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

Abstract

Population aging has created a growing need for scalable measures of health beyond chronological age. Existing approaches to biological age often rely on clinical or laboratory data, limiting their ability to capture real-world, continuously evolving health status. We propose *MotionAge*, a deep learning framework that learns a mortality-calibrated biological age directly from high-frequency wearable activity data. The approach combines deep sequence models with a wear-aware modeling strategy that explicitly encodes observation reliability, enabling robust representation learning from noisy and irregularly observed time series. In NHANES accelerometer data, MotionAge improves mortality discrimination over chronological age and established benchmarks. These results highlight the potential of wearable data to enable scalable, real-time assessment of aging in population settings.

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 1 billion in 2020 to over 2 billion by 2050, with older adults comprising an increasing proportion of the total population in many countries. As life expectancy rises, however, longevity is often accompanied by a growing burden of chronic diseases such as cardiovascular disease, diabetes, and neurodegenerative disorders, raising concerns not only about survival but also about quality of life. In this context, there is a growing need for a unified, quantitative measure that captures an individual's overall health status beyond chronological age, commonly referred to as biological 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). 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.

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). Recent work has further demonstrated that wearable signals can be used to construct aging-related measures, for example by predicting chronological age from photoplethys-

. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

mography (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 unsupervised biomarkers, with biological relevance inferred through associations with health outcomes. Moreover, such measures are often proposed as biomarkers of biological aging without systematic comparison to established clinical or epidemiological benchmarks.

Wearable data are inherently high-dimensional and possess complex temporal structure, which requires careful processing to avoid systematic bias and to preserve meaningful behavioral patterns. Advances in machine learning provide a natural framework for addressing these challenges. In particular, deep learning architectures designed for sequential data, such as long short-term memory (LSTM) networks (Hochreiter & Schmidhuber, 1997), gated recurrent units (GRU) (Cho et al., 2014), and transformers (Vaswani et al., 2017), are well-suited for modeling such data due to their ability to capture temporal dependencies and nonlinear relationships. These models enable the extraction of informative representations from high-frequency wearable data while preserving temporal structure.

In this study, 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. 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.

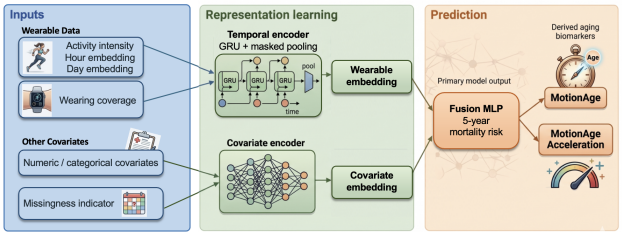


Figure 1. Overview of the proposed MotionAge. Wear-aware activity sequences and optional static baseline covariates are encoded separately, fused to predict 5-year mortality risk, and then mapped onto an age scale to yield a mortality-calibrated biological age estimate.

2. Methods

2.1. 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. Below, we give a general description of the preprocessing and model training procedure; additional details are provided in Appendix S2.

2.2. Wear-Aware Preprocessing and Sequence Construction

Our preprocessing pipeline preserves temporal structure while explicitly representing observation reliability (Figure S1). We first identified non-wear at the minute level using the Choi algorithm (Choi et al., 2011), yielding a binary wear indicator $w_t \in \{0, 1\}$. We then aggregated the minute-level series into 5-minute epochs. For each epoch e, we computed a wear-aware activity summary

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

i.e., the average activity intensity over wear minutes within epoch e. We also retained temporal context features (hour-of-day and day-of-week) and defined an epoch-level wearing coverage

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

with threshold $\tau = 0.2$.

Participant-level sequences were constructed by sliding a fixed-length window over the ordered epoch sequence, with window length treated as a tunable hyperparameter. Windows with insufficient overall wearing coverage were excluded. For retained windows, low-coverage epochs were not removed; their wearing coverage was carried forward as a reliability signal in the pooling stage. This preserves the regular temporal grid while allowing the model to distinguish between sparse observation and low activity.

2.3. 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, capturing both instantaneous activity and structured temporal variation. Given a sequence of length T, a sequence

encoder (GRU or Transformer) produced hidden representations $\mathbf{h}_1, \ldots, \mathbf{h}_T$.

To account for irregular observation and non-wear, we applied masked temporal pooling:

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

where $\varepsilon > 0$ ensures numerical stability. This yields a reliability-weighted summary in which epochs with insufficient wear contribute minimally.

