Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Biological aging measures for all-cause mortality prediction in two population-based cohorts

Shayan Mostafaei

shayan.mostafaei@ki.se

Karolinska Institutet

Jakob Lindh

Karolinska Institutet

Chenxi Qin

Karolinska Institutet

Jonathan K. L. Mak

Karolinska Institutet

Sara Hägg

Karolinska Institutet

Article

Keywords: All-cause mortality, biological aging, risk prediction, cohort study

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11252530/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Biological aging measures may improve mortality prediction beyond chronological age and conventional risk factors. We evaluated PhenoAge, the Klemera–Doubal method (KDM), homeostatic dysregulation (HD), frailty index (FI), and leukocyte telomere length (TL) in TwinGene (n = 9,617) and UK Biobank (UKB; n = 314,064). Separate multivariable Cox models evaluated each biological aging measure, adjusting for age, sex, body mass index, smoking, and education, with twin-pair-clustered robust variance estimation in TwinGene. Per 1-SD increase, mortality HRs for age-adjusted PhenoAge and KDM residuals, log HD, and FI ranged from 1.11–1.27 in TwinGene and 1.24–1.40 in UKB; TL HRs were 0.96 (95% CI: 0.93–0.99) and 0.94 (0.92–0.95), respectively. In 10-fold cross-validation, Cox prediction models adding PhenoAge residual, FI, and TL to the conventional model increased Harrell's C-index from 0.780 to 0.787 in TwinGene and from 0.720 to 0.746 in UKB. Incremental prognostic gains were larger in UKB.

Introduction

Chronological age is one of the strongest predictors of morbidity and mortality, yet people of the same age can differ markedly in physiological function, disease burden, and survival. Biological aging measures have been developed to capture this inter-individual heterogeneity by quantifying age-related changes in physiological or molecular systems that are not fully represented by chronological age alone [1–3]. Among the most widely used approaches are clinical biomarker-based algorithms such as PhenoAge, the Klemera–Doubal method (KDM), and homeostatic dysregulation (HD), which integrate information across multiple physiological domains [2–6]. Other aging-related measures capture complementary dimensions of the aging process, including the frailty index (FI), which quantifies accumulated health deficits, and leukocyte telomere length (TL), a marker of cellular replicative history [7–9]. Previous studies have linked accelerated biological aging to morbidity, functional decline, and mortality, supporting the potential value of these measures for identifying differences in health risk among individuals of similar chronological age [3,10,11].

For biological aging measures to contribute meaningfully to mortality-risk assessment, however, association with mortality alone is insufficient. A useful measure should provide prognostic information beyond that already captured by chronological age and established risk factors, and its predictive contribution should be evaluated across populations with different characteristics and measurement settings. This is particularly relevant for clinical biomarker-based aging algorithms because their implementation can depend on the biomarkers available in a given cohort. PhenoAge, KDM, and HD can be parameterized using different biomarker panels, and previous work has shown that biological aging measures derived from different physiological or molecular domains should not necessarily be regarded as interchangeable [6,12]. Differences in population age structure, follow-up, biomarker availability, and laboratory measurement may therefore influence observed associations and predictive performance. Demonstrating consistent incremental prognostic information across heterogeneous cohort settings

would provide stronger evidence that biological aging measures capture mortality-relevant variation rather than cohort-specific measurement characteristics.

In this study, we investigated whether commonly used biological aging measures provide prognostic information for all-cause mortality beyond chronological age and conventional risk factors in TwinGene and UK Biobank (UKB). We evaluated three clinical biomarker-based measures—PhenoAge, KDM, and HD—together with FI and TL as complementary indicators of biological aging. We applied a harmonized analytical framework independently in both cohorts to assess the consistency of findings across distinct populations and biomarker implementations. We examined each measure's association with all-cause mortality in a separate covariate-adjusted Cox model and evaluated the incremental contribution of PhenoAge residual, FI, and TL jointly to mortality prediction using survival-based measures of discrimination, prediction error, and calibration. We further evaluated whether associations varied by key demographic and lifestyle factors and assessed robustness in sensitivity analyses. This framework allowed us to assess whether mortality-relevant biological aging signals were consistent across cohort-specific implementations while quantifying their incremental prognostic value beyond conventional risk factors.

Results

Cohort characteristics and follow-up

The harmonized analytic samples included 9,617 participants in TwinGene and 314,064 in UKB (Fig. 1). TwinGene participants were older at baseline (64.77 ± 8.01 years) than UKB participants (56.39 ± 8.09 years), while the proportion of men was similar (47.3% vs 45.9%). Median follow-up was 16.76 years (IQR: 15.27–17.30) in TwinGene and 13.79 years (IQR: 13.10–14.45) in UKB. Mean body mass index (BMI) was 24.96 ± 3.28 kg/m² and 27.34 ± 4.70 kg/m², respectively, and high/intermediate education was more frequent in UKB (83.5%) than in TwinGene (56.6%). Ever smoking was recorded for 56.9% of TwinGene and 44.9% of UKB participants.

The raw biomarker-based aging measures also showed distinct cohort distributions. Mean PhenoAge was 56.30 ± 9.07 years in TwinGene and 47.65 ± 10.03 years in UKB, while mean KDM was 56.29 ± 8.15 and 54.12 ± 9.43 years, respectively. Mean HD was 2.72 ± 1.00 in TwinGene and 2.25 ± 1.00 in UKB, whereas mean FI values were similar (0.12 ± 0.08 and 0.12 ± 0.07, respectively). During follow-up, 3,090 deaths (32.1%) occurred in TwinGene and 25,084 deaths (8.0%) in UKB. TwinGene comprised 6,466 unique twin-pair identifiers, including 3,151 complete pairs corresponding to 6,302 participants (65.5% of the analytic sample); this within-pair structure was accounted for in subsequent survival analyses and cross-validation (Table 1).

Table 1
Baseline characteristics and biological aging measures in the harmonized TwinGene and UK Biobank analytic cohorts

Characteristics	TwinGene	UK Biobank
Participants, n	9,617	314,064
Deaths, n (%)	3,090 (32.1%)	25,084 (8.0%)
Follow-up, years, median [IQR]	16.76 [15.27, 17.30]	13.79 [13.10, 14.45]
Age at baseline, years	64.77 ± 8.01	56.39 ± 8.09
Male sex, n (%)	4,548 (47.3%)	144,249 (45.9%)
Body mass index, kg/m^2	24.96 ± 3.28	27.34 ± 4.70
Ever smoker, n (%)	5,472 (56.9%)	140,903 (44.9%)
High/intermediate education, n (%)	5,443 (56.6%)	262,279 (83.5%)
PhenoAge, years	56.30 ± 9.07	47.65 ± 10.03
Klemera-Doubal Method, years	56.29 ± 8.15	54.12 ± 9.43
Homeostatic dysregulation	2.72 ± 1.00	2.25 ± 1.00
Frailty index	0.12 ± 0.08	0.12 ± 0.07
Telomere length	0.04 ± 0.34	0.83 ± 0.13

Values are mean ± SD, median [IQR], or n (%), as appropriate. Smoking was harmonized as ever versus never smoking. High/intermediate education was defined as ≥ 10 years of education in TwinGene and as the combined intermediate and high qualification categories in UKB. PhenoAge, KDM, and HD are shown on their original cohort-specific scales for descriptive purposes. Frailty index is reported as a deficit-accumulation proportion, and telomere length as the cohort-specific adjusted leukocyte telomere length measure. For mortality analyses, PhenoAge and KDM were residualized on chronological age using natural splines, homeostatic dysregulation was log-transformed, and all aging measures were standardized within cohort. TwinGene pair dependence was accounted for in subsequent inferential analyses.

Correlations among biological aging measures

Correlations among the aging measures ranged from weak to strong in both cohorts (Fig. 2A, B). Chronological age was strongly correlated with raw PhenoAge and raw KDM in TwinGene (r = 0.95 and 0.86, respectively) and UKB (r = 0.85 for both).

Raw PhenoAge and raw KDM were also strongly correlated with each other (r = 0.86 in TwinGene and 0.90 in UKB). Correlations of raw HD with raw PhenoAge and raw KDM were weak in TwinGene (r = 0.19 for both) and stronger in UKB (r = 0.30 and 0.46, respectively). FI showed generally weak correlations with the biomarker-based aging measures, whereas cohort-specific adjusted TL was weakly inversely

correlated with chronological age, raw PhenoAge, and raw KDM in both cohorts ($r \approx -0.15$ to -0.13 in TwinGene and -0.19 to -0.17 in UKB).

Biological aging measures and all-cause mortality

Each of the five biological aging measures was associated with all-cause mortality in a separate Cox model adjusted for chronological age, sex, BMI, smoking, and education (Table 2; Fig. 3). The biological aging measures were not mutually adjusted. HRs and 95% CIs are reported per 1-SD higher measure, using age-adjusted residuals for PhenoAge and KDM and log-transformed HD. TwinGene models additionally used twin-pair-clustered robust variance estimation.

Table 2
Adjusted associations of biological aging measures with all-cause mortality in TwinGene and UK Biobank

Biological aging measure	TwinGene HR (95% CI)	*P*	FDR-adjusted *P*	UKB HR (95% CI)	*P*	FDR-adjusted *P*
PhenoAge residual	1.21 (1.18–1.25)	1.56×10^{-31}	3.89×10^{-31}	1.40 (1.38–1.42)	$< 1\times10^{-300}$	$< 1\times10^{-300}$
KDM residual	1.11 (1.07–1.15)	9.30×10^{-10}	1.16×10^{-9}	1.37 (1.34–1.39)	$< 1\times10^{-300}$	$< 1\times10^{-300}$
Homeostatic dysregulation	1.27 (1.22–1.32)	1.18×10^{-35}	5.90×10^{-35}	1.24 (1.22–1.26)	1.24×10^{-131}	1.55×10^{-131}
Frailty index	1.15 (1.11–1.19)	4.46×10^{-15}	7.43×10^{-15}	1.26 (1.24–1.28)	1.19×10^{-184}	1.99×10^{-184}
Telomere length	0.96 (0.93–0.99)	0.025	0.025	0.94 (0.92–0.95)	5.02×10^{-13}	5.02×10^{-13}

Hazard ratios (HRs) are reported per 1-SD higher biological aging measure. Models were adjusted for chronological age and body mass index using natural splines, together with sex, smoking, and education. PhenoAge and KDM were residualized on chronological age using natural splines with 3 degrees of freedom; homeostatic dysregulation was log-transformed before standardization. Frailty index and telomere length were standardized within cohort. TwinGene analyses included 9,617 participants and 3,090 deaths and used pair-clustered robust variance estimation; UK Biobank analyses included 314,064 participants and 25,084 deaths. *P*-values were adjusted separately within each cohort across the five prespecified biological aging measures using the Benjamini–Hochberg false-discovery-rate procedure. KDM: Klemera-Doubal Method

In TwinGene, higher PhenoAge residual was associated with increased mortality risk (HR: 1.21, 95% CI: 1.18–1.25), as were KDM residual (HR: 1.11, 95% CI: 1.07–1.15), HD (HR: 1.27, 95% CI: 1.22–1.32), and

FI (HR: 1.15, 95% CI: 1.11–1.19). Longer TL was associated with lower mortality risk (HR: 0.96, 95% CI: 0.93–0.99).

