Preprint — Not peer-reviewed

The Variance Is the Signal — Decomposing Epigenetic Clock Discordance into Meaningful Components

Steffen Uhlig[1 3 5 *], Carsten Uhlig[2 5], Karina Hettwer[1], Kapil Nichani[1], Robin Uhlig[1] and Kirsten Simon[3 4]

[1] *QuoData GmbH, Quality & Statistics, Berlin, Germany*

[2] *German Heart Centre Munich, Germany*

[3] *QD Labs LP, Frankford, Delaware, U.S.*

[4] *New Diagnostics GmbH, Dresden, Germany*

[5] *Cebio GmbH, Munich, Germany*

* Corresponding author: *steffen.uhlig@quodata.de*

May 2026

Abstract

Background. Epigenetic clocks applied to the same blood sample yield discordant age-acceleration estimates. The standard interpretation treats this discordance as a problem, which clock is right? And the question itself has no clean answer, because different clocks are trained on different outcomes and weight different methylation sites. We take a different route: we treat the discordance itself as structured information. This serves two purposes. At the population level, it tells us what the clocks have in common and what is private to each. At the individual level, it allows a differentiated interpretation of a person's clock readings, not just a single biological age, but a profile of distinct aging components.

Methods. We decompose multi-clock age acceleration into two latent components via maximum likelihood factor analysis: (1) a general epigenomic erosion factor (δ), reflecting stochastic methylation drift, and (2) a specific biological aging factor (θ), capturing pathway-specific alterations. Factor axes are oriented via orthogonal Procrustes rotation toward Tong et al. (2024) stochasticity estimates, separating data-driven estimation from theory-guided interpretation. The factor model is estimated subject to a lower bound on each clock's residual variance corresponding to the technical repeatability standard deviation reported in the literature. The

decomposition yields, for each individual, an erosion score δ_i, a pathway-aging score θ_i, and a vector of clock-specific residuals ε_{ik} that capture what each clock sees in that person beyond the two shared dimensions. Three cohorts: GSE40279 (n=656, USA), GSE51032 (n=845, Italy), GSE41169 (n=95, Netherlands), six clocks each.

Results. At the population level, the decomposition reveals what the clocks measure in common and what is private to each. Horvathv1 (93% erosion variance) and PCHorvath1 (49%) are dominated by stochastic drift. GrimAgeV1 (28% specific) loads almost exclusively on pathway-driven aging, with near-zero erosion contribution. PhenoAge and Hannum occupy intermediate positions. The loading pattern is consistent across the three cohorts. At the individual level, the same decomposition produces a three-component profile per person: two interpretable scores (δ_i, θ_i) shared across all six clocks, plus six clock-specific residuals (ε_{ik}) that flag clock-by-person interactions not predicted by the shared dimensions. We illustrate the per-person interpretation with three contrasting profiles. As a methodological validation, we recover the established slower epigenetic aging in Hispanic vs. Caucasian individuals (Horvath et al. 2016; Faul et al. 2023) and localise the effect specifically to the erosion dimension (β = −0.44, p = 1.0×10^{-6}), with no detectable effect on specific aging (p = 0.18).

Conclusions. Discordance between epigenetic clocks is not noise to be averaged away but a signature of distinct underlying processes. The two-factor decomposition provides two things at once: a structural answer to "what does each clock predominantly measure?", a property of the clock, and a differentiated three-component profile (δ_i, θ_i, ε_{ik}) for each individual that single-clock approaches cannot deliver. This is a structural rather than metrological contribution; a full measurement-uncertainty budget in the ISO 5725 sense requires interlaboratory data not available here and is the subject of a companion paper.

Keywords: *epigenetic clocks, DNA methylation, biological aging, factor analysis, variance decomposition, epigenomic erosion, stochastic aging, clock discordance, measurement uncertainty, reproducibility*

1. Introduction

1.1 The assessment paradox of epigenetic clocks

Epigenetic clocks based on DNA methylation patterns have emerged as the most widely used biomarkers of biological aging. Since the landmark publications of Horvath (2013) and Hannum et al. (2013), multiple generations of clocks have been developed, each trained on different outcomes: chronological age (first generation), mortality-associated phenotypes (second generation, including PhenoAge and GrimAge), and longitudinal rates of biomarker change (third generation, DunedinPACE). These clocks are routinely applied in aging research, clinical trials of longevity interventions, and epidemiological studies of disease risk. The present study is restricted to clocks that output a biological age estimate or an age-correlated quantity (in years or

comparable units), so that age acceleration — the difference between the clock's output and chronological age — is a meaningful residual. Rate-of-aging clocks such as DunedinPACE and its predecessors (Belsky et al., 2022, Belsky et al., 2020), which output a pace in years per year rather than a cumulative age estimate, are conceptually distinct: their output is nearly uncorrelated with chronological age, rendering the standard age-residualization step uninformative, and their scale is incommensurable with that of the age-level clocks. Including rate-based and level-based clocks in the same correlation matrix would conflate two fundamentally different measurement constructs. We therefore exclude DunedinPACE and its predecessor DunedinPoAm38 from the analysis. The remaining six clocks all output a biological age or an age-correlated quantity. Among these, DNAmTL estimates telomere length, where higher values indicate younger biology — the opposite direction to all other clocks. To align all clocks so that positive age acceleration consistently indicates more aging, we reverse the sign of DNAmTL before analysis; the resulting variable represents accelerated telomere attrition.

The standard approach to evaluating an epigenetic clock is to assess how accurately it predicts chronological age, typically quantified by correlation coefficients and mean absolute error (MAE). Horvath's clock achieves an MAE of approximately 3.6 years across tissues; Hannum's clock approximately 3.9 years in blood. These metrics are widely cited as indicators of clock quality. However, this approach contains a fundamental paradox, first identified in the work of Zhang et al. (2019): as the prediction of chronological age improves (by increasing training sample sizes), the association between the clock's age acceleration residual and mortality attenuates. In their study, the best-performing predictor of chronological age — trained on over 12,000 samples — showed no significant association with mortality at all.

This is the Zhang paradox: a perfect chronological-age clock would, by definition, have zero age acceleration residuals, and therefore no predictive value for any biological outcome. The clinical utility of epigenetic clocks does not reside in how well they match chronological age, but precisely in their deviation from it. The deviation — epigenetic age acceleration (EAA) — is the quantity of biological interest. A person whose epigenetic age exceeds their chronological age is aging faster than expected; conversely, a person whose epigenetic age is lower is aging more slowly.

1.2 Variability of clock outputs and the scope of this paper

Every epigenetic age estimate is the output of a measurement process and therefore carries some variability. At minimum, repeated measurement of the same DNA sample on the same platform produces non-identical results. This within-sample, within-laboratory variability, i.e., the repeatability standard deviation in the ISO 5725 vocabulary, often referred to as “technical noise” in the aging literature, has been characterised empirically by Higgins-Chen et al. (2022), who reported values on the order of 1–2 years for several established clocks. Sehgal et al. (2025) have since extended this work by separating technical from biological reliability across a broader panel of 18 clocks. In the factor analysis developed in this paper, we use these reported values as a lower bound on each clock's residual variance, to prevent the model from claiming a clock is more precise than its laboratory repeatability permits.

Apart from this lower bound, we make no further attempt to quantify measurement uncertainty in this paper. We proceed under the working assumption that the measurement-uncertainty component is small compared with the biological variability between individuals, an assumption that need not hold in every laboratory, on every platform, or for every clock (Kriukov et al., 2024). A formal validation of epigenetic clocks in the ISO 5725 sense would require interlaboratory comparison studies with shared reference samples, a topic addressed in a companion paper (Uhlig et al., 2026) and not pursued here.

This paper is therefore not about quantifying measurement uncertainty. It is about the variability of the clocks themselves — the structured pattern of agreement and disagreement between different clocks applied to the same individuals — and about what that structure reveals: that all these clocks have a lot in common, that they also have things that are not in common, and that both the shared and the clock-specific components have biological interpretations.

1.3 Inter-clock discordance as a window into aging biology

Different clocks yield systematically discordant age acceleration estimates for the same individual. A person may appear five years biologically older according to Horvath's clock but three years younger according to PhenoAge. The discordance is not random: some clocks correlate more strongly with each other than with others, in a pattern that turns out to reflect the biological processes they were trained to detect.