When static covariates were available, they were encoded via a multilayer perceptron $g_\phi(\mathbf{s})$, where $\mathbf{s}$ denotes participant-level baseline features. In the late-fusion model, the pooled wearable representation and static representation were concatenated and passed to a prediction network:

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

We also considered a residual fusion formulation, in which a wearable-derived prediction is augmented by an additive correction from static covariates:

$$z = z_{\text{wear}} + z_{\text{static}}.$$

The model outputs a scalar logit corresponding to the log-odds of 5-year mortality risk. A probability estimate is obtained via the sigmoid transformation, and the model is trained using a class-weighted binary cross-entropy loss. Because each participant contributes multiple windows, predictions are first obtained at the window level and then aggregated to the participant level by averaging predicted probabilities.

2.4. Interpretation via MotionAge

A central design goal is to obtain an interpretable phenotype rather than only a mortality score. To achieve this, we estimated a sex-specific relationship between chronological age A and predicted mortality risk in the training data. MotionAge was then defined via the inverse mapping, representing the age whose expected mortality risk matches the participant's predicted risk. Age acceleration was defined as the difference between MotionAge and chronological age.

2.5. Training Procedure

We defined model inputs using a wearable (Fitbit) branch and optional static covariates. The wearable branch consisted of temporal features derived from wear-aware preprocessing and was always included. Static baseline covariates were optionally incorporated as auxiliary inputs.

We evaluated three configurations: **F** (wearable only), **FA** (wearable + age), and **FRC** (wearable + routine covariates including age, sex, BMI, waist circumference, and blood pressure). This design allows assessment of both the standalone contribution of wearable signals and their incremental value when combined with standard covariates.

Hyperparameters were selected on a development split and fixed for evaluation. Candidate models were ranked using validation AUPRC. Additional optimization and implementation details are provided in the Appendix.

3. Evaluation Setup

3.1. 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 S1. 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 S3.

3.2. 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 $AA = \hat{A} - A$. We then fitted a logistic regression evaluation model on the training split using $\hat{A}$ or AA as the primary predictor, with sex included as a covariate when appropriate, and applied this model to the held-out test split to obtain predicted 5-year mortality probabilities. Performance was evaluated separately on each held-out test fold using the C-index and area under the receiver operating characteristic curve (AUROC); the final reported C-index and AUROC are the means of the fold-specific test-set estimates across the five folds.

Several sequence-model variants were implemented includ-

ing 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 S3.

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 the median. For LLMAge, we followed the prompting procedure described in the original paper to obtain the overall 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.

4. Result and Evaluation

Table 1 reports the main five-fold cross-validation result for 5-year mortality prediction in participants aged 40 years or older. MotionAge models achieved the strongest mortality 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). Even MotionAge-F, which used wearable activity alone without chronological age or additional covariates in the deep learning model, outperformed the chronological-age baseline and LLMAge in both C-index and AUROC. In Appendix S3, 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.

Table 1. Five-fold cross-validation evaluation summary for 5-year mortality prediction in the overall population and in participants aged 40 years or older. MotionAge was obtained using GRU-based models with late fusion. Values are means across complete outer folds. *Abbrev.: A = chronological age; S = sex; $\hat{A}$ = estimated biological age (MotionAge for learned sequence models, and PhenoAge/LLMAge for the PhenoAge/LLMAge benchmark); AA = age acceleration ($\hat{A}-A$); M = mortality outcome defined as death within 5 years of baseline assessment; F = Fitbit-only backbone; FA = Fitbit and chronological age; FRC = Fitbit and routine baseline covariates, including age, sex, household income, body mass index, waist circumference, systolic blood pressure, and diastolic blood pressure; C-index = concordance index; AUROC = area under the receiver operating characteristic curve. Mortality AUROC subcolumns indicate predictor sets used in the evaluation logistic model.*

Model	Mort C-index			Mort AUROC		
	C(M, A)	C(M, $\hat{A}$)	C(M, AA)	A+S	$\hat{A}$+S	AA+A+S
Age $\geq$ 40						
Age	**0.795**	-	-	**0.803**	-	-
LLMAge		0.797	0.516	-	0.806	0.806
PhenoAge	-	0.832	0.683	-	0.836	0.836
MotionAge-F	-	0.797	**0.752**	-	0.809	0.836
MotionAge-FA	-	0.823	0.673	-	0.829	0.832
MotionAge-FRC	-	**0.848**	0.740	-	**0.852**	**0.850**

5. Discussion

In this study, we propose MotionAge, a wearable-derived biological age constructed by modeling mortality risk from high-frequency activity data. Our results demonstrate that wearable-based representations can provide meaningful and scalable measures of aging, achieving improved mortality prediction compared with both a chronological age baseline and an established biomarker-based benchmark. Notably, these gains are obtained using a relatively limited set of routine covariates, highlighting the potential of wearable data to complement or partially substitute traditional clinical measurements. Beyond predictive performance, MotionAge captures interpretable behavioral patterns, with higher biological age associated with lower intensity and less dynamic activity profiles, as well as increased mortality risk within fixed chronological-age groups.

