MotionAge: A Deep Learning Framework for Biological Age Prediction from Wearable Activity

Yilin Song* YS4057@CUMC.COLUMBIA.EDU
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, USA

Jiaqi Yin* JACKIEYIN@OPENAI.COM
OpenAI, San Francisco, CA, USA

Yirou Hu YH3964@CUMC.COLUMBIA.EDU
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, USA

Tian Gu TG2880@CUMC.COLUMBIA.EDU
Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, USA

*These authors contributed equally.

Abstract

Population aging is a major public health challenge, driving increasing burden from chronic disease and mortality. To better characterize individual health beyond chronological age, there is growing interest in biological age as a quantitative measure for risk stratification and longitudinal monitoring. Existing biological age models largely rely on laboratory or structured clinical data, limiting scalability and their ability to capture dynamic aspects of health. We propose *MotionAge*, a deep learning framework that learns a mortality-calibrated biological age directly from high-frequency wearable activity data. The framework combines deep sequence models with a wear-aware modeling strategy that explicitly represents observation reliability, allowing the model to distinguish device non-wear from observed inactivity in noisy and irregularly observed time series. Unlike prior wearable-based methods that predict chronological age or construct unsupervised biomarkers, MotionAge is trained to predict 5-year mortality risk and subsequently maps that risk onto an interpretable age scale. In NHANES accelerometry data, MotionAge achieved stronger mortality discrimination over chronological age and established benchmarks. Within fixed chronological-age bands, higher MotionAge acceleration was associated with lower activity and worse survival, with the largest separation observed among older adults. These results demonstrate that reliability-aware deep learning of wearable time series can yield scalable and interpretable aging phenotypes aligned with clinically relevant health outcomes.

Keywords: Biological age, deep sequence models, wearable data, wear-aware modeling

1. Introduction

Population aging is rapidly reshaping the demographic structure of societies worldwide. According to the World Health Organization, the global population aged 60 years and older is expected to nearly double from about one billion in 2020 to over two billion by 2050, with older adults comprising an increasing proportion of the total population in many countries

Jiaqi Yin contributed to this work in a personal capacity. Views are the authors' own and do not necessarily represent OpenAI.

(World Health Organization, 2021). As life expectancy rises, however, longevity is often accompanied by a growing burden of chronic diseases. Conditions such as cardiovascular disease, diabetes, and neurodegenerative disorders account for a substantial proportion of morbidity and mortality in aging populations, placing increasing strain on healthcare systems (Lopez et al., 2006; GBD 2019 Risk Factors Collaborators, 2020). While chronological age is commonly used as a proxy for aging, it does not adequately capture heterogeneity in health status across individuals (Lichtenstein et al., 2000; Mitnitski et al., 2002). People of the same chronological age can differ substantially in physiological function, disease risk, and resilience to stressors. As a result, there is growing interest in biological age, a quantitative measure intended to reflect an individual's underlying physiological state relative to their chronological age.

A reliable estimate of biological age has important clinical implications. It can enable earlier identification of individuals at elevated risk for adverse outcomes, support personalized prevention and intervention strategies, and provide a meaningful endpoint for monitoring disease progression or treatment response. A substantial body of work has focused on estimating biological age using clinical and molecular data. Early approaches, including epigenetic clocks, were designed to predict chronological age with deviations interpreted as indicators of accelerated or decelerated aging (Horvath, 2013; Hannum et al., 2013). Subsequent work has shifted toward improving clinical relevance by incorporating mortality and morbidity risk into the target, leading to second-generation aging measures such as BioAge, PhenoAge and GrimAge (Levine, 2013; Levine et al., 2018; Lu et al., 2019). While these methods provide biologically meaningful and clinically predictive measures, they typically rely on laboratory assays, genetic profiling, or curated clinical variables that are costly, invasive, and collected infrequently. As a result, they are limited in their ability to capture dynamic changes in health status over time or to support continuous monitoring in real-world settings.

In contrast, the widespread adoption of wearable devices offers a scalable and non-invasive source of continuously collected behavioral and physiological data. Devices such as Fitbit, Apple Watch and Oura Ring capture detailed information on daily activity patterns, including step counts, activity intensity, heart rate, and sleep characteristics. These signals reflect the integrated output of multiple physiological systems, including cardiovascular fitness, metabolic efficiency, and circadian regulation, and therefore provide a proxy for overall health status that is difficult to obtain from sparse clinical measurements. Recent studies have shown that wearable-derived activity features are strongly associated with health status, functional decline, and risk of chronic disease (Phillips et al., 2018; Zhou et al., 2022; Golbus et al., 2023; Chen et al., 2023; Nagata et al., 2024). However, leveraging wearable data for biological age estimation presents substantial methodological challenges. Wearable data are inherently high-dimensional and exhibit complex temporal structure, with observations collected at fine time scales over extended periods. At the same time, these data are noisy and subject to irregular observation patterns due to device non-wear, varying user adherence, and measurement error. Simple aggregation of these signals into summary statistics can obscure important temporal dynamics and introduce bias, particularly when missingness is informative. Capturing meaningful aging-related patterns therefore requires models that can learn from high-frequency time series while accounting for data quality and temporal dependencies.

Advances in machine learning, particularly deep learning methods for sequential data, provide a natural framework for addressing these challenges. Models such as long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), gated recurrent units (GRU) (Cho et al., 2014), and transformer-based architectures (Vaswani et al., 2017) are well-suited for modeling complex temporal processes due to their ability to capture nonlinear relationships and long-range dependencies. These approaches enable the extraction of informative representations from raw wearable data, potentially uncovering latent patterns associated with physiological aging that are not accessible through handcrafted features or static summaries.

Despite these advances, existing work has not fully explored the use of high-frequency wearable time series to directly estimate biological age using clinically meaningful outcomes such as mortality rather than chronological age as the learning target. In this work, we propose a deep learning framework that leverages ActiGraph-derived wearable data to estimate biological age from high-resolution activity time series (Figure 1). By explicitly modeling data reliability alongside temporal dynamics, the proposed method extracts robust representations of behavioral and physiological patterns associated with aging. The primary contribution of this work is a framework for constructing and validating a mortality-calibrated biological age measure from wearable activity data, combining wear-aware preprocessing with deep sequence modeling. Our contributions are threefold:

1. We develop a wearable-based biological age model that directly targets mortality risk which is clinically meaningful, enabling improved alignment between predicted age and health risk.

2. We introduce a hierarchical, wear-aware aggregation strategy to address non-wear and irregular observation, improving robustness and interpretability of learned representations.

3. We demonstrate that the proposed approach achieves strong predictive performance for mortality over established benchmarks and that deviations between predicted biological age and chronological age are associated with clinically meaningful indicators of health status.

Generalizable Insights about Machine Learning in the Context of Healthcare. This work highlights several general principles for applying machine learning to real-world healthcare data. First, in high-frequency observational data such as wearable signals, missingness and data reliability are not nuisances but informative aspects of the data-generating process; explicitly modeling non-wear and observation quality can materially improve both interpretability and predictive performance. Second, aligning model training with clinically meaningful outcomes, rather than surrogate targets such as chronological age, leads to representations that better reflect actionable health risk. Third, temporally rich behavioral data can provide complementary information beyond static clinical or laboratory features when appropriate sequence models are used to capture structure over time. Finally, translating model outputs into clinically interpretable quantities facilitates integration into downstream decision-making and evaluation frameworks. Together, these observations suggest a general design pattern for healthcare machine learning: jointly model temporal

dynamics and data reliability, optimize for clinically relevant endpoints, and calibrate learned representations into interpretable phenotypes.

2. Related Work

2.1. Biological Age Estimation

Biological age modeling has evolved from early approaches based on predicting chronological age (Horvath, 2013; Hannum et al., 2013) to more recent methods that directly target clinically meaningful outcomes, such as mortality, including BioAge (Levine, 2013), PhenoAge (Levine et al., 2018), and GrimAge (Lu et al., 2019). By incorporating clinical outcome-based targets, these models demonstrate stronger associations with survival and disease risk and offer improved clinical interpretability. More recent works have expanded biological age modeling to organ-specific aging clocks, aiming to capture heterogeneous aging processes across physiological systems (Qiu et al., 2023; Wang et al., 2026). Additionally, advances in machine learning have enabled the use of large-scale clinical data, including multimodal electronic health records and large language model (LLM)-based representations, to construct aging-related measures (Li et al., 2025). Despite these advances, most existing biological age estimators rely on laboratory assays or structured clinical summaries, which are costly, sparsely observed, and provide a largely static view of health.

Our work is closest in spirit to second-generation clocks in that it uses mortality risk as an intermediate target, but differs fundamentally in its reliance on passively collected, high-frequency wearable behavior rather than clinical or laboratory features.

2.2. Deep Learning for Wearable and Health Time Series

Deep learning methods have been widely applied to sequential health data, including wearable device signals and electronic health record time series. Architectures such as long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997), gated recurrent units (GRU) (Cho et al., 2014), and transformer-based models (Vaswani et al., 2017) are well-suited for capturing nonlinear relationships and long-range temporal dependencies in high-dimensional time series. These approaches have achieved strong performance in tasks such as activity recognition, disease prediction, and risk modeling (Zhang et al., 2022).

In the context of wearable data, prior studies have shown that continuously collected behavioral and physiological signals are associated with health status, aging processes, and disease risk. Recent work has further demonstrated that wearable signals can be used to construct aging-related measures, for example by predicting chronological age from photoplethysmography (PPG) or activity data (McIntyre et al., 2021; Miller et al., 2025; Nie et al., 2025). Other approaches have proposed wearable-derived digital biomarkers of aging and longevity based on activity patterns or circadian rhythms, linking these measures to inflammation, biological age, and mortality risk (Shim et al., 2023). While these approaches leverage high-frequency physiological data, they are typically trained to predict chronological age or construct biomarkers whose biological relevance is inferred through associations with health outcomes. Consequently, they optimize fundamentally different objectives from our mortality-calibrated biological age model. In particular, Shim et al. (2023) develop wearable-derived biomarkers based on activity patterns rather than estimating biological

age, whereas Miller et al. (2025) predict chronological age from raw wearable PPG signals rather than processed activity intensity data with a mortality-calibrated target. For this reason, we do not include these methods as numerical baselines, as direct comparisons would not constitute a like-for-like evaluation.