Associations were directionally consistent in UKB. Higher PhenoAge residual (HR: 1.40, 95% CI: 1.38–1.42), KDM residual (HR: 1.37, 95% CI: 1.34–1.39), HD (HR: 1.24, 95% CI: 1.22–1.26), and FI (HR: 1.26, 95% CI: 1.24–1.28) were associated with higher mortality risk, whereas longer TL was inversely associated with mortality (HR: 0.94, 95% CI: 0.92–0.95). All five associations remained significant after FDR correction within each cohort.

Incremental mortality prediction beyond chronological age and conventional risk factors

The biological aging-enhanced model jointly added PhenoAge residual, FI, and TL to the conventional model containing chronological age, sex, BMI, smoking, and education (Table 3; Fig. 4). One clinical biomarker-derived measure (PhenoAge residual) was combined with measures of deficit accumulation and cellular aging to capture complementary domains. KDM residual and HD were evaluated separately in association analyses and as alternatives to PhenoAge residual in sensitivity prediction models.

Table 3
Cross-validated survival prediction performance in TwinGene and UK Biobank

Cohort	Model	Harrell C-index (95% CI)	5-year AUC	10-year AUC	5-year Brier score	10-year Brier score
TwinGene	Age	0.768 (0.759–0.777)	0.767	0.785	0.0406	0.0917
TwinGene	Conventional	0.780 (0.771–0.788)	0.782	0.797	0.0402	0.0900
TwinGene	Conventional + BA	0.787 (0.779–0.796)	0.793	0.805	0.0394	0.0877
UK Biobank	Age	0.692 (0.688–0.697)	0.667	0.688	0.0149	0.0410
UK Biobank	Conventional	0.720 (0.715–0.724)	0.700	0.720	0.0149	0.0405
UK Biobank	Conventional + BA	0.746 (0.742–0.751)	0.741	0.750	0.0147	0.0395

Prediction performance was estimated from 10-fold cross-validated out-of-fold predictions. The age model included chronological age modeled using a natural spline with 4 degrees of freedom. The conventional model additionally included sex, body mass index (BMI; modeled using a natural spline with 3 degrees of freedom), smoking, and education. The conventional + biological aging (BA) model further included PhenoAge residual, frailty index (FI), and telomere length (TL). PhenoAge residualization and standardization of PhenoAge residual, FI, and TL were performed within each training fold before application to held-out participants. In TwinGene, co-twins were retained within the same cross-validation fold. Harrell C-index summarizes overall discrimination; AUC denotes time-dependent area under the receiver operating characteristic curve; CI, confidence interval. Time-dependent AUC and Brier scores were estimated at 5 and 10 years using censoring-aware methods. Higher C-index and AUC and lower Brier scores indicate better predictive performance. Paired differences between the conventional and conventional + BA models, with 95% CIs, are reported in Supplementary Table S1.

In TwinGene, the Harrell's C-index increased from 0.768 (95% CI: 0.759–0.777) for chronological age alone to 0.780 (0.771–0.788) after addition of sex, BMI, smoking, and education, and to 0.787 (0.779–0.796) after further addition of the biological aging measures (PhenoAge residual, FI, and TL). The corresponding increase from the conventional model to the biological aging-enhanced model was $\Delta C = 0.0073$ (95% CI: 0.0052–0.0094). At 5 years, time-dependent AUC increased from 0.782 to 0.793 after adding biological aging measures, and at 10 years from 0.797 to 0.805. Brier scores decreased from 0.0402 to 0.0394 at 5 years and from 0.0900 to 0.0877 at 10 years.

In UKB, the Harrell's C-index increased from 0.692 (95% CI: 0.688–0.697) for chronological age alone to 0.720 (0.715–0.724) for the conventional model and to 0.746 (0.742–0.751) after addition of biological

aging measures, corresponding to ΔC = 0.0263 (95% CI: 0.0240–0.0286) beyond conventional risk factors. The 5-year AUC increased from 0.700 to 0.741, and the 10-year AUC from 0.720 to 0.750. Corresponding Brier scores decreased from 0.0149 to 0.0147 at 5 years and from 0.0405 to 0.0395 at 10 years. Paired uncertainty analyses supported improvements in both time-dependent AUC and Brier score at 5 and 10 years in both cohorts, with all corresponding 95% confidence intervals excluding zero (Supplementary **Table S1**).

In sensitivity analyses, replacing PhenoAge residual with KDM residual or log-transformed HD while retaining FI and TL also improved prediction beyond conventional risk factors. In TwinGene, the KDM- and HD-based models yielded C-indices of 0.782 and 0.786, 5-year AUCs of 0.783 and 0.790, and 10-year AUCs of 0.798 and 0.804, respectively. In UKB, the corresponding C-indices were 0.745 and 0.744, the 5-year AUCs were 0.736 and 0.730, and the 10-year AUCs were 0.747 and 0.741. Among the three evaluated enhanced models, the PhenoAge-based model had the highest discrimination and lowest prediction error in UKB, whereas differences among the three biological aging-enhanced models were smaller in TwinGene. Overall, the censoring-aware discrimination and prediction-error metrics consistently indicated incremental prognostic information from the biological aging measures beyond chronological age and conventional risk factors.

Calibration of the biological aging-enhanced model at 5 and 10 years in both cohorts is shown in Fig. 5. In TwinGene, predicted mortality exceeded observed mortality in the highest-risk decile at 5 years (21.79% versus 19.65%), whereas predicted and observed mortality were closer at 10 years (52.84% versus 52.08%). The second-highest decile showed overprediction at 10 years (24.44% predicted versus 21.73% observed). In UKB, the highest-risk decile showed underprediction at both horizons, with predicted versus observed mortality of 5.20% versus 5.52% at 5 years and 14.47% versus 15.30% at 10 years.

Exploratory decision curve analyses showed that the biological aging-enhanced model generally provided greater net benefit than the conventional model across most evaluated threshold probabilities at both 5 and 10 years in TwinGene and UKB (Supplementary **Fig. S1**). The separation was more pronounced in UKB, particularly at 10 years, whereas the incremental net benefit was smaller in TwinGene, consistent with the more modest improvements in discrimination observed in that cohort.

Effect modification

Effect modification by sex, BMI, smoking, and education was evaluated for each of the five biological aging measures, with 20 interaction tests per cohort and Benjamini–Hochberg correction applied separately within each cohort (Supplementary **Table S2**). In TwinGene, only the interaction between HD and sex remained significant after FDR correction (*P* for interaction = 0.0014; FDR-adjusted *P* = 0.027). The association of HD with mortality was stronger in men (HR = 1.34, 95% CI: 1.27–1.40) than in women (HR = 1.19, 95% CI: 1.12–1.26).

In UKB, eight interactions survived FDR correction: PhenoAge residual with smoking; KDM residual with sex, BMI, and smoking; HD with sex and BMI; FI with sex; and TL with sex. No interaction involving education survived correction in either cohort. Although the HD–sex interaction was supported in both cohorts, its direction differed: in UKB, HD was more strongly associated with mortality in women (HR = 1.32, 95% CI: 1.28–1.36) than in men (HR = 1.21, 95% CI: 1.19–1.24), opposite to the pattern in TwinGene. Thus, no interaction showed both FDR-supported evidence and a consistent direction across cohorts. Subgroup-specific HRs per 1-SD higher aging measure and their 95% CIs are presented in Supplementary **Table S3** and Supplementary **Fig. S2**.

Proportional hazards diagnostics and sensitivity analyses

Schoenfeld residual diagnostics indicated some departures from the proportional hazards assumption (Supplementary **Table S4**). In TwinGene, exposure-specific evidence of non-proportionality was observed for PhenoAge residual ($P = 0.0012$) and HD ($P = 0.0375$), whereas there was no evidence for KDM residual ($P = 0.678$), FI ($P = 0.079$), or TL ($P = 0.876$). In UKB, exposure-specific tests indicated departures from proportionality for all five aging measures, alongside significant global tests.

To characterize these departures, associations were estimated separately during 0–5, 5–10, and > 10 years of follow-up (Supplementary **Table S5**). In TwinGene, there was no evidence of variation in exposure effects across intervals for any aging measure (all P for variation over time > 0.05). In UKB, PhenoAge residual, KDM residual, HD, and FI showed significant time variation (all $P < 0.001$), with generally stronger associations earlier in follow-up and attenuation thereafter, whereas there was no evidence of variation in the TL association across intervals ($P = 0.420$). Accordingly, the primary Cox estimates, particularly in UKB, should be interpreted as average associations over follow-up rather than strictly time-invariant effects.

The primary associations were nevertheless robust to alternative analytic specifications. In a 2-year landmark analysis excluding participants who died or were censored within the first 2 years, associations were similar to those observed in the primary analysis (Supplementary **Table S6**). PhenoAge residual remained associated with mortality in TwinGene (HR: 1.21, 95% CI: 1.17–1.25) and UKB (HR: 1.37, 95% CI: 1.35–1.40). Corresponding estimates for KDM residual were 1.11 (95% CI: 1.07–1.15) and 1.34 (95% CI: 1.32–1.36), respectively. HD and FI also remained positively associated with mortality, whereas the TL estimate remained directionally inverse in TwinGene and inversely associated in UKB.