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 help establish wearable-derived biological age as a practical tool for population health monitoring and individualized risk assessment.

Data and Code

The NHANES data are publicly available at https://wwwn.cdc.gov/nchs/nhanes/default.aspx. Our code will be made available on Github upon acceptance.

Impact Statement

This work aims to advance the use of wearable data and machine learning for scalable health monitoring and biological age estimation. By enabling non-invasive and continuously updated assessment of aging and mortality risk, such approaches may support earlier detection of health decline and improved population health management. However, the use of wearable data also raises considerations related to data privacy, access, and potential disparities in device availability across populations. Future work should carefully address these issues to ensure equitable and responsible deployment of wearable-based health models.

References

Chen, M., Landré, B., Marques-Vidal, P., van Hees, V. T., van Gennip, A. C. E., Bloomberg, M., Yerramalla, M. S., Benadjaoud, M. A., and Sabia, S. 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.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. 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.

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

Golbus, J. R., Gosch, K., Birmingham, M. C., Butler, J., Lingvay, I., Lanfear, D. E., Abbate, A., Kosiborod, M. L., Damaraju, C. V., Januzzi, J. L., Spertus, J., and Nallamothu, B. K. 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.

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., Klotzle, B., Bibikova, M., Fan, J.-B., Gao, Y., Deconde, R., Chen, M., Rajapakse, I., Friend, S., Ideker, T., and Zhang, K. 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.

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

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

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

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

Li, Y., Huang, Q., Jiang, J., Du, X., Xiang, W., Zhang, S., Pan, Z., Zhao, L., Cui, Y., Ke, L., Yin, B., Liu, L., Feng, G., Yan, S., Gao, L., Liu, Y., Yuan, Y., Guo, Y., Yang, Y., Ma, W., Yang, Y., and Di, Q. 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.

Lu, A. T., Quach, A., Wilson, J. G., Reiner, A. P., Aviv, A., Raj, K., Hou, L., Baccarelli, A. A., Li, Y., Stewart, J. D., Whitsel, E. A., Assimes, T. L., Ferrucci, L., and Horvath, S. DNA methylation grimage strongly predicts lifespan and healthspan. *Aging*, 11(2):303–327, 2019.

McIntyre, R. L., Rahman, M., Vanapalli, S. A., Houtkooper, R. H., and Janssens, G. E. 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.

Miller, A. C., Futoma, J., Abbaspourazad, S., Heinze-Deml, C., Emrani, S., Shapiro, I., and Sapiro, G. A wearable-based aging clock associates with disease and behavior. *Nature Communications*, 16(1):9264, 2025. doi: 10.1038/s41467-025-64275-4.

Nagata, M., Komaki, S., Nishida, Y., Ohmomo, H., Hara, M., Tanaka, K., and Shimizu, A. 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.

Nie, G., Zhao, Q., Tang, G., Li, Y., and Hong, S. 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/.

Phillips, S. M., Cadmus-Bertram, L., Rosenberg, D., Buman, M. P., and Lynch, B. M. 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.

Qiu, W., Chen, H., Kaeberlein, M., and Lee, S.-I. 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.

Shim, J., Fleisch, E., and Barata, F. 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.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Wang, Y., Xiao, S., Liu, B., Jiang, R., Liu, Y., Hang, Y., Chen, L., Chen, R., Vitiello, M. V., Bennett, D., Wang, B., Lv, J., Yu, C., Haslam, D. E., Zheng, Q., Gerszten, R. E., Bao, Y., Shi, J., Xie, J., Lu, L., Li, L., van Duijn, C. M., Wang, D. D., Chen, Z., and Chan, A. T. 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.

Zhou, H., Zhu, R., Ung, A., and Schatz, B. 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.

S1. Demographics Summary Table

Table S1. Sample demographics and routine covariates among NHANES 2003–2006 participants with available 5yr-mortality information

Characteristic	Overall (N=8979)
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

S2. Epoch aggregation and sequence construction

S2.1. Wear-Aware Preprocessing and Sequence Construction

Non-wear detection. Non-wear intervals were identified using the Choi algorithm (Choi et al., 2011), defined as periods of at least 90 consecutive minutes of zero counts, allowing up to two non-consecutive spike minutes (0–100 counts) if flanked by at least 30 consecutive zero-count minutes. This yields a minute-level wear indicator $w_t \in \{0, 1\}$.