Horvath and Hannum, both first-generation clocks trained on chronological age, show high mutual correlation in their raw outputs ($r \approx 0.95$ for predicted age), while their correlation with PhenoAge is lower ($r \approx 0.90$). After residualizing on chronological age, these correlations drop substantially (e.g., $r = 0.60$ for Horvathv1–Hannum residuals in GSE40279), but the structured pattern persists: clocks based on similar methodologies agree more closely, while clocks targeting different biological constructs diverge.

1.4 Stochastic versus deterministic components of methylation change

A key insight into this structure comes from the work of Tong et al. (2024), who asked: how much of an epigenetic clock's predictive accuracy could be explained by a purely stochastic process of DNA methylation change? By simulating random methylation drift from known starting states, they demonstrated that 66–75% of Horvath's clock accuracy and up to 90% of the more precise Zhang clock could be accounted for by stochastic drift alone. In contrast, only 63% of PhenoAge's accuracy was explained by stochastic processes, suggesting that it captures a larger proportion of non-stochastic, biologically deterministic signal.

This distinction between stochastic methylation drift (the gradual, genome-wide accumulation of random methylation errors during cell division) and deterministic methylation changes (targeted alterations driven by specific biological processes such as inflammation, immune senescence, or metabolic stress) provides a conceptual framework for understanding inter-clock discordance. First-generation clocks, selected to predict chronological age, naturally favor CpG sites whose

methylation changes correlate most strongly with time — and stochastic drift, by its nature, correlates well with time. Second- and third-generation clocks, trained on mortality phenotypes or longitudinal biomarker changes, preferentially select CpG sites reflecting specific biological processes.

1.5 A factor model for inter-clock discordance

Factor analysis has been applied to multiple epigenetic clocks before: Faul et al. (2023) used it as a data-reduction technique in the Health and Retirement Study, finding that Horvath, Hannum, PhenoAge, and GrimAge load on one factor while DunedinPACE loads on a second, with both factors predicting mortality. Their two-factor solution differs from the one we obtain here in three respects that are worth making explicit. First, Faul et al. include DunedinPACE (a rate-of-aging clock) in the panel; for the reasons given in Section 1.1 we exclude it, and this changes the latent structure substantially, because the DunedinPACE-vs-rest contrast is what defines their second factor. Second, Faul et al. extract factors by an unrotated principal-components-style decomposition, without orienting the axes toward an external biological target; their factors are therefore the directions of largest variance, not directions chosen to correspond to known aging processes. Third, their analysis is outcome-oriented, compressing the panel into composite predictors of mortality, rather than construct-oriented: it does not ask what each clock measures and does not decompose inter-clock variance into biologically interpretable components. The factor structure we recover here (a stochastic-drift dimension separated from a pathway-aging dimension, both anchored in the level-based clocks) is therefore not contradicted by Faul's findings; it answers a different question, with a different panel and a different rotation, on the same residual correlation matrix.

In this paper, we take a different approach. We exploit the differential stochasticity profiles of epigenetic clocks to decompose multi-clock age acceleration into two latent components: a general epigenomic erosion factor (δ), reflecting stochastic methylation drift, and a specific biological aging factor (θ), capturing deterministic, pathway-specific alterations. We formalize this decomposition as a maximum likelihood factor model, estimated on the sample correlation matrix without any prior assumptions. The Tong et al. stochasticity estimates enter only through a post-estimation Procrustes rotation that orients the factor axes, cleanly separating model estimation (data-driven) from factor interpretation (theory-guided).

Crucially, the same decomposition operates at two complementary levels: it characterises what each clock predominantly measures (a property of the panel), and it yields for each individual a profile of three biologically interpretable components, two shared scores plus one vector of clock-specific residuals. The formal definition is given in Section 2.3; the empirical pattern across cohorts in Section 3; the per-person interpretation, with worked examples, in Section 3.5.

We apply the model to three independent cohorts and address three questions. (1) Is the loading pattern, which clocks measure predominantly erosion and which measure predominantly specific aging, consistent across populations? (2) Are the two latent components independent, or do

erosion and specific aging become coupled over the lifespan? (3) What does the per-person decomposition into (δ_i, θ_i, ε_{ik}) look like in practice, and what does it tell us about an individual that single-clock readings cannot?

2. Methods

2.1 Data preparation and correlation structure

Let y_{ik} denote the z-standardized age-acceleration residual of clock k (k = 1, …, K) for person i (i = 1, …, N), obtained by regressing each clock's output on chronological age and sex, then standardizing to mean zero and unit variance. With K = 6 clocks and N individuals, the data are summarized by the K × K sample correlation matrix S. Because the variables are z-standardized, the diagonal entries of S are all equal to 1: each clock has unit variance by construction. The off-diagonal entry S_{kl} is the Pearson correlation between the age-acceleration residuals of clocks k and l across all N individuals. A high value (e.g. S_{kl} = 0.76 for Horvathv1–PCHorvath1) means that individuals who appear accelerated by one clock also appear accelerated by the other; a low value (e.g. 0.19 for Horvathv1–GrimAge) means the two clocks disagree about who is accelerated.

The structure of S reveals what the clocks are measuring. Consider three limiting cases. First, if all clocks measured the same underlying quantity with negligible noise, every pair would correlate perfectly: S would be a matrix of ones. Second, if all clocks measured the same quantity but each with substantial independent measurement error, S would have ones on the diagonal and correlations below 1.0 on the off-diagonals, but the off-diagonal values would all be approximately equal, reflecting a single common factor attenuated by noise. Third — and this is the empirically observed case, S may show a structured pattern of high and low correlations, indicating that different clocks measure partly different biological constructs. For example, in GSE40279 the first-generation clocks (Horvathv1, PCHorvath1) correlate at 0.76 with each other but only at 0.19 with GrimAge. This asymmetry cannot be explained by a single latent factor plus noise; it requires at least two latent dimensions that load differently on different clocks.

To illustrate how a latent factor emerges from S, consider a simplified example with three clocks (A, B, C) that share a single common factor with loadings λ_A = 0.9, λ_B = 0.7, λ_C = 0.3. The model-implied off-diagonal correlations are the products of the loadings: r_{AB} = 0.9 × 0.7 = 0.63, r_{AC} = 0.9 × 0.3 = 0.27, r_{BC} = 0.7 × 0.3 = 0.21. The rank ordering of correlations directly reflects how strongly each clock loads on the shared dimension. Clocks A and B, which load strongly, correlate highly; clock C, which loads weakly, correlates poorly with both. The uniqueness of each clock — the fraction of variance not captured by the common factor — is $1 - \lambda^2$: 19% for A, 51% for B, and 91% for C. In this example, clock C is predominantly measuring something other than the shared factor. When we observe this same pattern in real data (Horvathv1 loading at 0.97 vs. GrimAge at 0.10 on the erosion factor), the interpretation is analogous: GrimAge captures variation that is largely independent of what Horvathv1 measures.

For each clock k, we partition its unit variance (after z-standardization) into a communality h_k^2 — the proportion attributable to shared aging dimensions — and a uniqueness $\psi_k = 1 - h_k^2$ — the proportion specific to that clock alone. The eigenvalue decomposition of S quantifies the dimensionality of the shared structure: each eigenvalue represents the variance explained by one linear combination of the K clocks. An eigenvalue greater than 1.0 means that this combination captures more variance than any single clock contributes on its own — the clocks are collectively tracking something. An eigenvalue below 1.0 means the opposite: no coherent shared pattern, just residual variation. The number of eigenvalues exceeding unity is one diagnostic for how many independent shared dimensions exist in the data.

How many shared dimensions are needed, and how the communalities and uniquenesses are estimated, is the subject of Section 2.2. What biological constructs these dimensions represent is the subject of Section 2.3.

Supplementary Box S1. A worked example deriving the covariance matrix of two hypothetical clocks from three independent aging dimensions is provided in Supplementary Box S1.

2.2 Maximum likelihood estimation and variance decomposition

In a two-factor model, the loading matrix Λ is identified from the data only up to orthogonal rotation: for any orthogonal matrix R (with R'R = I), the rotated loading matrix Λ* = ΛR produces the same model-implied covariance Λ*Λ*' = ΛΛ'. The data determine how many factors are needed and how much variance they explain, but not the orientation of the factor axes in the two-dimensional latent space.

