indices (which are insensitive to shifts or scaling), and unlike comparisons of estimated deaths (which only reflect population-level means).

The first plot (Figure 5) shows that overall, BAs offer little improvement over CA in estimating death counts. However, Calibrated PhenoAge—constructed specifically for this task—does provide small improvements, particularly for 10-year follow-up. The second plot (Figure 6) explores this further, comparing PhenoAge calibrated on

1-, 5-, and 10-year survival probabilities. The best results tend to occur when the calibration horizon matches the follow-up period, although the gains are modest and may reflect parallel shifts rather than true adaptation.

Figure 7 presents Brier Skill Scores. From this perspective, a clear winner emerges: the Random Forest model, which consistently outperforms all others, including both PhenoAge and its calibrated variants. All considered BAs improve upon CA, with the exception of the restricted LASSO-CA biomarker set, where some BAs perform worse. Interestingly, the original PhenoAge outperforms its calibrated counterparts in Brier skill. This suggests that while calibration may reduce bias in expected death counts, it does so at the expense of overall prediction accuracy, as measured by mean squared error.

In short, applying BAs to mortality tables improves individual-level performance, as demonstrated by Brier skill scores and concordance. Population-level results, however, are more mixed, showing only modest improvement.

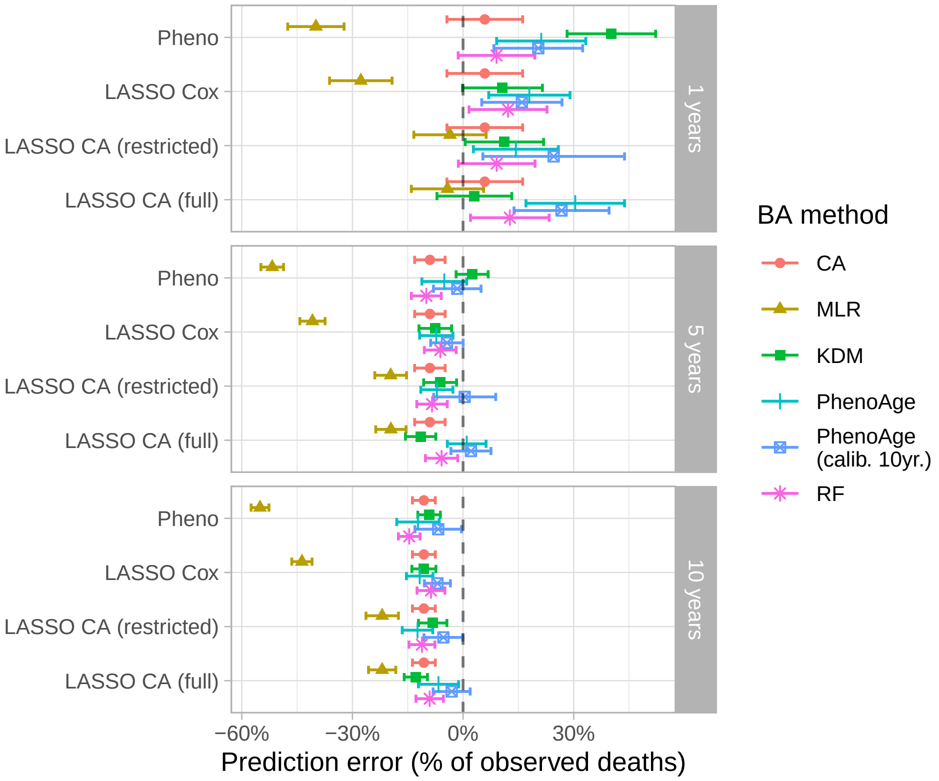


Fig. 5 Percentage error in predicted deaths compared to observed deaths within 1, 5, and 10 years of follow-up for considered BAs.

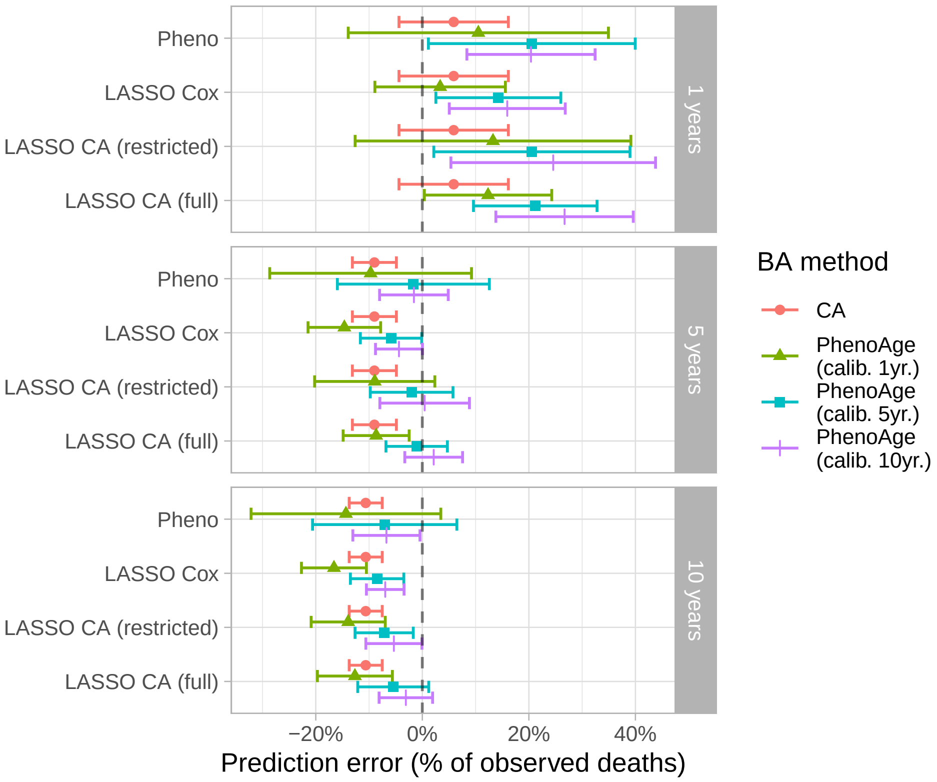


Fig. 6 Percentage error in predicted deaths compared to observed deaths within 1, 5, and 10 years of follow-up for PhenoAge calibrated on 1, 5 and 10 year survival probabilities.

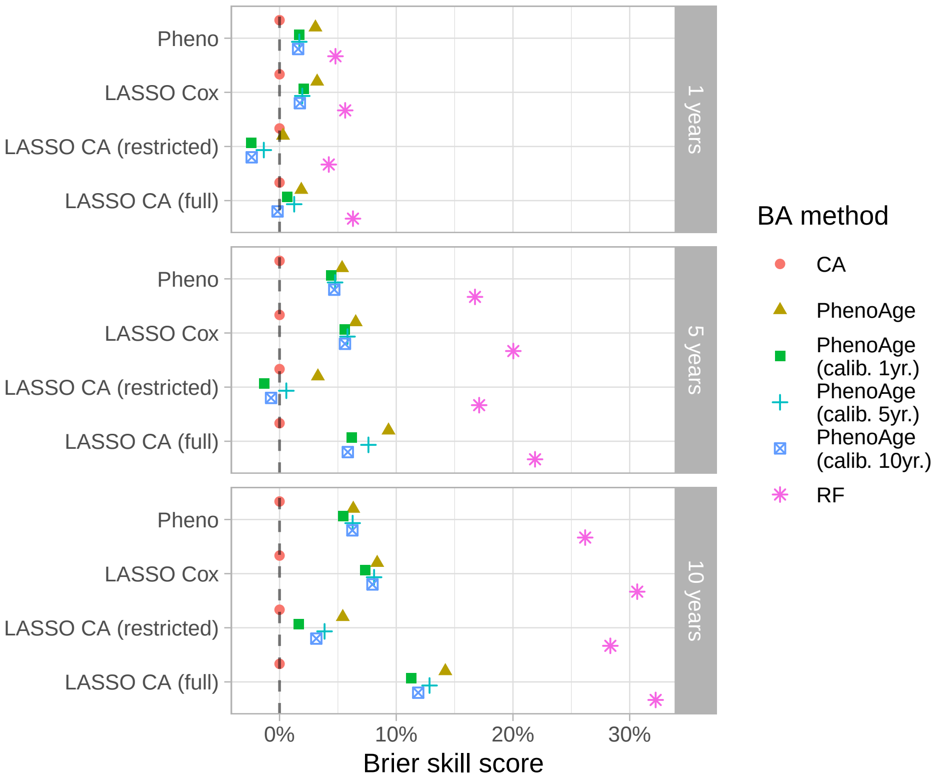


Fig. 7 Brier skill scores for observed deaths at 1-, 5-, and 10-year follow-up, based on PhenoAge calibrated to the corresponding survival horizon.

6 Discussion

In this study, we compared five biological age estimation methods—MLR, KDM, PhenoAge, Calibrated PhenoAge and Random Forest (RF)—using NHANES data and a wide range of biological, behavioral, and sociological biomarkers. Beyond methodological comparison, our aim was to evaluate how these BAs perform when applied to practical tasks such as predicting mortality or estimating the number of deaths using a standard life table. The results offer insights into how BA can be adapted for prevention-focused applications.

6.1 Summary of Findings

Most biological ages achieved improvements over chronological age in mortality prediction and disease association, while maintaining moderate-to-strong correlation with CA. In mortality prediction, Random Forest (RF) and PhenoAge consistently outperformed CA, KDM showed moderate gains, and MLR offered limited improvement, performing worse on the smallest marker set. For disease association, all BAs showed modest performance ($R^2 \approx 0.3$), with limited variation across methods. When used to estimate death counts, BAs showed little advantage at the population level, but all improved individual-level accuracy over CA, with RF being the most accurate overall.

We also examined associations with self-rated health, behavioral factors, and social determinants. These showed weak associations ($R^2 < 0.05$) with no BA substantially outperforming CA, suggesting that biological ages constructed from physiological markers alone may not capture these dimensions of health.

