Biological Age for Prevention in Insurance

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

[1]Laboratoire SAF EA2429, ISFA, University of Lyon, Université Claude Bernard Lyon 1, 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.

Abstract

Biological age (BA) offers a promising approach to encapsulate 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 NHANES data, we compare four BA calculation methods—multiple linear regression (MLR), Klemera-Doubal Method (KDM), PhenoAge, and Random Forest (RF). We include a practical application of estimating death counts from life tables.

Our findings reveal that PhenoAge and RF consistently outperform other methods in mortality prediction and provide a better match with observed death counts after calibration. 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 estiating 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

1.1 Prevention

Health-related risks are unique in that they can be significantly altered by medical intervention and lifestyle choices. The ability of a potential policyholder to reduce these risks—meaning the probability of occurrence of adverse health events—is referred to as self-protection. Many types of insurance, such as health coverage (medical expenses), life insurance, long-term care insurance, and disability coverage, cover just such risks. The seminal work of Ehrlich and Becker [1] explored the impact of self-protection on insurance demand, showing that the relationship between market insurance availability and self-protection is complex: the two can either complement or substitute each other depending on specific conditions. Notably, Ehrlich and Becker [1] demonstrated that the existence of an insurance market could, under certain conditions, increase the demand for self-protection. For this to occur, insurance premiums must be sensitive to risk reduction resulting from self-protection; if insurance pricing ignores self-protection efforts, it can create a moral hazard, where policyholders are less incentivized to reduce risk.

Even outside microeconomic analysis, insurers have a vested interest in promoting policyholder health. This is especially true for policies where insurers have limited control over policyholder selection, as is often the case with group insurance contracts, such as employer-sponsored health benefits. Here, the dynamics described by Ehrlich and Becker [1] extend to the employer level: employers benefit from a healthier workforce, which reduces the cost of market insurance and improves productivity. We refer to such collective action aimed at reducing risk as prevention.

There are a few pathways to implement a prevention program. First and most direct is by including the all information on the risk level into the calculation of the premium, thus incentivizing more self-protection as per theory. This could segment the market not only by biological measures but also by the willingness to participate in these programs, leading to concerns about discrimination. However, such programs should also result in a healthier population overall, it is therefore a balancing act from a public health perspective. Second possible approach is by rewarding preventive action (self-protecting action to be precise), rather than focusing solely on absolute risk levels, monetarily or otherwise. Encouraging self-protection directly does not necessarily lead to increased segmentation, as the willingness to participate and the ability to reduce risk are not a priori tied to absolute risk levels. While there is currently no clear data on whether higher-risk individuals are more likely to participate, this association could significantly impact the design and outcomes of prevention programs. Last approach is though partnerships with some service offering health-promoting activity, e.g., with

gyms, wearable manufacturers, fitness applications etc. This is essentially targeted advertisement for the numerous applications already available.

There are many examples of prevention programs by insurance companies that encourage a healthy lifestyle through one or a combination of the above approaches. For instance, "Generali Vitality," is a program targeting employer-sponsored policyholders, offers monetary rewards for consistent physical activity (e.g., walking), a balanced diet, and preventive healthcare (e.g., vaccinations, checkups). This app-based program provides statistics and recommendations based on user activity. Other similar programs are offered under the Vitality brand, notably in the UK and Southeast Asia, through a joint venture with AIA. UnitedHealthcare's UHC Rewards also targets employer-sponsored policyholders, providing monetary rewards for daily step goals and regular checkups. Another example is Dacadoo, a company that partners with insurers and employers to offer personalized health tracking. AON's Well One application, based on Dacadoo products, focuses on employee health and happiness tracking, albeit without direct monetary rewards. This highlights how such programs often fit into larger "workplace wellness" initiatives, particularly popular in the United States.

In addition to these prevention-focused initiatives, tools like the QaliDays calculator exemplify how biological age can be used to communicate risk levels effectively. QaliDays is designed to evaluate the need for long-term care contracts by synthesizing complex health information into a single metric that is accessible to both insurers and policyholders. This use of biological age emphasizes its dual role as a means of quantifying risk and simplifying communication.

For individual policyholders, as opposed to group or employer-sponsored policies, there appear to be fewer reward programs. Some examples include Ambetter's My Health Pays, Anthem's Smart Rewards, and Bright HealthCare Rewards. Additionally, the existence of innumerable health and fitness apps providing tracking and gamification features to encourage healthier lifestyles—without any direct ties to insurance—shows that individuals are often self-motivated to stay healthy.

Some programs reward specific activities directly, as seen in UHC Rewards, which attributes a dollar value to each day with at least 5000 steps. Others construct a health score, such as Dacadoo's "HealthScore" or the "Vitality Age" used by Vitality programs. This "Vitality Age" is an example of a biological age (BA) a concept well known in medical literature that seeks to quantify the aging process, or rather better represent the expressions of aging on the body. BAs are particularly useful in this context because they integrate measurable health information into a single interpretable value that can inform risk assessment and prevention strategies.

1.2 Biological age

Aging is a biological process that concerns almost all organisms. In humans, aging is associated with a higher incidence of disease, increased medical needs, and risk of death. This progressive degradation is called senescence, and many attempts have been made to measure it. Chronological age (CA), that is, the time since birth, is at best a proxy for the actual rate of senescence, with many factors, such as genetics, behavior,

and environment, impacting the actual rate of aging. Biological age (BA) in an alternative conception of age which is intended to reflect the underlying senescence process. Measuring BA, however, is challenging, and many methods have been proposed.

Beyond the general description given earlier, BA does not have a precise, universally accepted definition. Instead, several methods have been proposed to calculate an age that better reflects outcomes such as mortality. Following Klemera and Doubal [2]'s characterization, we define BA as a set of biological markers coupled with a statistical method to combine them into a quantity that can be interpreted as an age.

This characterization leads to a multitude of possible BAs. Together with the fact that the aging process itself has many causes and consequences, it is unlikely that a single BA could reflect all of these equally well. Thus, the whole exercise is best viewed as a pragmatic search for a combination of markers and methods that provide useful predictions for a quantity of interest, rather than a holistic measure of health [3]. There have been high hopes expressed about the potential of BA. A well-constructed BA could be used to measure the efficacy of treatment without having years of followup [4], or even quantify the slowdown or reversal of aging [5]. We further believe that there is great value in expressing the impact of various markers as an age due to ease of interpretation and communication that it affords.

Various choices have been proposed for both biological markers and the methods to combine them. The earliest attempts to construct a biological age linearly combined biological markers using linear regression or principal component (PC) analysis, in hopes of finding a dimension or a linear combination that best reflects aging. See, for example, Hollingsworth et al. [6] for an early proposal to construct a BA based on MLR, or Nakamura [7] for a PC-based analysis.

There have been efforts to provide a principled construction for BA, most notably in the work of Klemera and Doubal [2], which gave rise to the KDM BA. But the larger innovation came with the increased use of omics-based, i.e., genetic, markers to construct BA. Horvath [8] was the first to introduce a biological age based on DNA methylation. The proposed method still essentially uses a linear model but is based on methylation data as the biomarkers.

Levine [9] first explored the link between BA and mortality and then went on to introduce a BA based explicitly on predicting mortality [10]. This gave rise to the so-called 'second-generation epigenetic clocks.' McCrory et al. [11] provides a comparison of such epigenetic clocks.

In summary, most BA methods are based on regressing some outcome on a set of biomarkers. The main differences between methods are the:

1. outcome they are based on (e.g., chronological age or mortality),
2. biomarkers that are used (e.g., clinical measures or epigenetic data),
3. method used (e.g., how the first two points combine to create a BA).

More recently, efforts have been made to apply various machine learning algorithms to improve the quality of the models; see Li et al. [12] for an overview.

Considerable work has been done in cataloging various BAs as well as the biomarkers they are based on. Li et al. [12] provides an exhaustive list of various BAs calculated over time, together with the methods used, country of study, sample size, and the

number and type of biomarkers used (although epigenetic clocks are omitted). Kuiper et al. [13] instead focuses exclusively on epigenetic clocks and includes comparisons of associations with mortality and clinical frailty measures. McCrory et al. [11] also compares epigenetic clocks, focusing on their association with other age-related markers (e.g., chronological age, walking speed, grip strength). Bafei and Shen [14] describes various BA methods. Hastings et al. [15] compares multiple aging measures, including PhenoAge and KDM, in relation to multiple age-related outcomes, such as physical and cognitive functioning, using NHANES data. Cho et al. [16] compares BA methods available at the time, such as MLR, KDM, and HocM, on the Work Ability Index (WAI) using a small dataset. It also provides a detailed discussion of the considerations involved in calculating BA. Klemera and Doubal [2], besides introducing KDM, provides a useful discussion of BA methods in general.

This study complements prior works, particularly Bafei and Shen [14], by providing a quantitative comparison of methods for calculating BA. Unlike Hastings et al. [15], which focuses on pre-defined BAs, our study evaluates BA calculation methods themselves by comparing how these methods perform when applied to different biomarker sets. To ensure a fair and comprehensive assessment, we define and use four distinct biomarker sets to examine the consistency and robustness of BA calculations under various conditions. Our comparison is based on a range of outcomes, including mortality and associations with CA.

1.3 Biological ages for prevention

BA is a well-developed concept with existing practical applications in insurance and public health. However, actuarial literature provides little guidance on constructing BAs as a risk indicator for health-related risks, and the topic remains largely absent from discourse in the field.

In public health contexts, BA has been demonstrated as an effective tool for assessing health outcomes and guiding interventions. For example, Kang et al. [17] illustrated how BA metrics could be used to predict health outcomes and support public health initiatives in South Korea. Similarly, Petit Sarazin [18] proposed the calculation of an "aging score," emphasizing BA's ability to synthesize complex health information into a single, interpretable metric.

On the other hand, BA is directly useful as a risk metric. It can serve as a drop-in replacement for chronological age (CA) in premium calculations, while improving the predictability of risk. This facilitates the transition to risk assessments grounded in personalized health data. Beyond pricing, BA provides valuable insights into residual life expectancy, making it particularly useful for retirement contracts, where it can enhance the accuracy of provision estimates.

Finally, factors that influence BA, such as lifestyle and behavior, are often directly tied to the risk levels that determine premiums. This connection allows for the design of prevention programs that not only encourage healthier behaviors but also align financial incentives with reduced risk. It is precisely this dual property of being both easy to interpret for policyholders and a reliable risk estimator that makes BA a prime candidate to serve as the basis for prevention programs.

The key assumption underpinning these applications is that BA is easier to interpret than direct measures of risk, such as incidence probabilities. This assumption holds particularly true for risks strongly associated with age and with relatively low incidence rates. For example, in life insurance, the difference between a 0.26% and a 0.30% annual mortality risk might seem negligible, but presenting this difference as "biological age 35" versus "chronological age 30" is more striking and likely to resonate with policyholders. This framing is particularly advantageous for long-term, low-incidence risks, as it encourages prevention early, when interventions are most effective.

Constructing BA has unique data requirements that can, at least initially, be addressed using NHANES. The NHANES dataset offers a comprehensive range of biomarkers and demographic data, making it a valuable resource for constructing biological ages tailored to various contexts. Its breadth allows for flexibility in developing BAs for specific prevention goals or health outcomes. However, NHANES also presents certain challenges, such as incomplete joint observations and data limited to specific survey cycles. To address these limitations, we employ multiple imputation, ensuring data continuity and robust statistical comparisons. This approach not only strengthens the foundation for BA research but also provides a scalable framework for practical and actuarial applications. While NHANES is particularly useful for initial implementation, longer-term applications of BA in insurance or prevention programs will require more frequent and individualized data collection.

This article provides an introduction to constructing biological ages (BAs) as health indicators and explores their potential applications in prevention programs, particularly in insurance contexts. First, we review common BA methods, including their underlying principles and methodological differences. We then perform a quantitative comparison of these methods using multiply imputed NHANES data, evaluating their association with mortality and health outcomes. Special attention is given to the ability of BA methods to estimate death counts using a standard life table, where calibration techniques are applied to address biases. Finally, we discuss how BA methods can be adapted for prevention-focused applications, emphasizing the importance of aligning methods with practical needs, such as interpretability for policyholders and flexibility for insurers. These findings highlight the potential of BA to act as both a reliable risk estimator and a tool for incentivizing healthier behaviors.