We exploit this property to cleanly separate model estimation from biological interpretation. The factor model is estimated by maximum likelihood (ML) on the sample correlation matrix of the K z-standardized residuals — a closed-form optimization problem requiring no distributional assumptions on the latent variables. The ML estimates maximize the log-likelihood, or equivalently minimize the negative log-likelihood (up to a constant):

$$F = log|\Lambda\Lambda' + \Psi| + tr(S(\Lambda\Lambda' + \Psi)^{-1}) - log|S| - K$$

where S is the sample correlation matrix and Ψ = diag(ψ_1, …, ψ_k) contains the uniquenesses.

Number of factors.

The number of latent factors is determined by the eigenvalue decomposition of S. The eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k$ represent the variance explained by each principal component of the correlation matrix. Under the Kaiser criterion, factors with eigenvalues exceeding 1.0 are retained; however, this heuristic is known to undercount factors in structured correlation matrices with few variables. We therefore supplement it with a likelihood ratio test (LRT) comparing the one-factor and two-factor models, and with AIC and BIC. The absolute goodness-of-fit is assessed by the Bartlett-corrected chi-square test and the root mean square of residual correlations (RMSR). Model selection results for each cohort are reported in Section 3.

Estimation of clock-specific uniqueness.

The eigenvalues of S determine the number of factors, but the eigenvectors of S are not the final factor loadings. The ML factor model imposes a specific structure on S: S ≈ ΛΛ' + Ψ, where Λ is the K × 2 loading matrix and Ψ = diag(ψ_1, …, ψ_k) is constrained to be diagonal. This diagonal constraint is the central methodological choice: it requires that all correlation between clocks is mediated through the two latent factors. Any variance not captured by the factors must be entirely clock-specific, no residual correlations between clocks are permitted. The ML optimization finds Λ and Ψ jointly by minimizing the discrepancy between the model-implied covariance ΛΛ' + Ψ and the observed correlation matrix S. For each clock k, the uniqueness ψ_k = 1 − h_k^2 (with two-factor communality h_k^2 = $\lambda_{k\delta}^2$ + $\lambda_{k\theta}^2$) quantifies the fraction of that clock's variance not attributable to any shared dimension. The adequacy of the diagonal assumption is assessed by the RMSR.

Technical repeatability constraint.

Unconstrained ML factor analysis can produce boundary solutions (Heywood cases) in which the estimated uniqueness of a clock approaches zero, implying that the clock has no measurement error, a physically unrealistic result. We prevent this by imposing a clock-specific lower bound on each uniqueness: $\psi_k \geq (\sigma_{tech,k} / s_k)^2$, where $\sigma_{tech,k}$ is the within-sample technical repeatability standard deviation of clock k (obtained from replicate measurements of the same DNA sample) and s_k is the standard deviation of the age-acceleration residual. This constraint ensures that the model cannot claim a clock is more precise than its laboratory repeatability permits. For four of the five year-scale clocks, we use σ_{tech} values from Higgins-Chen et al. (2022): approximately 1.0 years for Horvathv1 and PCHorvath1, 1.2 years for Hannum, and 2.0 years for PhenoAge. For GrimAgeV1 and DNAmTL, where published replicate data are not available, a conservative default of ψ_{min} = 0.005 is used. The constrained optimization is implemented via reparametrization: ψ_k = $\psi_{min,k}$ + exp(γ_k), where γ_k is the unconstrained parameter estimated by L-BFGS-B. This formulation preserves the smooth optimization landscape of standard ML estimation while enforcing the lower bound.

The σ_{tech} values from Higgins-Chen et al. were obtained under specific laboratory conditions (Illumina 450K, single laboratory) and may not be directly transferable to other studies. However, the constraint is a lower bound, so the direction of potential error is conservative: if the true technical SD in GSE40279 is larger than assumed, the constraint is non-binding and the results are unaffected; if it is smaller, the constraint is slightly too strict, yielding a marginally overestimated minimum uniqueness. In practice, only Horvathv1 is at the constraint in the two large cohorts. For all other clocks, the estimated uniquenesses lie well above their respective ψ_{min} values (by factors of 3 to 140), rendering the exact choice of σ_{tech} immaterial for the factor structure reported here.

Rotation-invariant quantities.

The ML solution determines, for each clock k, the communality h_k^2 (the fraction of variance shared with the latent factors) and the uniqueness ψ_k = 1 − h_k^2 (the fraction not explained). These

quantities are invariant under any orthogonal rotation of the factor axes: for any rotation matrix R with R'R = I, the row norms of the loading matrix are preserved ($\|\Lambda_k R\|^2 = \|\Lambda_k\|^2$). What does depend on the rotation is the partition of each clock's communality between the two factors — and hence the labels "erosion" and "specific".

Implementation.

The ML factor model was estimated using scipy.optimize (L-BFGS-B), with PCA-based initialization. Model adequacy was assessed by Bartlett's test of sphericity and the RMSR. Bootstrap confidence intervals for loadings were computed from 500 resamples with Procrustes alignment.

2.3 Factor rotation and biological interpretation

Rotational indeterminacy

For any K × 2 loading matrix Λ and any 2 × 2 orthogonal rotation matrix R (satisfying R'R = RR' = I), the rotated loadings Λ* = ΛR yield identical model-implied covariance: Λ*Λ*' = ΛRR'Λ' = ΛΛ'. In two dimensions, R is parameterized by a single angle φ. The data cannot distinguish between any two rotations of the factor space. All rotations produce identical fit statistics, identical communalities h_k^2 = $\lambda_{\delta k}^2$ + $\lambda_{\theta k}^2$, and identical uniquenesses ψ_k = 1 − h_k^2. What changes is the partition of each clock's communality between the two factors and hence the biological labels attached to them.

Procrustes rotation

To orient the axes, we define a K × 2 target matrix T, where column 1 contains the Tong stochasticity coefficients π_k (normalized to unit length) and column 2 contains (1 − π_k) (likewise normalized). The orthogonal Procrustes problem finds the rotation R that minimizes the Frobenius norm:

$$R^* = arg\,min \|\Lambda R - T\|_F \quad subject\ to\ R'R = I$$

The Frobenius norm ‖A‖_F is the square root of the sum of all squared elements:

$$\|A\|_F = \sqrt{\sum_{k,j} {a_{kj}}^2}$$

Minimizing $\|\Lambda R - T\|_F$ therefore amounts to finding the rotation R that makes the rotated loading matrix ΛR as similar as possible to the target T in a least-squares sense. The solution is $R^* = VU'$, where $U\Sigma V'$ is the singular value decomposition of $\Lambda' T$. This rotation orients the first factor axis as closely as possible toward the theoretically expected stochastic-drift direction. Because R is orthogonal, all rotation-invariant quantities (communalities, uniquenesses) are preserved. The rotation affects interpretation (what δ and θ represent biologically), not the model fit.

The two-component model

After rotation, the factor model for individual i and clock k takes the explicit form:

$$y_{ik} = \lambda_{k\delta} \cdot \delta_i + \lambda_{k\theta} \cdot \theta_i + \varepsilon_{ik}$$

where δ_i is the individual's erosion score (stochastic methylation drift), θ_i is the individual's specific aging score (pathway-driven biological aging), $\lambda_{k\delta}$ and $\lambda_{k\theta}$ are the clock-specific loadings on each dimension, and ε_{ik} is the clock-specific residual with variance ψ_k. The erosion and specific dimensions are orthogonal by construction (Corr(δ, θ) = 0 in the population model), though the empirical Thompson scores may show a small sample correlation. The loading $\lambda_{k\delta}$ quantifies how strongly clock k reflects erosion: Horvathv1 (λ_δ = 0.97) is almost a pure erosion indicator, while GrimAge (λ_δ = 0.10) is nearly orthogonal to it. Conversely, $\lambda_{k\theta}$ quantifies the clock's sensitivity to specific aging pathways: GrimAge (λ_θ = 0.53) loads most strongly here, consistent with its training on mortality-associated plasma protein surrogates. The variance of each clock is fully decomposed: %Erosion = $\lambda_{\delta k}^2$, %Specific = $\lambda_{\theta k}^2$, %Residual = ψ_k, summing to 1.0 for each clock.