Positioning of this work. In contrast to prior work, we propose a framework that directly estimates biological age from high-frequency wearable time series incorporating mortality risk. Our approach builds on advances in deep sequence modeling while explicitly addressing challenges unique to wearable data, including irregular observation and non-wear. In particular, wearable activity data are only partially observed, and periods of non-wear may resemble true inactivity. We incorporate wear-time information directly into the modeling pipeline as a reliability signal through a hierarchical masking and aggregation strategy. By treating observation quality as part of the representation, our method jointly models temporal dynamics and data reliability, enabling the construction of a robust and clinically interpretable biological age measure.

3. Methods

3.1. Overview of the Proposed MotionAge

As shown in Figure 1, we first used a deep sequence model to predict each participant's 5-year mortality risk from preprocessed wear-aware activity sequences and optional static covariates. The predicted risk is then mapped onto an age scale using a sex-specific inverse calibration, yielding a mortality-calibrated estimate of biological age. We refer to this estimate as *MotionAge*, and define age acceleration as the difference between MotionAge and chronological age.

3.2. Wear-Aware Preprocessing and Sequence Construction

Our preprocessing pipeline preserves temporal structure while explicitly representing observation reliability. This is important because low activity may reflect either true inactivity or device non-wear. Conflating the two can bias activity summaries and cause the model to learn adherence patterns rather than behavior. We therefore use wear status to identify observed periods and quantify epoch-level reliability.

To achieve this, we first identified non-wear at the minute level using the Choi algorithm (Choi et al., 2011). Specifically, candidate non-wear intervals were defined as contiguous segments lasting at least 90 minutes, allowing at most two non-consecutive spike minutes with intensity in $(0, 100]$, and requiring at least 30 consecutive zero-count minutes on both sides of each spike. Minutes within such intervals were labeled as non-wear ($w_t = 0$), and all remaining minutes were labeled as wear ($w_t = 1$), yielding a binary *wearing masking* for each minute.

We then aggregated the minute-level series into 5-minute epochs. For each epoch e, we computed a wear-aware activity summary

$$x_e = \frac{\sum_{t \in e} w_t \, d_t}{\sum_{t \in e} w_t}, \tag{1}$$

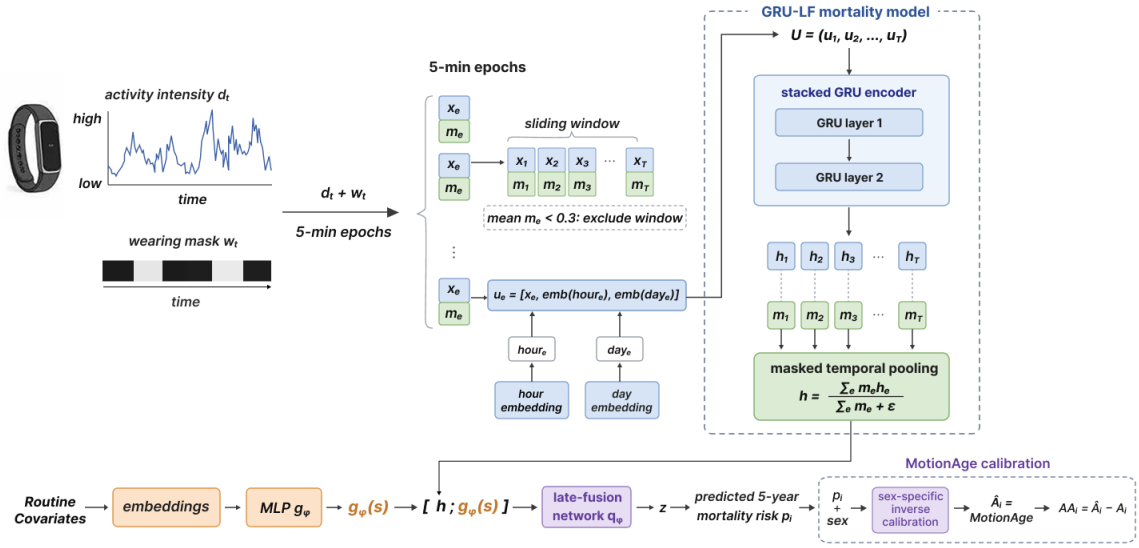


Figure 1: Overview of the proposed MotionAge. Wear-aware activity sequences and optional static baseline covariates are encoded separately, fused to predict 5-year mortality risk, and then mapped onto an age scale to yield a mortality-calibrated biological age estimate.

i.e., the average activity intensity over wear minutes within epoch e, computed when at least one wear minute is present, with w_t restricting the calculation to observed wearing periods and d_t denoting the minute-level intensity. We also retained the hour-of-day and day-of-week indices for each epoch. Let r_e denote the fraction of wear minutes in epoch e and m_e the corresponding binary epoch-validity indicator,

$$r_e = \frac{1}{|e|} \sum_{t \in e} w_t, \qquad m_e = \mathbb{I}\{r_e \geq \tau\},$$

with threshold $\tau = 0.2$.

Participant-level sequences were constructed by sliding a fixed-length window over the ordered epoch sequence; the primary analysis used $T = 1008$ consecutive 5-minute epochs. Windows were excluded when the mean of the binary indicators m_e fell below 0.3. Within retained windows, all temporal positions, including those with $m_e = 0$, remained in the encoder input; at masked pooling, hidden states at positions with $m_e = 0$ received zero direct weight. This preserves the regular temporal grid while distinguishing sparsely observed epochs from observed inactivity. Figure 2 shows the preprocessing steps described, and additional construction details and sensitivity analyses are provided in Appendix A2.

3.3. Sequence Model and Training Objective

For each retained window, the wearable input at epoch e consisted of activity intensity together with temporal context features, including hour-of-day and day-of-week indicators.

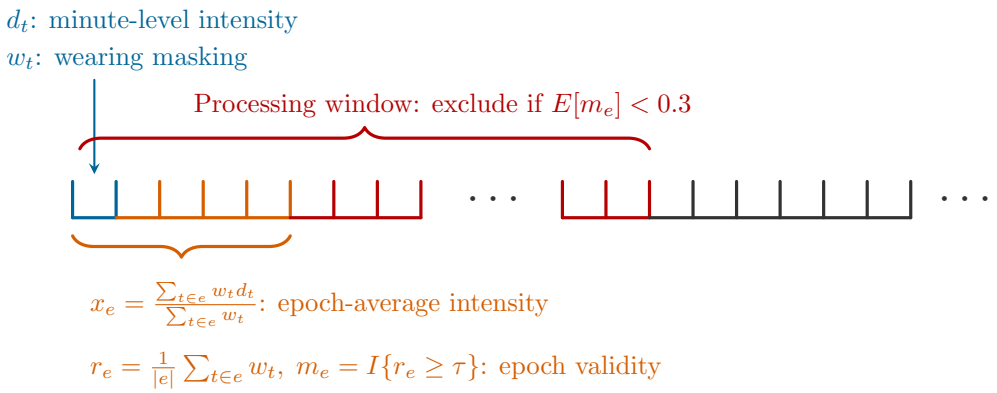


Figure 2: Minute-level activity intensity d_t and wear indicator w_t are aggregated into fixed-length epochs during preprocessing. For each epoch e, we compute a wear-aware average intensity x_e, wear fraction r_e, and binary validity indicator m_e. A whole window is excluded if its mean m_e is below 0.3; within retained windows, all epoch positions remain in the encoder input and m_e determines their direct weight in the masked mean.

These features represented both instantaneous activity and structured temporal variation (e.g., daily and weekly patterns). Given a sequence of length T, a sequence encoder, either GRU or Transformer, produced epoch-level hidden representations $\mathbf{h}_1, \ldots, \mathbf{h}_T$. To obtain a single representation for each window, we aggregated these hidden states using masked temporal pooling. Unlike ordinary mean pooling, which weights every epoch equally, masked pooling averages only over epochs with sufficient wear:

$$\overline{\mathbf{h}} = \frac{\sum_{e=1}^{T} m_e \mathbf{h}_e}{\sum_{e=1}^{T} m_e + \varepsilon},$$

where m_e is the binary epoch-validity indicator and $\varepsilon > 0$ is a small constant added for numerical stability. This yielded a reliability-weighted sequence summary: epochs with insufficient wear ($m_e = 0$) therefore received zero direct weight in the window-level representation, whereas valid epochs ($m_e = 1$) contributed equally. Because all temporal positions remained in the encoder input, invalid epochs could still indirectly influence the contextualized hidden states at valid positions.

When static covariates were available, we modeled them separately from the wearable sequence because the two data types encode different forms of information: the wearable branch captures time-varying behavioral patterns, whereas baseline covariates summarize relatively stable participant characteristics. The static covariates were encoded using a multilayer perceptron $g_\phi(\mathbf{s})$, where $\mathbf{s}$ denotes the vector of participant-level baseline features, producing a complementary representation of participant-level characteristics. In the late-fusion model, this static representation was concatenated with the pooled wearable representation and passed to a prediction network q_ψ:

$$z = q_\psi\big([\overline{\mathbf{h}}; g_\phi(\mathbf{s})]\big).$$

This feature-level fusion allows the prediction network to learn joint interactions between temporal activity patterns and baseline characteristics (Stahlschmidt et al., 2022; Wang et al., 2023).

As an alternative, we considered a residual fusion formulation inspired by residual learning and additive side-network adaptation (He et al., 2016; Zhang et al., 2020). The intuition is to preserve a wearable-based prediction as the primary signal and use static covariates only to refine that prediction. Specifically, we first obtained a wearable-derived $z_{\text{wear}} = q_{\psi_1}(\overline{\mathbf{h}})$, and then modeled an additive correction term based on static features, $z_{\text{static}} = q_{\psi_2}([\overline{\mathbf{h}}; g_\phi(\mathbf{s})])$. The final prediction was $z = z_{\text{wear}} + z_{\text{static}}$. This formulation allowed static covariates to capture incremental information that is not already represented by the wearable-derived prediction.