2 Methods

In this article, we use multiple BA methods to compute BA on multiply imputed NHANES data. These ages are then compared using a battery of criteria that evaluate the ages' associations with various outcomes, including mortality, CA, diseases, and self-rated health. To focus on the BA and gain insight into inner workings of various methods, we calculate BAs on the same four sets of biomarkers for all methods.

2.1 Biological ages

In this article we compare four BA methods: MLR, KDM, PhenoAge, and random forest based method (RF). These biological age calculation methods are chosen to

represent broad approaches to calculating biological ages, namely BAs based on CA, and mortality. In this section we formally introduce the methods for calculating BA. 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,
- as before, CA is the chronological age and BA is the biological age

All BAs are fit separately for each sex.

2.1.1 Multiple Linear Regression

The earliest proposed method for calculating BA was introduced in Hollingsworth et al. [6]. 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:

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

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,

$$BA := \sum_{j=1}^{J} \beta_j X_j. \tag{1}$$

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 valid 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" [4]. Regression to the mean may also result in bias in observations far from the global age average [19].

2.1.2 Klemera and Doubal Method

Klemera and Doubal [2] 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 defined as the age that minimizes

the distance between observed and expected values of biological markers:

$$BA := \text{argmax}_t Q(t|X) = \text{argmax}_t \sum_{j=1}^{J} \alpha_j (F_j(t) - X_j).$$

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 a certain age.

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

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

This is essentially a weighted average of the markers X_j and the CA and is the formula most commonly associated with KDM. s_{CA} is an estimate of a regression of CA on the hypothetical latent BA, in practice it provides a weight for CA.

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:

$$BA = \underbrace{\frac{-\sum_{j=1}^{J} q_j \frac{k_j}{s_j^2}}{\sum_{j=1}^{J} \left(\frac{k_j}{s_j}\right)^2 + \frac{1}{s_{\text{CA}}}}}_{\text{Constant}} + \underbrace{\sum_{j=1}^{J} \left(\frac{k_j / s_j^2}{\sum_{i=1}^{J} \left(\frac{k_i}{s_i}\right)^2 + \frac{1}{s_{\text{CA}}}} \right) X_j}_{\text{Linear combination of markers}} + \underbrace{\frac{1/s_{\text{CA}}}{\sum_{j=1}^{J} \left(\frac{k_j}{s_j}\right)^2 + \frac{1}{s_{\text{CA}}}} \text{CA}}_{\text{CA contribution}}. \tag{3}$$

As KDM assumes that markers are uncorrelated, thus 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.

2.1.3 PhenoAge

Levine et al. [10] introduced an epigenetic clock based on genetic markers, called DNAm PhenoAge (where DNAm stands for DNA methylation). DNA methylation data is high-dimentional making it difficult to analyze it directly. Levine et al. [10] simplify the problem by first constructing an intermediate BA, PhenotypicAge, and then fit linear model with this PhenotypicAge the outcome. This intermediate BA, PhenotypicAge, is constructed by matching 10-year survival probability as estimated by a full Gomperz proportional hazards model to the same probability in a reference

model. This BA is thus notable for choosing mortality as the criteria for its definition. In this study we focus on the intermediate PhenotypicAge (not the full DNAm PhenoAge) as it does not require DNA methylation data. We shall refer to it simply as PhenoAge.

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

$$h(t) = \alpha \exp\left(t/\beta + \gamma_0 \mathrm{CA} + X\gamma\right),$$

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

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

Here, α', β', and γ_0' represent the same parameters as their non-primed counterparts. PhenoAge is then defined as the age where the 10-year survival probability predicted by the full model is the same as the probability predicted by the reference model, that is, PhenoAge is the solution to:

$$S^{\text{Full model}}(10|age = \mathrm{CA}) = S^{\text{Ref. model}}(10|age = \text{PhenoAge}) \tag{4}$$

Finally, if the expression above is developed, we find that PhenoAge corresponds to a linear combination of biological markers and CA. The constant term in the formula depends on shape and rate parameters of both models and on the period chosen for matching survival probabilities, 10 years in the original article, and denoted below by t below. PhenoAge is thus also given by :

$$BA := \frac{\log\left(\frac{\alpha\beta\left(e^{t/\beta}-1\right)}{\alpha'\beta'\left(e^{t/\beta'}-1\right)}\right)}{\gamma_0'} + \frac{\gamma_0}{\gamma_0'}\mathrm{CA} + \frac{1}{\gamma_0'}\sum_{j=1}^{J}\gamma_j x_j.$$

In Levine et al. [10] 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 only requires changing the t parameter, which in turn 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.

2.1.4 Random Forest BA

The fourth and last BA method considered in this article is based on ENABL Age from Qiu et al. [20]. Qiu et al. [20] use Gradient Boosting Trees (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. [20] 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. [21]. Once a valid mortality score is obtained, it is transformed to a scale interpretable as age by applying an exponential curve :

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

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

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

Unlike other methods, this BA cannot be expressed as the 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 intrest.

2.2 Biological markers

The key choice for any BA method is the biological markers used to construct it. Indeed, the novelty of many biological clocks is the biological marker that they are based on. In this work, however, the focus is on the method of calculation of BA. To keep comparison fair, the variable selection is done independently of the BA method.

We define four sets of biological markers to calculate BAs. All four sets are chosen from a pool of over 100 biological markers available in the NHANES. This pool includes a wide gamut of variables, including a variety of blood markers, and physical examinations. Various socio-economic and dietary variables are excluded from the pool to stay comparable with existing literature on the subject. Practical applications should consider including these variables.

The first set of variables is a set of 9 blood biomarkers that was chosen in Levine et al. [10] though a penalized Cox model. We use this set to establish a baseline performance from a well-known BA. We refer to this marker set as “Pheno”. The markers included therein are listed in Table 1.

Table 1 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 2 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)

The second set selected using the same methodology, i.e., penalized Cox, but on the full NHANES 1999-2018 dataset (sample size of approximately 50,000 observation of which approximately 10,000 is reserved for testing, and with a wider gamut of biomarkers) used in this article, rather than NHANES III (1988-1994) (sample size approximately 10,000) as used in Levine et al. [10]. The larger sample size results in a slightly larger set of markers being selected. Moreover, of 9 markers from the first set, only 2 are present in this one. We call this marker set "LASSO Cox". The markers included in this set are listed in Table 2.

The third and fourth sets are also selected by a penalized regression, but instead of mortality it uses CA as the outcome. This should provide an interesting comparison for the behavior of BA methods. The "1se" rule for choosing optimal penalization level leads to many more variables being chosen than in first or the second set. Therefore, the third set contains the variables chosen by the penalized regression at the penalty level that leads to approximately the same number of variables as in the second set.

Table 3 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)

Table 4 The four selected marker sets, with corresponding names, number of biomarkers and the selection criterion.

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

We call this marker set "LASSO CA (restricted)". The markers included in this set are listed in Table 3. The fourth set contains the variables selected with "1se" rule, and is called "LASSO CA (full). The markers included in this set are listed in Table 6, in the supplementary material due to its size. Only one variable is present both in"LASSO Cox" and restricted "LASSO CA" marker sets, and none reappear from "Pheno" marker set.

Table 4 summarizes the marker sets used. The four marker sets can be naturally ordered by their size, and the criteria used to construct them : small to large and mortality to CA.

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. The optimal penalization level is chosen to be within one standard deviation of minimal error. Then, variables that were chosen in more than half of imputation replications were chosen. This corresponds to the "majority" strategy proposed in Brand [22].

2.3 Comparison criteria

We chose a range of comparisons criteria to both compare BA between each other illustrate the strengths and weaknesses of considered BAs. The primary criteria include :

- Association with CA
- Mortality as measured by concordance index
- Association with chronic diseases

Each of these criteria are of interest on their own sake, and can be directly linked to insurable risks, such as life insurance for mortality and long term care for chronic disease.

We also include secondary criteria which serve to contextualize BAs and show how that they capture various aspects of health :

- Association with self-rated health
- Association with various behaviors, such as smoking
- Association with sociological variables, such as ethnicity, income.

All metrics are calculated on the test dataset and are accompanied by 95% confidence intervals constructed based on estimated standard errors and an assumption of normality. R^2 confidence intervals are based on the Fisher transformation. All standard error estimates take into account the imputation procedures.

2.4 NHANES Data

The National Health and Nutrition Examination Survey (NHANES) collects health-related data on the resident civilian, non-institutionalized population of the United States. Although first conducted in 1971, this article focuses on the survey cycles between 1999 and 2018, the period when NHANES adopted its current form, before disruptions caused by the COVID-19 pandemic. Surveys conducted since 1999 are collectively known as Continuous NHANES, though we will refer to it simply as NHANES throughout this article.

NHANES encompasses a wide range of demographic, questionnaire, examination, and laboratory data. Subjects are initially interviewed at home, followed by further examination and testing at Mobile Examination Centers (MEC). NHANES provides a wide array of data, including dietary habits, laboratory blood work, cardiovascular stress tests, dental health, grip strength, and many other health indicators. This extensive range of indicators makes NHANES an attractive choice for studies on BA.

NHANES survey data is publicly available through a dedicated website [23]. The public-use data excludes information that could be potentially identifiable. Data is grouped into two-year cycles to ensure sufficient sample size for anonymity and robustness of estimates. Additionally, data files are organized by subject matter. For example, files prefixed with “SPX” contain spirometry test results, but spirometry was only conducted during the 2007-2012 cycles, so data related to it is missing for all other cycles.

This pattern of missing data presents analysts with a choice: either include spirometry data and restrict analysis to the 2007-2012 cycles, or exclude it entirely. This is the fundamental problem we address through multiple imputation.

NHANES itself does not follow up with participants beyond the survey. However, mortality data, including vital status and cause of death as of 2019, is publicly available through the NHANES linked mortality files [24]. Not all participants are eligible for linkage, with those under 18 excluded. Additionally, some individuals had insufficient information for linkage, resulting in missing values. Follow-up durations and causes of death were also perturbed in some cases to preserve anonymity.

The length of follow-up varies by cycle, with up to 20 years for the 1999 cycle. However, we believe this extended follow-up period makes it difficult to draw valid inferences about the impact of markers measured during examination on mortality. Mortality models like the Cox proportional hazards model implicitly assume that the observed marker values remain constant throughout the observation period, an assumption we consider untenable for long follow-up periods. We retain the full followup period in this study to keep results comparable to existing works. The exploration of the impact of the choice of followup period is an important area needing further work.

To produce nationally representative and reliable estimates for targeted groups, NHANES employs a complex survey design. The final data contains weights that account for both the survey design and non-response rates. To address the survey design in the context of imputation, we follow the methodology used for imputing NHANES III (1988-1994) [25], including the variables used in sampling in our imputation model. A detailed list of variables used in imputation is provided in the next section.

In this article, we only include individuals aged 20 to 79. This age range is motivated by several factors. First, linked mortality data is only available for individuals aged 18 and older. Moreover, those aged 80 and above are all top-coded as 80 in some cycles (some cycles top code starting from 85) to preserve anonymity. Additionally, we consider this to be the most useful age range for the downstream task of calculating BA with a focus on prevention. Ages under 20 should, ideally, exhibit lower variance in BA, while ages beyond 79 may be too late for effective preventive measures. Finally, the Gompertz hazard model, used for PhenoAge, is best suited for this age range, as it does not fit well with ages outside the selected range.

2.5 Multiple imputation

This study introduces an innovative approach by applying multiple imputation to NHANES data in the context of comparing BA models. Multiple imputation by chained equations (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 analyzes. These results are then pooled, yielding robust inferences for both point estimates and variance estimates.

First introduced by Rubin [26] 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 [27]. 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.