In plain terms, the three components of the per-person model are: δ — the shared stochastic-drift score, what first-generation clocks see most clearly; θ — the shared pathway-driven aging score, what mortality-trained clocks see most clearly; and ε — the clock-specific residual, everything in a given clock's reading that is not captured by δ and θ. The δ and θ scores are expected to be primarily biological. The residual ε contains, at minimum, within-laboratory technical noise (bounded from below by σ_tech,k; Section 2.2), but also any clock-specific biological signal not shared with the panel. Between-laboratory and between-platform reproducibility contribute additional variance components that are not quantified by the present within-platform analysis; see Section 4.5.

Supplementary Box S2. A worked example showing how three hypothetical individuals with distinct aging profiles produce different clock readings is provided in Supplementary Box S2.

Factor scores for each individual are computed via Thompson (regression) scoring, which yields the best linear unbiased predictor of the latent variables given the observed clock residuals and is more robust under heteroscedasticity than the Bartlett alternative:

$$f_i = \Lambda'(\Lambda\Lambda' + \Psi)^{-1} y_i$$

with reliability matrix (the Guttman reliability of the regression scores)

$$Var(f) = \Lambda'(\Lambda\Lambda' + \Psi)^{-1}\Lambda$$

and corresponding estimation-error covariance matrix

$$Var(f - f) = I - \Lambda'(\Lambda\Lambda' + \Psi)^{-1}\Lambda$$

An important property of Thompson scores is that they are systematically shrunk toward zero relative to the true latent values: the scoring coefficients are optimal in the mean squared error sense but produce attenuated estimates, particularly for individuals with extreme factor values.

Three components per individual

For each person i, the fitted model yields three sets of quantities. (i) The erosion score δ_i — a single number, in z-units, summarising the individual's position on the stochastic-drift dimension as integrated across all six clocks. (ii) The pathway-aging score θ_i — analogous, on the specific-aging dimension. (iii) The clock-specific residuals $\varepsilon_{ik} = y_{ik} - \lambda_{k\delta} \cdot \delta_i - \lambda_{k\theta} \cdot \theta_i$ for k = 1, …, 6 — six numbers per person, one per clock, capturing what each clock sees beyond the two shared dimensions. The pair (δ_i, θ_i) places the individual on the two-dimensional latent aging plane that is shared across clocks; the residual vector (ε_{i1}, …, ε_{i6}) is the clock-by-person interaction. A large positive $\varepsilon_{i,GrimAge}$ in a person with average δ and θ, for instance, identifies an individual whose GrimAge reading is elevated for reasons not shared with the other clocks — a finding that motivates targeted follow-up rather than panel-averaged action. The per-person profile (δ_i, θ_i, ε_{i1}, …, ε_{i6}) is the basic output of the framework for any individual analysis.

Table 1. Tong stochasticity coefficients used as Procrustes rotation target.

Clock	π_{stoch}	Source	Generation
Horvathv1	0.70	Tong et al.	1st
Hannum	0.72	Assigned	1st
PhenoAge	0.63	Tong et al.	2nd
GrimAgeV1	0.55	Assigned	2nd
PCHorvath1	0.65	Assigned	1st (PC)
DNAmTL	0.60	Assigned	Telomere

Tong et al. (2024) reported stochasticity coefficients for Horvathv1 and PhenoAge. For the remaining four clocks, π_{stoch} values were assigned based on clock generation and training characteristics: first-generation clocks trained on chronological age (Hannum, PCHorvath1) were assigned values near 0.70, mortality-trained clocks (GrimAgeV1) near 0.55, and DNAmTL near 0.60. These values are external prior assumptions about clock design, not estimates derived from the analysis data. Crucially, the Procrustes target affects only the orientation of the factor axes (which axis is labeled "erosion" and which "specific"), not the communalities, uniquenesses, or any other rotation-invariant quantity.

Covariate analyses

Covariate associations of δ and θ were assessed via nested linear regression models. For GSE40279, models included chronological age, ln(age), sex, and ethnicity (Hispanic vs. Caucasian). The ln(age) term captures the established logarithmic relationship between age and DNA methylation (Horvath, 2013). Stratified analyses by ethnicity were performed to assess effect modification. For GSE51032, associations with cancer status (ICD-10 codes), cancer type, time to diagnosis, and age at menarche were examined. All analyses used two-sided tests with α = 0.05.

2.4 Data

We analyzed three publicly available DNA methylation datasets from the Gene Expression Omnibus (GEO):

GSE40279 (Hannum et al., 2013): 656 whole blood samples from a US population (426 Caucasian, 230 Hispanic), age range 19–101 years, Illumina 450K array. This is the primary analysis cohort.

GSE51032 (EPIC-Italy): 845 whole blood samples from an Italian prospective cohort, age range 35–80 years, Illumina 450K array. Includes pre-diagnostic cancer data (305 incident cases, 540 controls), age at menarche, and time to diagnosis.

GSE41169 (Dutch Blood): 95 whole blood samples from a Dutch population, age range 18–65 years, Illumina 450K array.

For each dataset, we computed six epigenetic clocks using the Biolearn Python library (Ying et al., 2024): Horvathv1, Hannum, PhenoAge, GrimAgeV1, PCHorvath1, and DNAmTL. All six clocks are computed from the same Illumina 450K methylation array; the multi-clock panel requires no additional laboratory cost or sample material beyond the standard assay. The sign of DNAmTL was reversed so that positive values indicate accelerated telomere attrition, aligning the direction of all clocks. Age acceleration residuals were obtained by linear regression of each clock's output on chronological age and sex, followed by z-standardization.

For GrimAgeV1 and DNAmTL, this step is effectively a cohort-specific linear calibration against chronological age. Unlike the chronological-age-trained clocks (Horvathv1, Hannum, PCHorvath1), these two outputs are not anchored to the age scale by training: GrimAgeV1 is trained on mortality-associated plasma protein surrogates, and DNAmTL on telomere length (kilobases). Their raw output therefore carries a clock- and cohort-specific intercept and slope against chronological age. Without this calibration, the resulting residuals would retain an age-dependent systematic offset that would propagate into the correlation matrix and contaminate the latent factor structure. Regressing each clock on age (and sex) within each cohort removes this offset and brings all six clocks onto a common centered, z-scaled residual scale. Because the calibration is performed independently per cohort, cohort-level differences in mean and slope (population composition, laboratory, batch) are absorbed into the regression constants rather than into the latent factors.

3. Results

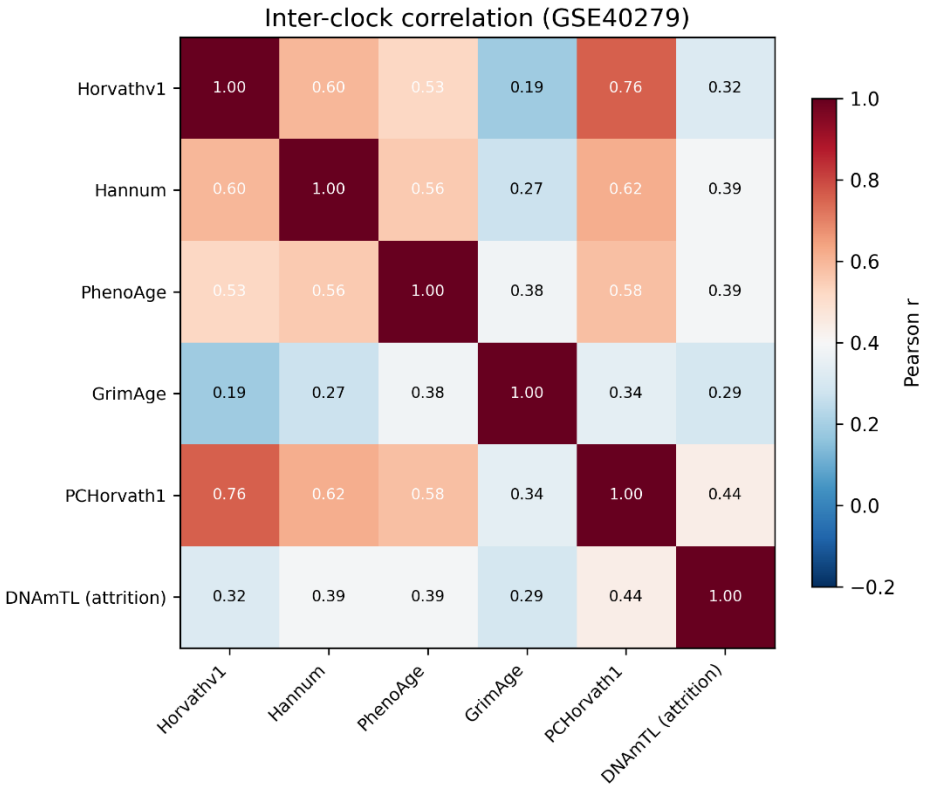
All primary analyses were conducted on GSE40279 (n = 656). Replication in GSE51032 (n = 845) and exploratory analysis in GSE41169 (n = 95) are reported in Section 3.3.