The final model outputs a scalar logit z, corresponding to the log-odds of 5-year mortality, which was transformed into a probability via the sigmoid transformation $\sigma(z) = \frac{1}{1+e^{-z}}$. Because 5-year mortality was relatively rare, the training data were class-imbalanced. We therefore used a positive-class-weighted binary cross-entropy loss with logits, which assigns greater weight to mortality cases and reduces the dominance of the majority survival class during training:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[\lambda\, y_i \log\sigma(z_i) + (1-y_i)\log\left(1-\sigma(z_i)\right)\right],$$

where $\lambda = N_-/N_+$ denotes the positive-class weight computed from the training split, with N_+ and N_- denoting the numbers of positive and negative training samples, respectively. The weighting increases the contribution of mortality cases during training, while computing the loss directly from logits improves numerical stability. Because each participant contributed multiple windows, the model produced multiple mortality estimates for the same individual. We aggregated the window-level predicted probabilities p_{iw} to obtain a single participant-level mortality risk, $p_i = \frac{1}{n_i}\sum_{w=1}^{n_i} p_{iw}$, where n_i is the number of retained windows for participant i.

We next converted the participant-level mortality risk into an interpretable age scale. The underlying intuition is that two individuals with the same predicted mortality risk should be assigned the same mortality-equivalent age within the same sex-specific reference population. Within outer fold k, calibration was estimated using only the fold-specific training partition $\mathcal{D}^{(k)}_{\text{train}}$ and separately for each sex group s. Chronological age was rounded to integer age bins a, and the representative risk within each sex–age bin was defined as

$$\widetilde{p}^{(k)}_{a,s} = \text{median}\left\{p_i : i \in \mathcal{D}^{(k)}_{\text{train}},\ \text{round}(A_i) = a,\ S_i = s\right\}, \tag{2}$$

where the median provides a robust estimate of the typical model-predicted risk at each age. After clipping these risks to $[\epsilon, 1-\epsilon]$, with $\epsilon = 10^{-4}$, we fitted the sex-specific weighted linear calibration

$$\text{logit}\left[\text{clip}(\widetilde{p}^{(k)}_{a,s}, \epsilon, 1-\epsilon)\right] = \alpha^{(k)}_s + \beta^{(k)}_s a + \varepsilon^{(k)}_{a,s},$$

using the number of training participants in each sex–age bin as the regression weight and requiring $\beta^{(k)}_s > 0$. MotionAge for a held-out participant was obtained by applying the fitted

inverse relationship,

$$\widehat{A}_i = \frac{\text{logit}[\text{clip}(p_i, \epsilon, 1-\epsilon)] - \alpha_{S_i}^{(k)}}{\beta_{S_i}^{(k)}}, \qquad AA_i = \widehat{A}_i - A_i. \tag{3}$$

Thus, $\widehat{A}_i$ is the MotionAge estimate, representing the age at which the fitted sex-specific reference curve yields the same predicted 5-year mortality risk as participant i, and AA_i measures whether that mortality-equivalent age is higher or lower than the participant's chronological age. Held-out participants were never used to estimate calibration parameters, and the primary analysis did not clamp $\widehat{A}_i$ to the training age range. Fold-specific parameters are reported in Appendix A3.1.

3.4. Training Procedure

The deep learning model always included a wearable branch consisting of temporal features derived from minute-level accelerometry after wear-aware preprocessing. Static baseline covariates were optionally incorporated through a separate auxiliary branch. We evaluated three input configurations: **F**, using wearable data alone; **FA**, using wearable data and chronological age; and **FRC**, using wearable data and routine covariates, including age, sex, household income, BMI, waist circumference, and systolic and diastolic blood pressure. These configurations allowed us to assess the predictive contribution of wearable signals alone and their incremental value when combined with standard baseline characteristics.

Before the main cross-validation analysis, we performed a one-time model-development step to select hyperparameters for each encoder family (GRU or Transformer), input configuration, and, where applicable, fusion strategy. Candidate models were tuned on a fixed development split using Optuna and ranked by validation AUPRC. The search covered sequence length and encoder capacity for the wearable branch, and, when static covariates were included, the dimensions and dropout of the covariate and fusion modules together with key optimization hyperparameters such as learning rate and weight decay. Training in this preliminary step used AdamW, class-weighted binary cross-entropy, a ReduceLROnPlateau scheduler, gradient clipping, and early stopping. The resulting hyperparameters were then fixed and reused throughout the subsequent evaluation study.

4. Evaluation Setup

4.1. Dataset

We used data from the 2003–2004 and 2005–2006 cycles of the National Health and Nutrition Examination Survey (NHANES), a nationally representative U.S. survey that combines interview, examination, and follow-up data. Minute-level accelerometer data were obtained from the physical activity monitor files (PAXRAW), which was linked to participant-level demographic and clinical variables using the unique participant identifier (`SEQN`) for training and evaluation. Mortality follow-up was used to define the primary prediction target: death within 5 years of baseline assessment. The analytic cohort ($N = 8{,}979$) included participants with available accelerometry data and mortality follow-up information sufficient to define a 5-year mortality outcome. A descriptive summary of the cohort is provided in Appendix A1.

We included all age groups rather than restricting to older adults. This design choice is motivated by (i) evidence that chronic disease onset is shifting toward younger populations, and (ii) the goal of learning a biological age model that generalizes across the full lifespan. Because short-term mortality is rare in younger adults, our primary discrimination analyses focus on the subgroup aged 40 years or older; results for the full population are provided in Appendix A3.

4.2. Evaluation Procedure

We evaluated model performance using 5-fold cross-validation. In each fold, the deep learning model was trained on the training split and used to derive MotionAge $\hat{A}$ via fold-specific calibration, from which age acceleration was computed as $AA = \hat{A} - A$. To evaluate 5-year mortality discrimination, we fitted separate logistic regression models in each outer-fold training partition using chronological age and sex ($A + S$), estimated biological age and sex ($\widehat{A} + S$), or age acceleration together with chronological age and sex ($AA + A + S$), and applied them to the corresponding held-out test partition. AUROC was calculated from the held-out predicted probabilities, while mortality C-indices were calculated directly from the corresponding age measures. Fold-specific AUROC and C-index estimates were summarized as the mean and sample standard deviation (SD) across the five outer folds in Table 1. Paired benchmark comparisons used a fold-structured, outcome-stratified bootstrap restricted to participants with predictions from both methods; complete procedures and results are reported in Appendix A3.3.

Several sequence-model variants were implemented including GRU- and Transformer-based encoders with late-fusion and residual-fusion strategies for different model inputs. Primary analyses focused on the GRU model with late fusion (GRU-LF), which gave the most consistent results. Full comparisons are reported in Appendix A3.

Benchmark Comparisons: We compared our method against three baseline approaches: (1) chronological age (**Age**); (2) **PhenoAge** (Levine et al., 2018); and (3) **LLMAge** (Li et al., 2025). Chronological age served as a simple reference predictor. PhenoAge was computed following the original formulation, with missing laboratory values imputed using medians estimated from the corresponding outer-fold training partition and applied unchanged to held-out participants. For LLMAge, we followed the prompting procedure described in the original paper to obtain the overall biological age estimate from participant-level clinical features. Missing values were explicitly indicated as "not available" within the prompt, and responses were generated using the `gpt-4o-mini` model. These benchmarks allowed us to situate the proposed method relative to both traditional biological age measures and recent large language model-based approaches.

5. Results

5.1. How Does MotionAge Compare with Chronological Age and Benchmark Aging Clocks?

Table 1 reports the main five-fold cross-validation result for 5-year mortality prediction in participants aged 40 years or older. MotionAge-FRC achieved the strongest mortality

Table 1: Five-fold cross-validation summary for 5-year mortality prediction in participants aged 40 years or older. MotionAge was obtained using GRU-based models with late fusion. Values are mean ± sample SD across the five outer folds. Boldface denotes the largest fold-mean point estimate in a column and does not by itself imply statistical separation. *Abbreviations: A = chronological age; S = sex; $\hat{A}$ = estimated biological age (MotionAge, PhenoAge, or LLMAge as applicable); $AA = \hat{A} - A$; F = wearable/activity only; FA = wearable data plus chronological age; FRC = wearable data plus routine covariates; C-index = concordance index; AUROC = area under the receiver operating characteristic curve.*

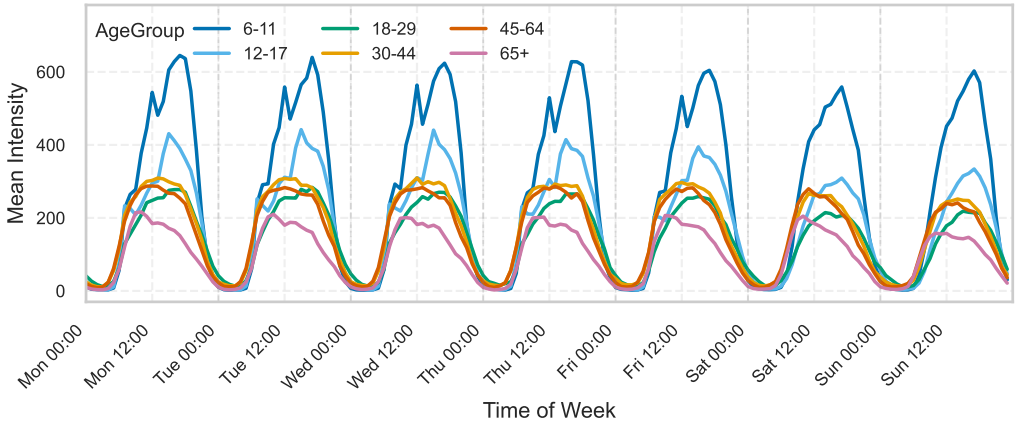
Model	Mortality C-index			Mortality AUROC		
	$C(M, A)$	$C(M, \hat{A})$	$C(M, AA)$	A+S	$\hat{A}$+S	AA+A+S
Age	**0.795 ± 0.019**	–	–	**0.803 ± 0.021**	–	–
LLMAge	–	0.797 ± 0.017	0.516 ± 0.015	–	0.806 ± 0.021	0.806 ± 0.020
PhenoAge	–	0.832 ± 0.020	0.683 ± 0.019	–	0.836 ± 0.021	0.836 ± 0.021
MotionAge-F	–	0.797 ± 0.031	**0.752 ± 0.037**	–	0.809 ± 0.034	0.836 ± 0.025
MotionAge-FA	–	0.823 ± 0.008	0.673 ± 0.089	–	0.829 ± 0.011	0.832 ± 0.012
MotionAge-FRC	–	**0.848 ± 0.013**	0.740 ± 0.023	–	**0.852 ± 0.015**	**0.850 ± 0.015**

discrimination among the evaluated approaches. In the age 40+ subgroup, MotionAge-FRC achieved the highest mortality C-index when using estimated biological age as the predictor (0.848) and the best AUROC under both evaluation settings that used the learned age representation (0.852 for $\hat{A} + S$ and 0.850 for $AA + A + S$). These values exceeded the chronological-age baseline (AUROC 0.803), LLMAge (AUROC 0.806), and PhenoAge (AUROC 0.836). On shared held-out participants, fold-structured paired analyses also favored MotionAge-FRC over Age, LLMAge, and PhenoAge (Appendix A3.3). In the stricter shared complete-case comparison with PhenoAge, however, the margin attenuated and the paired 95% confidence interval included zero (Appendix A3.4). The wearable-only MotionAge-F model yielded a fold-mean $\widehat{A} + S$ AUROC of 0.809; its small paired differences from Age and LLMAge had 95% confidence intervals that included zero (Appendix A3.3). In Appendix A3, GRU-based models consistently matched or exceeded Transformer-based alternatives, and the strongest overall-population performance was also obtained by a GRU late-fusion model.