Results were also consistent when attained age was used as the underlying timescale (Supplementary **Table S7**). PhenoAge residual remained associated with mortality in TwinGene (HR: 1.22, 95% CI: 1.18–1.26) and UKB (HR: 1.45, 95% CI: 1.43–1.47), as did KDM residual (HR: 1.11, 95% CI: 1.07–1.15 and HR: 1.40, 95% CI: 1.38–1.42, respectively). HD and FI remained positively associated in both cohorts, while TL remained inversely associated, although the association was weaker in UKB (TwinGene HR: 0.97, 95% CI: 0.94–0.99; UKB HR: 0.96, 95% CI: 0.95–0.98). Overall, these sensitivity analyses supported the

robustness of the primary mortality associations to early-event exclusion and alternative control for chronological age.

Discussion

In this study of two population-based cohorts, biological aging measures provided mortality-risk information beyond chronological age and conventional risk factors, although the magnitude of this additional information differed between settings. In multivariable Cox models evaluating each biological aging measure individually, higher PhenoAge residual, KDM residual, HD, and FI were associated with higher all-cause mortality hazards in both TwinGene and UKB, whereas higher TL was associated with lower mortality hazards. Adding PhenoAge residual, FI, and TL together to the conventional model containing chronological age, sex, BMI, smoking, and education increased Harrell's C-index from 0.780 to 0.787 in TwinGene and from 0.720 to 0.746 in UKB. Improvements were also observed for 5- and 10-year time-dependent AUC and Brier scores. These findings support the central premise that biological aging measures capture prognostic information that is not fully represented by chronological age, while also showing that their incremental predictive contribution is not uniform across populations and measurement settings.

The associations observed for the biomarker-based measures are consistent with their complementary conceptual foundations. PhenoAge was originally constructed to summarize a biomarker-derived mortality profile, whereas KDM estimates biological age from the relation between physiological biomarkers and chronological age, and HD quantifies multivariate deviation from a comparatively healthy physiological reference state [2–4,6]. In this study, PhenoAge residual, KDM residual, and HD were associated with mortality after adjustment for chronological age and conventional risk factors, but their relative effect sizes differed between cohorts: HD had the largest HR among these measures in TwinGene, whereas PhenoAge residual and KDM residual had larger HRs than HD in UKB. Thus, our findings do not support treating one algorithm as universally superior. Rather, they reinforce the view that different clinical biomarker algorithms capture partly overlapping but distinct dimensions of aging physiology. This interpretation is consistent with previous work showing that clinically derived biological aging measures predict age-related outcomes, including cancer and neurological disease, while differing in the magnitude of their associations across outcomes and populations[12,13].

The correlation structure further illustrates why these measures should not be considered interchangeable. Raw PhenoAge and raw KDM were strongly correlated with chronological age and with each other, whereas raw HD, FI, and cohort-specific adjusted TL showed weaker correlations with these age-scale measures. These descriptive correlations concern the raw measures, whereas the mortality models used age-adjusted PhenoAge and KDM residuals and log-transformed HD. Similar partial divergence among biological aging measures has been reported previously, suggesting that measures constructed from clinical physiology, deficit accumulation, and cellular aging need not reflect the same underlying process[1,10]. FI was associated with mortality after adjustment for chronological age and conventional risk factors in both cohorts, consistent with the deficit-accumulation model in which

increasing numbers of health deficits reflect declining physiological reserve and vulnerability; meta-analytic evidence has similarly demonstrated a graded association between higher FI and mortality[7,14]. TL showed smaller inverse associations with mortality, which is also consistent with previous evidence: a meta-analysis including Swedish Twin Registry cohorts found that shorter leukocyte telomeres were associated with higher all-cause mortality, while also showing heterogeneity according to age and measurement methodology[15]. Together, these results support a multidimensional view of biological aging rather than the assumption that a single measure captures all mortality-relevant aging processes.

A key contribution of the present study is the explicit evaluation of incremental prognostic value, rather than relying solely on associations with mortality. Strong hazard ratios do not necessarily translate into meaningful improvements in prediction when chronological age and established risk factors already discriminate risk well. An international study of more than 430,000 surgical patients found that PhenoAge was associated with postoperative outcomes after adjustment for chronological age and established surgical risk factors[16]. In TwinGene, the conventional model already achieved Harrell's C-index of 0.780, and adding biological aging measures increased this to 0.787 (ΔC = 0.0073). In UKB, the corresponding increase was larger, from 0.720 to 0.746 (ΔC = 0.0263). Time-dependent AUC and Brier scores showed the same overall pattern. Exploratory decision curve analyses generally showed higher net benefit for the biological aging-enhanced model than for the conventional model at both prediction horizons. However, because no specific clinical intervention or threshold for action was prespecified, these findings should be regarded as exploratory rather than evidence of established clinical utility. The smaller incremental improvement in TwinGene may partly reflect the stronger discrimination already provided by chronological age and conventional factors in this older cohort, leaving less opportunity for additional predictors to improve ranking. Differences in baseline age, mortality incidence, follow-up, biomarker panels, and measurement implementation may also contribute to the different magnitude of improvement, but these explanations were not formally tested. Accordingly, the findings should be interpreted as evidence that biological aging information adds to, rather than replaces, chronological age and established risk factors.

The directionally consistent mortality associations observed despite substantial differences between TwinGene and UKB support consistency of mortality associations across distinct cohort settings. TwinGene participants were older and had a higher mortality proportion and longer follow-up than UKB participants. The TwinGene biomarker-based aging algorithms used eight biomarkers, whereas the UKB implementation incorporated 18 clinical biomarkers spanning multiple physiological systems. Previous studies have emphasized that biological aging estimates depend on biomarker selection, reference populations, and algorithmic implementation and therefore should not automatically be assumed to be numerically interchangeable across datasets[6,10]. Our results extend this point by showing that directionally consistent mortality associations can coexist with differences in effect magnitude and predictive improvement across cohort-specific implementations. This distinction is important: the present analysis evaluates consistency of prognostic information across two settings rather than formal equivalence of biological aging measures or direct comparison of their absolute values.

Effect-modification analyses provided little evidence for a consistently reproducible modifier of biological aging–mortality associations. After correction for 20 interaction tests per cohort, only the HD-by-sex interaction remained significant in TwinGene, whereas eight interactions involving sex, BMI, or smoking survived correction in the substantially larger UKB sample. HD-by-sex was detected in both cohorts, but the subgroup pattern differed: HD was more strongly associated with mortality in men than women in TwinGene, whereas the reverse pattern was observed in UKB. This lack of directional consistency argues against interpreting the interaction findings as evidence of a stable sex-specific biological mechanism. More broadly, the greater number of significant interactions in UKB may partly reflect its much larger sample size and statistical power. We therefore regard these analyses as secondary evidence of possible context-dependent heterogeneity rather than as evidence supporting subgroup-specific risk thresholds.

Proportional hazards diagnostics indicated time-varying associations, particularly in UKB. In UKB, interval-specific analyses showed that associations for PhenoAge residual, KDM residual, HD, and FI were generally stronger earlier in follow-up and attenuated over time, whereas there was no evidence of variation in the TL association across the specified intervals (P = 0.420). The primary Cox estimates should therefore be interpreted as average associations over the full follow-up period rather than as strictly time-invariant effects. Nevertheless, the overall pattern of associations was broadly preserved across follow-up intervals, and the 2-year landmark and attained-age sensitivity analyses supported the main conclusions.

Several strengths should be considered when interpreting these findings. We evaluated the same overarching research question independently in two large population-based cohorts with markedly different age structures, follow-up profiles, mortality rates, and biomarker implementations. The analysis incorporated censoring-aware survival metrics rather than binary mortality ROC analyses, used 10-fold cross-validation with preprocessing estimated within training folds and predictions generated for held-out participants, retained co-twins in the same TwinGene folds, and accounted for within-pair dependence using cluster-robust inference. Multiple-testing correction was also applied to the primary biological aging associations and interaction analyses. These design features reduce several sources of optimistic performance estimation and directly address dependencies specific to the twin cohort.

Several limitations remain. First, biological aging and covariates were measured primarily at baseline; consequently, the analyses characterize baseline risk rather than longitudinal rates of biological aging, and causal interpretations cannot be made. Second, PhenoAge and KDM residuals were defined relative to cohort-specific age relationships, while biomarker panels and measurement procedures differed between TwinGene and UKB. The resulting measures therefore quantify relative biological aging within each setting, and their absolute values and effect magnitudes should not be interpreted as directly interchangeable. This consideration also applies to cohort-specific TL measurements. Third, both cohorts may differ from populations encountered in routine clinical practice. UKB in particular is affected by healthy-volunteer selection, with participants having lower mortality and generally more favorable health profiles than the UK population[17]. Complete-case selection and exclusion of extreme biological-

age values may have introduced selection bias and limit generalizability. Additional evaluation in more socioeconomically and ethnically diverse populations is therefore warranted. Fourth, cross-validation estimates internal predictive performance and does not establish transportability to an external population. Uncertainty estimates based on fixed out-of-fold predictions do not capture the full variability associated with refitting the models and their preprocessing pipelines. Although exploratory decision curve analyses suggested greater net benefit for the biological aging-enhanced model across a range of threshold probabilities, no specific clinical intervention or action threshold was prespecified. External validation and prospective impact studies are therefore required before these models can be considered for individual risk stratification or treatment decisions. Fifth, all-cause mortality combines heterogeneous causes of death and therefore does not identify which disease pathways account for the observed associations. Finally, we could not evaluate GOLD BioAge because several required biomarker inputs were unavailable in TwinGene, preventing harmonized computation across both cohorts; future work should compare these newer approaches under harmonized measurement conditions. GOLD BioAge has recently been proposed as a clinically simpler mortality-oriented biological aging framework and has been shown to predict morbidity and mortality in NHANES and UKB[18].

In conclusion, biological aging measures provided mortality-risk information beyond chronological age and conventional risk factors in both TwinGene and UKB. The findings support biological aging as a complementary prognostic dimension rather than a replacement for chronological age: associations were observed across all five measures, while the incremental predictive gain from the selected measures ranged from modest in TwinGene to more pronounced in UKB. Differences across measures and cohorts further emphasize that biological aging is not a single interchangeable quantity and that performance depends on population and measurement context. Further studies should evaluate these measures prospectively in additional populations, incorporate repeated measurements of biological aging, and determine whether the observed improvements translate into useful risk stratification or clinical decision benefit.