3.1 Inter-clock discordance

Before fitting the factor model, we characterize the raw discordance between clocks. The sample correlation matrix of the six z-standardized age-acceleration residuals in GSE40279 shows that

first-generation clocks (Horvathv1, PCHorvath1) correlate strongly with each other (r = 0.76) but weakly with mortality-trained clocks (Horvathv1 vs. GrimAgeV1: r = 0.19). This structured pattern of agreement and disagreement motivates the factor decomposition that follows.

To make the discordance tangible, we convert each pairwise correlation into an expected absolute discrepancy in years. For clocks k and l with EAA standard deviations s_k and s_l and correlation r_{kl}, the expected discordance is $E|EAA_k - EAA_l| = \sqrt{(s_k^2 + s_l^2 - 2\,r_{kl} \cdot s_k \cdot s_l)} \cdot \sqrt{(2/\pi)}$. For the five year-scale clocks, expected pairwise discrepancies range from approximately 3 years (Horvathv1 vs. PCHorvath1) to over 6 years (Horvathv1 vs. GrimAge). An individual reported as "5 years accelerated" by one clock may simultaneously appear "2 years decelerated" by another.



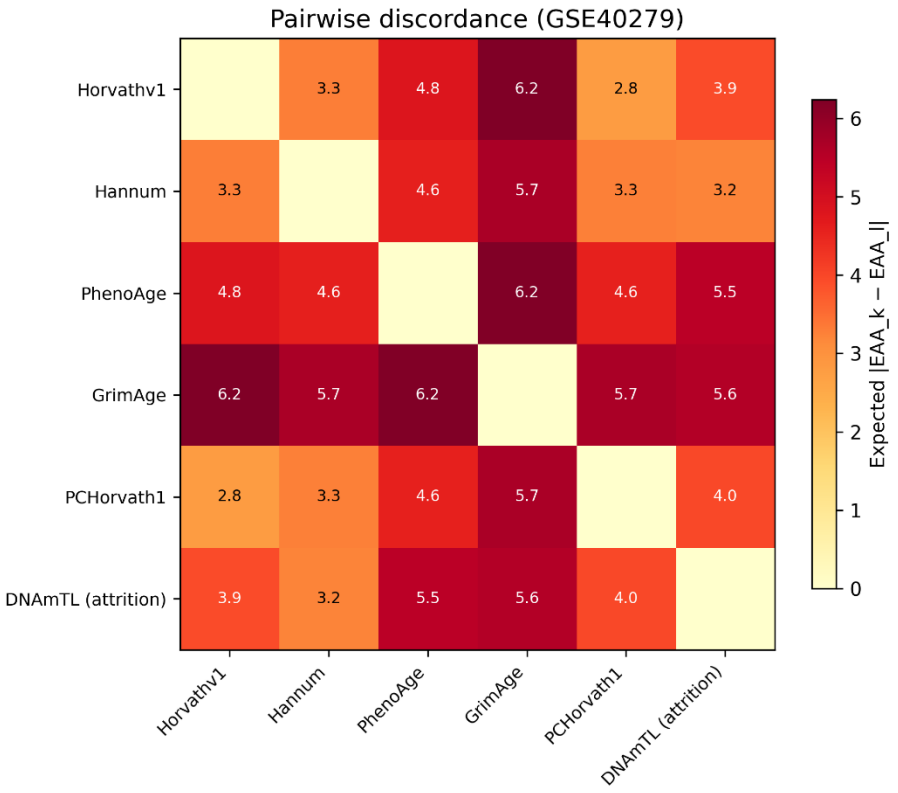


Figure 1. Inter-clock discordance in GSE40279. Left: Pearson correlation matrix of z-standardized age-acceleration residuals, showing structured agreement (Horvathv1–PCHorvath1: r = 0.76) and disagreement (Horvathv1–GrimAge: r = 0.19). Right: Expected pairwise absolute discordance in original units, ranging from 2.8 years (Horvathv1 vs. PCHorvath1) to 6.2 years (Horvathv1 vs. GrimAge).

3.2 Factor structure and variance decomposition

The two-factor ML model converged successfully. The first eigenvalue of the correlation matrix (3.29) substantially exceeds unity; the second (0.91) falls marginally below the Kaiser threshold but is decisively supported by the likelihood ratio test ($\Delta\chi^2$ = 101.0, Δdf = 5, p < 10^{-19}). Both AIC (48 vs. 137) and BIC (129 vs. 191) favor two factors. Together the two factors account for 70.0% of the total variance, specifically, the sum of the two retained eigenvalues divided by the trace of the K=6 correlation matrix, (3.29 + 0.91) / 6 = 0.70. This is a property of the eigenvalue spectrum of the correlation matrix as a whole, not an average of per-clock communalities; the latter quantity, computed as the unweighted mean of h_k^2 across the six clocks in Table 2, equals approximately 0.61, somewhat lower because not all the variance captured by the second factor is allocated to communalities in the constrained ML solution. The absolute goodness-of-fit test rejects the two-factor model as a perfect fit (Bartlett-corrected χ^2 = 11.9, df = 4, p = 0.018), but the practical significance of this misfit is negligible: the RMSR of 0.018 means the average unexplained off-diagonal correlation is less than 0.02, well below sampling variability at this sample size. After

Procrustes rotation toward the Tong target, the loading structure revealed a clear separation (Table 2).

Table 2. Factor loadings and variance decomposition (GSE40279, ML-FA).

Clock	λ_δ	λ_θ	%Erosion	%Specific	%Residual	SE(λ_δ)
Horvathv1	+0.965	+0.167	93.2%	2.8%	4.0%	0.037
PCHorvath1	+0.698	+0.494	48.7%	24.4%	26.9%	0.025
Hannum	+0.538	+0.493	29.0%	24.3%	46.7%	0.035
PhenoAge	+0.443	+0.586	19.6%	34.4%	46.0%	0.039
GrimAgeV1	+0.104	+0.526	1.1%	27.7%	71.2%	0.039
DNAmTL	+0.243	+0.516	5.9%	26.6%	67.5%	0.042

Erosion-dominant clocks: Horvathv1 (λ_δ = 0.97, 93% erosion variance) and PCHorvath1 (λ_δ = 0.70, 49%) loaded predominantly on the erosion factor.

Specific-dominant clocks: GrimAgeV1 (λ_θ = 0.53, 28% specific) loaded exclusively on the specific factor, with near-zero erosion contribution.

Mixed clocks: Hannum and PhenoAge loaded on both factors. Hannum showed a slight erosion dominance (29% vs. 24%), while PhenoAge was specific-dominant (34% vs. 20%).

DNAmTL: After sign reversal, DNAmTL (telomere attrition) loads positively on both factors, indicating that both erosion and specific aging are associated with accelerated telomere shortening.

Factor score estimation precision was SE_δ = 0.24 and SE_θ = 0.58, reflecting the greater number of clocks contributing to the erosion dimension.

Figure 2. Variance decomposition by clock. Left: GSE40279 (primary). Center and right: replication cohorts. Blue: erosion (δ), orange: specific (θ), grey: residual.