Epoch aggregation and wear-aware features. Minute-level data are partitioned into fixed-length epochs (e.g., 5-minute intervals). For each epoch e, we construct:

- a wear-aware activity summary
$$x_e = \frac{\sum_{t \in e} w_t \cdot \text{count}_t}{\sum_{t \in e} w_t},$$
defined as the average activity intensity over wear minutes only;
- temporal indices h_e and d_e encoding hour-of-day and day-of-week;
- a wear-coverage mask
$$m_e = \mathbb{I}\left\{ \left[\frac{1}{|e|} \sum_{t \in e} w_t \right] \geq \tau \right\},$$
where τ is a threshold (e.g., 0.2).

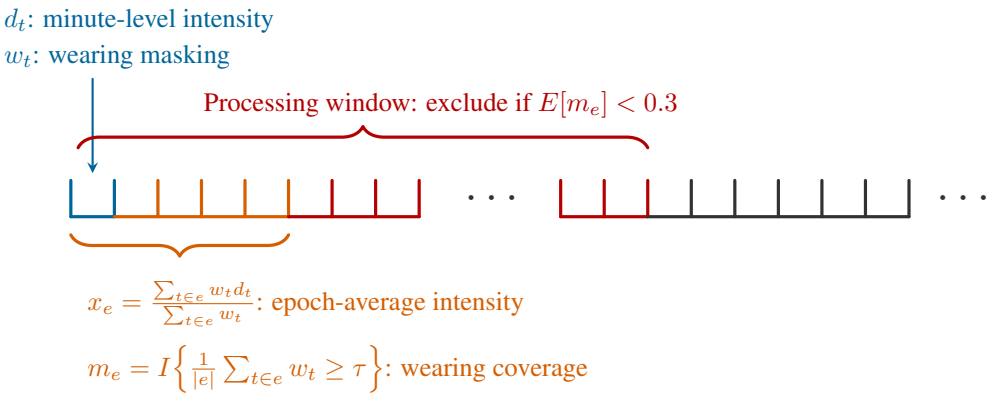


Figure S1. Minute-level activity intensity d_t and wear indicator w_t are aggregated into fixed-length epochs. For each epoch e, we compute a wear-aware average intensity x_e and a wearing coverage m_e. Epochs with insufficient wearing coverage are excluded, and the remaining epochs form the input sequence within each processing window.

This representation ensures that true inactivity (low activity during wear) is not conflated with missingness (non-wear).

Sequence construction with quality control. For each participant, we construct fixed-length sequences by extracting sliding windows of length `seq_len` with stride `stride` from the ordered epoch-level data. Each sequence consists of

$$\{(x_e, h_e, d_e, m_e)\}_{e=1}^{T}.$$

To ensure data quality, we apply hierarchical filtering:

- **Window-level coverage filtering:** windows with insufficient overall wear coverage (e.g., $\frac{1}{T}\sum_e m_e < \gamma$, with $\gamma \approx 0.3$) are excluded;
- **Masked sequence modeling:** within retained windows, low-coverage epochs are preserved with their masks, allowing the model to down-weight unreliable observations while retaining temporal structure.

This strategy preserves the regular temporal grid while propagating a reliability signal through the model.

S2.2. Modeling and Training Details

Residual fusion formulation. In addition to late fusion, we considered a residual fusion approach in which static covariates provide an additive correction to wearable-derived predictions:

$$z = z_{\text{wear}} + z_{\text{static}}.$$

Training objective. The model was trained using a class-weighted binary cross-entropy loss:

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

where $\lambda = N_-/N_+$ accounts for class imbalance.

Optimization details. Hyperparameters were optimized using Optuna on a development split. Training employed AdamW optimization, learning rate scheduling, gradient clipping, and early stopping.

S3. Result for all models

Table 1 summarizes model performance for 5-year mortality prediction across the overall population and the subgroup aged 40 years or older. In both populations, models that integrate wearable-derived features with routine covariates achieve the

Table S2. 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 means across complete outer folds. *Abbrev.: A = chronological age; S = sex; $\hat{A}$ = estimated biological age (MotionAge for learned sequence models, PhenoAge for the PhenoAge benchmark, and LLM-Overall-Age for the LLM-Age benchmark); AA = age acceleration ($\hat{A}-A$); M = mortality outcome defined as death within 5 years of baseline assessment; P_1 = stage-1 mortality probability; F = Fitbit-only backbone; FA = Fitbit and chronological age; FRC = routine baseline covariates, including age, sex, household income, body mass index, waist circumference, systolic blood pressure, and diastolic blood pressure; LF = late fusion; RF = residual fusion; Tr = Transformer; C-index = concordance index; AUROC = area under the receiver operating characteristic curve. Mortality AUROC subcolumns indicate predictor sets used in the evaluation logistic model.*