6.2 Methodological Insights

MLR's weak results underscore the challenge of building BAs without leveraging the strong signal contained in CA. By contrast, KDM incorporates CA and showed more stable, though still limited, performance. Mortality-based BAs like PhenoAge and RF offered the best results overall, reinforcing the idea that mortality is both a valuable and practical target for BA construction.

Between the two, RF slightly outperformed PhenoAge across most criteria and showed greater consistency across biomarker sets. RF's advantage likely stems from its modeling flexibility and robustness to overfitting. PhenoAge, in contrast, showed higher variance, especially across imputations and in sociological comparisons, suggesting potential overfitting due to its more complex construction. While RF requires more computation and post-hoc interpretation, PhenoAge's parametric structure makes it easier to understand and communicate.

Interestingly, all methods except RF and the calibrated PhenoAge variants can be reformulated as linear combinations of CA and other covariates. This suggests the possibility of defining a method that explicitly estimates the optimal coefficients with respect to a chosen criterion, rather than relying on existing indirect constructions.

6.3 Using BA to Estimate Death Counts

Applying BA measures to standard life tables revealed a disconnect between individual-level prediction accuracy and population-level bias. Simply substituting BA for CA did not improve death count estimates—but notably, it didn't worsen them either, even for BAs that diverge substantially from CA. However, these biases could not be corrected through simple rescaling and required a separate calibration step. Even then, calibration yielded only modest gains in aggregate accuracy and came at the cost of individual-level performance on other criteria.

Despite these challenges, RF consistently achieved the best Brier skill across all follow-up horizons, outperforming other methods. Interestingly, uncalibrated PhenoAge often outperformed its calibrated variants, suggesting a trade-off between correcting aggregate bias and maintaining individual-level discrimination.

In short, while mortality-based BAs can improve individual risk prediction over CA, using them to estimate population-level death counts from life tables demands additional calibration and careful handling.

6.4 Practical Applications and Prevention

We view BA primarily as a communication tool: expressing risk in "years of age" rather than abstract metrics makes it more tangible. Yet as shown, not all BAs are equal in this regard. Effective application requires tailoring BAs to specific use cases.

This adaptation rests on three components:

1. Target outcome: BAs should be constructed against outcomes relevant to the intended application. While mortality is a natural and versatile target, other outcomes (e.g., disease onset) may be more appropriate for specific prevention goals. Mortality-based methods like RF and PhenoAge are especially well-suited for such substitutions.
2. Variable choice: For prevention, variables should be meaningfully linked to self-protection. Some variables (like smoking) are directly modifiable, while others (like blood pressure) require external knowledge to link with interventions. NHANES offers a rich set of such variables and can serve as a guide for selecting biomarkers in prevention programs.
3. Method selection: PhenoAge and RF both performed well and offer complementary strengths. RF is flexible and accurate, while PhenoAge is easier to interpret. In contexts where explainability and transparency matter—such as insurance or public health messaging—the simplicity of PhenoAge may outweigh RF's marginal gains in accuracy. KDM also performed well relative to CA and MLR, making it our method of choice when mortality data is not available.

6.5 Conclusion

This study provides a structured comparison of biological age methods, evaluating their performance across multiple criteria relevant to prevention and insurance contexts. Our findings show that no single method dominates across all outcomes. PhenoAge and Random Forest (RF) consistently outperform in mortality prediction, while methods like KDM and MLR offer reasonable performance without requiring mortality data.

A key contribution of this work is to highlight trade-offs between mortality prediction and association with chronological age. In general, methods that improve mortality prediction tend to be less correlated with age, and vice versa. Moreover, RF's weaker performance on disease association suggests that different BA methods may capture distinct aspects of aging, and that optimizing for one outcome can limit performance on others. This underscores a recurring theme throughout the article: BA methods should be tailored to the application at hand.

We also introduce several methodological variants of existing BA models, which can serve as a basis for developing application-specific BAs. Some variants may help mitigate known limitations—for instance, extending PhenoAge to include non-linear effects and interactions could narrow the gap with RF in predictive performance. Similarly, MLR variants may improve its performance for smaller biomarker sets.

There are also entirely novel and promising approaches to modeling biological age. For example, Albrecher et al. [23] propose mortality models in which aging is represented as a Markov process over latent life-stage states. In this framework, biological age corresponds to progression along a hidden structure of physiological states. Such models offer a principled and flexible alternative to conventional BA estimation and present a valuable opportunity for actuarial science to contribute to the development of the biological age literature.

Finally, we explore how BA methods can be integrated into practical applications, such as estimating death counts from life tables. While improvements over CA are modest at the population level, BAs provide more accurate individual-level estimates, especially when calibrated.

Altogether, this work offers guidance for selecting BA methods based on the goals of a given application—whether the priority is mortality or morbidity prediction, accuracy, interpretability, or ease of communication. By clarifying the strengths and limitations of each approach, we aim to support the integration of BA into prevention strategies that align health outcomes with insurance incentives.

Acknowledgements. The authors are grateful to Christian Robert (ISFA—SAF Laboratory) for providing guidance on writing of this manuscript.

References

[1] Klemera, P., Doubal, S.: A new approach to the concept and computation of biological age. Mechanisms of Ageing and Development **127**(3), 240–248 (2006) https://doi.org/10.1016/j.mad.2005.10.004

[2] Wilson, D.L.: Aging hypothesis, aging markers and the concept of biological age. Experimental Gerontology **23**(4-5), 435–438 (1988) https://doi.org/10.1016/0531-5565(88)90049-6

[3] Hollingsworth, J.W., Hashizume, A., Jablon, S.: Correlations between tests of aging in Hiroshima subjects–an attempt to define "physiologic age". The Yale

Journal of Biology and Medicine **38**(1), 11–26 (1965)

[4] Nakamura, E.: A study on the basic nature of human biological aging processes based upon a hierarchical factor solution of the age-related physiological variables. Mechanisms of Ageing and Development **60**(2), 153–170 (1991) https://doi.org/10.1016/0047-6374(91)90128-M

[5] Horvath, S.: DNA methylation age of human tissues and cell types. Genome Biology **14**(10), 115 (2013) https://doi.org/10.1186/gb-2013-14-10-r115

[6] Levine, M.E., Lu, A.T., Quach, A., Chen, B.H., Assimes, T.L., Bandinelli, S., Hou, L., Baccarelli, A.A., Stewart, J.D., Li, Y., Whitsel, E.A., Wilson, J.G., Reiner, A.P., Aviv, A., Lohman, K., Liu, Y., Ferrucci, L., Horvath, S.: An epigenetic biomarker of aging for lifespan and healthspan. Aging **10**(4), 573–591 (2018) https://doi.org/10.18632/aging.101414

[7] Li, Z., Zhang, W., Duan, Y., Niu, Y., Chen, Y., Liu, X., Dong, Z., Zheng, Y., Chen, X., Feng, Z., Wang, Y., Zhao, D., Sun, X., Cai, G., Jiang, H., Chen, X.: Progress in biological age research. Frontiers in Public Health **11**, 1074274 (2023) https://doi.org/10.3389/fpubh.2023.1074274

[8] Kuiper, L.M., Polinder-Bos, H.A., Bizzarri, D., Vojinovic, D., Vallerga, C.L., Beekman, M., Dollé, M.E.T., Ghanbari, M., Voortman, T., Reinders, M.J.T., Verschuren, W.M.M., Slagboom, P.E., Van Den Akker, E.B., Van Meurs, J.B.J.: Epigenetic and Metabolomic Biomarkers for Biological Age: A Comparative Analysis of Mortality and Frailty Risk. The Journals of Gerontology: Series A **78**(10), 1753–1762 (2023) https://doi.org/10.1093/gerona/glad137

[9] McCrory, C., Fiorito, G., Hernandez, B., Polidoro, S., O'Halloran, A.M., Hever, A., Ni Cheallaigh, C., Lu, A.T., Horvath, S., Vineis, P., Kenny, R.A.: GrimAge Outperforms Other Epigenetic Clocks in the Prediction of Age-Related Clinical Phenotypes and All-Cause Mortality. The Journals of Gerontology. Series A, Biological Sciences and Medical Sciences **76**(5), 741–749 (2021) https://doi.org/10.1093/gerona/glaa286

[10] Hastings, W.J., Shalev, I., Belsky, D.W.: Comparability of biological aging measures in the National Health and Nutrition Examination Study, 1999-2002. Psychoneuroendocrinology **106**, 171–178 (2019) https://doi.org/10.1016/j.psyneuen.2019.03.012

[11] Cho, I.H., Park, K.S., Lim, C.J.: An empirical comparative study on biological age estimation algorithms with an application of Work Ability Index (WAI). Mechanisms of Ageing and Development **131**(2), 69–78 (2010) https://doi.org/10.1016/j.mad.2009.12.001

[12] Study Group QalyDays: QalyDays (2019). www.qalydays.com Accessed 2025-12-23

[13] Ehrlich, I., Becker, G.S.: Market Insurance, Self-Insurance, and Self-Protection. Journal of Political Economy **80**(4), 623–648 (1972). Publisher: The University of Chicago Press

[14] Center for Disease Control and Prevention: NHANES Questionnaires, Datasets, and Related Documentation (2024). https://wwwn.cdc.gov/nchs/nhanes/Default.aspx Accessed 2024-08-30

[15] Qiu, W., Chen, H., Kaeberlein, M., Lee, S.-I.: ExplaiNAble BioLogical Age (ENABL Age): an artificial intelligence framework for interpretable biological age. The Lancet Healthy Longevity **4**(12), 711–723 (2023) https://doi.org/10.1016/S2666-7568(23)00189-7

[16] Ingram, D.K.: Key questions in developing biomarkers of aging. Experimental Gerontology **23**(4-5), 429–434 (1988) https://doi.org/10.1016/0531-5565(88)90048-4