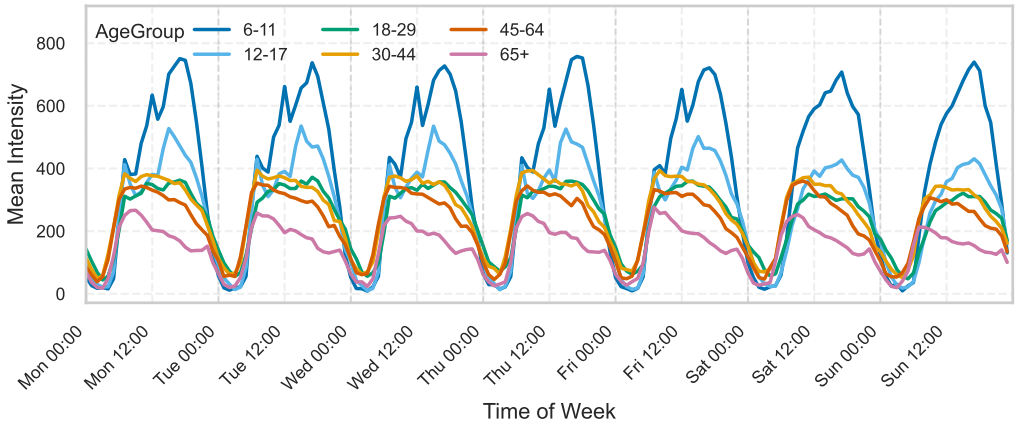
These results suggest that high-frequency behavioral signals contain prognostic information that is not captured fully by age alone or by static clinical summaries. The gains are especially notable because MotionAge-FRC uses a comparatively modest set of routine covariates rather than the richer laboratory feature sets used by PhenoAge or the broader structured-clinical inputs used by LLMAge. The comparison therefore supports the value of wear-aware temporal modeling as an additive source of clinically relevant information.

5.2. Does Wearing Coverage Recover Biologically Plausible Activity Patterns?

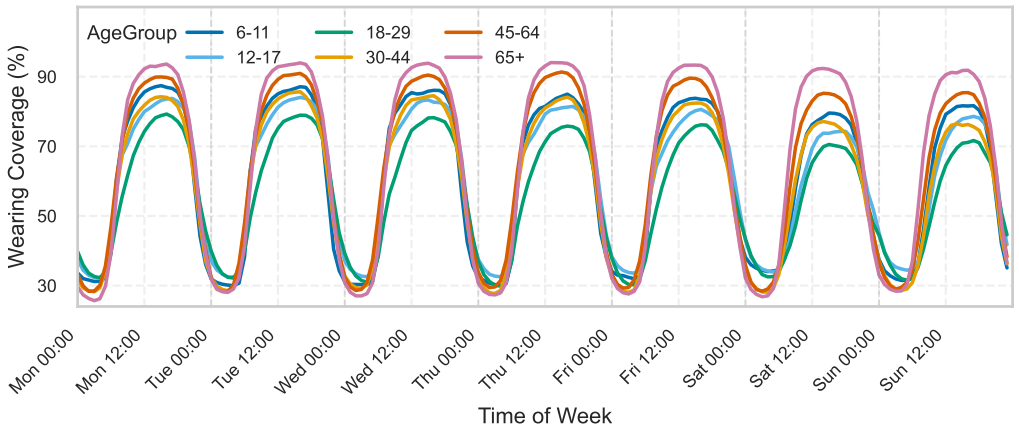
To assess whether explicit treatment of non-wear changes the learned input signal, we examined one-week activity trajectories aggregated from 5-minute epochs into hourly bins



(*a*) Weekly hourly mean activity intensity by age group before applying the wearing coverage.



(*b*) Weekly hourly mean activity intensity by age group after restricting to epochs retained by the wearing coverage.



(*c*) Weekly hourly wearing coverage by age group.

Figure 3: Wear-aware hourly activity diagnostics by age group. Panel (a) shows unmasked profiles; Panel (b) shows corrected separation; Panel (c) shows wearing coverage.

and stratified by age group. We compared the hourly mean intensity profile before and after applying the wearing coverage and also visualized hourly wearing coverage itself (Figure 3).

Before masking, the trajectories exhibited counterintuitive age ordering: the 18–29 group had lower overall mean intensity than the 45–64 group (143 vs. 156 intensity units) and fell below it in 87 of 168 weekly hour bins. The wearing coverage profiles suggested that this distortion might be driven by systematic non-wear rather than true behavioral differences. Wearing coverage varied strongly over time, dropping to 28.9%–35.1% overnight and rising to 72.2%–91.7% during the day, and also varied across age groups, from 56.1% in ages 18–29 to 65.3% in ages 65+. After applying the wearing coverage, the activity profiles became more interpretable: the age-order reversal disappeared (255 vs. 245 mean intensity in ages 18–29 and 45–64), the contrast between youngest and oldest groups widened (175 to 299 intensity units), and the temporal profiles became sharper, with younger groups peaking later in the day and older groups peaking earlier.

These findings show that wearing coverage is not just a cosmetic preprocessing step. In this dataset, it was necessary to recover biologically plausible differences in both the magnitude and timing of activity across age groups. This supports the broader claim that missingness mechanisms in wearable data must be modeled explicitly if downstream predictions are to be interpretable.

Preprocessing diagnostics across nearby non-wear durations, epoch thresholds, coverage cutoffs, and sequence lengths, together with an end-to-end sensitivity analysis for the window-coverage cutoff, are reported in Appendix A2.1. Under the stricter 0.4 cutoff, fold-mean AUROCs were 0.846 for $\widehat{A} + S$ and 0.844 for $AA + A + S$, compared with 0.852 and 0.850 under the primary 0.3 cutoff. Because the cutoff changed both model fitting and participant eligibility, this comparison is treated as a descriptive pipeline-level sensitivity analysis rather than evidence of equivalence.

5.3. Does MotionAge Capture Clinically Meaningful Heterogeneity?

To examine whether MotionAge reflected more than a re-expression of chronological age, we stratified participants within fixed chronological-age bands according to MotionAge acceleration and examined two downstream summaries: weekly activity profiles (Figure 4) and 5-year survival (Figure 5).

Within strata of chronological age, MotionAge-based groupings exhibited consistent and monotonic differences in both behavioral and survival outcomes. Specifically, individuals classified as biologically "older" relative to their age-matched peers demonstrated systematically lower activity intensity profiles across the weekly cycle compared to those classified as "younger," with the magnitude of separation increasing in older age bands. This divergence was modest in younger adults but became substantial in participants aged 45 and above, where biologically older individuals exhibited both reduced peak activity and attenuated daily and weekly variation. Among participants aged 45–64, the high-acceleration group had 38.2% lower mean activity and 22.9% lower weekly amplitude than the low-acceleration group; the corresponding differences among participants aged 65 years or older were 67.2% and 41.0% (Appendix A4.1). Because these same activity sequences contribute to MotionAge itself, the contrasts characterize the expression of the learned phenotype rather than providing independent validation or causal feature effects.

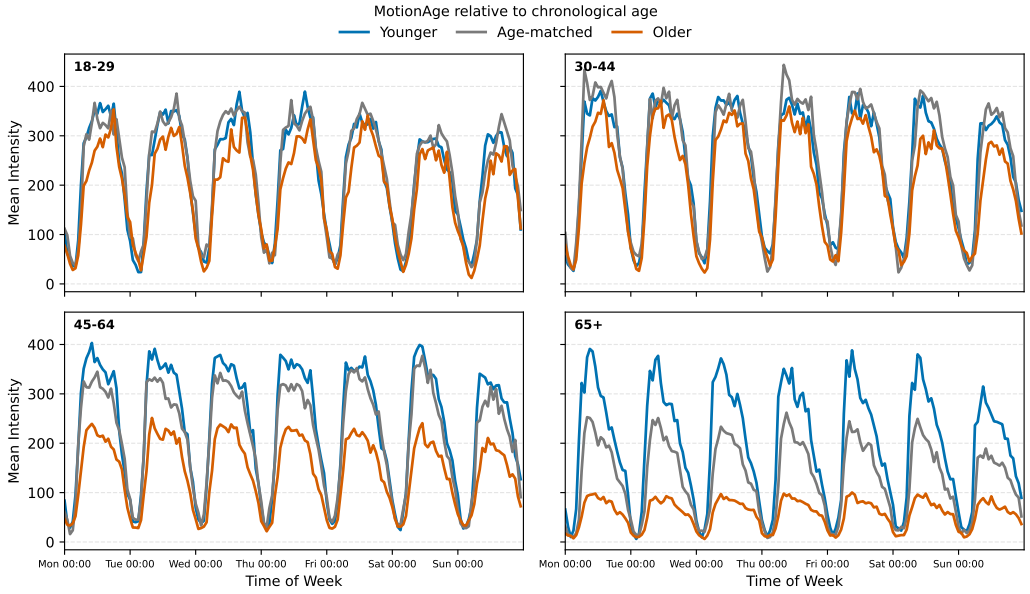


Figure 4: Weekly mean activity intensity within fixed chronological-age bands, stratified by relative MotionAge group. Participants were first divided into four chronological-age bands (18–29, 30–44, 45–64, and 65+ years). Within each band, MotionAge acceleration (MotionAgeAccel = MotionAge − chronological age) was used to define "Younger" (bottom 20%), "Age-matched" (40th–60th percentile), and "Older" (top 20%) groups relative to same-age peers. Each curve shows the weekly hourly mean intensity aggregated from 5-minute epochs retained by the wearing coverage. In every age band, the "Older" group exhibits lower activity than the "Younger" group, with the separation increasing markedly from ages 45 onward.