Model	Age C-index	Mort C-index			Mort 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+S
Overall population								
Base	-	**0.870**	-	-	**0.875**	-	-	-
PhenoAge	0.916	-	0.890	0.692	-	0.893	0.894	-
LLM-Age	**0.943**	-	0.871	0.588	-	0.877	0.878	-
GRU-F	0.618	-	0.819	0.709	-	0.832	0.889	0.890
GRU-FA-LF	0.933	-	0.882	0.664	-	0.887	0.887	0.885
GRU-FA-RF	0.901	-	0.887	0.713	-	0.892	0.891	0.890
GRU-FRC-LF	0.849	-	**0.899**	0.721	-	**0.902**	**0.901**	**0.902**
GRU-FRC-RF	0.833	-	0.883	0.627	-	0.885	0.880	0.875
Tr-F	0.655	-	0.823	0.688	-	0.833	0.884	0.886
Tr-FA-LF	0.938	-	0.886	0.710	-	0.891	0.890	0.889
Tr-FA-RF	0.922	-	0.890	0.661	-	0.894	0.892	0.892
Tr-FRC-LF	0.858	-	0.889	**0.726**	-	0.893	0.886	0.883
Tr-FRC-RF	0.857	-	0.889	0.702	-	0.894	0.888	0.884
Age $\geq$ 40								
Base	-	**0.795**	-	-	**0.803**	-	-	-
PhenoAge	0.865	-	0.832	0.683	-	0.836	0.836	-
LLM-Age	**0.918**	-	0.797	0.516	-	0.806	0.806	-
GRU-F	0.685	-	0.797	**0.752**	-	0.809	0.836	0.836
GRU-FA-LF	0.889	-	0.823	0.673	-	0.829	0.832	0.826
GRU-FA-RF	0.860	-	0.833	0.737	-	0.839	0.839	0.836
GRU-FRC-LF	0.816	-	**0.848**	0.740	-	**0.852**	**0.850**	**0.851**
GRU-FRC-RF	0.794	-	0.826	0.637	-	0.828	0.824	0.827
Tr-F	0.682	-	0.790	0.730	-	0.801	0.831	0.832
Tr-FA-LF	0.881	-	0.829	0.721	-	0.837	0.836	0.835
Tr-FA-RF	0.856	-	0.835	0.669	-	0.842	0.840	0.842
Tr-FRC-LF	0.822	-	0.835	0.749	-	0.840	0.834	0.839
Tr-FRC-RF	0.817	-	0.833	0.718	-	0.840	0.835	0.838

strongest mortality discrimination, with the GRU-RC-LF model consistently attaining the highest C-index and AUROC across evaluation settings. Relative to the baseline model using chronological age and sex, as well as established benchmarks such as PhenoAge and LLM-based biological age, the wearable-enhanced models provide systematic improvements, particularly when combining estimated biological age with covariates. While age-focused models (e.g., GRU-A-LF and Tr-A-LF) achieve the highest concordance with chronological age, their mortality performance is generally lower than models incorporating broader covariate information, highlighting that accurate age reconstruction and mortality prediction are related but distinct objectives. Overall, these results demonstrate that wearable-derived temporal features, when combined with routine covariates, yield robust and competitive predictors of mortality risk.

S4. Additional results

S4.1. 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 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 S2).

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.

S4.2. 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 S3) and 5-year survival (Figure S4)

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.

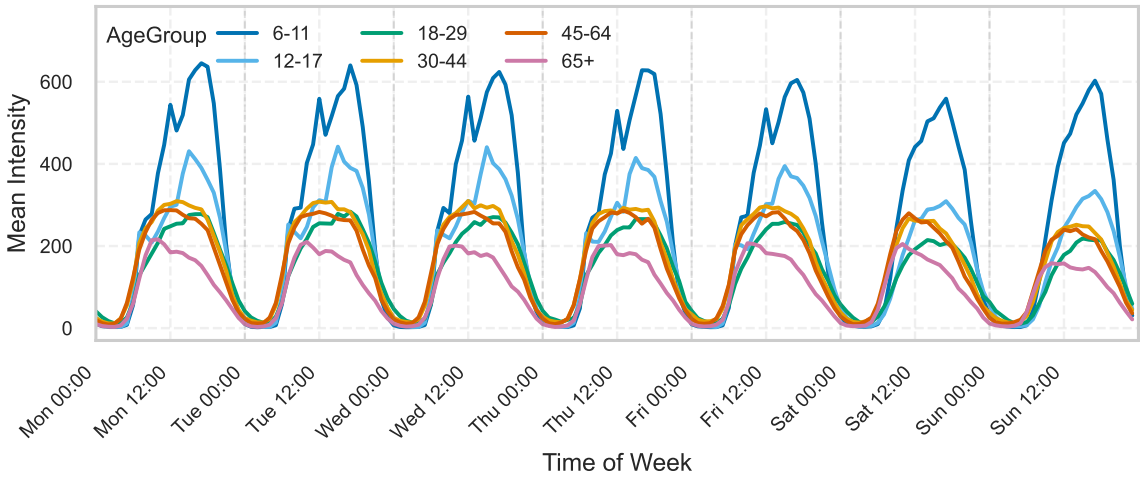
A similar ordering was observed in survival analyses. Within each chronological-age stratum, higher MotionAge acceleration was associated with progressively lower 5-year survival probabilities. While survival curves were nearly indistinguishable in the youngest group (18–29), clear separation emerged 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. 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.