This approach focuses on estimating the metrics of the BA method if fit on a hypothetical new dataset. If the goal was to construct the best BA possible based on NHANES data, then we would instead first aggregate the BA obtained for each individual, and only then compute the metrics. This what is done to calculate average BA for each CA for Figure 1.

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 [28].

Table 1 details the MICE Algorithm, from Van Buuren [29], 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.

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

- Biological markers of interest.
- Variables related to mortality (i.e., age at baseline, age at end of follow-up, vital status).
- Variables used in survey design, as per Liu et al. [30] (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.

Algorithm 1 MICE Algorithm

Input: Dataset $Y = \{Y_1, Y_2, \ldots, Y_p\}$ with missing values
Output: Multiple imputed 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^{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 imputed datasets

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.

2.6 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.

3 Biological Age comparison

There are clear differences between BAs in concordance scores as well as in correlation with CA. These two criteria lay the groundwork for our comparisons. Meanwhile, association with the number of disease and secondary metrics is much weaker and differences between BA hard to gauge. This is likely the result of neither BA explicitly aiming to maximize these quantities. To structure the analysis we first do a post-hoc analysis of the BA.

Among considered BA methods, RF and MLR methods both aim to estimate their respective criteria, mortality for RF, and CA for MLR. Unsurprisingly these methods tend to perform best with respect to their respective criteria. PhenoAge and KDM, however, both deviate from the straightforward evaluation of their respective criteria, mortality for PhenoAge and CA for KDM.

3.1 BA example

We illustrate the calculation of a BA on PhenoAge BA, fitted on the LASSO Cox marker set. As stated in the definition of PhenoAge, it can be expressed as a linear combination of CA and various markers. Table 5 provides the coefficients for this BA model. Coefficients are provided separately for males and females as models are fitted separately for both sexes. Data was centered and reduced before fitting, the coefficients therefore reflect change in BA in response to a 1 SD change in the underlying variable, except CA which was kept as-is.

Without proceeding to an analysis of each individual effect, we remark upon some notable points. First the intercept is positive for both males and females. The coefficient attributed to CA is in both cases slightly less than one. Let us take an example of a 40-year-old male with average value for each marker (an unrealistic scenario, as any individual is unlikely to have average values for all the markers). For this case the model simply attributes a BA of $7.21 + 0.83 \cdot 40 = 40.4$. Any deviation from the average marker values would be added to this age.

Table 5: PhenoAge BA model coefficients and corresponding standard errors, separate for each sex. Coefficients express change in BA per change of 1 standard deviation of marker.

Variable	Female Coef.	Male Coef.
RIDAGEYR	0.94	0.83
(Intercept)	2.12	7.21
BMX_WACR	1.25	2.45
LBDSALSI	-2.97	-2.05
LBDSCRSI	0.99	0.59
LBDSGBSI	0.23	1.58
LBDSGLSI	0.76	0.99
LBXBAP	1.06	0.31
LBXP1	NA	0.59
LBXRDW	1.41	2.04
LBXSAPSI	0.36	0.91
LBXSCLSI	-0.38	0.16
LBXSGTSI	0.71	0.51
LBXSNASI	-0.27	-1.01
SPX_FEV5h3	0.00	0.00
SPX_PEF5h3	0.00	0.00
URXUMA	0.54	0.27

3.2 Chronological age

We now examine the relationship of various BAs with CA. CA is the obvious reference for any quantity interpretable as age. As we shall see some BAs deviate too much from CA to be useful, whereas others are too close to it to provide any new information. Figure 1 compares the average BA for each method and for each age. Further, Figure

2 compares correlations of BAs with chronological age, measured using R^2. Most BAs achieve an R^2 of 0.6 or above.

MLR-BA is the notable exception, showing too little variation with age, especially for smaller marker sets. This is reflected with an R^2 as low as 0.21. MLR is constrained to using markers only, and not CA itself; otherwise, MLR would ignore available markers and coincide with CA. With this in mind, MLR's R^2 logically increases from the lowest value on the smallest "Pheno" marker set up to an $R^2 \approx 0.83$ on the largest full "LASSO CA" marker set.

KDM, on the other hand, closely follows the CA line, acheiving a higher correlation still, with all R^2 values above 0.82. Like MLR, KDM's correlation is largest for the largest marker set at 0.99—the largest R^2 obtained. Such high association with CA is due to KDM not being constrained to exclude CA, yet still being based on CA, though indirectly. This high correlation brings it too close to CA, as we shall see, hurting its performance on other criteria.

PhenoAge is also strongly correlated with CA. But its R^2 decreases with marker set size. It stars out around 0.9 and go down as low as 0.73.

A visual inspection of RF's curves in Figure 1 shows that unlike all other BAs, RF's curves have a distinctive shape : RF starts with a plateau for ages under 30, and proceeds to underestimate CA into the early 60s. This is reflected in RF's overall lower correlation with CA. The differences in R^2 between markers sets for this method is the simplest to interpret as RF has only one objective: predict mortality. In this respect, RF is correlated with CA only to the extent that mortality is as well. So RF is least correlated when the marker set is best able to predict mortality, i.e., "LASSO Cox" and the full "LASSO CA" marker sets, and most correlated when the marker set is least associated with mortality.

RF's distinctive shape, combined with its respectable performance suggests, that there the rate of aging, or the rate of increase of risk, is slower for ages under about 35. However, the systematic deviation from CA can be viewed as a drawback, as it will assign a BAs lower than CA to a large part of the population. Indeed, the comparisons done in the article are insensitive to shifts by a constant and would therefore not capture such a bias. All in all, this calls for some adjustments to the method before it can be put into practice.

3.3 Mortality

We now consider BA's association with mortality, measured via time–on-study concordance. Both all-cause mortality and cause-specific mortality is considered. The death causes examined are those available in NHANES linked mortality data, namely : Diseases of heart, Malignant neoplasms, Cerebrovascular diseases, Chronic lower respiratory diseases, Diabetes, Alzheimer's disease, Accidents, Nephritis, Influenza and pneumonia. The choice of timescale is important as time-on-study timescale concordance leads to much higher scores.

Figure 3 compares the concordance score of all BAs. For methods based on mortality, i.e., PhenoAge and RF, concordance is closely tied to the optimization criteria for these methods, resulting in good performance, with RF performing the best, followed by PhenoAge. KDM and MLR follow, and either tie with or trail behind CA

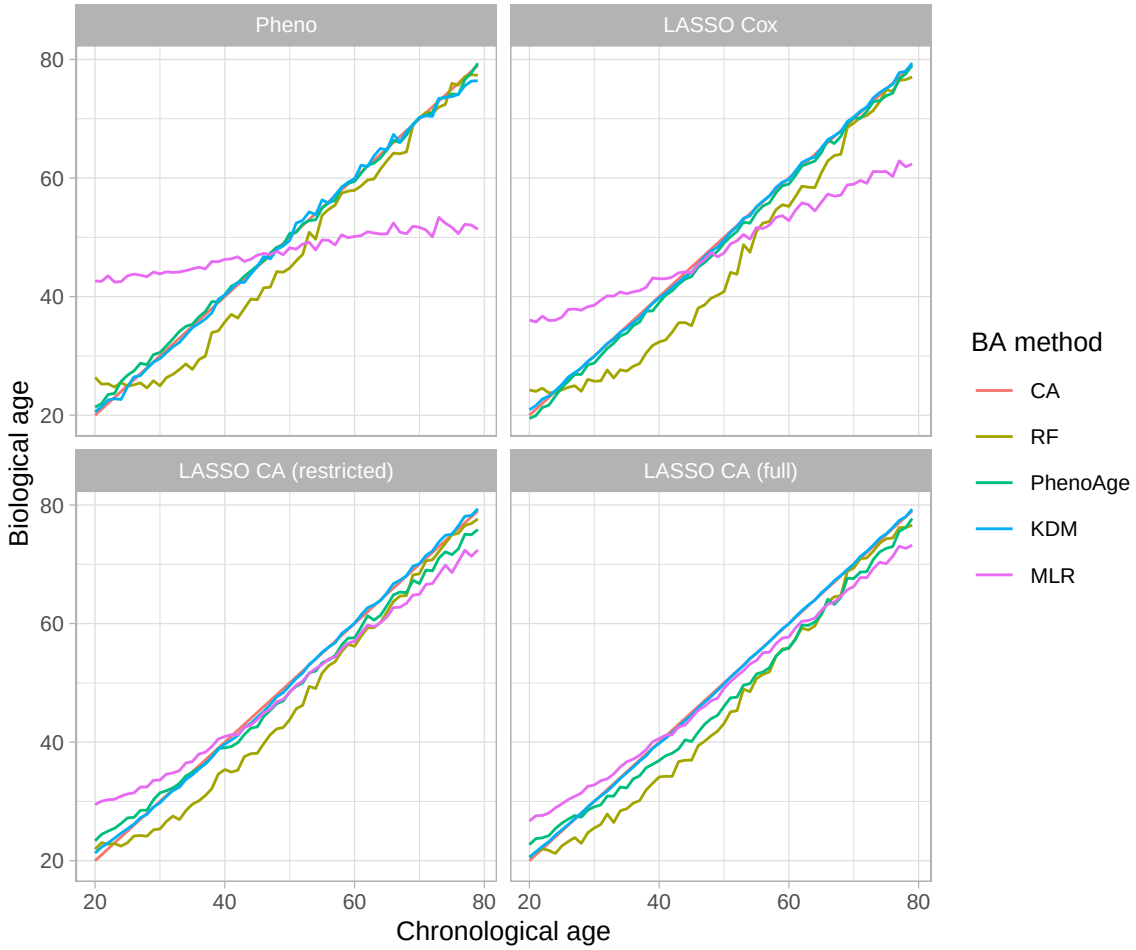

Fig. 1 Average BA with age, for each marker set.

depending on the marker set. Only RF and PhenoAge are significantly better than CA.

MLR-BA performance primarily depends on the method for selecting markers and is the only BA showing significant variation across marker sets. For "Pheno" marker set, MLR-BA's concordance is significantly below that of CA. For all other marker sets it is close to CA's concordance of 0.81.

KDM shows concordance scores that are slightly lagert that that of CA, but never significantly so. Being strongly correlated with CA, it is unable to achieve a better result.

PhenoAge, explicitly based on a survival model, should capture the impact of biological markers on mortality well. Indeed, for all marker sets achieves better performance than CA, MLR and KDM. Somewhat surprisingly, PhenoAge does not achieve a concordance significantly higher than that of KDM, although it is substantially

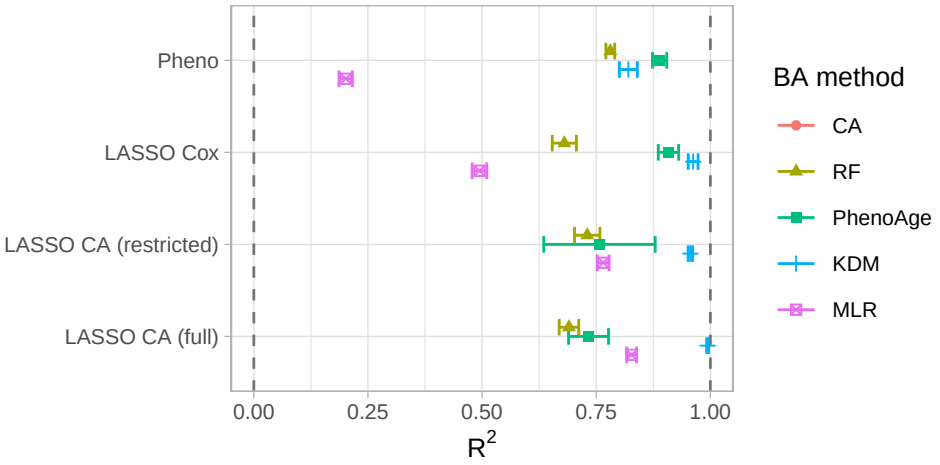

Fig. 2 R^2 vs chronological age.