[17] Dubina, T.L., Mints, A.Y., Zhuk, E.V.: Biological age and its estimation. III. Introduction of a correction to the multiple regression model of biological age in cross-sectional and longitudinal studies. Experimental Gerontology **19**(2), 133–143 (1984) https://doi.org/10.1016/0531-5565(84)90016-0

[18] Levine, M.E.: Modeling the Rate of Senescence: Can Estimated Biological Age Predict Mortality More Accurately Than Chronological Age? The Journals of Gerontology Series A: Biological Sciences and Medical Sciences **68**(6), 667–674 (2013) https://doi.org/10.1093/gerona/gls233

[19] Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. The Annals of Applied Statistics **2**(3) (2008) https://doi.org/10.1214/08-AOAS169

[20] Center for Disease Control and Prevention: NCHS Data Linkage - Mortality Data - Public-Use Files (2022). https://www.cdc.gov/nchs/linked-data/mortality-files/ Accessed 2024-08-30

[21] Brand, J.P.L.: Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. PhD Thesis, Erasmus University Rotterdam (April 1999)

[22] Social Security Administration: Period Life Table, 2010, as used in the 2014 Trustees Report (2014). https://www.ssa.gov/oact/STATS/table4c6_2010_TR2014.html

[23] Albrecher, H., Bladt, M., Bladt, M., Yslas, J.: Mortality modeling and regression with matrix distributions **107**, 68–87 https://doi.org/10.1016/j.insmatheco.2022.08.001

[24] Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys, 1st edn. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ (1987). https://doi.org/10.1002/9780470316696 . https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316696

[25] Schafer, J.L.: Multiple Imputation Models and Procedures for NHANES III (2001). https://wwwn.cdc.gov/Nchs/Data/Nhanes3/7a/doc/mimodels.pdf

[26] Rammon, J., He, Y., Parker, J.D.: Multiple imputation to account for linkage ineligibility in the NHANES-CMS Medicaid linked data: General use versus subject specific imputation models. Statistical Journal of the IAOS **35**(3), 443–456 (2019) https://doi.org/10.3233/SJI-180470

[27] Buuren, S., Groothuis-Oudshoorn, K.: mice : Multivariate Imputation by Chained Equations in R. Journal of Statistical Software **45**(3) (2011) https://doi.org/10.18637/jss.v045.i03

[28] van Buuren, S.: Flexible Imputation of Missing Data, Second Edition, 2nd edn. Chapman and Hall/CRC, Second edition. Boca Raton, Florida : CRC Press, [2019] (2018). https://doi.org/10.1201/9780429492259 . https://www.taylorfrancis.com/books/9780429492259

[29] Liu, B., Yu, M., Graubard, B.I., Troiano, R.P., Schenker, N.: Multiple imputation of completely missing repeated measures data within person from a complex sample: application to accelerometer data in the National Health and Nutrition Examination Survey. Statistics in Medicine **35**(28), 5170–5188 (2016) https://doi.org/10.1002/sim.7049

A Supplementary material

A.1 Cause-specific mortality

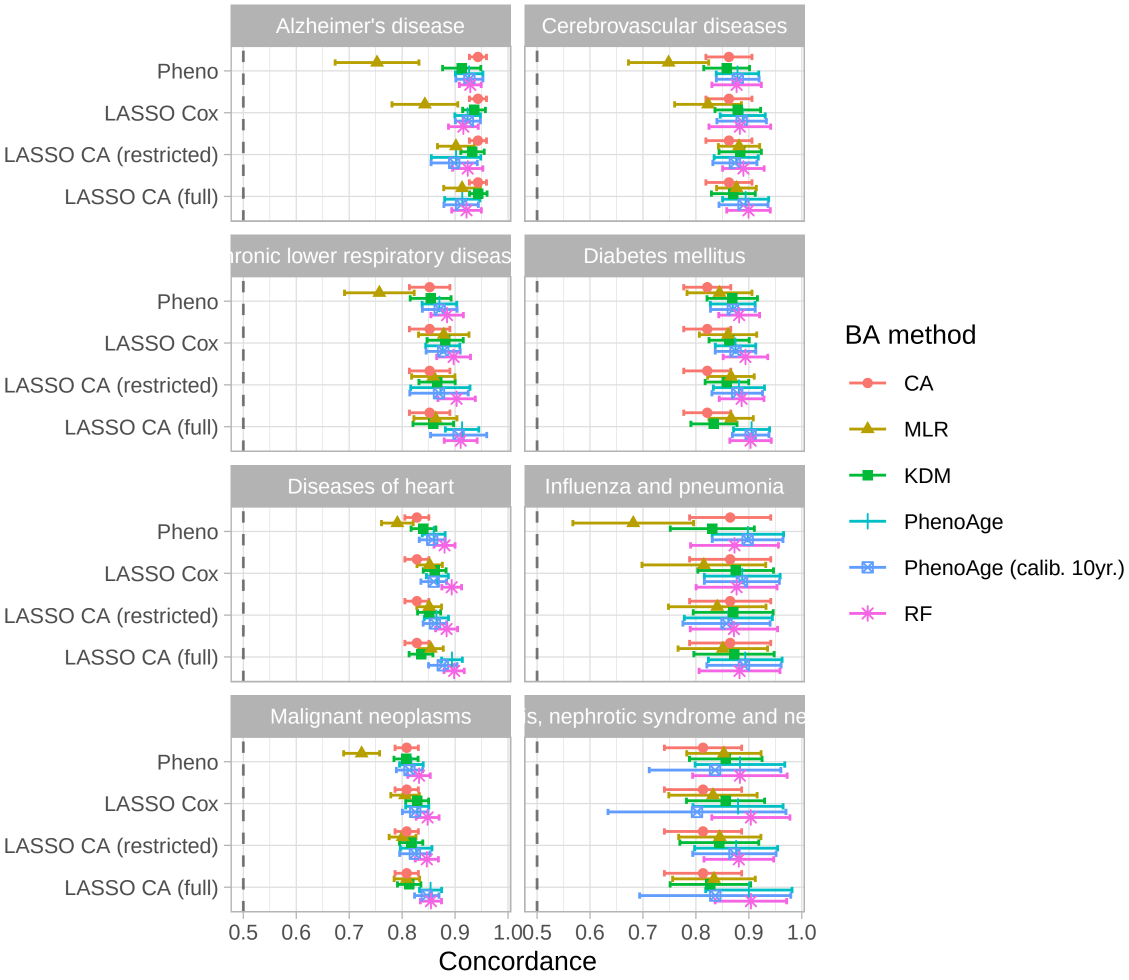


Fig. 8 Concordance with respect to age of death (cause-specific).

A.2 Disease-specific comparison

Figure 9 shows the correlation between BAs and the presence of specific diseases. Most correlations are modest, ranging from near zero (e.g., for asthma or obesity) to about 0.25 (e.g., for hypertension).

In general, BAs perform comparably to CA. CA slightly outperforms BAs for cancer, hyperlipidemia, and arthritis, and significantly outperforms RF for hyperlipidemia. Conversely, BAs—especially RF—perform significantly better than CA for kidney disease and diabetes. MLR, surprisingly, also shows a relatively strong correlation with diabetes.

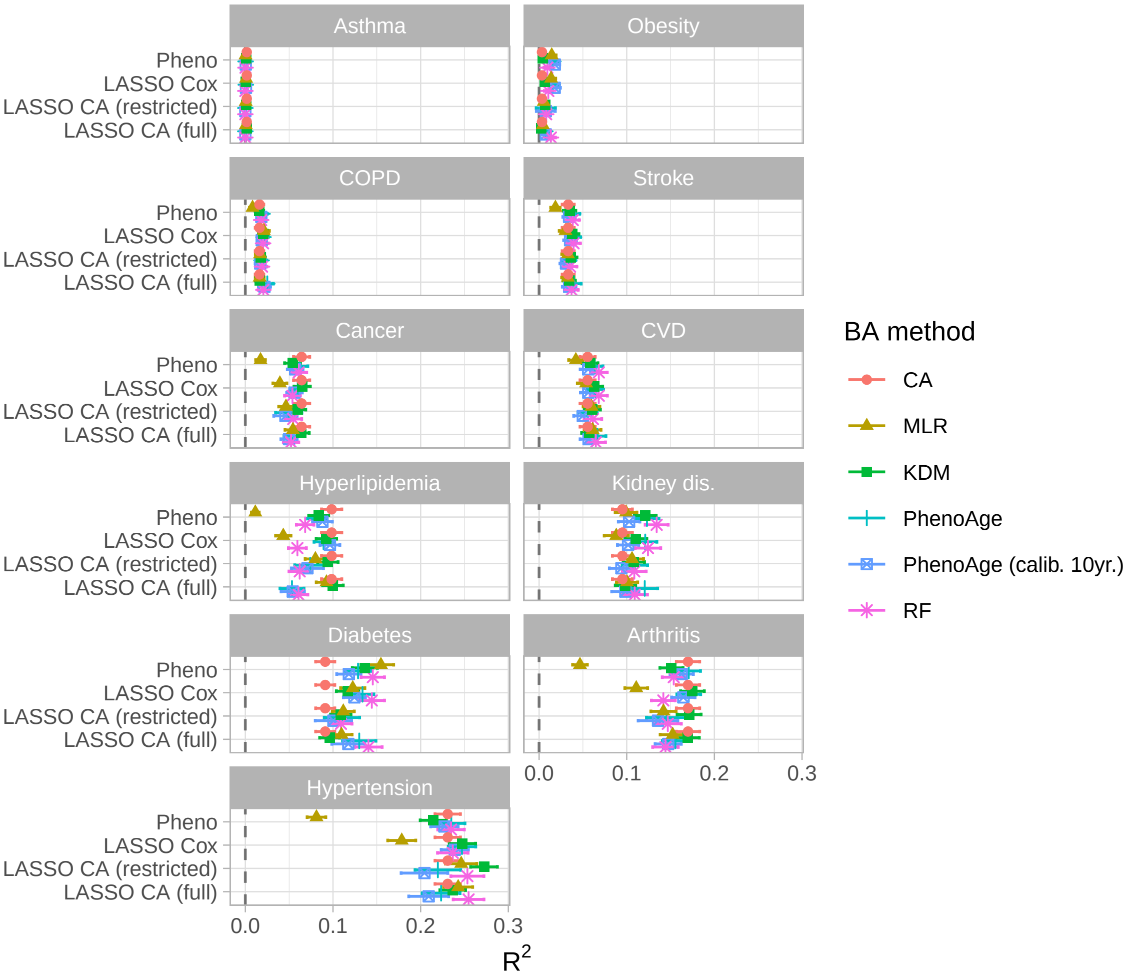


Fig. 9 Variance in Disease Presence Explained by Biological Age (R^2).