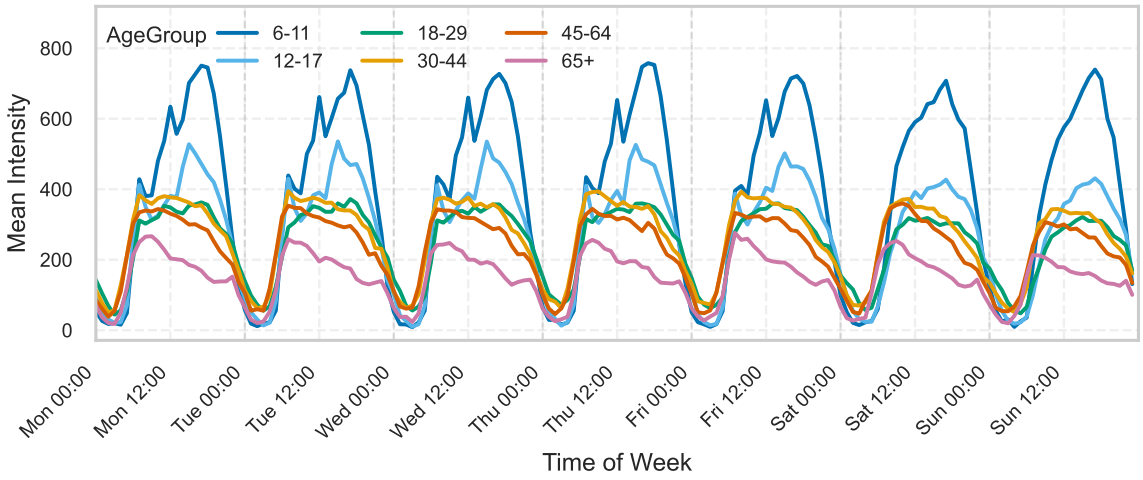
S5. Representative Subject-Level Examples

To complement the population averages in Figure S3, Figure S5 shows one matched low- versus high-MotionAgeAccel pair per chronological-age band from the same GRU-RC-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-masked 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.

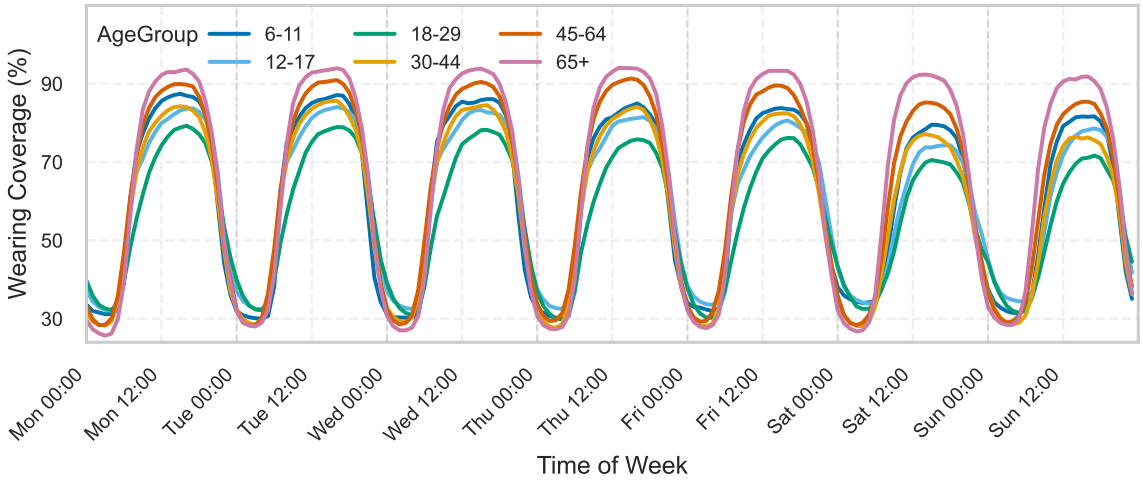
The matched examples mirror the population-level trends while preserving realistic subject-level variability. In the younger matched pairs (ages 24 and 35 years), the high-acceleration exemplar had 55.6 and 52.8 fewer mean-intensity units than the low-acceleration exemplar, respectively. In the older 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.



(a) Weekly hourly mean activity intensity by age group before applying the wear-coverage mask.