Methods

Study design and populations

We conducted parallel analyses in two population-based cohorts, TwinGene and UKB, to evaluate whether commonly used biological aging measures provide prognostic information for all-cause mortality beyond chronological age and conventional risk factors. The same overarching analytical framework was applied independently within each cohort.

TwinGene is a population-based study within the Swedish Twin Registry (STR). Between 2004 and 2008, approximately 12,600 twins born in 1958 or earlier who had previously participated in the Screening Across the Lifespan Twin Study (SALT) were recruited for health assessment and blood collection[19,20]. Blood samples were obtained following overnight fasting and linked to questionnaire, health examination, and nationwide register data.

Among 12,256 TwinGene participants, 9,895 had complete data for the eight biomarkers required to derive the biomarker-based aging measures. After excluding participants with extreme biological-age values (> 5 SD from the mean), 9,617 participants remained. These participants also had complete mortality follow-up, FI, TL, and covariate information and constituted the final analytic sample. These participants represented 6,466 unique twin-pair identifiers; 3,151 pairs contributed both co-twins and 3,315 contributed one participant. Familial dependence was explicitly incorporated into the statistical analysis rather than removing one member of complete twin pairs. TwinGene recruitment and biomarker collection have been described previously[19,21].

UKB is a population-based prospective study of more than 500,000 adults aged approximately 37–73 years recruited from 22 assessment centers in England, Wales, and Scotland between 2006 and 2010. Participants provided questionnaire information, physical measurements, biological samples, and consent for linkage to national health records [22]. For the present phenotype-based analyses, 331,699 participants had complete data for the 18 biomarker inputs used to derive the biomarker-based aging measures. Of these, 331,601 had PhenoAge, KDM, and HD estimates. One participant with an extreme age-scale biological-age value (> 5 SD for PhenoAge or KDM) was excluded, leaving 331,600 participants. After additionally requiring valid mortality follow-up, complete FI, TL, and complete information on age, sex, BMI, smoking, and education, the final complete-case analytic sample comprised 314,064 participants.

Ethics statement

TwinGene participants provided written informed consent, and the study was approved by the relevant regional ethics review boards in Sweden. The present project was approved by the Swedish Ethical Review Authority (Dnr. 2022-06634-01). UKB received approval from the North West Multi-Centre Research Ethics Committee, and all participants provided written informed consent. The present analyses were conducted under UKB Application 22224. Data available to the investigators were de-identified. The study was conducted in accordance with the Declaration of Helsinki and is reported in accordance with the STROBE statement for observational studies and the applicable TRIPOD recommendations for prediction-model studies.

Biological aging measures

Biological aging was characterized using five complementary measures: PhenoAge, KDM, HD, FI, and TL. PhenoAge and KDM were calculated using cohort-specific biomarker panels and implementations trained in the National Health and Nutrition Examination Survey (NHANES) III within the BioAge framework; HD was calculated relative to an NHANES III reference sample[2–4,6]. KDM estimates an age-equivalent physiological state from age-related biomarker patterns; PhenoAge maps a biomarker-based mortality score onto an age scale; and HD quantifies multivariate physiological deviation from a comparatively healthy reference state using Mahalanobis distance. Unlike PhenoAge and KDM, HD is not itself expressed as an age in years.

TwinGene biomarker-based aging measures

The TwinGene implementations of PhenoAge, KDM, and HD used eight clinical biomarkers: serum glucose, HbA1c, creatinine, cystatin C, triglycerides, total cholesterol, LDL cholesterol, and C-reactive protein (CRP). The biomarker panel was defined from measures available in TwinGene using previously specified criteria of ≤ 20% missingness and an absolute correlation with chronological age > 0.1.

The algorithms were parameterized in NHANES III using the R package BioAge and subsequently applied to TwinGene[6]. KDM parameters were estimated separately in women and men among nonpregnant participants aged 30–75 years. PhenoAge was parameterized using a Gompertz mortality model incorporating the selected biomarkers and chronological age among participants aged 20–84 years, without sex-specific fitting. For HD, sex-specific reference distributions were derived from nonpregnant participants aged 20–30 years with BMI < 30 kg/m² and selected biomarker values within clinically acceptable ranges. NHANES survey weights were not used. PhenoAge training used the mortality outcome and follow-up supplied in the BioAge NHANES III dataset, derived from linked mortality data through 2015 with follow-up capped at 240 months. The fitted parameters and reference distributions were then used to calculate the corresponding measures in TwinGene.

Logarithmic transformations were applied to glucose (mmol/L), HbA1c (%), creatinine (μmol/L), triglycerides (mg/dL), and CRP before calculation of the biological aging measures. Total and LDL cholesterol were expressed in mg/dL and cystatin C in mg/L.

Clinical chemistry measurements in TwinGene have been described previously[21,23]. Glucose, total cholesterol, LDL cholesterol, and triglycerides were measured using routine clinical chemistry methods on semi-automated Beckman Coulter analyzers, while HbA1c was measured using ion-exchange chromatography. High-sensitivity CRP was measured using an immunoturbidimetric assay. Creatinine was quantified enzymatically using Architect c8000/c16000 analyzers, and cystatin C using a particle-reinforced immunoturbidimetric assay on an Architect ci8200 analyzer (Abbott Laboratories). The distributions of the eight biomarkers in the final analytic sample are shown in Supplementary **Fig. S3**.

UKB biomarker-based aging measures

The NHANES III–trained model objects described by Mak et al. were applied to UKB using the corresponding biomarker units and transformations[12]. The three algorithms incorporated 18 age-associated clinical measures spanning metabolic, renal, hepatic, inflammatory, hematological, cardiovascular, anthropometric, and pulmonary domains: forced expiratory volume in 1 second (FEV_1), systolic blood pressure, blood urea nitrogen, HbA1c, total cholesterol, creatinine, serum glucose, waist circumference, red cell distribution width, albumin, alkaline phosphatase, triglycerides, mean cell volume, uric acid, lymphocyte percentage, red blood cell count, C-reactive protein, and diastolic blood pressure. As described by Mak et al., candidate biomarkers were selected in NHANES III based on ≤ 20% missingness and an absolute correlation with chronological age > 0.1. KDM was trained in non-pregnant participants aged 30–75 years with complete biomarker data, and PhenoAge in participants aged 20–84 years with complete biomarker data. The HD reference comprised non-obese participants aged 20–30 years whose biomarkers were within age- and sex-specific normal ranges. All three algorithms were

parameterized in NHANES III using the BioAge framework and subsequently projected to UKB, as described previously[12].

UKB serum chemistry biomarkers were measured using standardized high-throughput assays, with most chemistry measures quantified on the Beckman Coulter AU5800 platform. HbA1c was measured in red blood cells using high-performance liquid chromatography on the Bio-Rad VARIANT II Turbo platform. Hematological measures were obtained from fresh EDTA whole blood using automated hematology analyzers, while waist circumference, blood pressure, and spirometric measures were collected using standardized UKB assessment protocols. The distributions of the 18 inputs in the final analytic sample are shown in Supplementary **Fig. S3**.

Frailty index and telomere length

FI was treated as a complementary deficit-accumulation measure of biological aging rather than as a biomarker-derived age estimate. FIs were constructed following the standard deficit-accumulation approach in which health deficits are coded between 0 and 1 and summed relative to the number of evaluable deficits[7].

In TwinGene, FI was based on 44 self-reported health deficits collected through SALT and spanning multiple health domains. This 44-item SALT FI has previously been evaluated in relation to mortality[24]. In UKB, the FI was based on the previously validated 49-item measure developed from baseline questionnaire data[25]. FI was retained on its original deficit-proportion scale for descriptive analyses and standardized within cohort for association and prediction analyses.

Leukocyte TL was included as a molecular measure of cellular aging. In TwinGene, TL was measured from peripheral-blood DNA using quantitative polymerase chain reaction (qPCR), and the cohort-specific adjusted TL measure was used in the analyses. Previous STR studies have described the measurement and mortality associations of TL in TwinGene[15]. In UKB, leukocyte TL was measured using a multiplex qPCR method and expressed relative to a single-copy gene as a T/S ratio; we used the available adjusted UKB TL measure accounting for technical sources of variation[26]. Because TL measurement and adjustment procedures differed between cohorts, absolute TL values were treated as cohort-specific rather than directly comparable.

Age adjustment and transformation of biological aging measures

Raw PhenoAge, KDM, and HD values were retained for descriptive summaries and correlation analyses. For mortality association and prediction analyses, PhenoAge and KDM were processed following the approach used in previous clinical biological-aging studies. Within each cohort, raw PhenoAge and KDM were regressed on chronological age using a natural spline with 3 degrees of freedom, and the resulting residuals were retained as PhenoAge residual and KDM residual, respectively. Positive residuals indicate a biological-age estimate above that expected for chronological age. This spline-based residualization

reduces both linear and nonlinear dependence on chronological age[12]. HD was not residualized on chronological age, because it is defined as physiological deviation from a reference state rather than as an age measured in years. Owing to its right-skewed distribution, HD was natural-log transformed, consistent with previous applications[12]. PhenoAge residual, KDM residual, log-transformed HD, FI, and TL were subsequently standardized to mean 0 and SD 1 within each cohort so that effect estimates corresponded to a 1-SD difference in each aging measure. For association analyses, residualization and standardization were performed within each full analytic cohort. For prediction analyses, residualization models and standardization parameters were estimated within training folds and applied to the corresponding held-out folds.

Outcome definition and follow-up

The primary outcome was all-cause mortality. Follow-up time was calculated from the baseline assessment to death from any cause or administrative censoring, whichever occurred first.

For TwinGene, mortality was identified through linkage of STR participants to Swedish national population and mortality registers using personal identity numbers. Follow-up for the present analysis extended through 1 February 2024. For UKB, deaths were ascertained through linked national mortality registers. Follow-up time was calculated from the baseline assessment date to death from any cause or administrative censoring on 30 December 2022, whichever occurred first. Time since baseline was used as the primary survival timescale.

Covariates and harmonization

The primary covariate set was deliberately harmonized across the two cohorts and comprised chronological age, sex, BMI, smoking, and education. Age and BMI were modeled flexibly in the primary Cox models using natural splines rather than assuming strictly linear relationships.