A similar ordering was observed in survival analyses. Within each chronological-age stratum, higher MotionAge acceleration was associated with progressively lower 5-year survival probabilities. While survival curves were nearly indistinguishable in the youngest group (18–29), clear separation emerged in middle age and became pronounced in older adults, particularly in the 65+ group, where biologically older individuals showed markedly steeper declines in survival over follow-up.

These qualitative results strengthen the interpretation of MotionAge as a clinically meaningful phenotype rather than merely a transformed risk score. Here, interpretability refers to understanding the learned MotionAge phenotype rather than attributing individual predictions to specific input features. Specifically, MotionAge is interpretable in two complementary ways. First, it is calibrated onto a biological age scale, allowing predictions to be expressed in units that are familiar to clinicians. Second, Figure 4 characterizes the behavioral phenotype associated with MotionAge: within each chronological-age band, individuals with higher MotionAge acceleration consistently exhibit lower activity levels and weaker weekly rhythmicity than their age-matched peers. These analyses characterize the phenotype represented by MotionAge rather than providing feature-attribution explanations.

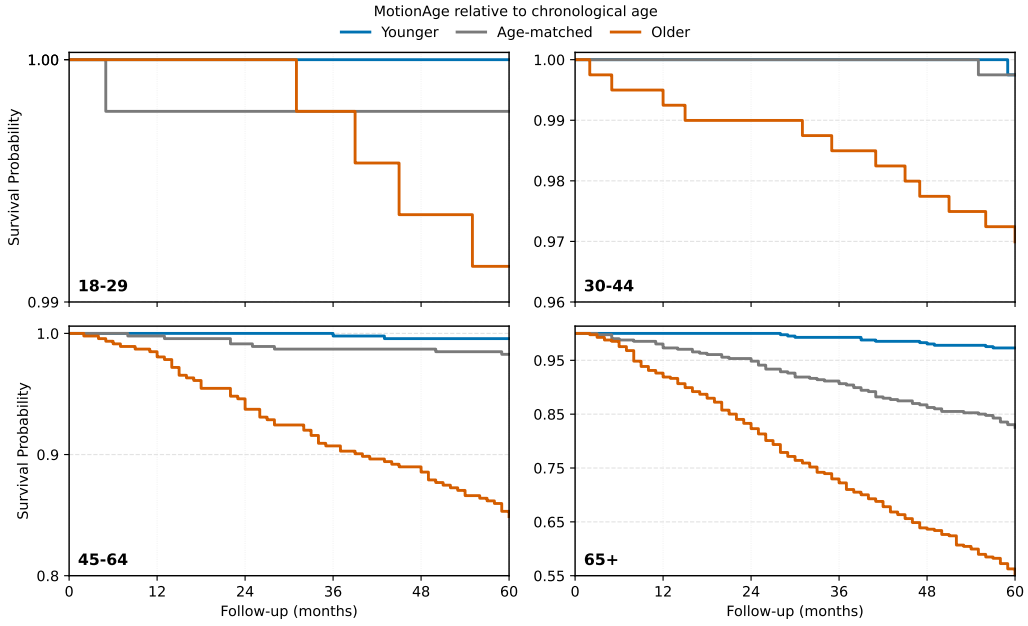


Figure 5: Five-year survival within fixed chronological-age bands, stratified by relative MotionAge group. Kaplan–Meier curves were administratively censored at 60 months and used the same within-band MotionAge acceleration strata as Figure 4: "Younger" (bottom 20%), "Age-matched" (40th–60th percentile), and "Older" (top 20%). Because strata were defined separately within each age band, each panel compares participants of similar chronological age. Survival separation is minimal in ages 18–44, but becomes pronounced in ages 45–64 and largest in ages 65+, where the "Older" group shows substantially lower 60-month survival than same-age peers in the "Younger" group.

Within-age-group separation in both behavior and survival suggests that MotionAge captures heterogeneity not explained by chronological age alone, and that the learned representation tracks differences in real-world functioning that align with downstream health outcomes.

6. Discussion

MotionAge addressed a practical question in machine learning for healthcare: can high-frequency wearable behavior be translated into a clinically interpretable aging phenotype without relying on invasive laboratory testing? Our findings suggest that this is feasible when the observational challenges of wearable data are modeled explicitly. In particular, device non-wear cannot be treated as equivalent to inactivity. Ignoring wear status distorted even basic age-stratified activity profiles, whereas wear-aware preprocessing recovered more biologically plausible differences in both activity magnitude and timing across age groups.

From a clinical perspective, MotionAge offers a scalable complement to existing biological age measures. Unlike clocks built from laboratory assays or static clinical summaries,

wearable-derived measures can capture free-living behavioral patterns at much finer temporal resolution and may support repeated assessment over time. MotionAge-FRC showed stronger mortality discrimination than chronological age and LLMAge and modestly stronger discrimination than PhenoAge in the primary analysis. Overall, the results suggest that free-living activity contains mortality-relevant information that is not fully represented by chronological age or static clinical measures.

Although prospective validation is required before MotionAge could inform clinical decisions, it may eventually serve as a risk-stratification measure that identifies individuals who warrant further assessment of frailty, functional decline, or other aging-related health risks. It may also be useful for population-level health monitoring, longitudinal tracking of aging trajectories, and integration into wearable or digital health platforms for continuous assessment. At present, however, MotionAge should be viewed primarily as a research phenotype rather than a validated clinical decision instrument. Its clinical utility, appropriate decision thresholds, and implications for downstream intervention remain to be established. MotionAge should also not be interpreted as a direct mechanistic measure of biological aging. Rather, it is a mortality-calibrated phenotype whose value lies in summarizing wearable-derived variation associated with clinically important outcomes. Its interpretation necessarily depends on the reference population, sensing device, observation protocol, and mortality horizon used for model development. Therefore, the strongest claim supported by this study is not that MotionAge measures "true" biological aging, but that it provides an interpretable age-scale representation of wearable-derived risk.

Because MotionAge acceleration is defined as the difference between estimated biological age and chronological age, it may be affected by regression-to-the-mean. In particular, estimated ages may shrink toward the center of the training-age distribution, producing systematically positive acceleration at younger ages and negative acceleration at older ages. Consequently, extreme positive or negative age-acceleration values should be interpreted with caution. Importantly, our primary conclusions are based on population-level mortality discrimination and within-age-group comparisons, rather than the interpretation of individual extreme residuals. Future work should investigate age-specific calibration and normalization strategies to further improve the interpretation of age acceleration.

Several limitations should be noted. First, the study used NHANES accelerometry from a specific survey period and device configuration, so external validity to modern consumer wearables and other populations remains to be established. Second, while the wear-aware pipeline reduces bias from device non-wear, it cannot eliminate all measurement error or misclassification of observation quality. Third, the mapping from mortality risk to an age scale is population-calibrated and may require recalibration in cohorts with different baseline risk structures.

Future work will focus on validating MotionAge across diverse datasets and wearable platforms, including devices with different sensing modalities and sampling frequencies. Extending the framework to integrate multi-modal wearable data (e.g., activity, heart rate, sleep, and physiological signals) may further improve robustness and predictive performance. In addition, incorporating longitudinal data and repeated measurements could enable the study of within-individual aging trajectories and dynamic changes in biological age over time. Together, these directions may help establish wearable-derived biological age as a practical tool for population health monitoring and individualized risk assessment.

Data and Code Availability

The NHANES data are publicly available at https://wwwn.cdc.gov/nchs/nhanes/default.aspx. All analyses in this study were conducted using publicly available NHANES data; no restricted-access NHANES data were used. To facilitate reproducibility, the complete MotionAge implementation is publicly available at https://github.com/jackie-jiaqi-yin/MotionAge.

References

Mathilde Chen, Benjamin Landré, Pedro Marques-Vidal, Vincent T. van Hees, April C. E. van Gennip, Mikaela Bloomberg, Manasa S. Yerramalla, Mohamed Amine Benadjaoud, and Séverine Sabia. Identification of physical activity and sedentary behaviour dimensions that predict mortality risk in older adults: development of a machine learning model in the whitehall ii accelerometer sub-study and external validation in the colaus study. *EClinicalMedicine*, 55:101773, 2023. doi: 10.1016/j.eclinm.2022.101773. URL https://www.thelancet.com/journals/eclinm/article/PIIS2589-5370(22)00502-8/fulltext.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

Leena Choi, Zhouwen Liu, Charles E. Matthews, and Maciej S. Buchowski. Validation of accelerometer wear and nonwear time classification algorithm. *Medicine and Science in Sports and Exercise*, 43(2):357–364, 2011.

GBD 2019 Risk Factors Collaborators. Global burden of 87 risk factors in 204 countries and territories, 1990–2019. *The Lancet*, 396(10258):1223–1249, 2020. doi: 10.1016/S0140-6736(20)30752-2.

Jessica R. Golbus, Kensey Gosch, Mary C. Birmingham, Javed Butler, Ildiko Lingvay, David E. Lanfear, Antonio Abbate, Mikhail L. Kosiborod, C V Damaraju, James L. Januzzi, John Spertus, and Brahmajee K. Nallamothu. Association between wearable device measured activity and patient-reported outcomes for heart failure. *JACC: Heart Failure*, 2023. doi: 10.1016/j.jchf.2023.05.033.

Gregory Hannum, Justin Guinney, Li Zhao, Li Zhang, Gentry Hughes, Srini Sadda, Brian Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, Ricardo Deconde, Ming Chen, Indika Rajapakse, Stephen Friend, Trey Ideker, and Kun Zhang. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, 49(2):359–367, 2013. doi: 10.1016/j.molcel.2012.10.016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14 (10):R115, 2013.