A.3 Self-Rated Health

Self-rated health is an important piece of information known to be an independent predictor of mortality, even after controlling for other factors such as functional limitation. It is commonly examined in biological age studies, making it a useful benchmark for comparison with prior work.

Figure 10 compares the variance in self-rated health explained by each BA method. All BAs show statistically significant but weak associations with self-rated health ($R^2 \approx 0.03 \pm 0.02$), only modestly exceeding CA ($R^2 = 0.015$). RF performs best, achieving significantly higher R^2 than CA across all marker sets and outperforming MLR and KDM on the largest marker set. PhenoAge shows intermediate performance, exceeding CA on three of four marker sets but not significantly different from other BA methods. MLR and KDM do not significantly improve upon CA for any marker set.

These weak associations suggest that self-rated health captures dimensions of health status not strongly reflected in either chronological age or mortality-based biological ages.

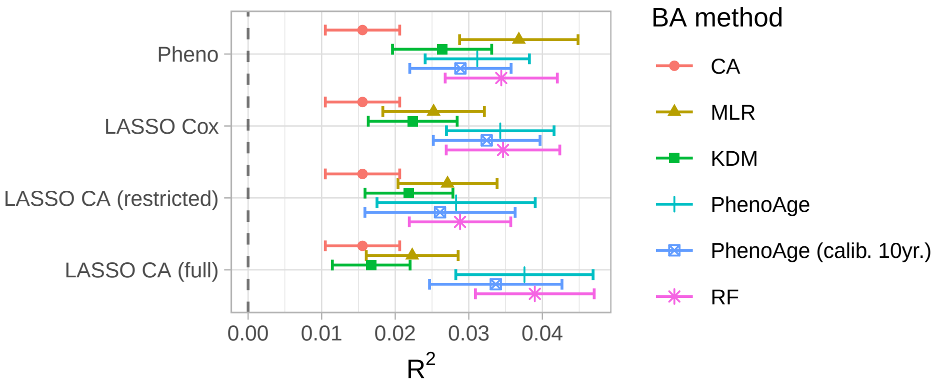


Fig. 10 Variance in self-rated health explained by Biological Age (R^2).

A.4 Marker sets

A.4.1 Pheno marker set

A.4.2 LASSO Cox marker set

A.4.3 Restricted LASSO CA marker set

A.4.4 Full LASSO CA marker set

Table 6: Biological markers from the full ”LASSO CA” marker set.

Variable	Description
BAX_balance	Combined failure time for all trials (seconds)
BIDECF	Estimated extracellular fluid volume (L)
BIDPFAT	Estimated percent body fat
BID_ECFpct	Estimated extracellular fluid volume (%)
BID_WaterFFM	Estimated total water body volume to fat-free mass ratio (l/kg)
BMXARMC	Arm Circumference (cm)
BMXHT	Standing Height (cm)
BMXLEG	Upper Leg Length (cm)
BMXTHICR	Thigh Circumference (cm)
BMXWAIST	Waist Circumference (cm)
BMX_WACR	Ratio between waist and arm circumferences
BMX_invBMI	The inverse of BMI
BMX_logLBXCRP	Log transform of LBXCRP
BPXDAR	DBP average reported to examinee
BPXPLS	60 sec. pulse (30 sec. pulse * 2)
BPXPULS	Pulse regular or irregular?
BPXSAR	SBP average reported to examinee
BPX_invsyspress	Inverted systolic blood pressure
CVDESVO2	Estimated VO2max (ml/kg/min)
DRXTCARB	Carbohydrate (gm)
DRXTFIBE	Dietary fiber (gm)
DRXTKCAL	Energy (kcal)
DRXTPROT	Protein (gm)
DRXTSUGR	Total sugars (gm)
DRXTTFAT	Total fat (gm)
DXDTOBMC	Total Bone Mineral Content (g)
DXDTOFAT	Total Fat (g)
DXD_TOBMCpctT	Total Bone Mineral Content to weight ratio
DXXSATM	Subcutaneous fat mass
DXXVFATM	Visceral adipose tissue mass
LBDBANO	Basophils number (1000 cells/uL)
LBDHDD	Direct HDL-Cholesterol (mg/dL)
LBDLYMNO	Lymphocyte number (1000 cells/uL)
LBDNENO	Segmented neutrophils num (1000 cell/uL)
LBDSBUSI	Blood Urea Nitrogen (mmol/L)
LBDSCASI	Total Calcium (mmol/L)
LBDSCHSI	Cholesterol, refrigerated serum (mmol/L)
LBDSGBSI	Globulin (g/L)
LBDSGLSI	Glucose, refrigerated serum (mmol/L)
LBDSIRSI	Iron, refrigerated serum (umol/L)
LBDSPHSI	Phosphorus (mmol/L)
LBDSTBSI	Total Bilirubin (umol/L)
LBDSUASI	Uric acid (umol/L)
LBXBAP	Bone alkaline phosphotase (ug/L)

LBXBAPCT	Basophils percent (%)
LBXFER	Ferritin (ng/mL)
LBXGH	Glycohemoglobin (%)
LBXLYPCT	Lymphocyte percent (%)
LBXMC	Mean Cell Hgb Conc. (g/dL)
LBXMCVSI	Mean cell volume (fL)
LBXME	Measles
LBXMOPCT	Monocyte percent (%)
LBXMPSI	Mean platelet volume (fL)
LBXP1	Total prostate specific antigen (ng/mL)
LBXPLTSI	Platelet count (1000 cells/uL)
LBXRDW	Red cell distribution width (%)
LBXSC3SI	Bicarbonate (mmol/L)
LBXSCLSI	Chloride (mmol/L)
LBXSGTSI	Gamma Glutamyl Transferase (GGT) (IU/L)
LBXSKSI	Potassium (mmol/L)
LBXSNASI	Sodium (mmol/L)
LBXSOSSI	Osmolality (mmol/Kg)
LBXTSH1	Thyroid stimulating hormone (uIU/mL)
LBXTT3	Triiodothyronine (T3), total (ng/dL)
LBXTT4	Thyroxine, total (T4) (ug/mL)
LBXVIDMS	25OHD2+25OHD3 (nmol/L)
MGX_PFkg_model	Grip peak force estimated from height, sex, and grip strength. (kg)
SPXNEV	Baseline Extrapolated Volume (mL)
SPXNF257	Baseline FEF 25-75% (mL/s)
SPXNFET	Baseline Forced Expiratory Time (s)
SPXNFEV5	Baseline FEV 0.5 (mL)
SPXNFVC	Baseline FVC (mL)
SPXNPEF	Baseline PEF (mL/s)
SPX_FEV5h3	Forced expiratory volume in 5 seconds to height cubed ratio (N/m^3)
SSKLOTH	Klotho (pg/ml)
TELOMEAN	Mean T/S ratio
TELOSTD	Asso. Std. Dev. of Mean Telomere Length
URXUCR	Creatinine, urine (mg/dL)
URXUMA	Albumin, urine (ug/mL)

Table 3 Biological markers from the "Pheno" marker set.

Variable	Description
LBDSALSI	Albumin, refrigerated serum (g/L)
LBDSCRSI	Creatinine, refrigerated serum (umol/L)
LBDSGLSI	Glucose, refrigerated serum (mmol/L)
LBXCRP	C-reactive protein(mg/dL)
LBXLYPCT	Lymphocyte percent (%)
LBXMCVSI	Mean cell volume (fL)
LBXRDW	Red cell distribution width (%)
LBXSAPSI	Alkaline Phosphatase (ALP) (IU/L)
LBXWBCSI	White blood cell count (1000 cells/uL)

Table 4 Biological markers from the "LASSO Cox" marker set.

Variable	Description
BMX_WACR	Ratio between waist and arm circumferences
LBDSALSI	Albumin, refrigerated serum (g/L)
LBDSCRSI	Creatinine, refrigerated serum (umol/L)
LBDSGBSI	Globulin (g/L)
LBDSGLSI	Glucose, refrigerated serum (mmol/L)
LBXBAP	Bone alkaline phosphotase (ug/L)
LBXP1	Total prostate specific antigen (ng/mL)
LBXRDW	Red cell distribution width (%)
LBXSAPSI	Alkaline Phosphatase (ALP) (IU/L)
LBXSCLSI	Chloride (mmol/L)
LBXSGTSI	Gamma Glutamyl Transferase (GGT) (IU/L)
LBXSNASI	Sodium (mmol/L)
SPX_FEV5h3	Forced expiratory volume in 5 seconds to height cubed ratio (N/m^3)
SPX_PEF5h3	Peak expiratory flow in 5 seconds to height cubed ratio (L/m^3)
URXUMA	Albumin, urine (ug/mL)

Table 5 Biological markers from the restricted ”LASSO CA” marker set.