Smoking was harmonized as ever versus never smoking. Education was harmonized as high/intermediate versus low education. In TwinGene, high/intermediate education was defined as ≥ 10 years of education. In UKB, education was based on the highest self-reported qualification and classified as low, intermediate, or high; for harmonization, the intermediate and high categories were combined and contrasted with the low category. Alcohol consumption was available in UKB but was not included in the primary model because an equivalent harmonized variable was not available for TwinGene.

In both cohorts, we included all participants who met the eligibility criteria and had complete information on the cohort-specific biomarker inputs, all five biological aging measures, mortality follow-up, and covariates required for the primary analyses. No missing values were imputed during biomarker processing or construction of the final analytical datasets.

Statistical analysis

Descriptive and correlation analyses

Baseline characteristics were summarized as mean ± SD, median [interquartile range], or number (%), as appropriate. Raw PhenoAge, raw KDM, raw HD, FI on its deficit-proportion scale, and cohort-specific adjusted TL were used for descriptive analyses. Pairwise Pearson correlation coefficients were calculated among chronological age and these measures separately within TwinGene and UKB.

Primary mortality association analyses

Associations between each biological aging measure and all-cause mortality were estimated using Cox proportional hazards regression with time since baseline as the underlying timescale. Each aging measure was evaluated individually in a multivariable Cox model containing the conventional covariates; the aging measures were not mutually adjusted. Models were adjusted for chronological age using a natural spline with 4 degrees of freedom, sex, BMI using a natural spline with 3 degrees of freedom, ever-smoking status, and education.

Hazard ratios (HRs) with 95% confidence intervals (CIs) are reported per 1-SD higher biological aging measure. In TwinGene, all Cox models used pair-clustered robust variance estimation based on the twin-pair identifier to account for within-pair dependence while retaining both members of complete pairs. UKB participants were treated as independent observations.

Tests of the biological aging–mortality associations and multiplicative interactions used two-sided Wald tests. *P*-values for the five primary biological aging-measure associations were adjusted separately within each cohort using the Benjamini–Hochberg FDR procedure[27]. FDR-adjusted $P < 0.05$ was considered statistically significant. No formal statistical tests compared HR magnitudes between cohorts.

Cross-validated mortality prediction

Incremental prognostic performance was evaluated using three prespecified nested Cox models. The age model included chronological age modeled with a 4-degree-of-freedom natural spline. The conventional model additionally included sex, BMI modeled with a 3-degree-of-freedom natural spline, smoking, and education. The biological aging-enhanced model further included PhenoAge residual, FI, and TL. To construct a parsimonious multidomain model, we retained one clinical biomarker-derived aging algorithm alongside conceptually complementary measures of deficit accumulation (FI) and cellular aging (TL). PhenoAge was selected because its original construction incorporates mortality-related information; its strong association with chronological age in both cohorts supported using its age-residualized form. KDM residual and HD remained part of the individual association analyses but were excluded from the joint prediction model to limit overlapping biomarker information and potential multicollinearity. The enhanced model therefore evaluated one selected multidomain predictor set rather than exhaustively comparing all possible combinations of biological aging measures. In sensitivity analyses, PhenoAge residual was replaced in turn by KDM residual or log-transformed HD, while FI and TL were retained. These alternative models used the same cross-validation folds, covariates, preprocessing procedures, and performance metrics as the primary biological aging-enhanced model.

The sensitivity analysis was intended to assess robustness to the choice of biomarker-derived aging algorithm rather than to perform exhaustive model selection.

Prediction performance was estimated using 10-fold cross-validation with out-of-fold predictions. A single 10-fold partition was used within each cohort and was held fixed across all three prediction models. In TwinGene, the existing pair-grouped cross-validation partition was retained, with co-twins assigned to the same fold to prevent information from one member of a pair entering the training set while the co-twin was held out. In UKB, a single mortality-balanced 10-fold partition was generated using a fixed random seed and reused for all models. Fold assignments were fixed throughout the analyses; uncertainty estimates therefore condition on the observed fold allocation and do not incorporate variability arising from alternative cross-validation partitions.

To avoid preprocessing leakage, transformations requiring parameter estimation were performed independently within each training fold. PhenoAge and KDM were separately residualized on chronological age using natural splines with 3 degrees of freedom fitted in the training data; the fitted functions were then applied unchanged to the corresponding held-out participants. HD was log-transformed. PhenoAge residual, KDM residual, log-transformed HD, FI, and TL were standardized using means and SDs estimated exclusively from the training fold and applied unchanged to the held-out fold. Thus, each participant's prediction was generated using preprocessing and model fitting that did not use that participant's data for parameter estimation.

Overall discrimination was summarized using Harrell's C-index with 95% CIs. Incremental discrimination was quantified as the within-cohort difference in Harrell's C-index between the biological aging-enhanced and conventional models. Harrell's C-index was calculated from the final out-of-fold linear predictors using survival::concordance. In TwinGene, uncertainty in ΔC was estimated using 1,000 twin-pair cluster-bootstrap replicates, resampling pair IDs with co-twins retained together while keeping the original folds and out-of-fold predictions fixed. In UKB, uncertainty in ΔC was derived from the covariance of the paired concordance estimates[28]. For the alternative biomarker-based aging sensitivity models, performance was summarized using the same Harrell's C-index, 5- and 10-year time-dependent AUCs, and Brier scores. Differences relative to the conventional model and to the PhenoAge-based enhanced model were examined as sensitivity contrasts rather than as formal selection of a superior biological aging algorithm.

Because mortality is a time-to-event outcome subject to censoring, prediction at specific horizons was evaluated using censoring-aware methods rather than ordinary binary ROC curves. Time-dependent AUCs at 5 and 10 years were calculated using inverse-probability-of-censoring weighting (IPCW), and prediction error was evaluated using Brier scores at the same horizons[29,30]. Higher Harrell's C-index and AUC indicate better discrimination, whereas lower Brier scores indicate lower prediction error. Uncertainty in incremental predictive performance was evaluated using paired comparisons of the final out-of-fold predictions. In TwinGene, 95% confidence intervals were estimated using 1,000 twin-pair cluster-bootstrap replicates, retaining co-twins within the same resampled cluster. In UKB, paired

influence-function inference was used for differences in time-dependent AUC and Brier score, while uncertainty in the Harrell's C-index difference was derived from the covariance of the paired concordance estimates. These uncertainty estimates were conditional on the fixed out-of-fold predictions and did not include refitting the prediction models or their preprocessing pipelines.

Calibration of the biological aging-enhanced model was evaluated separately at 5 and 10 years by comparing cross-validated predicted mortality risk with censoring-adjusted observed risk across deciles of predicted risk. The primary objective was to assess within-cohort incremental prediction; predictive performance was therefore not subjected to formal statistical comparison between TwinGene and UKB.

As an exploratory assessment of potential decision utility, decision curve analysis[31] was performed at 5 and 10 years using the final out-of-fold absolute-risk predictions from the conventional and biological aging-enhanced models. Net benefit was compared with treat-all and treat-none strategies across illustrative threshold probabilities of 0.5–20% for 5-year mortality and 1–30% for 10-year mortality. The biological aging-enhanced model included PhenoAge residual, FI, and TL in addition to the conventional predictors. Because no specific clinical intervention or action threshold was prespecified, the decision-curve analyses were considered exploratory and were not interpreted as establishing clinical utility. Censoring before each evaluation horizon was accounted for using Kaplan–Meier survival estimates within the time-to-event decision curve analysis. No smoothing was applied.

Effect-modification analyses

Effect modification was evaluated for sex, BMI, smoking, and education by adding a multiplicative interaction term between each modifier and each of the five biological aging measures to the adjusted Cox model. This yielded 20 interaction tests per cohort. TwinGene interaction models retained pair-clustered robust variance estimation. Each interaction was evaluated using a one-degree-of-freedom Wald test of the multiplicative interaction coefficient.

Sex, smoking, and education were modeled categorically. For BMI interaction analyses, BMI was standardized within cohort and modeled continuously. Benjamini–Hochberg FDR correction was applied separately within each cohort across the 20 interaction tests. Only interactions with FDR-adjusted $P < 0.05$ were interpreted further. For FDR-supported categorical interactions, subgroup-specific HRs and 95% CIs were estimated; for BMI interactions, HRs were estimated at the cohort-specific mean and at − 1 and + 1 SD of BMI.

Proportional hazards diagnostics and sensitivity analyses

The proportional hazards assumption was evaluated using Schoenfeld residual diagnostics for each primary Cox model[32]. Because such tests can be highly sensitive in very large samples, exposure-specific and global tests were considered together with the robustness of effect estimates under alternative model specifications. To further characterize departures from the proportional hazards assumption, follow-up was partitioned into 0–5, 5–10, and > 10 years, corresponding to the principal

prediction horizons used in the study. Cox models allowing separate biological aging coefficients within each follow-up interval were fitted using interval-stratified baseline hazards while retaining the covariate specification of the primary models. Equality of the three interval-specific exposure coefficients was evaluated using a robust Wald test with two degrees of freedom. TwinGene models used twin-pair-clustered robust variance estimation, whereas UKB interval-split models used participant-clustered robust variance estimation.

Two additional sensitivity analyses were performed. First, to reduce the potential influence of severe baseline illness and imminent death, a 2-year landmark analysis was conducted among participants who remained alive and under observation beyond 2 years after baseline. Follow-up was restarted at the 2-year landmark, and age at the landmark was used in the adjustment model.

Second, the primary Cox analyses were repeated using attained age as the underlying timescale in a left-truncated start–stop formulation, with age at baseline as entry time and age at death or censoring as exit time. Chronological age was therefore not additionally included as a model covariate in this sensitivity analysis. Other covariates and TwinGene pair-clustered robust inference were retained.

We considered the recently proposed GOLD BioAge as a contemporary comparator[18]. The full GOLD BioAge algorithm incorporates clinical biomarkers including hematological, hepatic, renal, metabolic, and inflammatory measures. Several required inputs were not available within the harmonized TwinGene biomarker panel; consequently, GOLD BioAge could not be constructed consistently in both cohorts and was not included in head-to-head analyses.

Declarations

Software and reproducibility: Analyses were conducted in R version 4.5.1. Cox regression and concordance analyses used the “survival” package [33] and censoring-aware prediction assessment used “riskRegression” package[34], subgroup-specific effects following supported interactions were estimated using “emmeans”, clinical biological-aging algorithms were implemented using “BioAge” [6], decision curve analysis used the “dcurves” package, and figures were generated using “ggplot2” package[35]. Survival prediction analyses used out-of-fold predictions throughout.