Morgan E Levine. Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age? *The Journals of Gerontology: Series A, Biological sciences and medical sciences*, 68(6):667–674, 2013.

Morgan E. Levine, Ake T. Lu, Andrew Quach, Brian H. Chen, Themistocles L. Assimes, Stefania Bandinelli, Li Hou, Andrea A. Baccarelli, Joshua D. Stewart, Yuan Li, Eric A. Whitsel, James G. Wilson, Alex P. Reiner, Abraham Aviv, Kurt Lohman, Yongmei Liu, Luigi Ferrucci, and Steve Horvath. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)*, 10(4):573–591, 2018. doi: 10.18632/aging.101414.

Yanjun Li, Qi Huang, Jin Jiang, Xusheng Du, Wenxin Xiang, Shiqi Zhang, Zean Pan, Liyuan Zhao, Yuyan Cui, Limei Ke, Bo Yin, Linfeng Liu, Guoqing Feng, Shouyi Yan, Liangcai Gao, Yang Liu, Yujuan Yuan, Yanying Guo, Yuqing Yang, Weizhi Ma, Yining Yang, and Qian Di. Large language model-based biological age prediction in large-scale populations. *Nature Medicine*, 31(9):2977–2990, 2025. doi: 10.1038/s41591-025-03856-8.

Paul Lichtenstein, Niels V. Holm, Pia K. Verkasalo, Anastasia Iliadou, Jaakko Kaprio, Markku Koskenvuo, Eero Pukkala, Axel Skytthe, and Kari Hemminki. Environmental and heritable factors in the causation of cancer. *New England Journal of Medicine*, 343 (2):78–85, 2000. doi: 10.1056/NEJM200007133430201.

Alan D. Lopez, Colin D. Mathers, Majid Ezzati, Dean T. Jamison, and Christopher J.L. Murray. Global and regional burden of disease and risk factors, 2001: systematic analysis. *The Lancet*, 367(9524):1747–1757, 2006. doi: 10.1016/S0140-6736(06)68770-9.

Ake T Lu, Audrey Quach, James G Wilson, Alex P Reiner, Abraham Aviv, Kumar Raj, Li Hou, Andrea A Baccarelli, Yuan Li, Joshua D Stewart, Eric A Whitsel, Themistocles L Assimes, Luigi Ferrucci, and Steve Horvath. DNA methylation grimage strongly predicts lifespan and healthspan. *Aging*, 11(2):303–327, 2019.

Rebecca L. McIntyre, Mizanur Rahman, Siva A. Vanapalli, Riekelt H. Houtkooper, and Georges E. Janssens. Biological age prediction from wearable device movement data identifies nutritional and pharmacological interventions for healthy aging. *Frontiers in Aging*, 2:708680, 2021. doi: 10.3389/fragi.2021.708680.

Andrew C. Miller, Joseph Futoma, Salar Abbaspourazad, Christina Heinze-Deml, Saba Emrani, Ian Shapiro, and Guillermo Sapiro. A wearable-based aging clock associates with disease and behavior. *Nature Communications*, 16(1):9264, 2025. doi: 10.1038/s41467-025-64275-4.

Arnold B. Mitnitski, Alex J. Mogilner, and Kenneth Rockwood. Accumulation of deficits as a proxy measure of aging. *The Scientific World Journal*, 2:323–336, 2002. doi: 10.1100/tsw.2002.46.

Masatoshi Nagata, Shohei Komaki, Yuichiro Nishida, Hideki Ohmomo, Megumi Hara, Keitaro Tanaka, and Atsushi Shimizu. Influence of physical activity on the epigenetic

clock: evidence from a Japanese cross-sectional study. *Clinical Epigenetics*, 16(1):142, 2024. doi: 10.1186/s13148-024-01756-1.

Guangkun Nie, Qinghao Zhao, Gongzheng Tang, Yaxin Li, and Shenda Hong. Artificial intelligence-derived photoplethysmography age as a digital biomarker for cardiovascular health. *Communications Medicine*, 5:481, 2025. doi: 10.1038/s43856-025-01188-9. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC12630966/.

Siobhan M Phillips, Lisa Cadmus-Bertram, Dori Rosenberg, Matthew P Buman, and Brigid M Lynch. Wearable technology and physical activity in chronic disease: Opportunities and challenges. *American Journal of Preventive Medicine*, 54(1):144–150, 2018. doi: 10.1016/j.amepre.2017.08.015.

Wei Qiu, Hugh Chen, Matt Kaeberlein, and Su-In Lee. ExplaiNAble BioLogical Age (ENABL Age): an artificial intelligence framework for interpretable biological age. *The Lancet Healthy Longevity*, 2023. doi: 10.1016/S2666-7568(23)00189-7.

Jinjoo Shim, Elgar Fleisch, and Filipe Barata. Wearable-based accelerometer activity profile as digital biomarker of inflammation, biological age, and mortality using hierarchical clustering analysis in NHANES 2011–2014. *Scientific Reports*, 13(1):9326, 2023. doi: 10.1038/s41598-023-36062-y. URL https://www.nature.com/articles/s41598-023-36062-y.

Sören Richard Stahlschmidt, Benjamin Ulfenborg, and Jane Synnergren. Multimodal deep learning for biomedical data fusion: a review. *Briefings in Bioinformatics*, 23(2):bbab569, 2022. doi: 10.1093/bib/bbab569.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Weinan Wang, Pedram Mohseni, Kevin L. Kilgore, and Laleh Najafizadeh. Demographic information fusion using attentive pooling in CNN-GRU model for systolic blood pressure estimation. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–4, 2023. doi: 10.1109/EMBC40787.2023.10340926.

Yunhe Wang, Sihao Xiao, Bowen Liu, Rongtao Jiang, Yuxi Liu, Yian Hang, Li Chen, Runsen Chen, Michael V. Vitiello, Derrick Bennett, Baihan Wang, Jun Lv, Canqing Yu, Danielle E. Haslam, Qianyan Zheng, Robert E. Gerszten, Yanping Bao, Jie Shi, Junqing Xie, Lin Lu, Liming Li, Cornelia M. van Duijn, Dong D. Wang, Zhengming Chen, and Andrew T. Chan. Organ-specific proteomic aging clocks predict disease and longevity across diverse populations. *Nature Aging*, 6(1):162–180, 2026. doi: 10.1038/s43587-025-01016-8.

World Health Organization. Decade of healthy ageing: Baseline report, 2021. URL https://www.who.int/publications/i/item/9789240017900.

Jeffrey O. Zhang, Alexander Sax, Amir Zamir, Leonidas Guibas, and Jitendra Malik. Side-tuning: A baseline for network adaptation via additive side networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV*

2020, volume 12348 of *Lecture Notes in Computer Science*, pages 698–714. Springer, 2020. doi: 10.1007/978-3-030-58580-8_41.

Shibo Zhang, Yaxuan Li, Shen Zhang, Farzad Shahabi, Stephen Xia, Yu Deng, and Nabil Alshurafa. Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors*, 22(4):1476, 2022. doi: 10.3390/s22041476.

Haowen Zhou, Ruoqing Zhu, Anita Ung, and Bruce Schatz. Population analysis of mortality risk: Predictive models from passive monitors using motion sensors for 100,000 UK Biobank participants. *PLOS Digital Health*, 1(10):e0000045, 2022. doi: 10.1371/journal.pdig.0000045. URL https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000045.

Appendix A1. Cohort Summary

Table A1: Sample demographics and routine covariates among NHANES 2003–2006 participants with available 5-year mortality information.

Characteristic	**Overall (N = 8,979)**
Mortality, n (%)	555 (6.2%)
Age, years (mean ± SD)	46.1 ± 20.3
Sex	
Male, n (%)	4,279 (47.7%)
Female, n (%)	4,700 (52.3%)
Annual household income	
\$0 to \$4,999, n (%)	206 (2.3%)
\$5,000 to \$9,999, n (%)	444 (4.9%)
\$10,000 to \$14,999, n (%)	756 (8.4%)
\$15,000 to \$19,999, n (%)	681 (7.6%)
\$20,000 to \$24,999, n (%)	709 (7.9%)
\$25,000 to \$34,999, n (%)	1,167 (13.0%)
\$35,000 to \$44,999, n (%)	889 (9.9%)
\$45,000 to \$54,999, n (%)	803 (8.9%)
\$55,000 to \$64,999, n (%)	536 (6.0%)
\$65,000 to \$74,999, n (%)	459 (5.1%)
\$75,000 and over, n (%)	1,765 (19.7%)
Over \$20,000, n (%)	113 (1.3%)
Under \$20,000, n (%)	36 (0.4%)
Refused, n (%)	31 (0.3%)
Don't know, n (%)	84 (0.9%)
Missing, n (%)	300 (3.3%)
BMI (mean ± SD)	28.4 ± 6.6
Waist circumference, cm (mean ± SD)	97.3 ± 15.7
Diastolic blood pressure, mmHg (mean ± SD)	68.5 ± 13.7
Systolic blood pressure, mmHg (mean ± SD)	123.7 ± 20.1

Appendix A2. Wear-Aware Sequence Construction and Sensitivity

Minute-level data were partitioned into fixed-length 5-minute epochs. For each epoch e, we stored three pieces of information: a wear-aware activity summary x_e, temporal indices for hour of day and day of week, and a binary epoch-validity indicator m_e. For participant i, the resulting sequence window can be written as

$$\{(x_{ie}, h_{ie}, d_{ie}, m_{ie})\}_{e=1}^{T}. \tag{A1}$$

Quality control was applied hierarchically. At the window level, sequences with insufficient mean epoch validity were excluded. Within retained windows, positions with $m_e = 0$

remained in the encoder input, but their hidden states received zero direct weight in the final masked mean. This design preserved alignment of the regular temporal grid while preventing unreliable observations from being interpreted as ordinary inactivity.

A2.1. Preprocessing sensitivity

We examined the four preprocessing choices that define the wear-aware analysis: the Choi minimum non-wear duration, the epoch-level wear threshold τ, the window-level coverage cutoff, and sequence length. Unless varied, settings were fixed at the primary values (90 minutes, $\tau = 0.2$, coverage cutoff 0.3, and $T = 1008$ epochs). The first two choices had little effect on retention or the age-ordering diagnostic used in the main text (Table A2). In particular, $\tau = 0.1$ and $\tau = 0.2$ were equivalent after aggregation to 5-minute epochs because one observed wear minute corresponds to a wear proportion of 0.2.