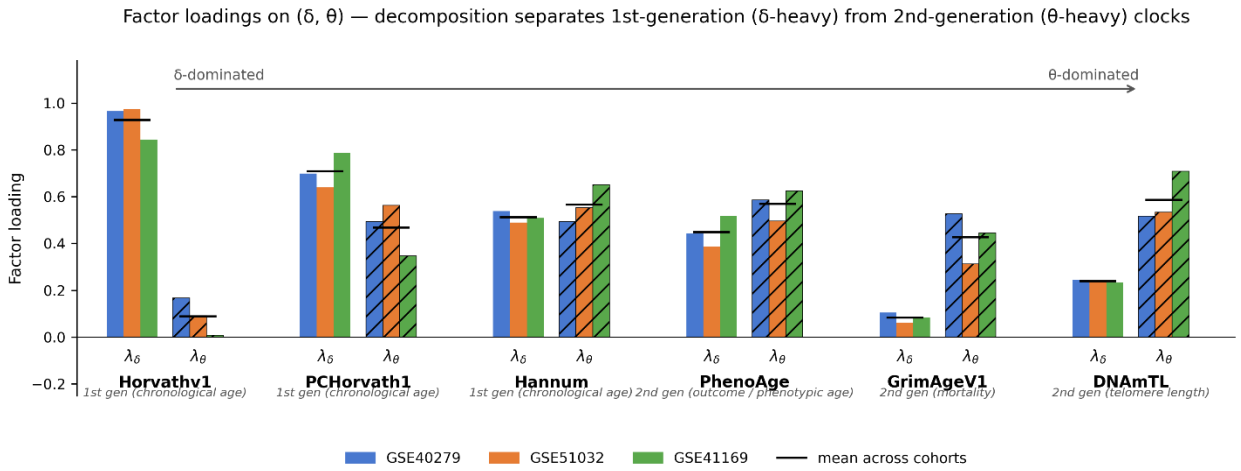


Figure 2 (loadings). Factor loadings (λ_δ, λ_θ) per clock, GSE40279. First-generation clocks (Horvathv1, PCHorvath1, Hannum) load primarily on δ; mortality-trained clocks (GrimAge, PhenoAge) and DNAmTL load primarily on θ.

3.3 Replication and generalization across cohorts

The two-factor model was applied independently to GSE51032 (n = 845) and GSE41169 (n = 95). GSE51032 serves as a full replication cohort; GSE41169, with its small sample size and narrow age range, is treated as an exploratory third dataset where individual loading estimates carry substantial uncertainty (several bootstrap CIs crossing zero). In both cohorts, the likelihood ratio test decisively rejected the one-factor model: $\Delta\chi^2$ = 62.0 (p = 4.7 × 10^{-12}) for GSE51032 and $\Delta\chi^2$ = 31.7 (p = 6.7 × 10^{-6}) for GSE41169. Both AIC and BIC favored two factors in each cohort. Notably, GSE41169 is the only cohort where the second eigenvalue exceeds unity (λ_2 = 1.11), satisfying the Kaiser criterion directly despite the small sample size.

The loading structure replicated consistently across all three cohorts (Figure 2). The ranking of clocks by erosion content was identical in every cohort: Horvathv1 (71–95% erosion) and PCHorvath1 (41–62%) loaded predominantly on the erosion factor; GrimAgeV1 (0.4–1%) loaded exclusively on the specific factor; Hannum and PhenoAge occupied intermediate positions; and DNAmTL (sign-reversed) loaded predominantly on the specific dimension (27–50%). The Horvathv1 uniqueness was at the technical repeatability constraint in both large cohorts (ψ = 0.04 in GSE40279, ψ = 0.04 in GSE51032) but not in GSE41169 (ψ = 0.29), where the smaller sample size yields greater estimation uncertainty. GrimAgeV1 showed progressively higher uniquenesses across cohorts (0.71, 0.90, 0.80), suggesting that its private signal — variance not shared with the other clocks in the panel — is consistently large but variable in magnitude.

The overall δ–θ correlation was weak and positive in all cohorts: r = 0.10 (GSE40279), r = 0.07 (GSE51032), r = 0.16 (GSE41169). Age-decade analysis showed a significant coupling only in the 61–70 decade of GSE40279 (r = 0.24, p = 0.002) and, unexpectedly, in the 51–60 decade of GSE51032 (r = 0.17, p = 0.0006). GSE41169, with only two evaluable age groups, showed no significant coupling.

Regression analyses in GSE51032 and GSE41169 showed no significant effects of age, ln(age), or sex on δ or θ (all p > 0.19). This is expected given the narrower age ranges and smaller sample sizes, which limit statistical power for detecting non-linear age effects. The RMSR values (0.031 in GSE51032, 0.046 in GSE41169) are slightly higher than in GSE40279 (0.018), consistent with noisier estimation in the smaller or more homogeneous cohorts, but remain well within acceptable ranges for a two-factor model.

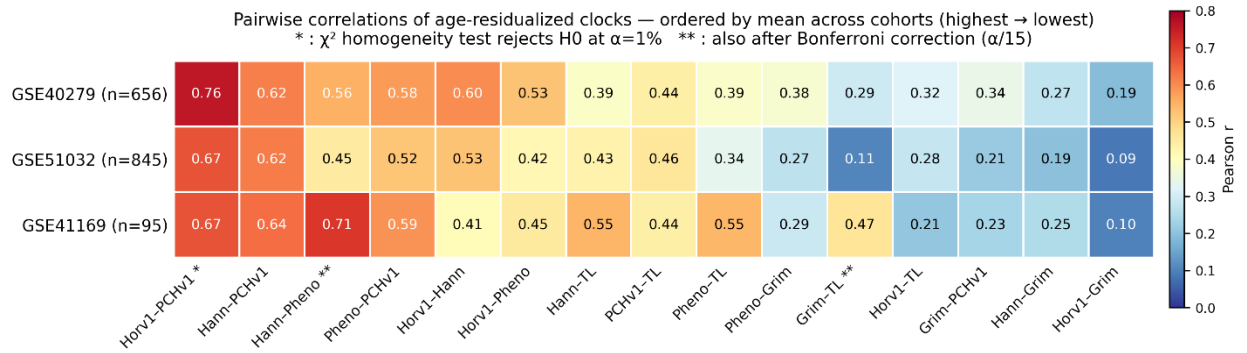


Figure 2b. Cohort consistency of the 15 pairwise inter-clock correlations. Cross-cohort correlation of the 15 r-values is r = 0.95 between the two large cohorts; 13 of 15 pairs are homogeneous across cohorts (χ², α = 1 %, Bonferroni).

3.4 Clock-specific residual variability

The two-factor decomposition partitions the variance of each clock's age-acceleration residual into three components: the share attributable to erosion ($\lambda^2_{\delta k}$), the share attributable to specific aging ($\lambda^2_{\theta k}$), and the share that is clock-specific (ψ_k). Table 2 reports these percentages. The clock-specific share ψ_k is what remains after the two latent dimensions are removed: it captures whatever variability in clock k is not shared with any of the other five clocks in the panel.

The interpretation of ψ_k is not unique. Conceptually, it contains at minimum the within-laboratory technical noise (bounded from below by $\sigma_{tech,k}$, see Section 2.2), variation from cell-type composition shifts in bulk blood, and any clock-specific biological signal not captured by erosion or specific aging. Recent work by Sehgal et al. (2025) makes this decomposition more concrete by distinguishing *technical* reliability (test–retest on the same blood sample) from *biological* reliability (repeated collections from the same individual over short timeframes, capturing day-to-day fluctuations from meals, stress, or circadian variation). In their analysis of 18 clocks, most achieve excellent technical reliability but show substantially worse biological reliability, meaning a large fraction of what we capture in ψ_k is plausibly short-timescale biological variation rather than laboratory noise. This is consistent with our interpretation that ψ_k is not 'measurement error' in the narrow sense. The data available here do not allow separating these components, and we do not attempt to interpret $\sqrt{\psi_k} \cdot s_k$ as a formal measurement uncertainty in the ISO 5725 sense, that would require interlaboratory data not included in this study (see Section 4.5). What ψ_k does provide, unambiguously, is a ranking of how much of each clock's output is unshared with the others in the panel.

By that ranking, Horvathv1 sits at the technical-repeatability constraint (ψ = 0.04), with virtually all of its variance explained by the two shared factors. PCHorvath1 (ψ = 0.27), Hannum (ψ = 0.47), and PhenoAge (ψ = 0.46) occupy intermediate positions. GrimAgeV1 (ψ = 0.71) has by far the largest clock-specific share among the year-scale clocks: most of what it measures is independent of the other five. This is consistent with its training on mortality-associated plasma protein surrogates, which capture biological processes not strongly weighted by chronological-age or telomere-trained clocks. Importantly, a large ψ_k does not mean a clock is noisy — it means the clock's signal is private to that clock, and the panel cannot cross-validate it.