Acknowledgements: This research was conducted using the UK Biobank Resource under Application 22224. We thank the UKB participants for their contribution to this work and acknowledge the Swedish Twin Registry for access to data. The Swedish Twin Registry is managed by Karolinska Institutet and receives funding through the Swedish Research Council under the grant no 2021-00180. This study was supported by the Swedish Research Council (Sara Hägg: 2022-01608) and the National Institute on Aging (Sara Hägg: R01AG067996-01A1). The funders had no role in the study design, data analysis, interpretation of results, preparation of the manuscript, or decision to submit the work for publication.

Author contributions: S.M., J.L., and S.H. contributed to conceptualization and methodology. S.M. and J.L. performed software development, formal analysis, and data curation. J.K.L.M. and C.Q. contributed

to data preparation and investigation. S.H. provided resources and supervision. S.M. led visualization and project administration. S.M. and S.H. wrote the original draft. All authors reviewed and edited the manuscript, approved the final version, and accept responsibility for the work.

Competing interests: The authors declare no competing interests.

Data availability: The UK Biobank data are available to registered researchers through the UK Biobank online portal (www.ukbiobank.ac.uk). The TwinGene data are part of the Swedish Twin Registry and are not publicly available due to ethical and legal restrictions. Access to TwinGene data requires approval from the Swedish Ethical Review Authority and the Steering Committee of the Swedish Twin Registry. Researchers interested in accessing TwinGene data may contact the registry (str-research@meb.ki.se) for further information on the application process.

Code availability: Analysis code used to generate the results reported in this study is available on GitHub. The repository identifies the scripts, software dependencies, and analysis sequence used for TwinGene and UK Biobank.

Supplementary material: Supporting information is available.

Generative AI and AI-assisted tools: We used ChatGPT (OpenAI) for language editing and to improve clarity of the manuscript text, and GitHub Copilot for assistance during R code development. All analyses, results, and interpretations were performed by the authors, who verified the correctness of the code and take full responsibility for the content of the manuscript.

References

1. Jylhävä, J., Pedersen, N. L. & Hägg, S. Biological age predictors. *EBioMedicine* 21, 29–36 (2017).
2. Klemera, P. & Doubal, S. A new approach to the concept and computation of biological age. *Mechanisms of ageing and development* 127, 240–248 (2006).
3. Liu, Z. *et al.* A new aging measure captures morbidity and mortality risk across diverse subpopulations from NHANES IV: a cohort study. *PLoS medicine* 15, e1002718 (2018).
4. Cohen, A. A. *et al.* A novel statistical approach shows evidence for multi-system physiological dysregulation during aging. *Mechanisms of ageing and development* 134, 110–117 (2013).
5. Cohen, A. A. *et al.* Cross-population validation of statistical distance as a measure of physiological dysregulation during aging. *Experimental gerontology* 57, 203–210 (2014).
6. Kwon, D. & Belsky, D. W. A toolkit for quantification of biological age from blood chemistry and organ function test data: BioAge. *Geroscience* 43, 2795–2808 (2021).
7. Searle, S. D., Mitnitski, A., Gahbauer, E. A., Gill, T. M. & Rockwood, K. A standard procedure for creating a frailty index. *BMC geriatrics* 8, 1–10 (2008).
8. Gao, X., Zhang, Y., Mons, U. & Brenner, H. Leukocyte telomere length and epigenetic-based mortality risk score: associations with all-cause mortality among older adults. *Epigenetics* 13, 846–857

(2018).

9. Vaiserman, A. & Krasnienkov, D. Telomere Length as a Marker of Biological Age: State-of-the-Art, Open Issues, and Future Perspectives. *Front Genet* 11, 630186, doi:10.3389/fgene.2020.630186 (2020).
10. Hastings, W. J., Shalev, I. & Belsky, D. W. Comparability of biological aging measures in the National Health and Nutrition Examination Study, 1999–2002. *Psychoneuroendocrinology* 106, 171–178, doi:10.1016/j.psyneuen.2019.03.012 (2019).
11. Zhang, J. J. *et al.* Changes in accelerated aging and risk of cardiovascular disease and mortality: three cohort studies. *BMC Med* 23, 533, doi:10.1186/s12916-025-04365-8 (2025).
12. Mak, J. K. *et al.* Clinical biomarker-based biological aging and risk of cancer in the UK Biobank. *British Journal of Cancer* 129, 94–103 (2023).
13. Mak, J. K., McMurran, C. E. & Hägg, S. Clinical biomarker-based biological ageing and future risk of neurological disorders in the UK Biobank. *Journal of Neurology, Neurosurgery & Psychiatry* 95, 481–484 (2024).
14. Kojima, G., Iliffe, S. & Walters, K. Frailty index as a predictor of mortality: a systematic review and meta-analysis. *Age and Ageing* 47, 193–200, doi:10.1093/ageing/afx162 (2018).
15. Wang, Q., Zhan, Y., Pedersen, N. L., Fang, F. & Hägg, S. Telomere Length and All-Cause Mortality: A Meta-analysis. *Ageing Research Reviews* 48, 11–20, doi:https://doi.org/10.1016/j.arr.2018.09.002 (2018).
16. Scarpa, J. R., Reid, M. C., Elemento, O. & Krumsiek, J. Biological aging increases risk of postoperative morbidity and mortality: an international, multi-cohort study. *npj Aging*, doi:10.1038/s41514-026-00441-2 (2026).
17. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *American Journal of Epidemiology* 186, 1026–1034, doi:10.1093/aje/kwx246 (2017).
18. Hao, M. *et al.* Gompertz Law-Based Biological Age (GOLD BioAge): A Simple and Practical Measurement of Biological Ageing to Capture Morbidity and Mortality Risks. *Advanced Science* 12, e01765, doi:https://doi.org/10.1002/advs.202501765 (2025).
19. Magnusson, P. K. E. *et al.* The Swedish Twin Registry: Establishment of a Biobank and Other Recent Developments. *Twin Research and Human Genetics* 16, 317–329, doi:10.1017/thg.2012.104 (2013).
20. Zagai, U., Lichtenstein, P., Pedersen, N. L. & Magnusson, P. K. E. The Swedish Twin Registry: Content and Management as a Research Infrastructure. *Twin Research and Human Genetics* 22, 672–680, doi:10.1017/thg.2019.99 (2019).
21. Chen, X. *et al.* Genetic and Environmental Contributions to the Covariation Between Cardiometabolic Traits. *Journal of the American Heart Association* 7, e007806, doi:10.1161/JAHA.117.007806.
22. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Medicine* 12, e1001779, doi:10.1371/journal.pmed.1001779 (2015).

23. Li, X. *et al.* Clinical biomarkers and associations with healthspan and lifespan: Evidence from observational and genetic data. *EBioMedicine* 66, 103318, doi:10.1016/j.ebiom.2021.103318 (2021).
24. Li, X. *et al.* The frailty index is a predictor of cause-specific mortality independent of familial effects from midlife onwards: a large cohort study. *BMC Medicine* 17, 94, doi:10.1186/s12916-019-1331-8 (2019).
25. Williams, D. M., Jylhävä, J., Pedersen, N. L. & Hägg, S. A Frailty Index for UK Biobank Participants. *The Journals of Gerontology: Series A* 74, 582–587, doi:10.1093/gerona/gly094 (2019).
26. Codd, V. *et al.* Measurement and initial characterization of leukocyte telomere length in 474,074 participants in UK Biobank. *Nature Aging* 2, 170–179, doi:10.1038/s43587-021-00166-9 (2022).
27. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)* 57, 289–300, doi:https://doi.org/10.1111/j.2517-6161.1995.tb02031.x (1995).
28. Therneau, T. A package for survival analysis in R. *R package version* 2, 2014 (2015).
29. Blanche, P., Dartigues, J.-F. & Jacqmin-Gadda, H. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Statistics in Medicine* 32, 5381–5397, doi:https://doi.org/10.1002/sim.5958 (2013).
30. Gerds, T. A. & Schumacher, M. Consistent Estimation of the Expected Brier Score in General Survival Models with Right-Censored Event Times. *Biometrical Journal* 48, 1029–1040, doi:https://doi.org/10.1002/bimj.200610301 (2006).
31. Vickers, A. J. & Elkin, E. B. Decision curve analysis: a novel method for evaluating prediction models. *Medical Decision Making* 26, 565–574 (2006).
32. SCHOENFELD, D. Partial residuals for the proportional hazards regression model. *Biometrika* 69, 239–241, doi:10.1093/biomet/69.1.239 (1982).
33. Therneau, T. M. & Lumley, T. Package 'survival'. *R Top Doc* 128, 28–33 (2015).
34. Ozenne, B., Sørensen, A. L., Scheike, T., Torp-Pedersen, C. & Gerds, T. A. riskRegression: predicting the risk of an event using Cox regression models. (2017).
35. Wickham, H. & Sievert, C. *ggplot2: elegant graphics for data analysis*. Vol. 10 (springer New York, 2009).