Table A2: Sensitivity of the wear mask at the primary window settings. Retained participants and deaths are evaluated after window construction. The age-ordering diagnostic indicates whether mean wear-retained activity remained higher in ages 18–29 than in ages 45–64.

Variant	Choi, min	τ	Retained n	Deaths	Age ordering preserved
Choi 60 min	60	0.2	8,676	543	Yes
Epoch threshold 0.1	90	0.1	8,682	543	Yes
Primary	90	0.2	8,682	543	Yes
Epoch threshold 0.3	90	0.3	8,680	543	Yes
Choi 120 min	120	0.2	8,691	543	Yes

The coverage cutoff and sequence length primarily controlled the number of usable windows and participants (Table A3). Retention declined under the stricter coverage cutoff and the longest sequence, as expected, but all settings retained more than 8,400 participants and at least 524 deaths.

Table A3: Window-retention sensitivity. Each group varies one parameter while holding the other at its primary value.

Varied parameter	Setting	Retained windows	Retained participants	Deaths
Coverage cutoff	0.2	17,063	8,806	550
Coverage cutoff	0.3	16,539	8,682	543
Coverage cutoff	0.4	15,561	8,452	526
Sequence length	576	24,449	8,776	546
Sequence length	1,008	16,539	8,682	543
Sequence length	2,016	8,437	8,437	524

To connect these diagnostics to mortality discrimination, we refit MotionAge-FRC from scratch at the stricter 0.4 coverage cutoff while holding the architecture, training hyperparameters, and outer-fold labels fixed. In participants aged 40 years or older, the stricter cutoff retained 4,959 held-out participants and 510 deaths, compared with 5,041 participants and 527 deaths under the primary cutoff. The $\widehat{A} + S$ AUROC was 0.846 rather

than 0.852, and the $AA + A + S$ AUROC was 0.844 rather than 0.850. Across the five fold labels, the differences (0.4 minus 0.3) were -0.0058 ± 0.0049 and -0.0061 ± 0.0063 (mean $\pm$ fold SD), respectively. Because the training partitions overlap and the cutoff changes both model inputs and participant eligibility, these estimates describe pipeline-level sensitivity rather than a participant-paired treatment effect or an equivalence test.

Appendix A3. MotionAge Calibration and Extended Evaluation

A3.1. Fold-specific MotionAge calibration

The calibration in Equations 2–3 used the same settings in all outer folds: the fit partition was the fold-specific training set, the representative sex–age-bin probability was the median, $\epsilon = 10^{-4}$, and regression weights were the sex–age-bin participant counts. The inverse-calibrated age was not clamped to the training age range. Table A4 reports the fitted intercepts and slopes for the primary MotionAge-FRC model. Both sex-specific slopes were positive in every fold; all fits spanned ages 18–85 and used 68 integer-age bins.

Table A4: Fold- and sex-specific calibration parameters for the primary MotionAge-FRC model. Sex codes follow NHANES coding (1, male; 2, female). N_{fit} is the number of training participants contributing participant-level risks to the calibration.

Fold k	Sex s	$\alpha_s^{(k)}$	$\beta_s^{(k)}$	N_{fit}
0	1	−5.912275	0.092758	3,278
0	2	−7.227223	0.096201	3,669
1	1	−6.122209	0.098533	3,330
1	2	−7.571703	0.107435	3,616
2	1	−5.729312	0.091179	3,327
2	2	−7.730077	0.110681	3,617
3	1	−6.467591	0.101797	3,293
3	2	−7.722024	0.109869	3,646
4	1	−6.116713	0.093212	3,340
4	2	−7.333964	0.099928	3,612

A3.2. Full model comparison

Table A5: Five-fold cross-validation evaluation summary for 5-year mortality prediction in the overall population and in participants aged 40 years or older. Values are fold mean ± sample SD across the five outer test folds. *Abbreviations: A = chronological age; S = sex; $\hat{A}$ = estimated biological age; $AA = \hat{A} - A$; M = mortality within 5 years; P_1 = stage-1 mortality probability; F = wearable/activity only; FA = wearable data plus chronological age; FRC = wearable data plus routine covariates; LF = late fusion; RF = residual fusion; Tr = Transformer; C-index = concordance index.*

Model	Age C-index	Mortality C-index			Mortality AUROC			
	$C(\hat{A}, A)$	$C(M, A)$	$C(M, \hat{A})$	$C(M, AA)$	A+S	$\hat{A}$+S	AA+A+S	P_1+A
Overall population								
Age	–	**0.870 ± 0.013**	–	–	**0.875 ± 0.015**	–	–	–
PhenoAge	0.916 ± 0.002	–	0.890 ± 0.014	0.692 ± 0.017	–	0.893 ± 0.015	0.894 ± 0.015	–
LLMAge	**0.943 ± 0.001**	–	0.871 ± 0.013	0.588 ± 0.011	–	0.877 ± 0.015	0.878 ± 0.015	–
GRU-F	0.618 ± 0.011	–	0.819 ± 0.025	0.709 ± 0.036	–	0.832 ± 0.029	0.889 ± 0.019	0.890 ±
GRU-FA-LF	0.933 ± 0.033	–	0.882 ± 0.006	0.664 ± 0.065	–	0.887 ± 0.009	0.887 ± 0.009	0.885 ±
GRU-FA-RF	0.901 ± 0.020	–	0.887 ± 0.015	0.713 ± 0.039	–	0.892 ± 0.018	0.891 ± 0.019	0.890 ±
GRU-FRC-LF	0.849 ± 0.009	–	**0.899 ± 0.011**	0.721 ± 0.022	–	**0.902 ± 0.011**	**0.901 ± 0.011**	**0.902 ±**
GRU-FRC-RF	0.833 ± 0.019	–	0.883 ± 0.007	0.627 ± 0.034	–	0.885 ± 0.005	0.880 ± 0.006	0.875 ±
Tr-F	0.655 ± 0.013	–	0.823 ± 0.018	0.688 ± 0.020	–	0.833 ± 0.023	0.884 ± 0.021	0.886 ±
Tr-FA-LF	0.938 ± 0.015	–	0.886 ± 0.015	0.710 ± 0.036	–	0.891 ± 0.017	0.890 ± 0.019	0.889 ±
Tr-FA-RF	0.922 ± 0.008	–	0.890 ± 0.013	0.661 ± 0.019	–	0.894 ± 0.016	0.892 ± 0.016	0.892 ±
Tr-FRC-LF	0.858 ± 0.007	–	0.889 ± 0.008	**0.726 ± 0.024**	–	0.893 ± 0.009	0.886 ± 0.008	0.883 ±
Tr-FRC-RF	0.857 ± 0.016	–	0.889 ± 0.008	0.702 ± 0.026	–	0.894 ± 0.007	0.888 ± 0.006	0.884 ±
Age ≥ 40								
Age	–	**0.795 ± 0.019**	–	–	**0.803 ± 0.021**	–	–	–
PhenoAge	0.865 ± 0.006	–	0.832 ± 0.020	0.683 ± 0.019	–	0.836 ± 0.021	0.836 ± 0.021	–
LLMAge	**0.918 ± 0.002**	–	0.797 ± 0.017	0.516 ± 0.015	–	0.806 ± 0.021	0.806 ± 0.020	–
GRU-F	0.685 ± 0.011	–	0.797 ± 0.031	**0.752 ± 0.037**	–	0.809 ± 0.034	0.836 ± 0.025	0.836 ±
GRU-FA-LF	0.889 ± 0.046	–	0.823 ± 0.008	0.673 ± 0.089	–	0.829 ± 0.011	0.832 ± 0.012	0.826 ±
GRU-FA-RF	0.860 ± 0.018	–	0.833 ± 0.020	0.737 ± 0.047	–	0.839 ± 0.024	0.839 ± 0.025	0.836 ±
GRU-FRC-LF	0.816 ± 0.002	–	**0.848 ± 0.013**	0.740 ± 0.023	–	**0.852 ± 0.015**	**0.850 ± 0.015**	**0.851 ±**
GRU-FRC-RF	0.794 ± 0.005	–	0.826 ± 0.007	0.637 ± 0.042	–	0.828 ± 0.007	0.824 ± 0.011	0.827 ±
Tr-F	0.682 ± 0.007	–	0.790 ± 0.020	0.730 ± 0.025	–	0.801 ± 0.025	0.831 ± 0.023	0.832 ±
Tr-FA-LF	0.881 ± 0.037	–	0.829 ± 0.023	0.721 ± 0.029	–	0.837 ± 0.025	0.836 ± 0.029	0.835 ±
Tr-FA-RF	0.856 ± 0.015	–	0.835 ± 0.019	0.669 ± 0.025	–	0.842 ± 0.023	0.840 ± 0.024	0.842 ±
Tr-FRC-LF	0.822 ± 0.002	–	0.835 ± 0.015	0.749 ± 0.024	–	0.840 ± 0.014	0.834 ± 0.016	0.839 ±
Tr-FRC-RF	0.817 ± 0.012	–	0.833 ± 0.010	0.718 ± 0.015	–	0.840 ± 0.010	0.835 ± 0.010	0.838 ±

The full comparison reinforces the pattern from the main text. Across both the overall cohort and the age-40-plus subgroup, the strongest mortality discrimination was obtained by GRU late-fusion models that combined wearable sequences with routine covariates. Transformer-based alternatives were competitive but did not outperform the best GRU configuration in any primary evaluation setting.

A3.3. Paired benchmark uncertainty

Paired uncertainty was estimated on shared held-out participants so that both scores in each comparison were evaluated on the same individuals. Within each outer test fold, deaths and survivors were resampled separately with replacement, and identical resampled participant indices were used for the two methods. For each of 10,000 bootstrap replicates, we computed the AUROC difference within each fold and averaged the five fold-level differences. Percentile intervals from this bootstrap distribution provide the 95% confidence intervals. Thus, the estimand is the paired difference in fold-mean AUROC, matching the aggregation used in Table 1. These intervals condition on the fitted models and observed fold assignment; they do not include variation from retraining the sequence models or redefining the outer folds.