(b) Weekly hourly mean activity intensity by age group after restricting to epochs retained by the wear-coverage mask.



(c) Weekly hourly wear-coverage mask by age group, shown as the percentage of 5-minute epochs retained in each hourly bin.

Figure S2. Wear-aware hourly activity diagnostics by age group. Panel a shows the unmasked hourly intensity profile, which exhibits attenuated and partially reversed age ordering in some time windows. Panel b shows the same summary after restricting to epochs retained by the wear-coverage mask, yielding clearer separation across age groups. Panel c shows hourly wear coverage itself, indicating strong time-of-day and age-dependent differences in retained observations.

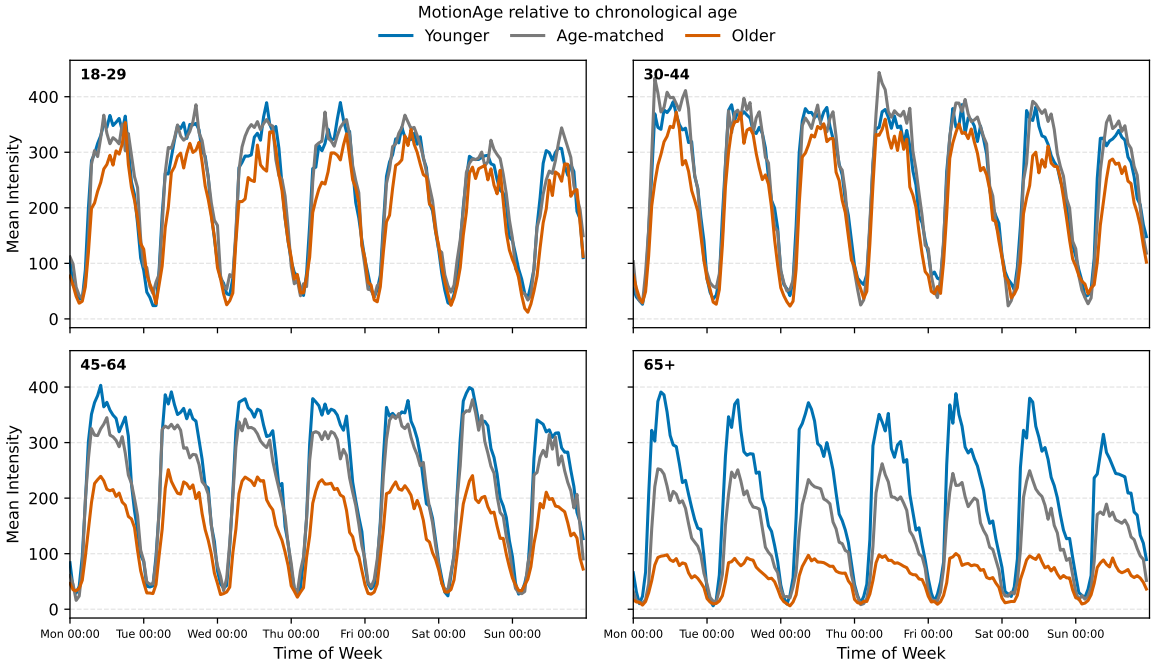


Figure S3. Weekly mean activity intensity within fixed chronological-age bands, stratified by relative MotionAge group. Participants were first divided into four chronological-age bands (18–29, 30–44, 45–64, and 65+ years). Within each band, MotionAge acceleration ($\mathrm{MotionAgeAccel} = \mathrm{MotionAge} - \mathrm{chronological\ age}$) 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.