Figures

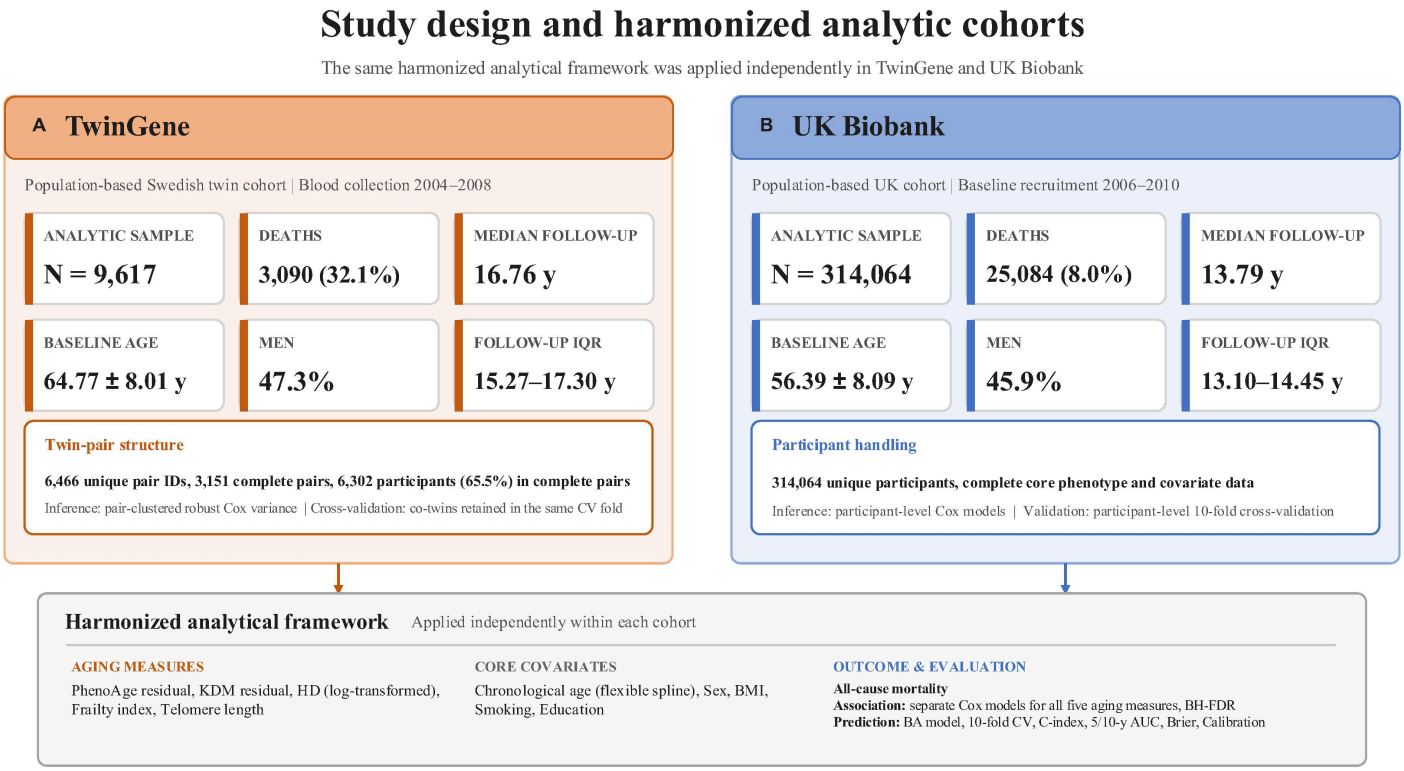


Figure 1

Study design and harmonized analytic cohorts in TwinGene and UK Biobank.

The final analytic samples included 9,617 participants in TwinGene and 314,064 participants in UKB, with 3,090 and 25,084 deaths, respectively. TwinGene comprised 6,466 unique twin-pair identifiers, including 3,151 complete pairs corresponding to 6,302 participants (65.5% of the analytic sample). A harmonized analytical framework was applied independently within each cohort. Biological aging was characterized using PhenoAge, KDM, homeostatic dysregulation (HD), frailty index (FI), and telomere length (TL). All five measures were evaluated separately in mortality-association analyses. The biological aging-enhanced prediction model included PhenoAge residual, FI, and TL. Primary mortality models adjusted for chronological age, sex, BMI, smoking, and education. TwinGene survival models used pair-clustered robust variance estimation, and co-twins were retained within the same cross-validation fold.

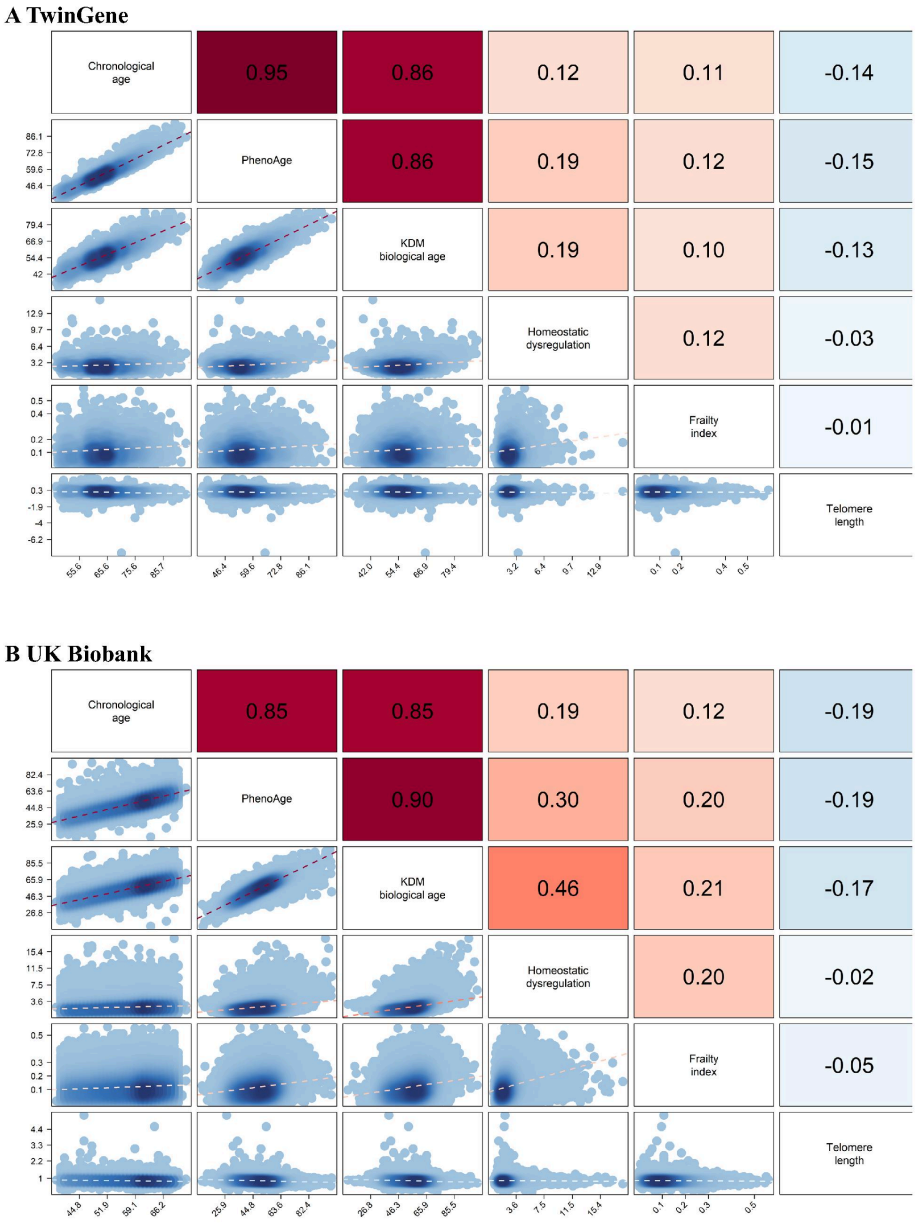


Figure 2

Correlations among biological aging measures in TwinGene and UK Biobank. Pairwise relationships among chronological age, raw PhenoAge, raw KDM, raw HD, FI, and adjusted TL are shown for (A) TwinGene (n= 9,617) and (B) UK Biobank (n= 314,064). The upper triangle displays Pearson correlation coefficients (*r*), while the lower triangle shows pairwise distributions and associations. PhenoAge, KDM, and HD values were used for this descriptive analysis; age-adjusted PhenoAge and KDM residuals and

log-transformed HD were used in subsequent mortality analyses. Correlation colors indicate direction and magnitude, with the same scale applied to both cohorts. Biomarker-derived measures are shown on cohort-specific scales, and absolute values should therefore not be interpreted as directly comparable between cohorts.

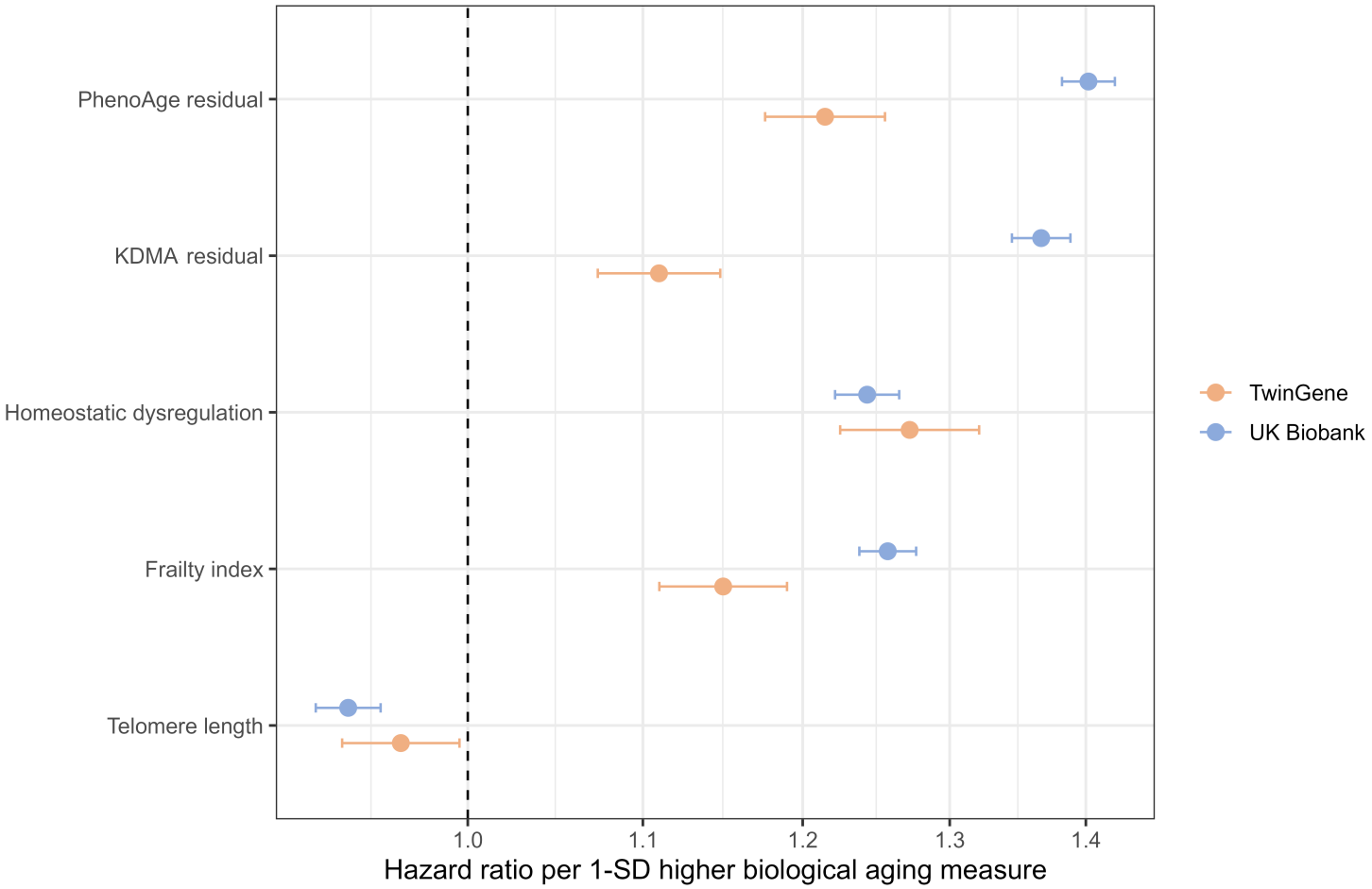


Figure 3

Adjusted associations of biological aging measures with all-cause mortality in TwinGene and UK Biobank.