The primary comparisons used 5,041 shared participants aged 40 years or older, including 527 deaths within 5 years (Table A6). Because Table 1 uses all participants available to each individual method, its separately computed point estimates need not differ by exactly the paired estimates reported here.

Table A6: Primary fold-structured paired AUROC comparisons for MotionAge-FRC. Differences are MotionAge-FRC minus the comparator on the shared held-out cohort.

Comparator	N	Deaths	ΔAUROC	95% CI
PhenoAge	5,041	527	0.0181	[0.0047, 0.0320]
LLMAge	5,041	527	0.0482	[0.0344, 0.0622]
Age	5,041	527	0.0522	[0.0384, 0.0664]

Table A7 extends the same paired analysis to all six rows of Table 1. Each contrast is the first method minus the second; Age uses the $A + S$ prediction and all biological-age rows use $\widehat{A} + S$. The three MotionAge-FRC benchmark contrasts use the same canonical bootstrap results as Table A6.

A3.4. PhenoAge missing-data sensitivity

We assessed whether the PhenoAge comparison depended on its laboratory missing-data handling. The primary benchmark estimated a separate median for each PhenoAge input from the outer-fold training partition and applied those medians unchanged to held-out participants. We additionally used training-fold medians stratified by sex and 10-year age band, with training-fold sex-specific and global medians as fallbacks, and evaluated a stricter complete-case cohort. The complete-case analysis changes the target population by conditioning on laboratory availability and is therefore a sensitivity analysis rather than a replacement for the primary benchmark.

Table A7: Pairwise fold-mean AUROC differences with 95% fold-structured paired-bootstrap confidence intervals.

Contrast	N	Deaths	ΔAUROC [95% CI]
LLMAge − Age	5,135	539	$0.0034\ [-0.0026, 0.0093]$
PhenoAge − Age	5,135	539	$0.0330\ [0.0202, 0.0460]$
PhenoAge − LLMAge	5,135	539	$0.0295\ [0.0179, 0.0418]$
MotionAge-F − Age	5,041	527	$0.0095\ [-0.0111, 0.0300]$
MotionAge-F − LLMAge	5,041	527	$0.0056\ [-0.0146, 0.0258]$
MotionAge-F − PhenoAge	5,041	527	$-0.0245\ [-0.0433, -0.0058]$
MotionAge-FA − Age	5,041	527	$0.0299\ [0.0204, 0.0395]$
MotionAge-FA − LLMAge	5,041	527	$0.0260\ [0.0158, 0.0366]$
MotionAge-FA − PhenoAge	5,041	527	$-0.0041\ [-0.0172, 0.0085]$
MotionAge-FA − MotionAge-F	5,041	527	$0.0204\ [0.0067, 0.0345]$
MotionAge-FRC − Age	5,041	527	$0.0522\ [0.0384, 0.0664]$
MotionAge-FRC − LLMAge	5,041	527	$0.0482\ [0.0344, 0.0622]$
MotionAge-FRC − PhenoAge	5,041	527	$0.0181\ [0.0047, 0.0320]$
MotionAge-FRC − MotionAge-F	5,041	527	$0.0427\ [0.0289, 0.0564]$
MotionAge-FRC − MotionAge-FA	5,041	527	$0.0222\ [0.0123, 0.0324]$

Table A8: Pre-imputation missingness among the nine PhenoAge biomarker inputs in the main age ≥ 40 held-out evaluation cohort. The denominator ($N = 5{,}135$) is the aggregate of the five outer held-out cross-validation test folds. For the primary benchmark, each biomarker's median was estimated from the corresponding outer-fold training partition only and applied unchanged to held-out participants.

PhenoAge biomarker	Missing, n	Missing, %
Albumin (`LBDSALSI`)	236	4.60
Creatinine (`LBDSCRSI`)	236	4.60
Glucose (`LBDSGLSI`)	236	4.60
Alkaline phosphatase (`LBXSAPSI`)	236	4.60
log CRP (`log_crp`)	202	3.93
Lymphocyte percentage (`LBXLYPCT`)	183	3.56
MCV	161	3.14
RDW (`LBXRDW`)	161	3.14
WBC (`LBXWBCSI`)	161	3.14

The complete-case PhenoAge benchmark retained 4,872 held-out participants and 483 deaths before intersection with available MotionAge-FRC predictions; the paired analysis used 4,786 participants and 472 deaths. On this shared subset, the point estimate remained slightly higher for MotionAge-FRC, but the confidence interval included zero. The primary leakage-safe comparison and the age/sex-stratified imputation sensitivity support a positive MotionAge-FRC difference in their evaluation cohorts, whereas the complete-case restriction narrows the margin and increases its uncertainty.

Table A9: Paired MotionAge-FRC versus PhenoAge sensitivity analyses among participants aged 40 years or older. AUROCs and differences are evaluated on the shared held-out participants in each row.

PhenoAge handling	N	Deaths	PhenoAge	MotionAge-FRC	ΔAUROC	95% CI
Training-fold median	5,041	527	0.8335	0.8516	0.0181	[0.0047, 0.0320]
Training-fold age/sex median	5,041	527	0.8356	0.8516	0.0161	[0.0028, 0.0296]
Shared complete cases	4,786	472	0.8479	0.8526	0.0047	[−0.0087, 0.0180]

Appendix A4. Activity-Pattern Interpretation and Representative Examples

A4.1. Activity-pattern interpretation

Table A10 quantifies the activity-pattern contrasts visualized in Figure 4. Within each chronological-age band, "Younger" and "Older" denote the bottom and top 20% of MotionAge acceleration, respectively. Mean intensity and weekly peak-to-trough amplitude were first computed per participant from wear-retained hourly profiles and then averaged within each group. The reported contrasts are high minus low acceleration, so negative values indicate lower activity in the high-acceleration group.

Table A10: Descriptive high- versus low-MotionAge-acceleration contrasts in wear-retained activity profiles within chronological-age bands. Percentage differences are relative to the low-acceleration group.

Age band	N per stratum	Mean intensity, %	Weekly amplitude, %
18–29	469	−15.9	−1.7
30–44	399	−13.8	−9.5
45–64	463	−38.2	−22.9
65+	407	−67.2	−41.0

Relative to the low-acceleration group, the high-acceleration group had lower mean intensity in every age band, with the largest contrasts after age 45. Weekly amplitude showed the same age-graded pattern. Because the activity sequences contribute to MotionAge itself, these same-sample summaries characterize a MotionAge-associated activity phenotype; they do not provide independent validation or identify which temporal inputs causally determine an individual prediction.

A4.2. Representative subject-level examples

To complement the population averages in Figure 4, Figure A1 shows one matched low- versus high-MotionAgeAccel pair per chronological-age band from the same GRU-FRC-LF fold. Within each age band, participants were first assigned to the low- and high-acceleration strata using the same bottom-20% and top-20% thresholds as in the main text. Candidate exemplars were then ranked by how closely their wear-aware weekly trajectory, mean intensity, weekly amplitude, and active-hour count matched the corresponding stratum

median. From the 25 most representative candidates in each stratum, we selected the best age- and sex-matched low/high pair. These trajectories are therefore intended as representative illustrations rather than extreme cases.

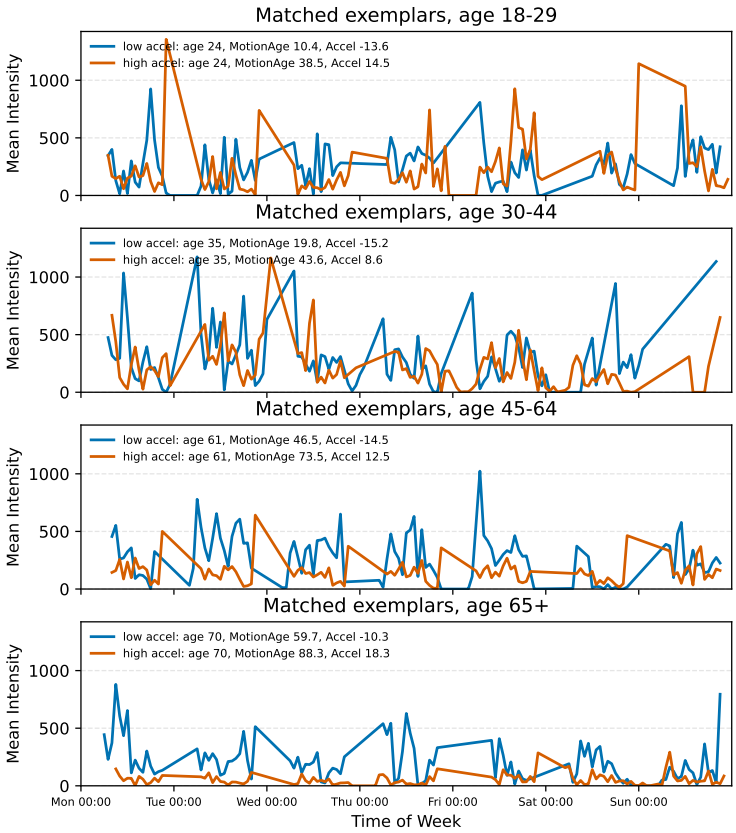


Figure A1: Representative matched exemplar pairs with contrasting MotionAge acceleration within the same chronological-age band. Each panel shows two participants selected to be similar in chronological age and sex but different in MotionAgeAccel, using only candidates whose wear-aware weekly trajectories were among the most representative of their respective strata. The low-acceleration exemplar corresponds to a younger estimated MotionAge relative to peers of the same chronological age, whereas the high-acceleration exemplar corresponds to an older estimated MotionAge.

The matched examples mirror the population-level trends while preserving realistic subject-level variability. In the younger matched pairs (ages 24 and 35 years), the high-acceleration exemplar had 55.6 and 52.8 fewer mean-intensity units than the low-acceleration exemplar, respectively. In the older matched pairs (ages 61 and 70 years), the corresponding gaps widened to 123.5 and 141.5 units. The weekly peak-to-trough amplitude showed the same direction of separation: for the age-61 pair, amplitude was 1938.0 in the high-acceleration exemplar versus 3097.0 in the low-acceleration exemplar, and for the age-70

pair it was 2084.0 versus 2819.2. These examples should not be interpreted as standalone evidence, but they make the learned pattern concrete: within the same chronological-age band, individuals with higher MotionAgeAccel tend to express a weaker and less sustained weekly activity rhythm.