Variable	Description
BAX_balance	Combined failure time for all trials (seconds)
BMXTHICR	Thigh Circumference (cm)
BPXPLS	60 sec. pulse (30 sec. pulse * 2)
BPXSAR	SBP average reported to examinee
BPX_invsyspress	Inverted systolic blood pressure
DXXVFATM	Visceral adipose tissue mass
LBDSBUSI	Blood Urea Nitrogen (mmol/L)
LBXMC	Mean Cell Hgb Conc. (g/dL)
LBXMCVSI	Mean cell volume (fL)
LBXME	Measles
LBXSC3SI	Bicarbonate (mmol/L)
LBXSOSSI	Osmolality (mmol/Kg)
LBXTT3	Triiodothyronine (T3), total (ng/dL)
SPXNF257	Baseline FEF 25-75% (mL/s)
SPXNFET	Baseline Forced Expiratory Time (s)
SPXNFVC	Baseline FVC (mL)
SPX_FEV5h3	Forced expiratory volume in 5 seconds to height cubed ratio (N/m^3)

A.5 Multiple imputation of NHANES

To address the block-missingness pattern in NHANES data, we apply multiple imputation by chained equations (MICE). MICE generates several plausible versions of a dataset by replacing missing values with estimates, thus addressing the uncertainty caused by missing data. Each imputed dataset is complete, allowing for separate complete-case analyses. These results are then pooled, yielding robust inferences for both point estimates and variance estimates. This approach allows us to include all survey cycles and variables of interest, rather than being forced to choose between breadth (more cycles, fewer variables) and depth (fewer cycles, more variables) as would be required under complete-case analysis.

Multiple imputation was first introduced by Rubin [24] and has since been applied in various contexts, including NHANES. For example, NHANES III (1988-1994) is available as a multiply imputed dataset, as described by Schafer [25]. Continuous NHANES also contains multiply imputed data, but only for specific datasets, such as accelerometer and Dual-Energy X-ray Absorptiometry (DEXA) data. Multiple imputation has also been used in studies using NHANES, for example, to impute Medicaid enrollment status [26]. However, no study has yet applied multiple imputation to NHANES data for the purpose of comparing BAs.

Most analyses of NHANES data assume that data are missing at random (MAR) and focus on complete-case analysis. This approach is reasonable because non-random missingness—mainly non-response—is accounted for in the subject weights. However, when applied to multiple cycles of NHANES, complete-case analysis forces analysts to choose between breadth and depth. Many variables are only measured in some cycles, and under complete-case analysis, analysts must choose between including more variables but fewer cycles, or including more cycles but fewer variables. Multiple imputation resolves this issue by allowing us to include all cycles and variables that would otherwise be missing.

In this study, we applied MICE to impute missing values across NHANES cycles and maximize available information. Each BA model was fit and evaluated on each imputed dataset independently, with the results then combined to form an aggregate estimate that accounts for missingness.

The MICE algorithm iterates over all variables in the dataset, fitting a model and imputing missing values from the posterior distribution at each step. With each iteration, the imputed values become more plausible. This process is analogous to a Gibbs sampler, with conditional distributions specified by the model [27].

Table 1 details the MICE Algorithm, from van Buuren [28], Section 4.5.2.

We apply the MICE algorithm for 30 iteration and over 10 replications. The number of iterations had to be set somewhat high for the marginal distribution to stabilize, this is due to large number of correlated variables included in the model.

We applied diagnostic procedures recommended by van Buuren [28], including convergence checks for MICE chains and comparison of variable distributions by year before and after imputation. The choice of 30 iterations and variable exclusions were informed by these diagnostics to ensure stable imputation behavior.

The following categories of variables were included into set of variables to be imputed:

Algorithm 1 MICE Algorithm

```
Input: Dataset $Y = \{Y_1, Y_2, \ldots, Y_p\}$ with missing values
Output: Multiple complete-case datasets
for each variable $Y_j$, $j = 1, \ldots, p$ do
    Specify imputation model $p(Y_j^{\text{mis}} | Y_j^{\text{obs}}, Y_{-j})$
    Initialize missing values $\dot{Y}_j^0$ with random draws from $Y_j^{\text{obs}}$
end for
for iteration $t = 1, \ldots, m$ do
    for each variable $Y_j$ where $j = 1, \ldots, p$ do
        Define $\dot{Y}_{-j}^t = (\dot{Y}_1^t, \ldots, \dot{Y}_{j-1}^t, \dot{Y}_{j+1}^{t-1}, \ldots, \dot{Y}_p^{t-1})$ ▷ Updated complete data for all variables up to $Y_j$ at iteration $t$.
        Draw parameters $\dot{\phi}_j^t \sim p(\phi_j^t | Y_j^{\text{obs}}, \dot{Y}_{-j}^t)$
        Draw imputations $\dot{Y}_j^t \sim p(Y_j^{\text{mis}} | Y_j^{\text{obs}}, \dot{Y}_{-j}^t, \dot{\phi}_j^t)$
    end for
end for
Repeat the process to create multiple complete-case datasets
```

- Biological markers of interest,
- Variables related to mortality (i.e., age at the time of examination, age at end of follow-up, vital status),
- Variables used in survey design, as per Liu et al. [29] (i.e., age, gender, ethnicity, and the masked variance pseudo-PSU, a proxy for the primary sampling unit not available in public-use data),
- Questionnaire items describing general health status (as per Schafer [25]).
- Variables needed to compute all the comparison criteria.

To reduce computational load, only variables with an absolute correlation greater than 10% are used in the imputation model for a given variable.

Conditional distributions depend on the type of variable: continuous variables were imputed using Predictive Mean Matching (PMM) due to its robustness distributional specification, binary variables via logistic regression, and categorical variables with multinomial regression for multicategory cases.

Article Review Follow-up

Dear Editor,

Thank you for the thorough and constructive review of our manuscript. We appreciate the reviewer's recognition that our work "could offer a nuanced understanding grounded in objective methods" and could be "of interest to the readership of the European Actuarial Journal."

We have carefully considered all comments and have made substantial revisions to address the concerns raised. Please find below a detailed point-by-point response to the reviewer's comments.

1. Temporal and Population Context

> The NHANES database is reported to be based on 'survey cycles between 1999 and 2018'. Which year is used as the base year for the selected biological markers?

We use data from all NHANES survey cycles between 1999 and 2018. The biological markers are measured across all these cycles, with biological age models trained on the combined data (excluding the test set). Mortality follow-up extends through 2019, resulting in varying follow-up periods for participants across different survey cycles.

> It is reported on page 13 that 'characteristics measured at the time of examination remain constant'. What are the potential implications of assuming the underlying covariates remain unchanged?

This assumption is a necessary limitation of the data structure. We measure biological markers at a single point in time during the survey examination, then assess their association with mortality outcomes that may occur years later. Essentially, we are asking: "How do markers measured at age 30 predict mortality risk from age 30 onwards (up to 20 years of follow-up in this dataset)?" This differs from a hypothetical repeated-measures design where we would ask: "How do current marker values affect current mortality risk?"—which would require longitudinal covariate measurements at multiple time points.

This limitation does not undermine our objective of assessing whether biological age can predict mortality in general, rather than estimating specific effects of individual markers. We chose not to censor follow-up times to maintain comparability with existing NHANES-based literature, particularly Levine et al. (2018) and Qiu et al. (2023).

> In Section 3, the calculated BAs are presented and compared across different methods. The underlying population/cohort and relevant time are not clear. For example, are the results for all individuals or stratified by sex, and are the BAs relevant to a single year, e.g. 2019, or averaged over 1999–2018?

Biological age models are fitted separately for each sex using all available NHANES cycles (1999-2018), excluding the test subset. These sex-specific models are then applied to estimate biological ages for the entire test subset, and all metrics presented in Section 3 (now called Section 4 “Evaluation of Biological Age Methods”) are calculated on this test set. The models thus capture patterns across the entire time period rather than representing a single year.

> Linked to the previous two comments, the interpretation of the results may change. For example, challenges regarding the interpretability of modelling outputs are noted on page 15 due to smaller BA numbers compared to calendar ages (CAs). In more recent years, with better health care, diet, and exercise, BAs would be expected to be lower—indicating a younger and healthier biological state—than the corresponding CAs, compared with earlier years. A comparison of trends over time would help put the discussion into context. Has this been already considered in the analysis? A discussion on this would be useful.

While examining temporal trends in biological age as a population health indicator is certainly valuable—and historically one of BA’s first applications—we deliberately chose not to pursue this analysis for two reasons. First, NHANES purposefully oversamples certain population segments to address specific policy questions, making it non-representative of the general US population without survey-weight adjustments. Since our goal is to demonstrate BA’s usefulness as a general methodological approach rather than to estimate health indicators for the US population specifically, we treat NHANES as an abstract population and do not apply survey weights. Second, trends in US population health over this period are complex; while life expectancy increased, health-adjusted indicators have stagnated in the 21st century, making it unclear a priori what trends we should expect in BA. These considerations led us away from this line of inquiry.

We adjusted the phrasing describing the NHANES data to better explain the ‘constant covariates’ assumption, clarifying that biological markers are measured once during survey examination and used to predict mortality throughout follow-up. We also added a short summary in the Biological Age Comparison section specifying that BA models are fitted on NHANES cycles 1999-2018, stratified by sex, with metrics computed on the test set.

2. Biomarkers Consistency Across Methods

> On page 7, it is stated that ‘we calculate BAs on the same four sets of biomarkers for all methods’. Is this different from the biomarkers reported in Tables 1 and 2? Or, perhaps, what is meant are the response variables, such as CAs, all-cause mortality, and so on?

The “four sets of biomarkers” refers to the four covariate sets described in Section 3.2 (Biological Markers). We have clarified this passage to avoid confusion—these are the predictor variables (biological markers/covariates), not the response variables (CA or mortality).

> Related to this, would it not be useful to see the model outputs based on identical biomarkers across the different methods—even in the Supporting Information? Although each method may have an optimal set of biomarkers (as listed in Tables 1 and 2), using the same biomarkers could provide a fairer comparison in terms of the predictive power of these methods.