Table 3. Standard deviation of age-acceleration residuals s_k by age cohort (GSE40279). Descriptive only; reflects total residual variability, not a decomposed component.

Clock	Unit	Overall	≤30	31–50	51–60	61–70	>70	r(age)
Horvathv1	years	5.0	3.7	4.2	5.0	5.4	5.0	0.918
PCHorvath1	years	5.1	3.2	3.9	4.5	5.8	5.4	0.887
Hannum	years	4.1	3.8	3.7	4.3	4.3	3.9	0.946
PhenoAge	years	6.9	5.7	5.6	6.4	7.5	7.2	0.852
GrimAgeV1	years	7.0	6.1	6.5	7.6	7.4	6.7	0.798
DNAmTL	kb	0.18	0.19	0.15	0.18	0.20	0.19	−0.758

s_k = SD of age-acceleration residuals (EAA) after regressing clock output on chronological age and sex. r(age) = Pearson correlation of raw clock output with chronological age. DNAmTL is sign-reversed (attrition); its SD is in kilobases.

3.5 Per-person decomposition: three contrasting profiles

To make the per-person interpretation concrete, we contrast three real individuals from GSE40279. Person A (sample #346, female, age 52, Hispanic) is erosion-dominant. Person B (sample #219, female, age 69, Caucasian) carries a comparably strong erosion signal but essentially zero pathway-aging component and small clock-specific residuals — a clean shared-signal profile. Person C (sample #632, male, age 69, Caucasian) has negligible erosion, moderate pathway-aging, and two large clock-specific residuals concentrated in the mortality-trained clocks. The three were not constructed; they were selected from the cohort to illustrate the three qualitatively distinct ways in which the (δ, θ, ε) decomposition can differ across individuals.

Figure 3 (panels A–C). Per-person decomposition for three real GSE40279 individuals. Each three-panel block shows (i) the position on the (δ, θ) plane in years, (ii) the six clock-

specific residuals ε in their native units, and (iii) the additive split of each clock reading into δ-contribution, θ-contribution, and ε. Persons A (#346), B (#219), C (#632).

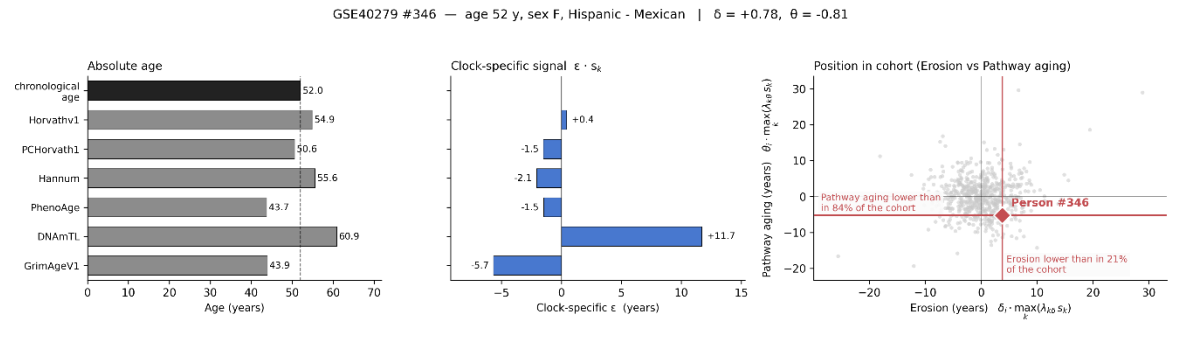


Figure 3A. Person A — sample #346 (F, 52, Hispanic). δ = +0.78, θ = −0.81.

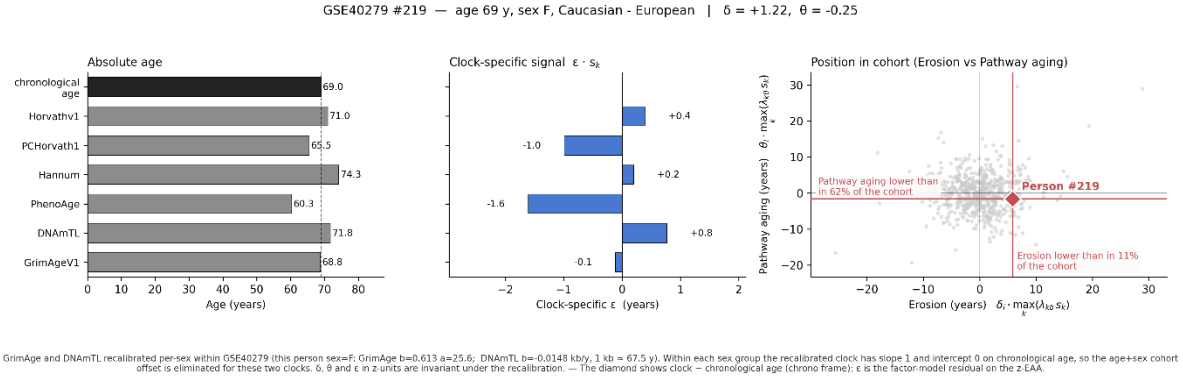


Figure 3B. Person B — sample #219 (F, 69, Caucasian). δ = +1.22, θ = −0.25; all |ε| < 1.7 years across the six clocks.

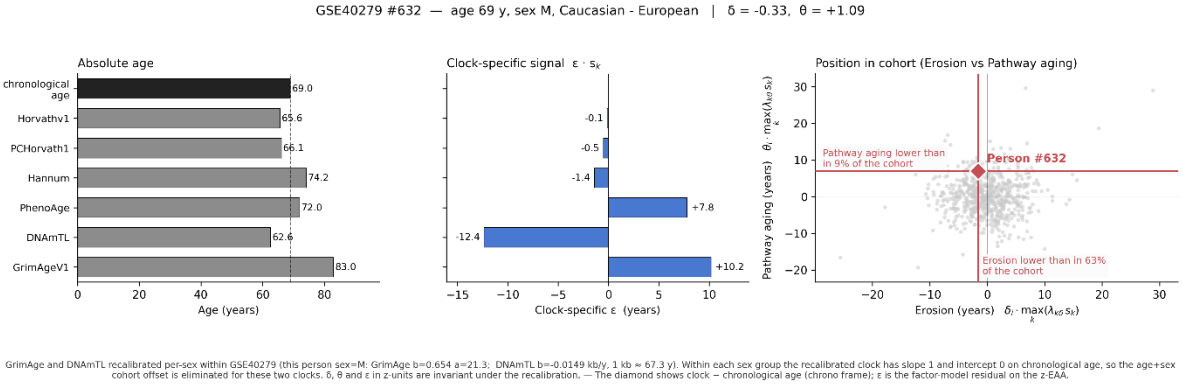


Figure 3C. Person C — sample #632 (M, 69, Caucasian). δ = −0.33, θ = +1.09; PhenoAge ε ≈ +8.6 y and GrimAge ε ≈ +9.4 y dominate.

Person A (#346): erosion-dominant with a moderate negative pathway component. δ = +0.78 places this individual above average on the stochastic-drift dimension; θ = −0.81 places them below average on pathway-aging. Reading across the clocks (Figure 3A): the δ-contribution is large and positive for the erosion-loaded clocks (Horvathv1 +3.8 y, PCHorvath1 +2.8 y), while the θ-contribution is negative and grows in clocks that load on pathway aging (GrimAge −3.0 y, PhenoAge −3.3 y). The two contributions partially cancel for the chronological-age-trained clocks

but reinforce each other for GrimAge, whose final reading is EAA ≈ −7 years. The same person, on the same blood sample, looks accelerated by Horvath and decelerated by GrimAge — and Figure 3A shows this is not an inconsistency but a consequence of the two clocks weighting the same two underlying processes differently.$\delta_i \theta_i$

Person B (#219): a clean shared-signal profile. δ = +1.22 (strong erosion), θ = −0.25 (negligible pathway-aging), and all six clock-specific residuals are small (|ε| < 1.7 years; Figure 3B). The δ-contribution alone drives the panel: Horvathv1 reads +5.3 y, PCHorvath1 +1.6 y, GrimAge +0.6 y — the cross-clock spread of roughly 5 years is generated almost entirely by the loadings multiplied by a single positive δ. Because θ ≈ 0, the second axis is silent; because the residuals are small, the clock-specific axis is silent. The (δ, θ, ε) profile collapses to one number (δ), and the panel reads consistently elevated across all clocks — exactly the pattern that a single-clock or panel-mean approach would describe correctly.