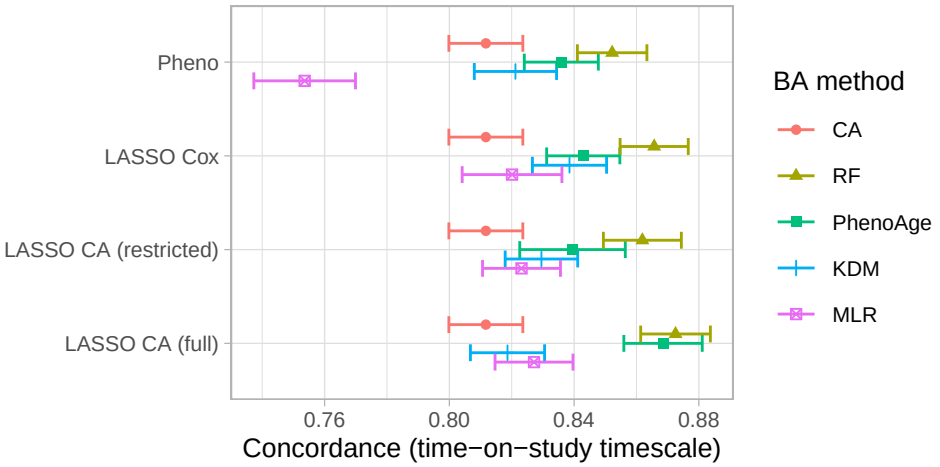

Fig. 3 Concordance with respect to age of death (all cause).

always higher than other BAs. Moreover, its concordance drops for the two LASSO CA marker sets. PhenoAge achieves the highest concordance on the "LASSO Cox" marker set, suggesting the importance the choice of variables.

RF, on the other hand, shows the best concordance for every marker set across both concordance measures. Its best performance is 0.873 on the full "LASSO CA" marker set, likely due to the large number of available markers. The second-best performance, only slightly lower at 0.866, is achieved on the much smaller "LASSO Cox" marker set. The good performance for "LASSO Cox", again underscoring the importance of the choice of variables. In spite of this, RF is not able to achieve concordance significantly different from PhenoAge.

The NHANES linked mortality data also includes information on specific causes of death. Figure 13 in the supplementary materials compares concordance scores for each

available cause. The same trends observed in all-cause mortality concordance reappear here for cause-specific mortality, although with considerably more uncertainty due to lower number of observed cause-specific deaths. Concordance scores vary across different diseases, with RF and PhenoAge generally lead in predicting survival time. Curiously, for Alzheimer's disease, no BA outperformed CA.

3.4 Chronic disease

In this section we examine the relationship between BA and the presence of various chronic diseases : cardiovascular disease, COPD, chronic kidney disease, asthma, arthritis, cancer, stroke, hypertension, hyperlipidemia, diabetes, and obesity. Chronic is often of interest on its own, but is also strongly associated with other risks such as disability and the need for long term care. The correlation between BA and each disease is measured using an R^2. Additionally, we compare correlation with the total number of chronic diseases. This allows for a comparison of how well different BA methods capture an individual's overall health status based on chronic disease prevalence.

Figure 4 compares the correlations between BAs and the number of comorbidities. The overall correlation between BAs and the number of diseases turned out to be a weak metric for distinguishing between BAs, as all methods hovered just above 0.3.

There are some exceptions. MLR, like its correlation with CA, starts significantly lower for the "Pheno" marker set but catches up with other methods by the "LASSO CA" marker set. PhenoAge also shows noticeably, though not significantly, worse performance for the "Lasso Cox" and restricted "LASSO CA" marker sets, with its dip in performance masked by large confidence intervals. These larger intervals likely stem from the uncertainty introduced by imputed missing values.

Only RF and PhenoAge significantly outperform CA, and only for the "Pheno" dataset. KDM also significantly surpasses CA, but only for the full "LASSO CA" marker set.

Overall, the differences in R^2 across marker sets were minimal, suggesting that BAs add little new information about the overall disease count that isn't already captured by CA.

Figure 5 compares the correlations between the presence of various diseases and BAs. The disease-specific correlations vary considerably, ranging from near zero for asthma and obesity to approximately 0.25 for hypertension.

Overall, the correlations for all BAs are close to those of the reference, CA. In fact, CA slightly outperforms other BAs in its association with cancer, hyperlipidemia, and arthritis, with CA significantly outperforming RF for hyperlipidemia. For other diseases, BAs tend to have slightly higher correlations than CA, but these differences are generally not significant. However, BA methods significantly outperform CA for kidney disease and diabetes. RF, in particular, significantly outperforms CA for kidney disease across three marker sets, while PhenoAge only achieves this with one marker set. For diabetes, all BAs show significantly higher correlations than CA in at least one marker set, with MLR surprisingly showing a competitive correlation for diabetes.

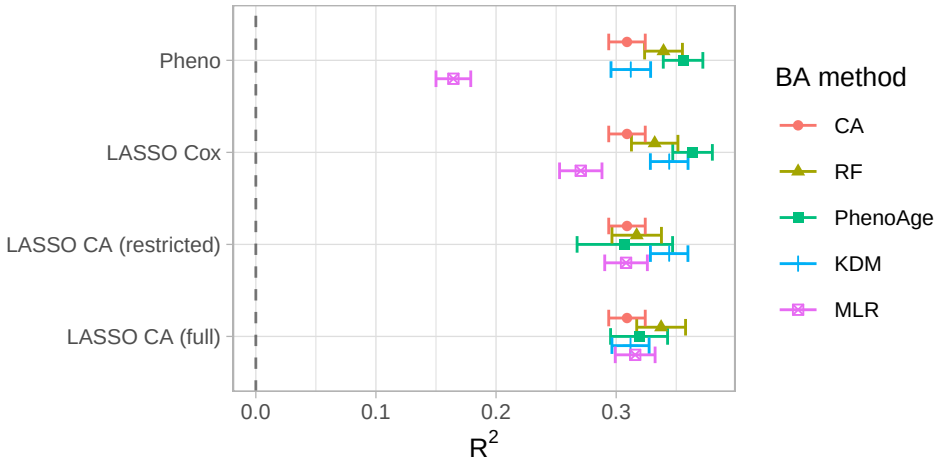

Fig. 4 R^2 vs the number of diseases.

3.5 Secondary metrics

We believe association with CA, mortality, and diseases to be the main criteria for evaluating a BA. However, there are a number of other criteria that may be of interest, but that we consider secondary. Such criteria are considered here.

3.5.1 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. Moreover, it is strongly correlated with CA, generally worsening with age. Figure 6 compares the R^2 values for the correlation between self-rated health and BAs.

All BAs are significantly associated with self-rated health, though the overall associations are weak, with $R^2 \approx 0.03 \pm 0.02$, and CA at $R^2 = 0.015$.

MLR and KDM do not achieve an R^2 significantly larger than CA for any marker set. PhenoAge's correlation is significantly higher than CA's for three out of four marker sets. However, PhenoAge itself is not significantly better than either MLR or KDM for any marker set. RF, meanwhile, is significantly different from CA for all marker sets, while also being significantly better than both MLR and KDM on the largest marker set.

3.5.2 Behavior

Figure 7 compares the R^2 values for the correlation between various behavioral variables and BA. The three behaviors examined are the number of alcoholic drinks consumed per week, the presence of any physical activity, and tobacco consumption.

All BAs show significant associations with the three behaviors considered here, yet none appear to be systematically different from the R^2 achieved by CA.

For physical activity, which is coded as 1 if any activity was declared and 0 otherwise, only PhenoAge for the full "LASSO CA" marker set achieves an R^2 significantly different from that of CA. However, this result is not statistically better than for any

other BA. In general, the various BAs yield slightly higher R^2 values than CA, but the differences are not significant. MLR for the "Pheno" marker set performs significantly worse in this case.

In terms of alcohol consumption, all BAs show similar associations to CA, with most BAs having R^2 values slightly lower than CA, though not significantly so.

For tobacco consumption, coded as "current," "past," or "never," only MLR for the "Pheno" marker set performs significantly worse than CA. All other BAs exhibit R^2 values slightly lower than that of CA, but again, without statistical significance.

3.5.3 Social

In this section we consider association of BAs with sociological factors that are known to have an impact on health. Included here are education, ethnicity, income level, and marital status. Marriage, for example, tends to prolong life. Many of these factors depend on age. The proportion of unmarried persons, for example, drops sharply until about 40, whereupon it stabilizes. Due to this association with age, we focus on the difference between BA and CA. This quantity should reflect the relative health of the individual for that age group. In fact, this quantity is often of interest on its own and called "BA acceleration" and defined as CA − BA. Figure 7 compares the R^2 values for the regression of BA acceleration on various behavioral variables.

Education

All BA accelerations are significantly associated with at least one marker set. The strongest association is observed for the "LASSO Cox" marker set ($R^2 \approx 0.02$) and the full "LASSO CA" marker set ($R^2 \approx 0.03$). MLR achieves a significant but small correlation. All other marker sets are not significantly different from each other but perform significantly better than MLR. PhenoAge, due to its large confidence intervals, is only significant for the smallest and the largest marker sets. KDM performs best for the last three marker sets, although the difference between PhenoAge and RF is not significant. RF is consistently significant and performs well across all marker sets.

Ethnicity

Almost all BA acceleration-marker set combinations are significantly associated with ethnicity. The two exceptions are MLR for the largest marker set and PhenoAge for the restricted "LASSO CA" marker set, where large confidence intervals hinder significance. KDM and MLR performed relatively poorly, with $R^2 \approx 0.01$ for all marker sets. PhenoAge likely performed better, but it is not significantly different from KDM. RF performed significantly better than MLR and KDM for three out of four marker sets. For the fourth marker set (restricted "LASSO CA"), all four BA accelerations performed similarly.

Income

All BA accelerations are significantly associated with income level, without exception. MLR seems to perform the worst, showing performance similar to KDM for the first two marker sets but dropping in performance thereafter. KDM, PhenoAge, and RF

performed similarly across all marker sets, with the only significant difference being between KDM and RF for the smallest marker set.

Marital Status

All BA accelerations are significantly associated with marital status, without exception. MLR starts with a strong correlation, but its performance declines as the size of the marker set increases, likely due to the strong correlation between MLR-BA acceleration and CA. KDM performs poorly, with the smallest R^2 for all marker sets, although it is never significantly different from the second worst. PhenoAge meanders from an R^2 close to that of KDM for the smallest marker set to an R^2 similar to that of RF for the third marker set. For the fourth marker set, PhenoAge is not significantly different from either RF or KDM. RF consistently performs well, with $R^2 \approx 0.03$ for all marker sets, frequently occupying first place or being tied for it.

4 Using BA to Estimate Death Counts

The previous section evaluated BA methods based on their associations with various outcomes using abstract measures such as linear correlation and concordance. While useful, these metrics are insensitive to scale. As we have seen, some BAs deviate significantly and systematically from CA. To explore the practical implications of these deviations, we evaluate BAs on a more concrete task: estimating the number of deaths. This analysis reveals that raw BA estimates are significantly biased. We attempt three approaches to align BAs with mortality data more closely: linearly adjusting ages, scaling the dispersion around the mean BA, and integrating the life table into PhenoAge's construction. Only the last approach provides satisfactory results.

In this section, we explore the use of BA to estimate the number of deaths. Rather than directly estimating deaths from the NHANES data, we reference an external life table—the 2010 period table for the U.S. Social Security population (source: Social Security Administration). This life table serves as an approximate representation of the observed population, with 2010 chosen as a midpoint of the NHANES observation period. The goal is not precise death estimation but rather an assessment of what happens when BA replaces CA for determining mortality probabilities. It is important to note that the life table may not perfectly align with NHANES mortality data due to NHANES being a non-random sample of the U.S. population (and the absence of sampling weight adjustments) and the broader time span of NHANES observations.

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 as follows:

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

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

$$ {}_nq_x = 1 - \prod_{k=x}^{x+n-1} 1 - q_k $$

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.