Indeed, all biological age methods are applied to all four biological marker sets, allowing us to both (i) compare how each method performs across different covariate regimes, and (ii) compare methods head-to-head on the same marker sets. This is why all biological age comparison plots contain 24 data points: 6 methods (5 BA methods + chronological age as reference) × 4 marker sets.

We revised the passage introducing the biological age comparison to clarify that “biomarkers” refers to the four covariate sets described in Section 3.2, not to response variables or method-specific optimal marker selections.

3. Data Imputation and Completeness

> Could you please comment on the completeness of the data and the percentage of missing values in the underlying cohort used for the main results?

NHANES data exhibits a specific missingness pattern often referred to as “block-missingness”: the set of biological markers measured varies across survey cycles. Some markers are only available starting from certain cycles, others are discontinued, and some are measured intermittently (measured, then absent for several cycles, then measured again). Within-cycle missingness for measured variables is relatively small, typically due to survey design decisions (e.g., grip strength only measured for those aged 40+).

To provide concrete numbers: the overall missingness rate for variables used in the models is 34%. Variables appear in an average of 7.8 NHANES cycles (out of 10). The within-cycle missingness rate, considering only cycles where the variable is measured, is 19%.

> Have you checked the distribution of related cohort(s) by age and year before and after imputation? This is important to ensure that model bias is avoided.

Yes, we followed the diagnostic procedures recommended by van Buuren (2018), including checks for convergence of MICE chains and comparison of means and standard deviations of variables by year before and after imputation. Some modeling choices were informed by these diagnostics, notably the chain length (30 iterations) and the exclusion of certain variables that behaved erratically or caused instability in other variables.

However, we note two important points: First, changes in distribution are not necessarily problematic—when missingness is not random, the distribution of missing items may legitimately differ from observed items. Second, it may not be appropriate to speak of "bias" in the traditional sense, as our goal is not to create an imputed dataset representative of the United States population, but rather to generate plausible datasets that allow us to compare BA methods under realistic conditions.

We added concrete missingness statistics to the Multiple Imputation section, clarifying the block-missingness pattern in NHANES data (34% overall missingness, 19% within-cycle). We also documented the diagnostic procedures followed to validate the imputation process, including convergence checks and distributional comparisons across years.

4. Readability and Flow

Introduction

> The introduction is detailed, but too long and somewhat repetitive. - I suggest focusing on a more concise introduction, with greater emphasis on the contribution of this study, where the narrative could be centred on BA and its value as a preventative measure. - A brief summary of the following sections could also be useful. - Section 1.1 could be shortened and combined with Section 1.3. Some of the arguments in Section 1.1 can be moved to Discussion, e.g. Section 5.4, to tighten the concluding remarks.

We restructured the introduction into a single, flowing narrative without subsections, reducing its length by approximately 30% while emphasizing this study's contribution more prominently. We streamlined the prevention program discussion while retaining essential motivating examples.

> The NHANES dataset is first mentioned in Section 1.3. It is not clear which population this dataset covers or what NHANES stands for.

We now introduce NHANES earlier in the revised introduction, providing its full name (National Health and Nutrition Examination Survey) and specifying that it is a large-scale U.S. health examination survey.

> Incidence rates and BA are compared in terms of interpretation, but an example on mortality is provided. Why not use an example more aligned with the argument?

Mortality is the primary outcome we consider in this study. The example uses mortality rates to illustrate that when absolute risk remains low (as it does for mortality throughout most of the lifespan), expressing differences as biological age better communicates the relative magnitude of risk change than probability differences alone.

Data and Methods Organization

> The NHANES dataset is introduced under Section 2, which primarily covers methods. I suggest introducing it in a separate section, e.g. Section 3 titled 'Data'. Furthermore, a data imputation method, MICE, is implemented. This is important and should be explained more clearly in the main text. Currently, there are references to 'imputation' but no clear description of how missing data are handled. I note a separate section in the Supplementary Information. Perhaps the method and its purpose could be briefly explained in the main text, with full details left for the Supplementary Information.

We have restructured the manuscript to place NHANES data in a separate "Data" section (now Section 3), distinct from the methods sections. In this section, we now provide key missingness statistics (as detailed in our response to your earlier comment) and briefly explain that we use multiple imputation to address this missingness, with the motivation that the relatively high missingness rate requires propagating uncertainty rather than using simpler imputation approaches. Full methodological details of the MICE implementation remain in the Supplementary Materials.

Results Section

> Section 3 reports the numerical results. A broader title can be more appropriate than 'Biological Age Comparison'.

We have renamed the sections to better reflect their scope. Section 4 is now titled "Evaluation of Biological Age Methods" (previously "Biological Age Comparison") and Section 5 remains "Using BA to Estimate Death Counts." This structure maintains the distinction between methodological evaluation and practical application.

Methods Overview

> There are four methods listed in the abstract but five methods explained in Section 2. It would be useful to provide a table that clearly maps the different methods to their names and acronyms.

We have added a table (Table 1) that clearly maps all five BA calculation methods to their names, acronyms, primary outcomes, and source references. Moreover, the ‘Methods’ section has been renamed ‘Biological Ages’ to reflect its new focus.

Tables and Supporting Information

> Table 1 could be moved to the Supporting Information. It would also be helpful to have a look-up table in the main text summarising which table in the Supporting Information corresponds to each method.

We have moved the detailed marker lists to the Supplementary Materials. In the main text, the table providing a summary of the four marker sets with their names, number of variables, selection criteria now also includes explicit references to the corresponding detailed tables in the Supplementary Materials.

Terminology and Statistical Measures

> There are comments and captions based on R^2, such as ‘Association with the number of diseases (R^2)’. The wording here sounds imprecise. Would it be possible to use more accurate terminology to put this into context? For example, is this about how much variability in the response variable is explained by BA?

We have revised figure captions to use more precise terminology. Captions now explicitly state that R^2 represents the variance in the outcome variable (chronological age, disease count, etc.) explained by biological age.

Self-rated Health and Other Variables

> Self-rated health is mentioned twice in the main text, including Section 2, but the related outcomes do not contribute to the narrative. A brief mention in the Discussion would be sufficient. I also note that the corresponding R^2 values are below 5%. It is therefore difficult to view this as a ‘modest’ contribution, as suggested on page 23. A similar comment applies to behavioural and social factors.

We have removed behavioral and social factors from the manuscript entirely. Self-rated health is retained in the Supplementary Materials as it is commonly examined in biological age studies and provides a useful benchmark for comparison with prior work. However, we now provide only a brief discussion of these weak associations ($R^2 < 0.05$) in the Discussion section, focusing the narrative on mortality and disease associations where biological ages show substantive improvements over chronological age.

5. Technical Corrections and Proofreading

> The manuscript requires proofreading and copy-editing. - Acronyms must be defined before use and then used consistently thereafter, without repeated re-definitions. - References do not include publication years. Please double-check the entries. Notation is not always coherent or intuitive. For example, t is often used for time, not age. It would be clearer to use more intuitive notation, such as a for age. Then, in $F_j(t)$ on page 8, t is referred to as 'a certain age'. Is this CA or one of the candidate BAs? Also, s_j is defined as the residual standard error for the jth marker on page 8, but then s_{CA} is referred to as 'an estimate from a regression on CA'. This is confusing and should be clarified. - The Supporting Information appears to continue the numbering of the main text, which is not standard practice.

Thank you for this feedback. We went over the manuscript making numerous adjustments. Notably, we slightly reformulated equations as well as their descriptions, and removed redefinitions of acronyms. These points as well as other typographical errors have been addressed.

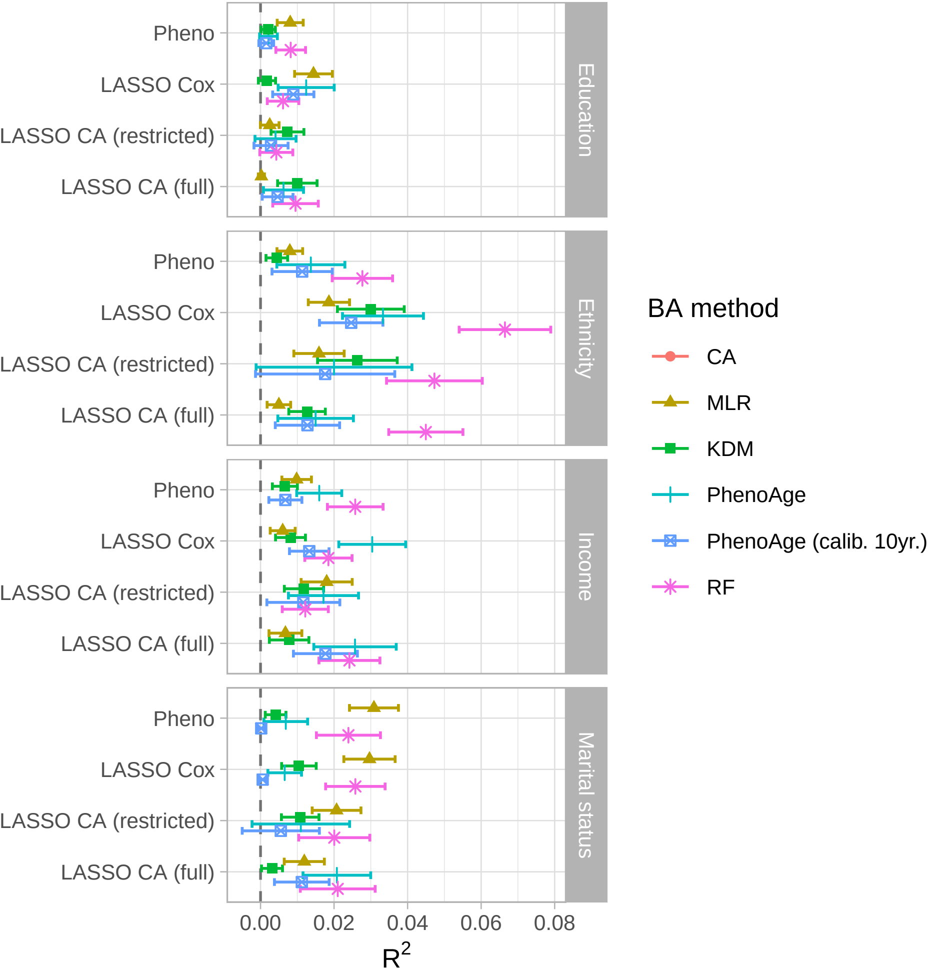
We believe these revisions have substantially improved the manuscript and address all the reviewer's concerns. We hope the revised manuscript now meets the standards for publication in the European Actuarial Journal and look forward to your decision.