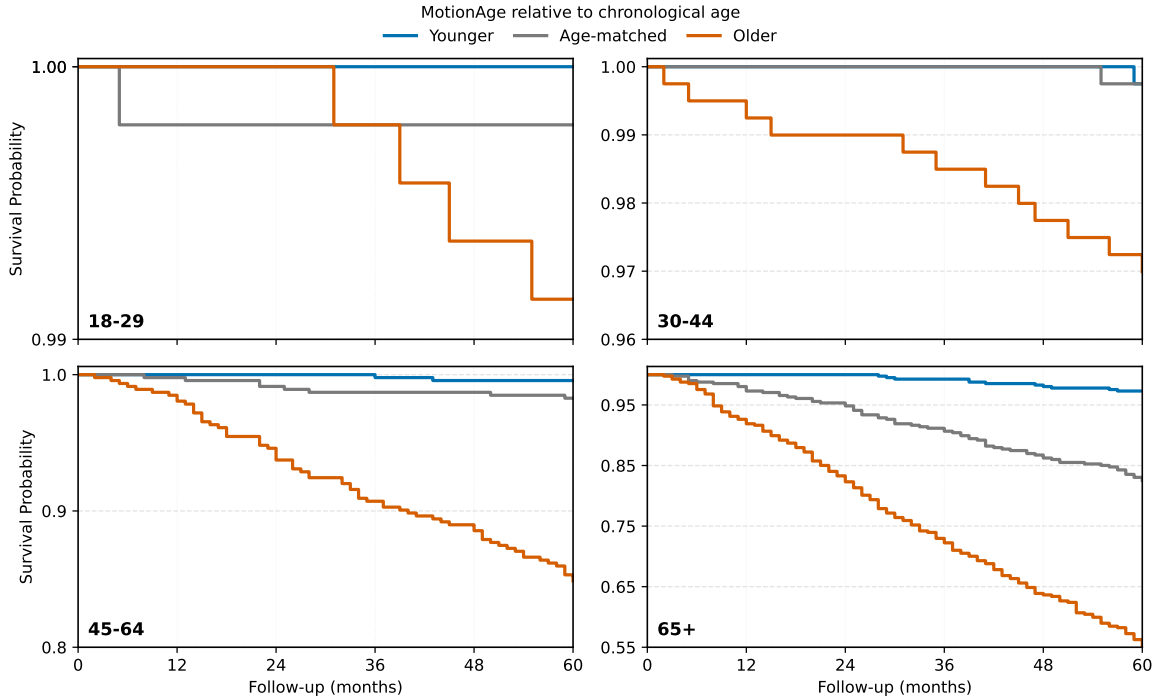


Figure S4. Five-year survival within fixed chronological-age bands, stratified by relative MotionAge group. Kaplan–Meier curves were administratively censored at 60 months and used the same within-band MotionAge acceleration strata as Figure S3: "Younger" (bottom 20%), "Age-matched" (40th–60th percentile), 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.

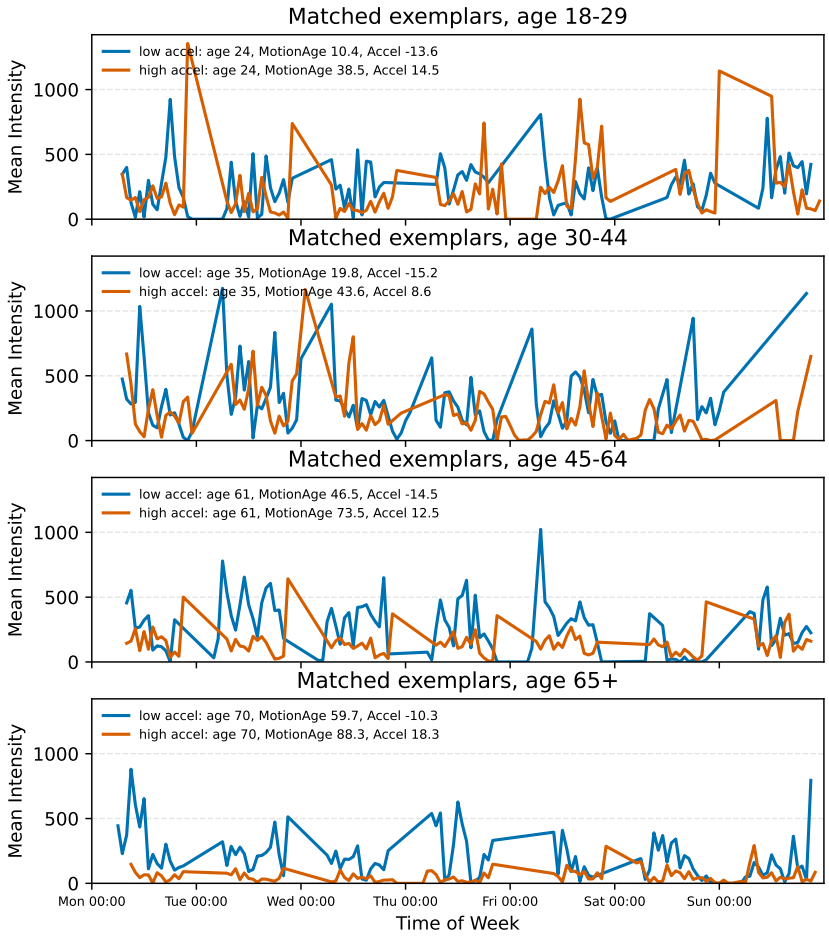


Figure S5. Representative matched exemplar pairs with contrasting MotionAge acceleration within the same chronological-age band. Each panel shows two participants selected to be similar in chronological age and sex but different in MotionAgeAccel, using only candidates whose wear-masked weekly trajectories 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 examples illustrate the same qualitative pattern as Figure S3: higher MotionAgeAccel is generally associated with lower sustained activity and flatter weekly dynamics, with the contrast becoming more pronounced in the older age bands.