4.1 Death counts

Figure 9 compares predicted and observed deaths as a percentage of observed deaths (where 0% indicates perfect prediction). We evaluate predictions over three time horizons: deaths within 1, 5, and 10 years. Notably, there are relatively few deaths within one year of examination (around 400). Furthermore, because BA models are trained without restrictions on the delay between examination and death, longer horizons may better reflect long-term survival and align more closely with BA methods. For 5- and 10-year follow-ups, we exclude later NHANES cycles from the dataset to ensure complete survival data. Specifically, for 5-year estimates, we exclude cycles 2013-2014 and 2015-2016, while for 10-year estimates, we further exclude 2009-2010, 2011-2012, and 2013-2014.

The results initally suggest that various BAs provide little advantage over CA in estimating the number of deaths. In fact, deviations from CA appear counterproductive, as methods and marker combinations most correlated with CA tend to yield the best results. For longer follow-ups, the predictions shift downward, underestimating deaths. This shift benefits random forest (RF) and PhenoAge, which tended to overestimate death counts for the 1-year follow-up. However, for 10-year follow-ups, all methods significantly underestimate the number of deaths.

Despite this, BAs have shown greater predictive power for mortality. We therefore interpret this discrepancy between abstract concordance scores and biases in predicted death numbers as a calibration issue. Here, calibration refers to the consistency between predicted probabilities and observed outcomes, akin to the calibration challenges in machine learning models. While BAs can discriminate relative mortality risk, they fail to assign reasonable absolute mortality risks. This shortcoming is unsurprising, as most BA methods do not explicitly align the age scale with specific mortality levels.

In this section we explore various approaches to realign BAs with the life table used. Of the methods considred, exploiting PhenoAge's ability to take into account a reference survival distribution results in best estimates, with similar adjustments possible for other methods.

4.2 Linear Adjustment

To explore whether scale misalignment contributes to biases in predicted death numbers, we re-regress the obtained BAs on CA to create the closest linear approximation of CA. This adjustment does not affect MLR CA, as it is already the closest linear

approximation. However, as shown in Figure 10, the results indicate that this adjustment worsens overall performance. Random Forest (RF) and PhenoAge now severely underestimate the number of deaths. The primary beneficiary of this transformation is KDM, which, due to its high correlation with CA, becomes almost indistinguishable from it after the linear adjustment.

These findings suggest that the issue is not merely one of BA scale. For longer follow-up periods (5 and 10 years), the linear adjustment provides marginally better results. However, since CA itself significantly underestimates the number of deaths over these longer horizons, a simple linear adjustment is unlikely to resolve the underlying issue.

4.3 Adjusting the Dispersion of BA

The linear adjustment of BA did not fundamentally improve its compatibility with mortality tables. We therefore consider another approach: adjusting the dispersion of BA around the average BA for each age group. Specifically, we rescale BAs within each age group by introducing a scaling factor, α. Let x represent a given age, and α be the scaling factor. The adjusted BA is defined as:

$$BA_i = \alpha \cdot BA_i + (1 - \alpha) \cdot \frac{1}{|\{i : x_i = x\}|} \sum_{i:x_i=x} BA_i$$

For $\alpha = 1$, the BA remains unchanged. For $\alpha < 1$, the within-age dispersion decreases, while for $\alpha > 1$, the within-age dispersion increases.

Figure 11 illustrates how the overall results change as a function of the scaling factor. There is a positive relationship between the scaling factor and the number of predicted deaths. This can be explained by the fact that the predicted number of deaths increases primarily for individuals with very large ages. As the scaling factor increases, more individuals are pushed into this high-age bracket, thereby increasing the predicted deaths.

By the intermediate value theorem, there exists a scaling factor that results in a perfect fit to observed deaths. However, this should not be used as a basis for estimation—much like scaling regression predictions to achieve desired results is not a valid statistical practice. With these caveats in mind, the plot suggests that both RF and PhenoAge are overdispersed and could achieve better predictions with reduced dispersion.

In contrast, KDM is largely indistinguishable from CA, rendering the scaling factor almost irrelevant. On the other hand, MLR is far removed from CA and benefits from increased dispersion, which aligns with its tendency to vastly underestimate the number of deaths.

4.4 Calibrated PhenoAge

Among the considered BAs, PhenoAge has the notable property of defining BA as the best age that aligns with a given reference survival distribution. When the reference is

the life table we use, this property should result in better alignment with the mortality estimates provided by the life table.

As before, Figure 12 compares the estimated and observed number of deaths over 1-, 5-, and 10-year follow-up periods. Although the default PhenoAge uses 10-year survival probabilities, we adjust this duration to match the follow-up period for each analysis.

Calibrated PhenoAge provides a more accurate estimate of the number of deaths than CA. The estimated number of deaths is closer to the observed value, and the observed death count almost always falls within the confidence intervals of the estimate.

In retrospect, this result is not entirely surprising. The mortality model underlying PhenoAge—a Gompertz proportional hazards model—likely provides a reasonable estimate of survival probabilities. However, the chosen mortality table does not perfectly align with the NHANES population for two reasons: first, NHANES is a small, non-random sample of the general population; second, the NHANES data covers a broader time period than the table. In this context, PhenoAge effectively serves as an adjustment, bridging the gap between observed mortality in the NHANES sample and the mortality predicted by the life table. Nonetheless, the mortality table appears to ground PhenoAge's estimates, as the uncalibrated PhenoAge did not achieve comparable accuracy.

It is also worth noting that predictions from Random Forest (RF) or any other model capable of incorporating an arbitrary number of covariates can be adapted using this calibration approach.

5 Discussion

In this work, we present and compare four biological age (BA) methods, exploring their use as tools for prevention. While BAs are already used in practice to communicate an individual's overall health, this study introduces them to an actuarial audience, building upon the comprehensive overview of BA methods provided in Bafei and Shen [14]. We present four well-known BA methods and provide some analysis on their construction. These BA are then assess based on their performance on a number of criteria and under diverse conditions. This comparison is done on the NHANES data, and includes a wide range of markers for BA construction and criteria for evaluation, ensuring robust and comprehensive analyses. Moreover, we employ multiple imputation to address missing data, preserving flexibility in the choice of variables. We procede to apply these BAs to the task of estmating the number of deaths using a standard life table. Finally, we offer guidance on choosing BA methods for self-protection and prevention-focused applications.

5.1 Summary of BA comparison

All BAs are strongly correlated with CA, with KDM achieving the highest correlation, while MLR had low correlations for smaller marker sets. Mortality prediction results were unsurprising, with RF and PhenoAge consistently outperforming CA, while KDM

showed an improvement over CA that was not large enough to be statistically significant. MLR failed to show meaningful improvements. In terms of associations with disease, BAs did not achieve much improvement over CA, with some exceptions for various marker-disease combinations. For self-rated health, mortality-based methods significantly outperformed CA, while MLR showed significant improvement for the two mortality-based marker sets. Although KDM improved upon CA, it was not significant. Similarly, for behavioral variables, BAs did not significantly improve upon CA. Sociological factors showed associations with various BA accelerations, but with little difference between BAs, except for MLR, which performed poorly.

RF generally emerged as the best performer or was tied for the best across most of the criteria considered, with PhenoAge following closely. KDM generally performed worse than RF but similarly to MLR, with a few notable exceptions: KDM showed strong associations with some diseases, behaviors, and education. MLR, unfortunately, consistently ranked last, especially for smaller marker sets, where its weak correlation with CA caused it to perform worse than CA.

5.2 Discussion of BA methods

MLR's underwhelming performance highlights that CA itself remains informative even in the presence of many markers. KDM's more consistent performance further supports the importance of CA. However, using CA as a criterion for defining BA comes with significant challenges. Either one must completely renounce the use of CA, as MLR does, or use CA indirectly, as KDM does. Mortality, as an alternative criterion examined in this study, appears to offer a more practical target. Mortality is not only valuable in its own right, but its unobserved nature prevents it from being used as a marker, thereby avoiding the biomarker paradox. Additionally, mortality can easily be substituted with another outcome.

The strong performance of PhenoAge and RF supports the case for using mortality as a criterion. This is further supported by the fact that the two smallest marker sets, both based on mortality, performed and better than their larger, CA-based counterparts. The main drawback of mortality-based BA is the longitudinal mortality data required for it.

Although both PhenoAge and RF are based on mortality, they exhibit different behaviors. PhenoAge generally performed slightly worse than RF, which can be explained by the more flexible model underlying RF. PhenoAge also exhibited less consistency, with marked drops in performance for certain marker sets, as for example for disease associations. PhenoAge also exhibited larger confidence intervals, most noticeably in comparisons with sociological variables. We interpret this as a sign of overfitting in PhenoAge, where variations in the markers used for each imputation replication result in variance in the final prediction. If overfitting is indeed the cause of these issues, it could be addressed through penalization or other methods designed to prevent overfitting. By contrast, RF is designed to be robust to overfitting. Furthermore, RF is conceptually simple: it calculates mortality scores and then transforms them into an age-like scale. In contrast, PhenoAge requires multiple models and applies an intuitive, yet ultimately arbitrary, probability-matching scheme to generate an age.

More broadly, three out of the four methods considered are, in fact, linear combinations of CA and markers, though neither KDM nor PhenoAge are explicitly formulated as such. For PhenoAge especially, an explicit reformulation using a linear combination could greatly simplify the construction, potentially through the use of a Cox proportional hazards model. We believe that such variants may prove useful, as the interpretability provided by parametric or semi-parametric models like Cox is essential for understanding the reasons behind individual BA estimations. While methods such as the Shapley-value-based approach in ExplainableAge (upon which RF is based) can provide post-hoc explanations, these require additional computation, further increasing RF's already considerable computational demand. Given RF's modest improvement upon PhenoAge, we suspect PhenoAge's performance can be improved by including a few interactions and non-linearities, retaining its interpretability.

5.3 Adapting Biological Ages for Prevention

We see the exercise of calculating BA, rather than simply using the underlying markers directly, as primarily a communication tool. The ability to quantify health in terms of years, rather than abstract hazard ratios, is a powerful means of conveying information.

The second leg of BAs is their risk-predicting ability. However, as we have seen, the theorical performance of BA does not necessarily reflect its ability to estimate death counts from a life table, with death count estinations rsullting from plugging in BA into a life table were worse than CA. Neither simple transformation or scailing helpped the problem. Instead, a supplimental step of calibration is required, where either the BA are adjusted to fit the life table used, or a new life table is constructed, based on a BA.

Thus any practical application of BA must be tailored to the specific needs and objectives of the intended use. A BA requires three critical components: (1) an outcome they are based, (2) the variables that are used, and (3) a method. These elements must be adapted to the context for which a BA is developed for. In this section we discuss some considerations to be taken into account when adapting a BA for prevention.

5.3.1 Target Outcome

The target outcome defines what the BA is meant to measure. Existing litterature focuses on a general notion of health and uses CA or mortality as the outcome for BA construction For prevention-focused applications, however, the target outcome can be directly substituted with a specific risk of interest, such as the incidence of a chronic disease or functional limitation.

Mortality-based methods, such as PhenoAge and random forest (RF), offer significant flexibility in substituting the target outcome. This flexibility makes them particularly suited to be adapted to other context. Substituting the target outcome is straightforward for these methods, whereas methods like MLR and KDM, which are more closely tied to CA, may become unrecognizable when adapted to different outcomes.

For crosssectional, as opposed to longitudinal data, RF is the more adaptable method. Random forest model underlying RF BA can be substituted by any other risk measure and then transformed to an age scale.

5.3.2 Variable Selection

The choice of variables for BA construction depends on the desired application. In this study, we used generalist variables representative of those typically collected during an extensive health checkup. However, for prevention-focused BAs, special attention must be paid to the relationship between variables and potential self-protective measures. This relationship can be direct, as for example, variables such as smoking status, which are directly modifiable through behavior changes; or indirect, such as biological markers like blood glucose levels, which can be linked to interventions (e.g., dietary improvements) through external knowledge.