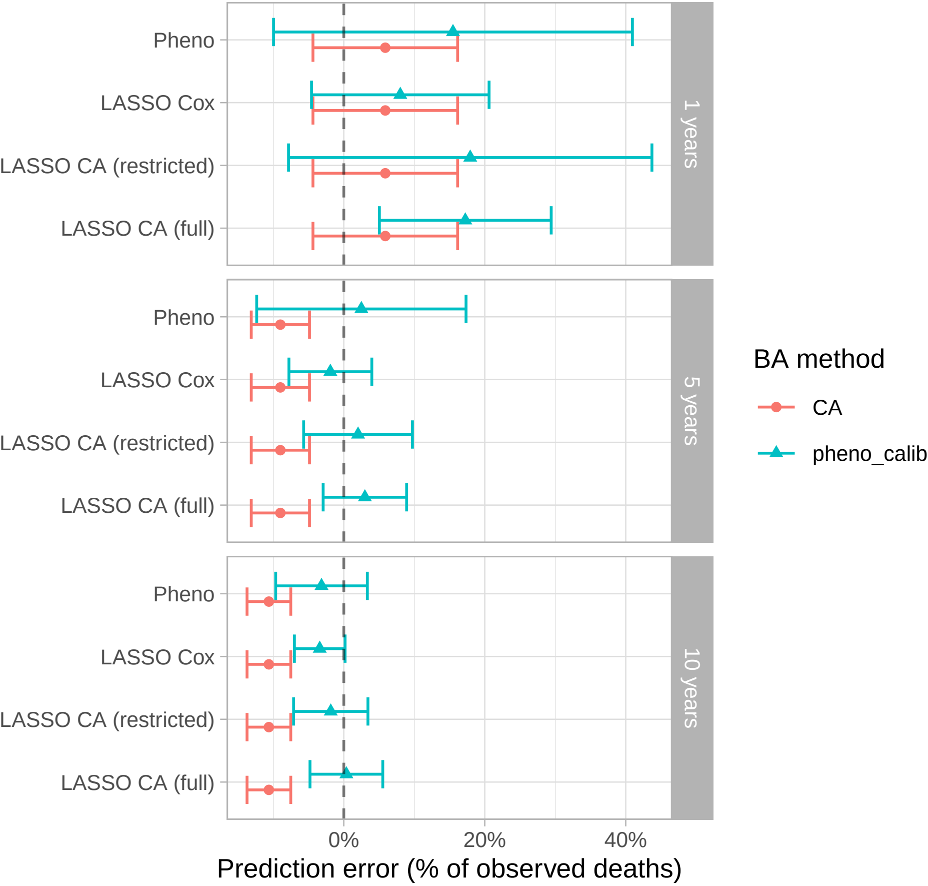
References

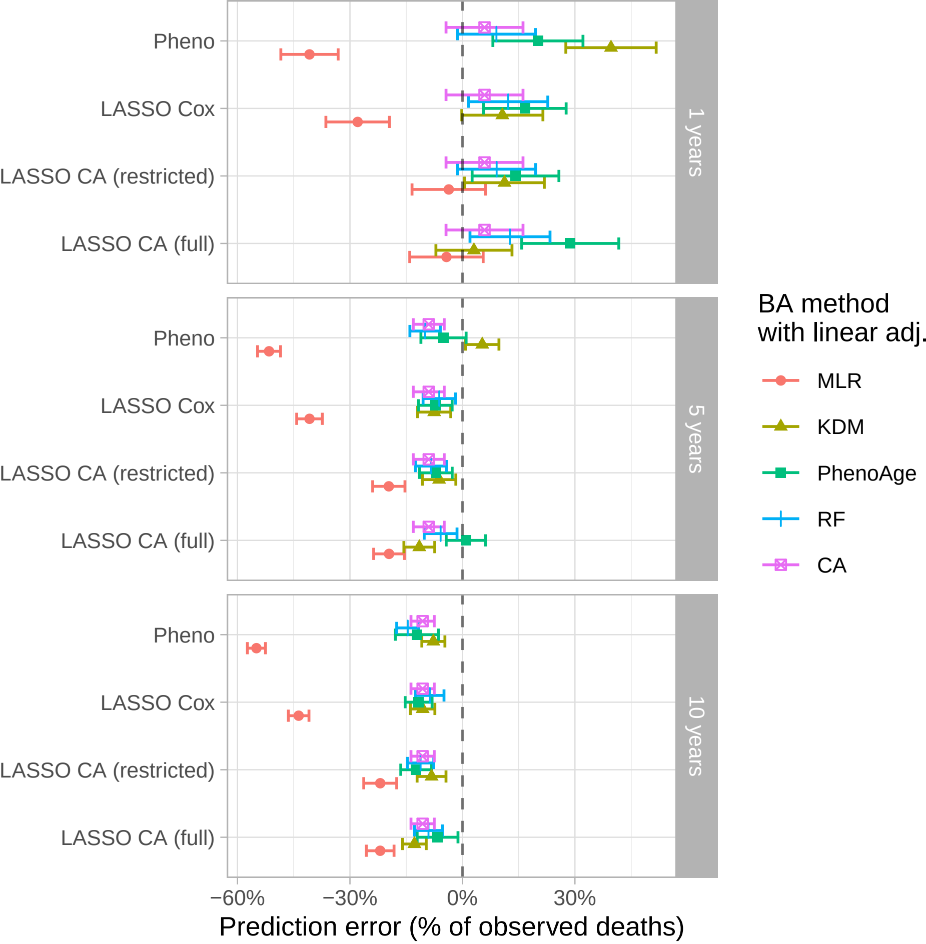
Levine, Morgan E., Ake T. Lu, Austin Quach, Brian H. Chen, Themistocles L. Assimes, Stefania Bandinelli, Lifang Hou, et al. 2018. "An Epigenetic Biomarker of Aging for Lifespan and Healthspan." *Aging* 10 (4): 573–91. https://doi.org/10.18632/aging.101414.

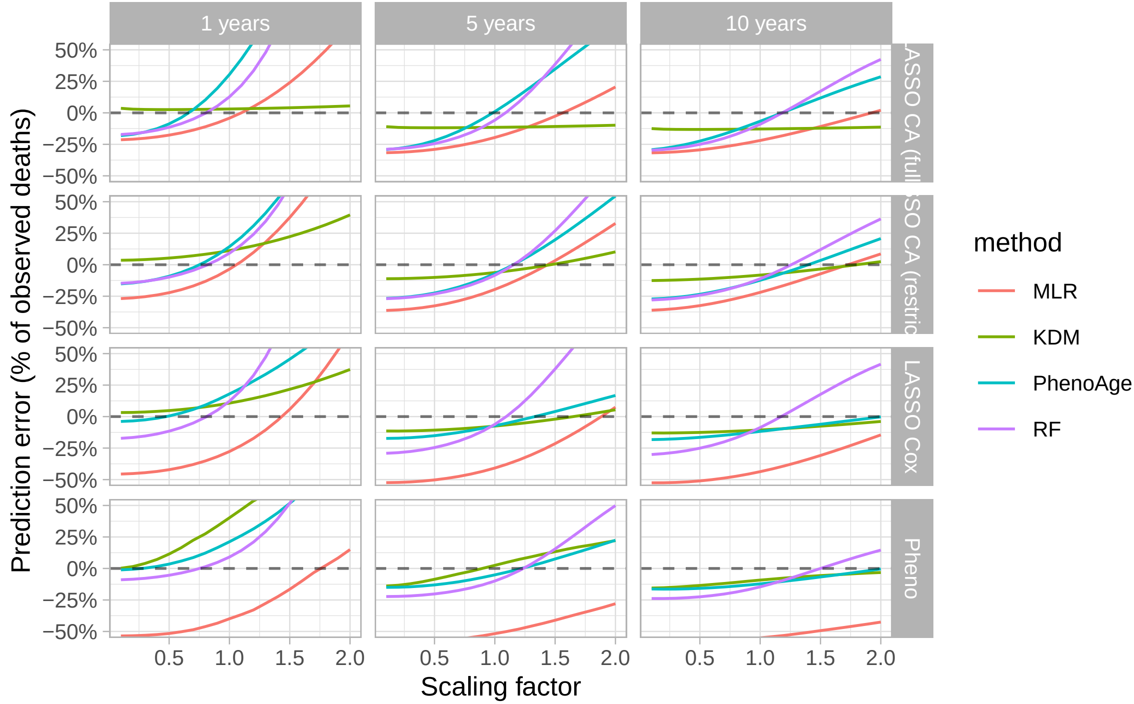
Qiu, Wei, Hugh Chen, Matt Kaeberlein, and Su-In Lee. 2023. "ExplaiNAble BioLogical Age (ENABL Age): An Artificial Intelligence Framework for Interpretable Biological Age." *The Lancet Healthy Longevity* 4 (12): e711–23. https://doi.org/10.1016/S2666-7568(23)00189-7.