Forest plot showing hazard ratios (HRs) and 95% confidence intervals for all-cause mortality per 1-SD higher biological aging measure. PhenoAge and KDM were residualized on chronological age using natural splines with three degrees of freedom, HD was log-transformed, and all five aging measures were standardized within cohort. Models were adjusted for chronological age using a four-degree-of-freedom natural spline, sex, BMI using a three-degree-of-freedom natural spline, smoking, and education. TwinGene models additionally used pair-clustered robust variance estimation. The dashed vertical line indicates HR= 1. FDR-adjusted *P*-values for the five biological aging measures within each cohort are reported in Table 2.

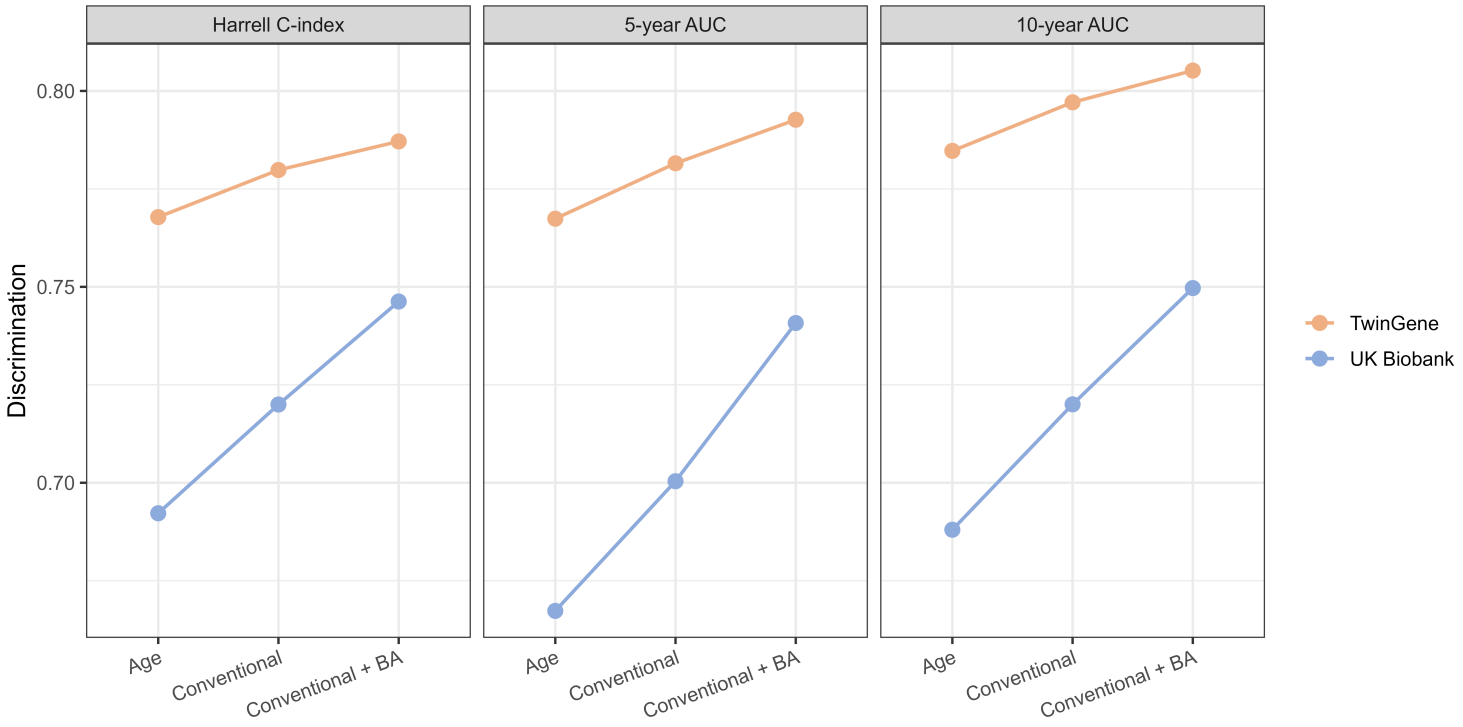


Figure 4

Cross-validated discrimination for all-cause mortality prediction in TwinGene and UK Biobank.

Discrimination is shown for three Cox prediction models: chronological age alone, chronological age plus conventional risk factors, and the conventional model plus biological aging measures. Conventional risk factors included sex, BMI, smoking, and education; biological aging measures included PhenoAge residual, frailty index, and telomere length. Performance was evaluated from 10-fold out-of-fold predictions using Harrell's C-index and censoring-aware time-dependent AUC at 5 and 10 years. PhenoAge residualization and standardization of PhenoAge residual, FI, and TL were performed within each training fold before application to held-out participants. In TwinGene, co-twins were assigned to the same fold. Higher Harrell's C-index and AUC indicate better discrimination.

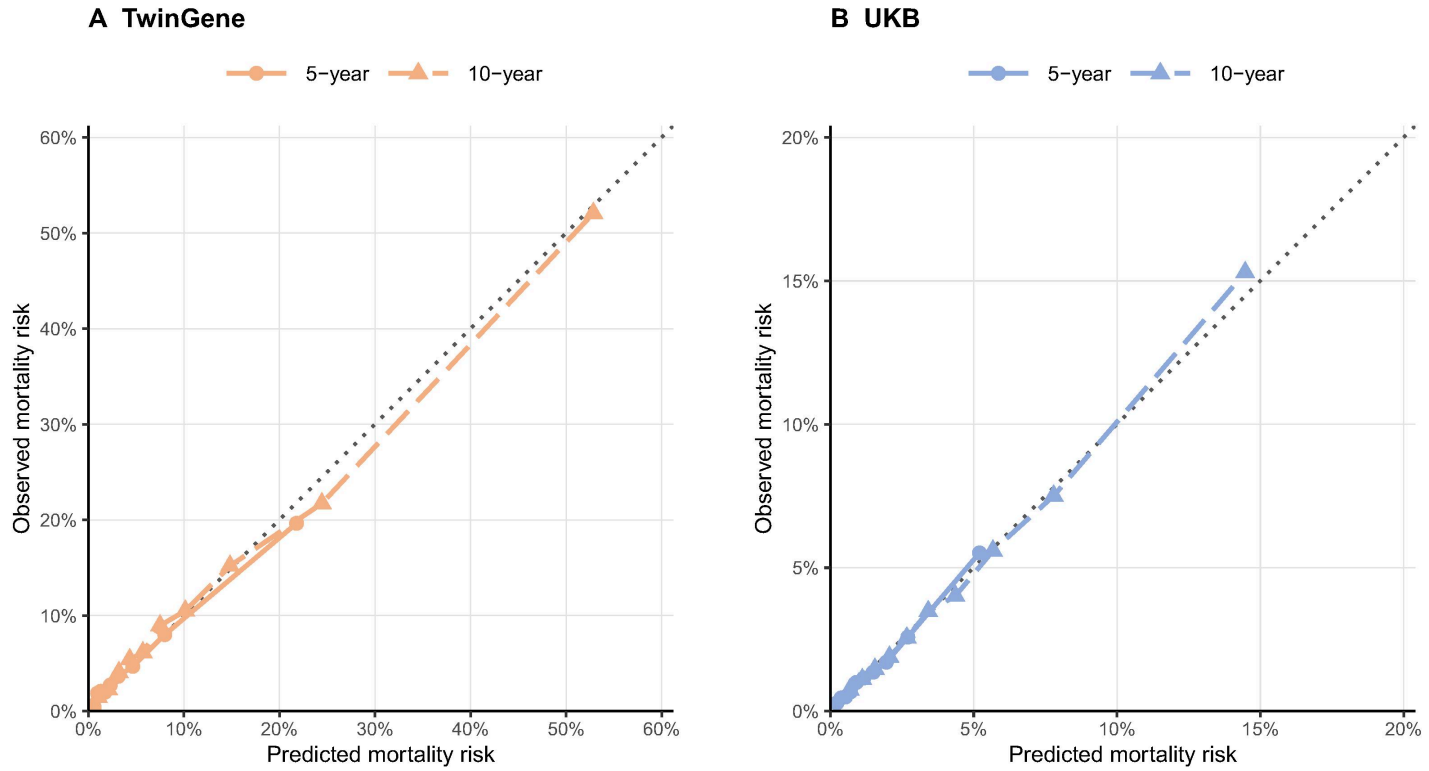


Figure 5

Calibration of the biological aging-enhanced mortality prediction model in TwinGene and UK Biobank.

Calibration at 5 and 10 years is shown in (A) TwinGene (orange) and (B) UKB (blue). Circles connected by solid lines represent 5-year predictions; triangles connected by dashed lines represent 10-year predictions. Predicted mortality risks were obtained from 10-fold out-of-fold predictions and compared with censoring-adjusted observed risks across deciles of predicted risk. The conventional model included chronological age, sex, BMI, smoking, and education; the biological aging-enhanced model additionally included PhenoAge residual, frailty index, and telomere length. PhenoAge residualization and standardization of biological aging measures were performed within each training fold before application to held-out participants. In TwinGene, co-twins were retained within the same cross-validation fold. The dotted diagonal denotes perfect calibration. Axis ranges differ between cohorts, with identical x- and y-axis limits within each panel.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.zip
- Supplementary.docx