Practical implementation may face challenges due to limited data availability, especially for insurers who typically rely on datasets collected during policy pricing. NHANES data, used in this study, offers significant flexibility in variable selection and a respectable sample size. It can serve as a useful resource to guide initial variable choices for prevention programs.

5.3.3 Choice of Method

The choice of BA method depends on the specific context and priorities of the application. The choice of method is somewhat constrained by the outcome of interest. With mortality-based methods being the obvious choice if mortality is the focus. Mortality-based methods are also more flexible and can adapt to various outcomes. MLR and KDM may still be of interest if the focus is on general health instead.

PhenoAge and RF, generally provided the most consistent and performant results, making them strong candidates for prevention-focused applications. These two methods are also conceptually the most flexible, and it should be straightforward to adapt these to any downstream application, as we did for death counts.

Interpretability is also an important consideration. For example, PhenoAge may be preferred over RF due to its ease of interpretation, with each variable assigned a linear coefficient. This transparency is particularly valuable when communicating BA results to policyholders. There is a trade-off between a precision of a BA in predicting a risk, and the ease of communication, as the impact of various covariables may inherently be complex with non-linearities and interactions. This trade-off must be carefully considered : simpler models, though less performant, may be more effective in encouraging self-protective behaviors.

5.4 Concluding remarks

Our findings demonstrate that it is possible to construct biological ages (BAs) that combine robust risk prediction with ease of interpretation. This positions BA as a cornerstone metric for prevention programs, enabling more personalized risk assessments and incentivizing healthier behaviors among policyholders. In this work, we have presented various BA methods and addressed practical issues that may arise when

applying these methods. By combining methodological insights with practical applications, this study provides a foundation for integrating BA into prevention-focused initiatives and insurance contexts.

Despite their potential, examples of BAs being integrated into prevention programs remain scarce, with limited literature exploring their application in this context. Further research could provide valuable insights into the effectiveness of widely adopted prevention strategies, such as regular exercise, balanced diets, and routine health checkups. Such work is essential to determine whether existing prevention programs fully incentivize best practices to optimize population health.

NHANES provides a robust foundation for advancing these efforts, offering data on insurance status, healthcare access, accelerometer readings (akin to fitness application data), and detailed dietary information. While this research may not uncover entirely new pathways to better health, it is critical for quantifying both risks and the potential reductions achieved through interventions—essential steps in designing rational prevention programs. By basing these analyses on biological ages, the impact of risk reduction can be communicated more effectively to the target population, further enhancing the success of prevention initiatives.

6 Supplementary material

6.1 Cause-specific mortality

6.2 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)

References

[1] Ehrlich, I., Becker, G.S.: Market insurance, self-insurance, and self-protection **80**(4), 623–648. Publisher: The University of Chicago Press. Accessed 2024-11-25

[2] Klemera, P., Doubal, S.: A new approach to the concept and computation of biological age **127**(3), 240–248 https://doi.org/10.1016/j.mad.2005.10.004 . Accessed 2024-08-01

[3] Wilson, D.L.: Aging hypothesis, aging markers and the concept of biological age **23**(4), 435–438 https://doi.org/10.1016/0531-5565(88)90049-6 . Accessed 2024-10-04

[4] Ingram, D.K.: Key questions in developing biomarkers of aging **23**(4), 429–434 https://doi.org/10.1016/0531-5565(88)90048-4 . Accessed 2024-10-07

[5] Fahy, G.M., Brooke, R.T., Watson, J.P., Good, Z., Vasanawala, S.S., Maecker, H., Leipold, M.D., Lin, D.T.S., Kobor, M.S., Horvath, S.: Reversal of epigenetic aging and immunosenescent trends in humans **18**(6), 13028 https://doi.org/10.1111/acel.13028 . Accessed 2024-10-17

[6] Hollingsworth, J.W., Hashizume, A., Jablon, S.: Correlations between tests of aging in hiroshima subjects–an attempt to define "physiologic age" **38**(1), 11–26

[7] 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 **60**(2), 153–170 https://doi.org/10.1016/0047-6374(91)90128-M . Accessed 2024-10-04

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

[9] Levine, M.E.: Modeling the rate of senescence: Can estimated biological age predict mortality more accurately than chronological age? **68**(6), 667–674 https://doi.org/10.1093/gerona/gls233 . Accessed 2024-08-01

[10] 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 **10**(4), 573–591 https://doi.org/10.18632/aging.101414 . Accessed 2024-08-01

[11] 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 **76**(5), 741–749 https://doi.org/10.1093/gerona/glaa286

[12] 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 **11**, 1074274 https://doi.org/10.3389/fpubh.2023.1074274

[13] 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 **78**(10), 1753–1762 https://doi.org/10.1093/gerona/glad137 . Accessed 2024-09-24

[14] Bafei, S.E.C., Shen, C.: Biomarkers selection and mathematical modeling in biological age estimation **9**(1), 13 https://doi.org/10.1038/s41514-023-00110-8 . Accessed 2024-08-02

[15] Hastings, W.J., Shalev, I., Belsky, D.W.: Comparability of biological aging measures in the national health and nutrition examination study, 1999-2002 **106**, 171–178 https://doi.org/10.1016/j.psyneuen.2019.03.012

[16] 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) **131**(2), 69–78 https://doi.org/10.1016/j.mad.2009.12.001 . Accessed 2024-10-07

[17] Kang, Y.G., Suh, E., Lee, J.-w., Kim, D.W., Cho, K.H., Bae, C.-Y.: Biological age as a health index for mortality and major age-related disease incidence in koreans: National health insurance service – health screening 11-year follow-up study **Volume 13**, 429–436 https://doi.org/10.2147/CIA.S157014 . Accessed 2024-11-25

[18] Petit Sarazin, M.: Elaboration D'un Score de Vieillissement : Propositions Théoriques. https://theses.hal.science/tel-00994941/ Accessed 2024-11-25

[19] 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 **19**(2), 133–143 https://doi.org/10.1016/0531-5565(84)90016-0

[20] Qiu, W., Chen, H., Kaeberlein, M., Lee, S.-I.: ExplaiNAble BioLogical age (ENABL age): an artificial intelligence framework for interpretable biological age **4**(12), 711–723 https://doi.org/10.1016/S2666-7568(23)00189-7 . Accessed 2024-09-30

[21] Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests **2**(3) https://doi.org/10.1214/08-AOAS169 . Accessed 2024-10-07

[22] Brand, J.P.L.: Development, Implementation and Evaluation of Multiple Imputation Strategies for the Statistical Analysis of Incomplete Data Sets

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

[24] NCHS Data Linkage - Mortality Data - Public-Use Files. https://www.cdc.gov/nchs/data-linkage/mortality-public.htm Accessed 2024-08-30

[25] Schafer, J.L.: Multiple Imputation Models and Procedures for NHANES III. https://wwwn.cdc.gov/Nchs/Data/Nhanes3/7a/doc/mimodels.pdf Accessed 2024-08-30

[26] Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys, 1st edn. Wiley Series in Probability and Statistics. Wiley. https://doi.org/10.1002/9780470316696 . https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316696 Accessed 2024-08-30

[27] 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 **35**(3), 443–456 https://doi.org/10.3233/SJI-180470 . Accessed 2024-08-30

[28] Buuren, S.V., Groothuis-Oudshoorn, K.: mice : Multivariate imputation by chained equations in r **45**(3) https://doi.org/10.18637/jss.v045.i03 . Accessed 2024-08-30

[29] Van Buuren, S.: Flexible Imputation of Missing Data, Second Edition, 2nd edn. Chapman and Hall/CRC. https://doi.org/10.1201/9780429492259 . https://www.taylorfrancis.com/books/9780429492259 Accessed 2024-11-15

[30] 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 **35**(28), 5170–5188 https://doi.org/10.1002/sim.7049 . Accessed 2024-08-30

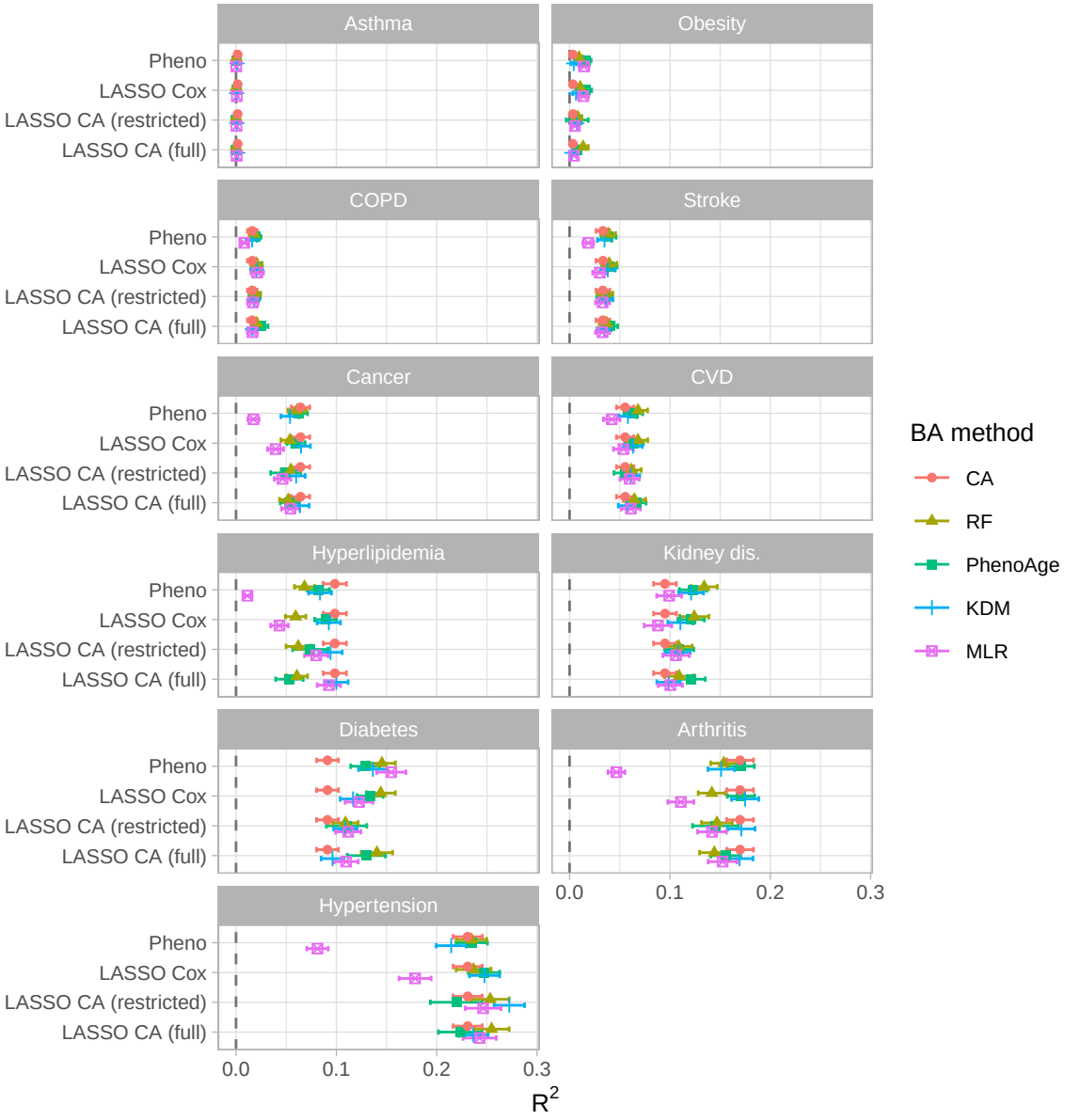

Fig. 5 R^2 vs the presence of disease.