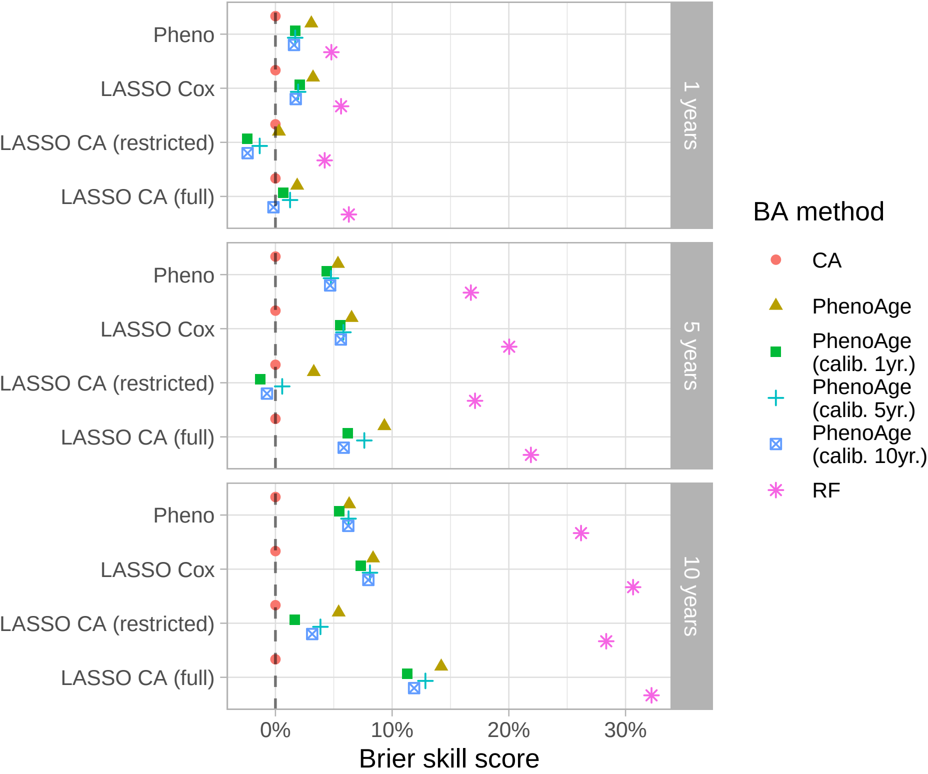
van Buuren, Stef. 2018. *Flexible Imputation of Missing Data, Second Edition.* 2nd ed. Second edition. Boca Raton, Florida : CRC Press, [2019]: Chapman; Hall/CRC. https://doi.org/10.1201/9780429492259.

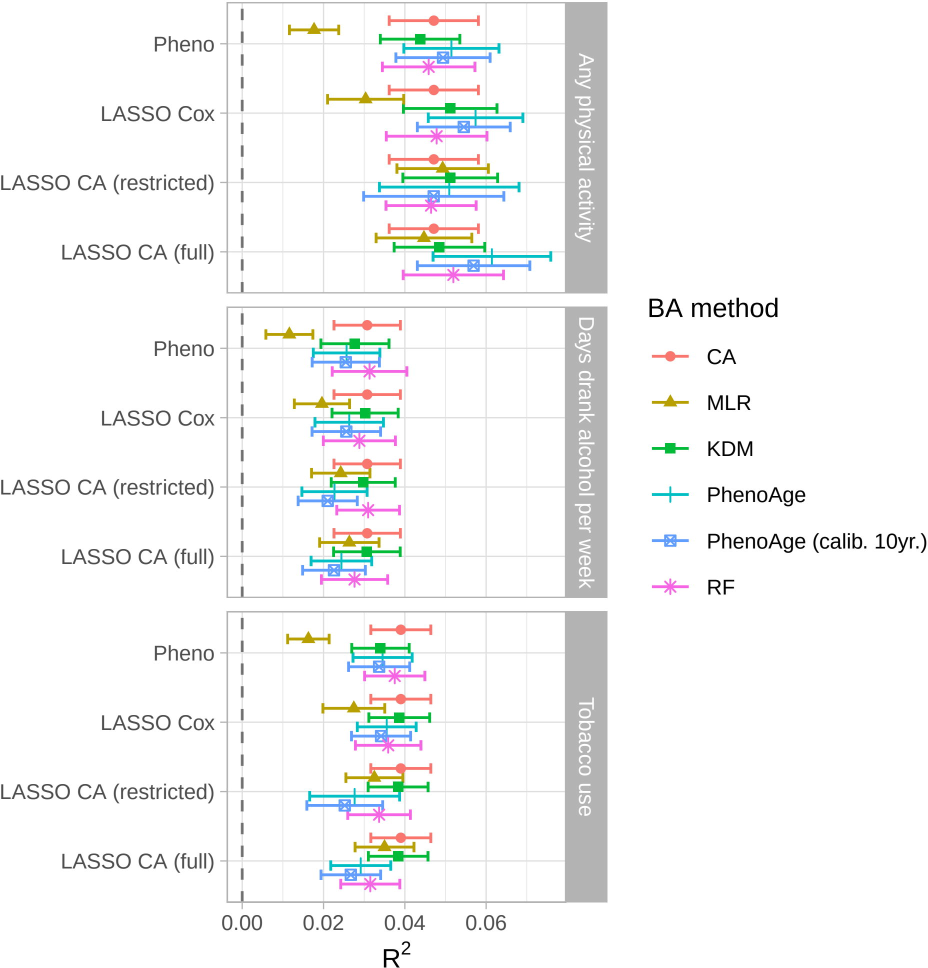

Pheno
LASSO Cox
LASSO CA (restricted)
LASSO CA (full)
Education
Ethnicity
Income
Marital status
0.00
0.02
0.04
0.06
0.08
R^2
BA method
CA
MLR
KDM
PhenoAge
PhenoAge (calib. 10yr.)
RF

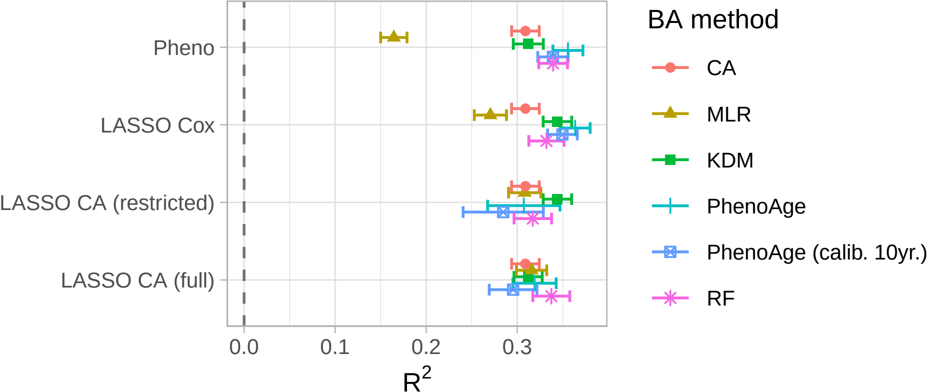

Pheno
LASSO Cox
LASSO CA (restricted)
LASSO CA (full)
1 years
5 years
10 years
0%
20%
40%
Prediction error (% of observed deaths)
BA method
CA
pheno_calib


Pheno
LASSO Cox
LASSO CA (restricted)
LASSO CA (full)
1 years
5 years
10 years
−60%
−30%
0%
30%
Prediction error (% of observed deaths)
BA method
with linear adj.
MLR
KDM
PhenoAge
RF
CA

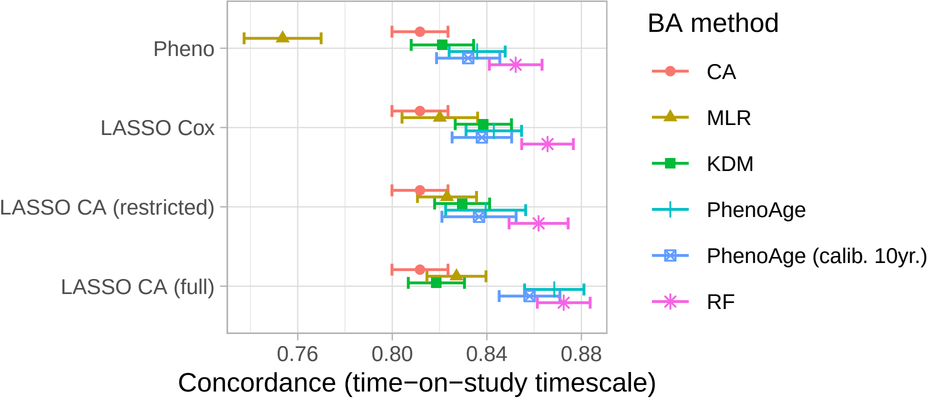

1 years
5 years
10 years
LASSO CA (full
LASSO CA (restric
LASSO Cox
Pheno
50%
25%
0%
−25%
−50%
Prediction error (% of observed deaths)
0.5
1.0
1.5
2.0
Scaling factor
method
MLR
KDM
PhenoAge
RF


Pheno
LASSO Cox
LASSO CA (restricted)
LASSO CA (full)
1 years
5 years
10 years
0%
10%
20%
30%
Brier skill score
BA method
CA
PhenoAge
PhenoAge (calib. 1yr.)
PhenoAge (calib. 5yr.)
PhenoAge (calib. 10yr.)
RF


Pheno
LASSO Cox
LASSO CA (restricted)
LASSO CA (full)
Any physical activity
Days drank alcohol per week
Tobacco use
0.00
0.02
0.04
0.06
R^2
BA method
CA
MLR
KDM
PhenoAge
PhenoAge (calib. 10yr.)
RF



Pheno
LASSO Cox
LASSO CA (restricted)
LASSO CA (full)
0.0
0.1
0.2
0.3
R^2
BA method
CA
MLR
KDM
PhenoAge
PhenoAge (calib. 10yr.)
RF


Pheno
LASSO Cox
LASSO CA (restricted)
LASSO CA (full)
0.76
0.80
0.84
0.88
Concordance (time−on−study timescale)
BA method
CA
MLR
KDM
PhenoAge
PhenoAge (calib. 10yr.)
RF