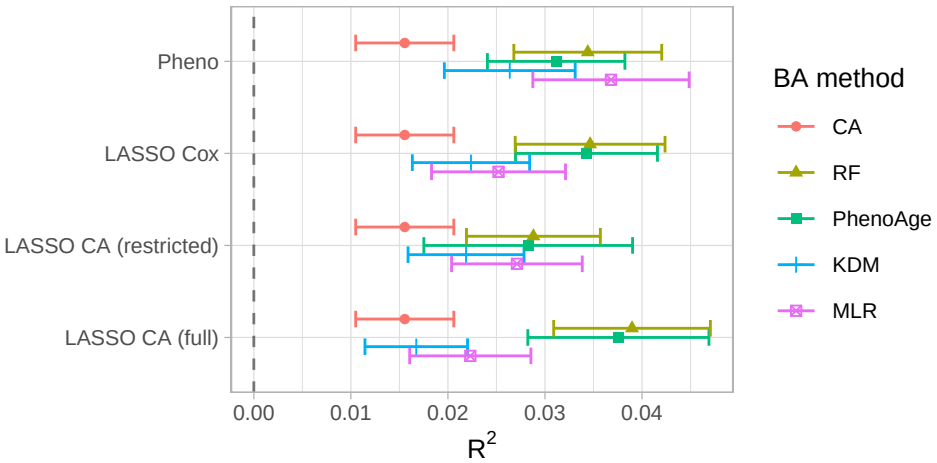

Fig. 6 R^2 vs self-rated health.

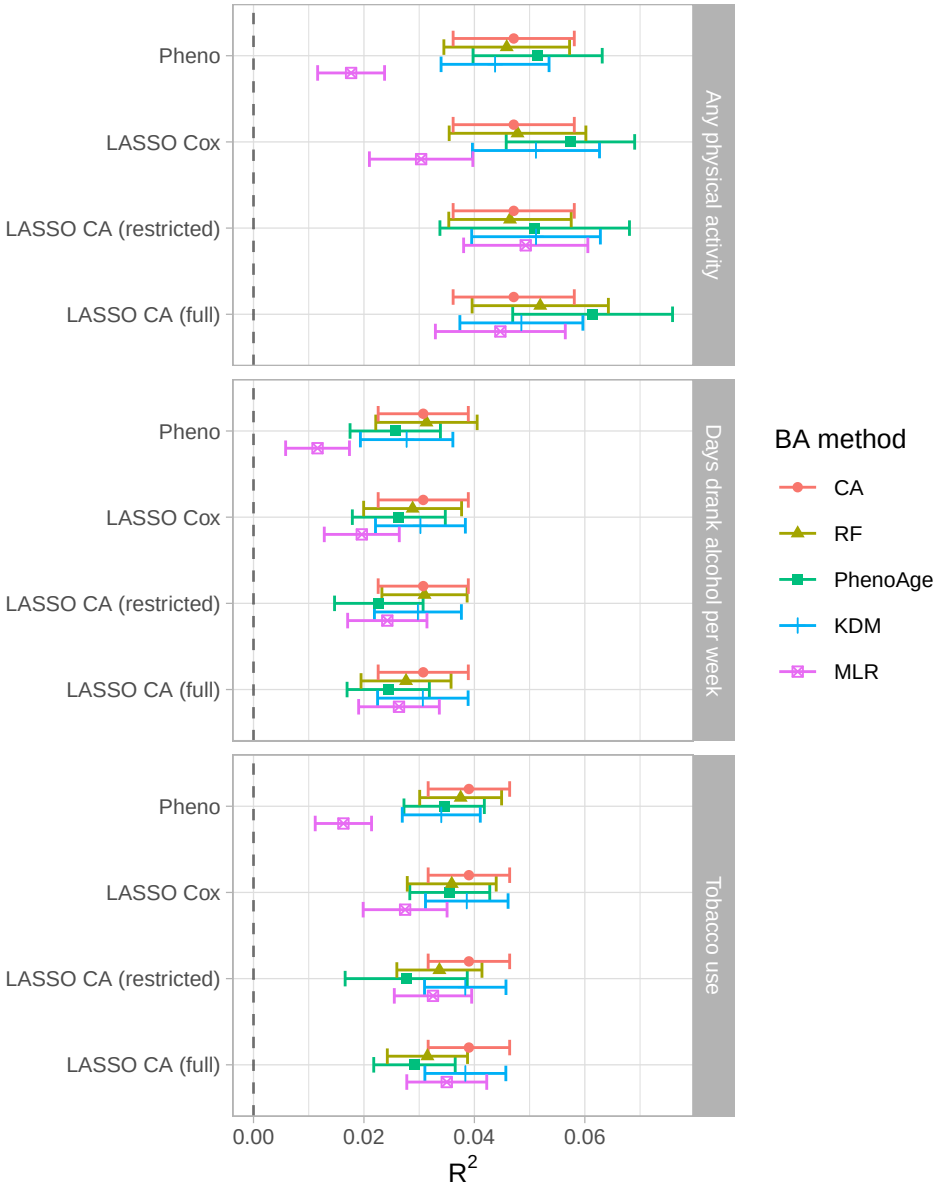

Fig. 7 R^2 vs behavioral variables.

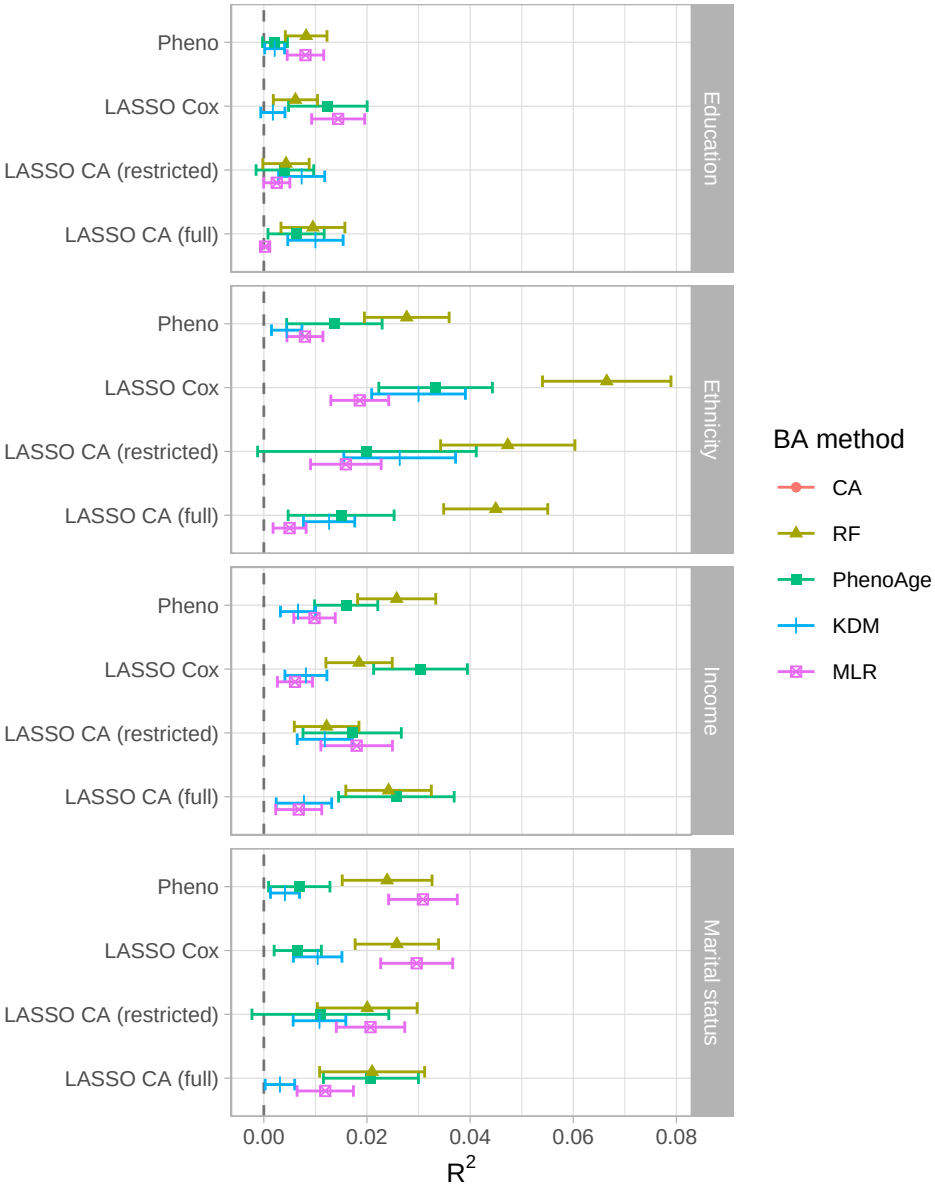

Fig. 8 R^2 vs sociological variables.

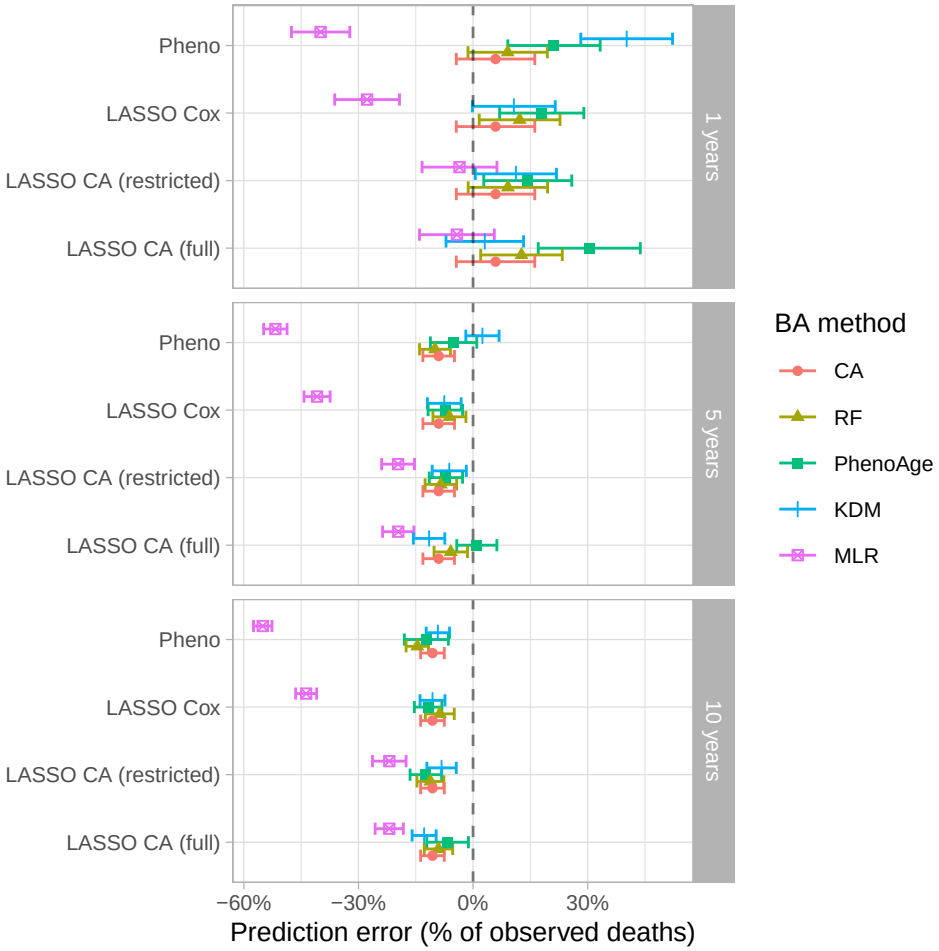

Fig. 9 Percentage error in predicted deaths compared to observed deaths within 1, 5, and 10 years of follow-up for considered BAs.

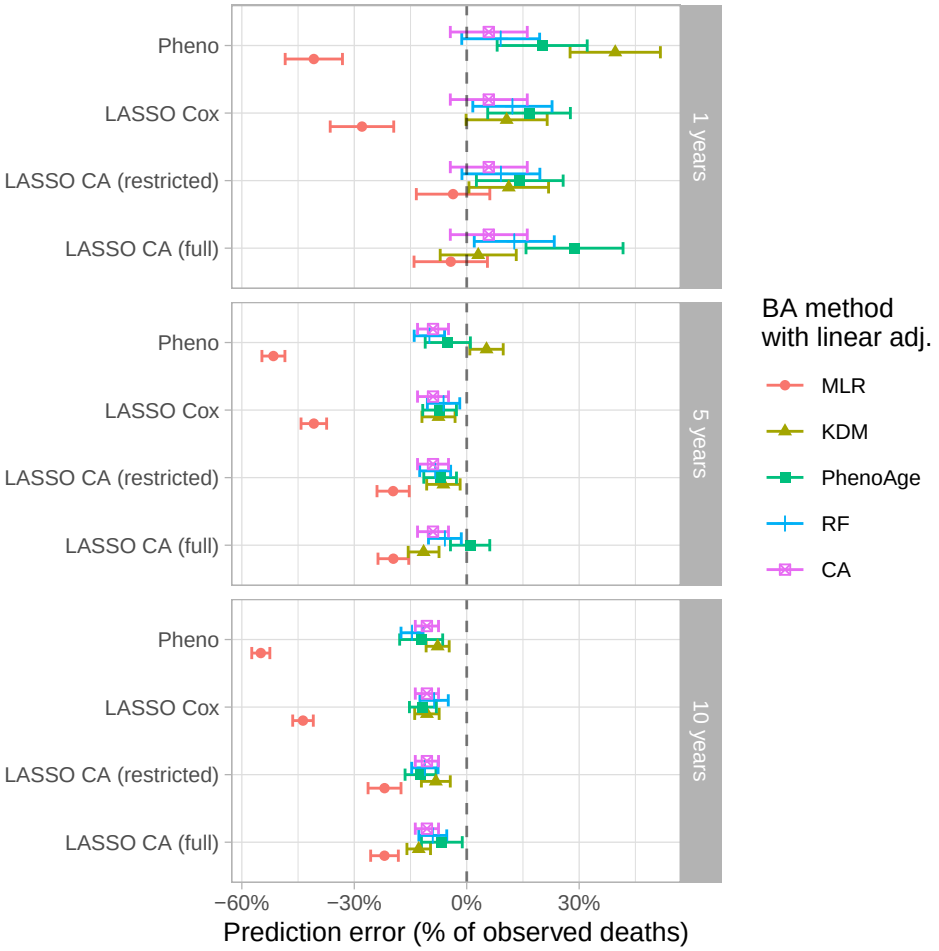

Fig. 10 Percentage error in predicted deaths compared to observed deaths within 1, 5, and 10 years of follow-up for linearly adjusted BAs.

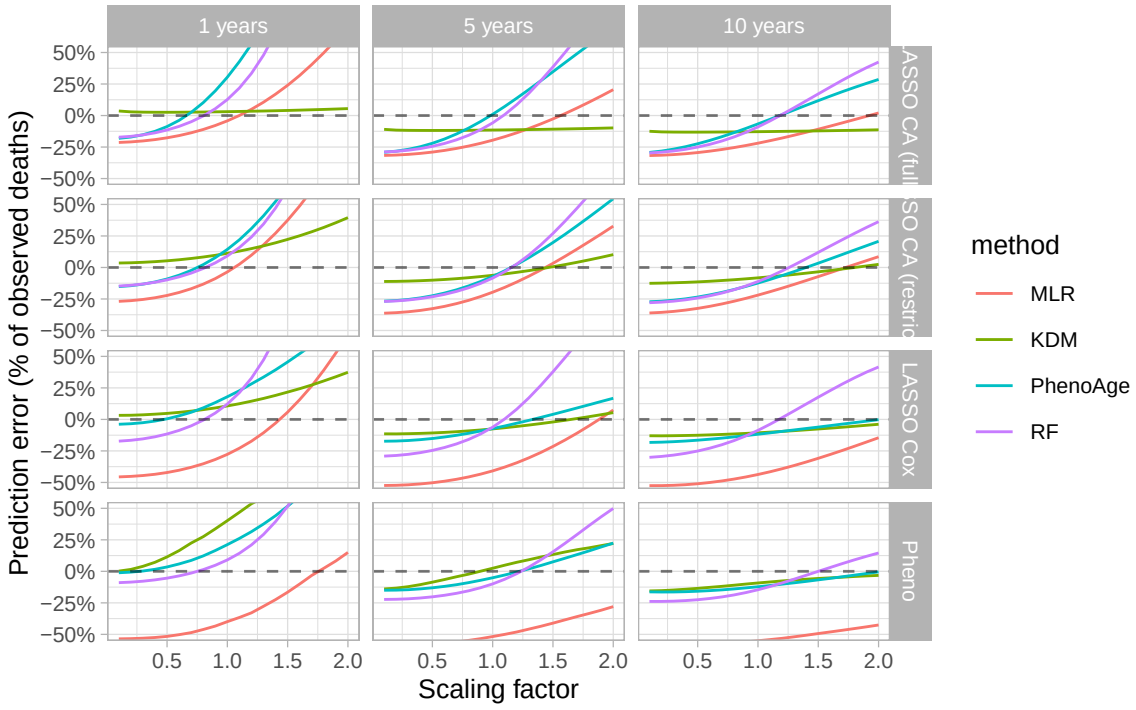

Fig. 11 Percentage error in predicted deaths within one year of examination as a function of the within-age scaling factor.

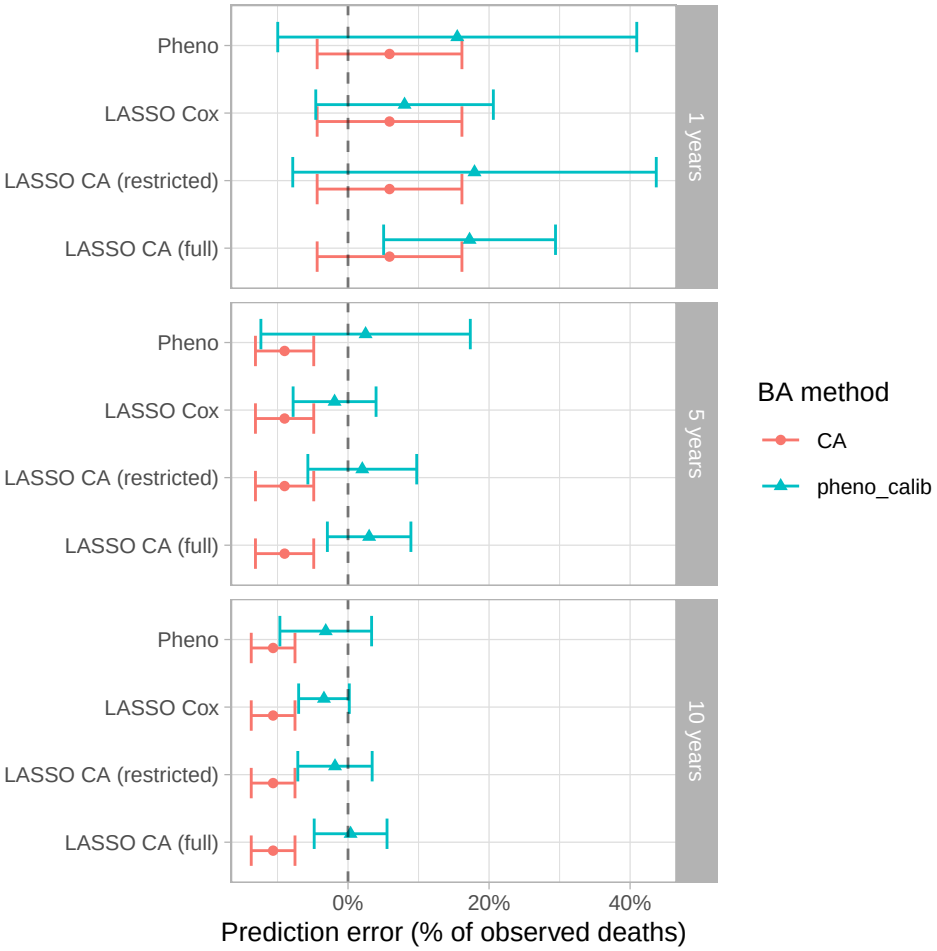

Fig. 12 Percentage error in predicted deaths compared to observed deaths within 1, 5, and 10 years of follow-up for calibrated PhenoAge and CA.

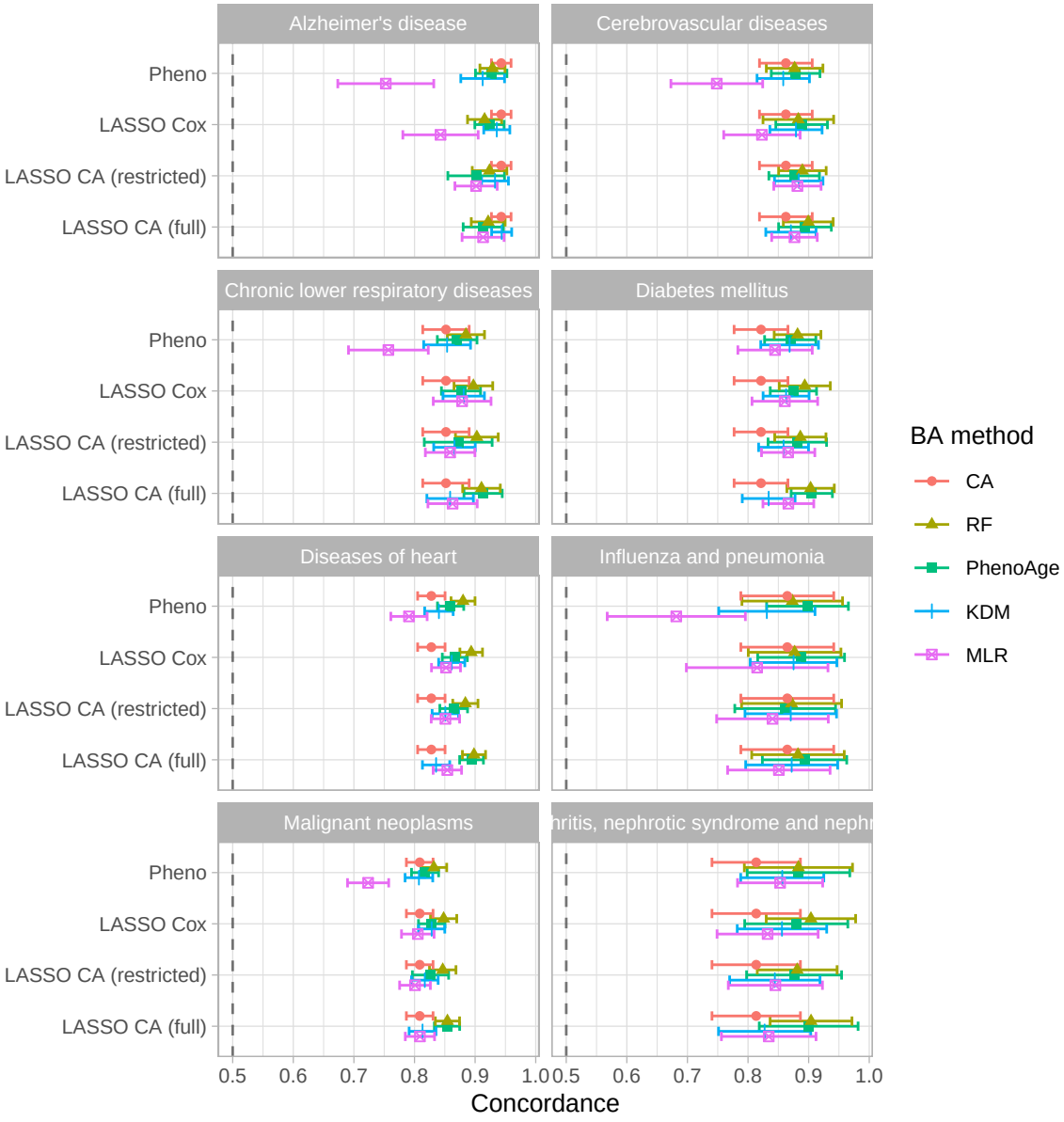

Fig. 13 Concordance with respect to age of death (cause-specific).