Person C (#632): the opposite case. δ = −0.33 (slightly below average), θ = +1.09 (well above average), and two of the six clock-specific residuals are very large: PhenoAge ε ≈ +8.6 y and GrimAge ε ≈ +9.4 y, while the other four clocks have residuals of one year or less (Figure 3C). The shared scores alone would predict EAA values of roughly +2 to +4 years for the mortality-trained clocks; the observed values are +12 and +13. Two clocks are seeing something in this individual that none of the other clocks see, and that the shared erosion–pathway model does not predict. The fact that the signal localises specifically to PhenoAge and GrimAge — rather than spreading across the panel — suggests an outcome-relevant biological process captured by their mortality-oriented training but not represented in the two shared dimensions. A report that stopped at (δ, θ) would have missed this case entirely; the clock-specific residuals are what flag it.

Three observations follow from this comparison. First — visible by comparing persons A and B — the shared scores (δ, θ) determine the structural part of the multi-clock signal: the apparent discordance across clocks is generated by the loading matrix acting on (δ, θ), not by any extra property of the person. Second — visible from person C — the residual vector ε is not a passive remainder. When it is large, it carries information that the shared dimensions cannot capture, and it can localise to specific clocks in a way that is biologically interpretable (here: the mortality-trained pair). Third, the three components are mechanistically distinct: shared erosion (δ), shared pathway aging (θ), and clock-specific residuals (ε) can each dominate independently, and the per-person interpretation depends on knowing which of the three is doing the work in any given individual.

A fourth observation concerns the panel mean as a summary statistic. For person A, the mean EAA across the five year-scale clocks is −0.7 years — close to zero, despite the fact that this person carries substantial signal on both shared dimensions (δ in the upper quintile of the population, θ in the bottom decile). The panel mean is near zero precisely because δ and θ point in opposite directions and partially cancel when averaged across clocks that load on both. A single-number summary — whether a single-clock reading or an unweighted multi-clock composite — necessarily projects the per-person profile onto one axis, and information orthogonal

to that axis is lost in the projection. The decomposition recovers what the mean averages away: person A is not biologically unremarkable, even though their panel mean would suggest so.

3.6 Methodological validation: the Hispanic effect localises to erosion

A natural test of whether the decomposition is biologically meaningful is to ask whether known population-level differences in epigenetic aging localise to specific factors. The slower epigenetic aging of Hispanic vs. Caucasian individuals — sometimes referred to as the "Hispanic paradox" of epigenetic aging — has been documented repeatedly in the literature, first by Horvath et al. (2016) using intrinsic and extrinsic age acceleration on the Horvath clock, and subsequently in larger and more diverse panels (Faul et al., 2023, in the Health and Retirement Study; see also recent meta-analyses of epigenetic age acceleration by world population). What is generally observed is that first-generation clocks (Horvath, Hannum) show slower aging in Hispanics, while second-generation clocks (PhenoAge, GrimAge) show smaller or absent effects.

This pattern is naturally explained by our decomposition: under the Tong-guided Procrustes rotation that defines what we call the "erosion" and "specific" axes, if the Hispanic effect operates predominantly on the stochastic-drift dimension δ, then it should appear most strongly in clocks that load heavily on δ (Horvathv1, PCHorvath1) and only weakly in clocks that load predominantly on θ (GrimAge). Regression of the two factor scores on ethnicity in GSE40279 (controlling for chronological age, ln(age), and sex) confirms this prediction: δ is reduced in Hispanic vs. Caucasian individuals by $\beta = -0.44$ standard deviations ($p = 1.0 \times 10^{-6}$), while θ shows no significant effect ($\beta = -0.10$, $p = 0.18$). The published cross-clock pattern is therefore not a curious property of certain clocks but a mechanistic localisation of the underlying effect: under the orientation chosen here, the Hispanic advantage in epigenetic aging — whatever its biological or environmental cause — appears to operate on the dimension that captures stochastic methylation drift, not on the dimension that captures pathway-driven aging. A different rotation (e.g. Varimax) would partition the same total variance differently between the two factors and could yield a different verbal interpretation of which axis carries the effect; what does not change under rotation is the underlying observation that the effect is concentrated in clocks loading heavily on the first principal component of the residual correlation matrix.

We emphasise that this is a methodological validation, not a new biological claim. The cross-clock differences in the Hispanic effect were already established in the literature; what the decomposition adds is a compact explanation in terms of two latent dimensions rather than six clock-specific effects. As a check on the framework, it is informative that the decomposition recovers a known signal and assigns it to a specific factor with the expected pattern. As a substantive contribution about Hispanic epigenetic aging, the finding is interpretive and subject to the well-known confounders of cross-sectional ethnicity comparisons (socioeconomic status, immigration generation, sample selection); we do not pursue those interpretations here.

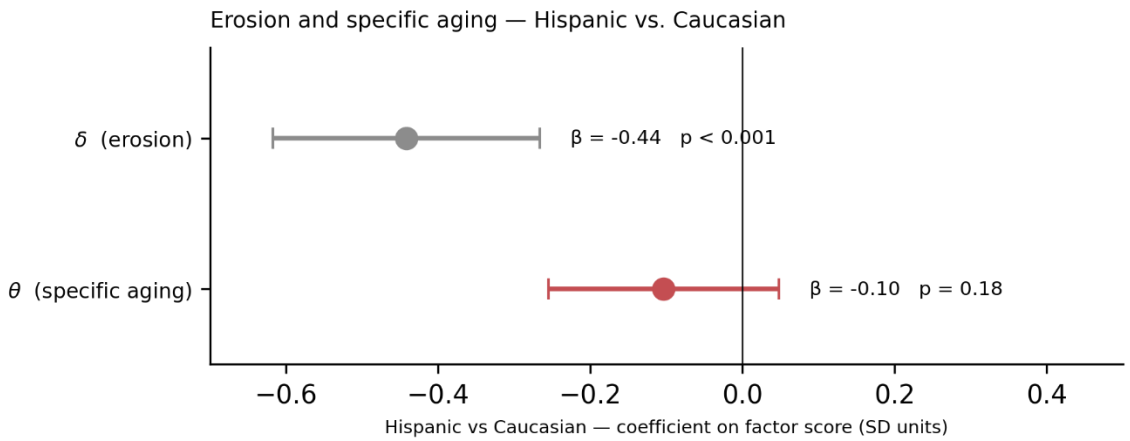


Figure 4. The Hispanic effect localises to δ. Regression of (δ, θ) on ethnicity in GSE40279 (n = 656), adjusted for age, ln(age) and sex. β_δ = −0.44 (p = 1 × 10⁻⁶); β_θ = −0.10 (p = 0.18, n.s.).

3.7 External verification: θ tracks the DunedinPoAm38 pace-of-aging clock

The two shared dimensions identified by the factor model are estimated from within the six-clock panel. To check whether they correspond to biologically meaningful constructs rather than internal axes of the correlation matrix, we relate (δ, θ) to an external pace-of-aging estimate that is not part of the factor model: DunedinPoAm38 (Belsky et al., 2020). DunedinPoAm38 is a rate-of-aging clock, its output is essentially uncorrelated with chronological age (r ≈ 0.09; see Section 1.1) and therefore lies on a different scale from the six panel clocks. For exactly this reason it was excluded from the factor model, which makes it a strict external validator.

Across all three cohorts, the pathway-aging score θ correlates positively with DunedinPoAm38 (r = 0.17–0.39, bootstrap 95 % CIs excluding zero in each cohort), while the erosion score δ shows essentially zero correlation (CIs cross or touch zero). The pattern is consistent with the biological reading of the two dimensions: pathway-driven, outcome-relevant aging, what θ is intended to capture, should track an independent pace-of-aging measure, while stochastic methylation drift should not. Crucially, this external check uses information that did not enter the factor model and is therefore not a tautology.

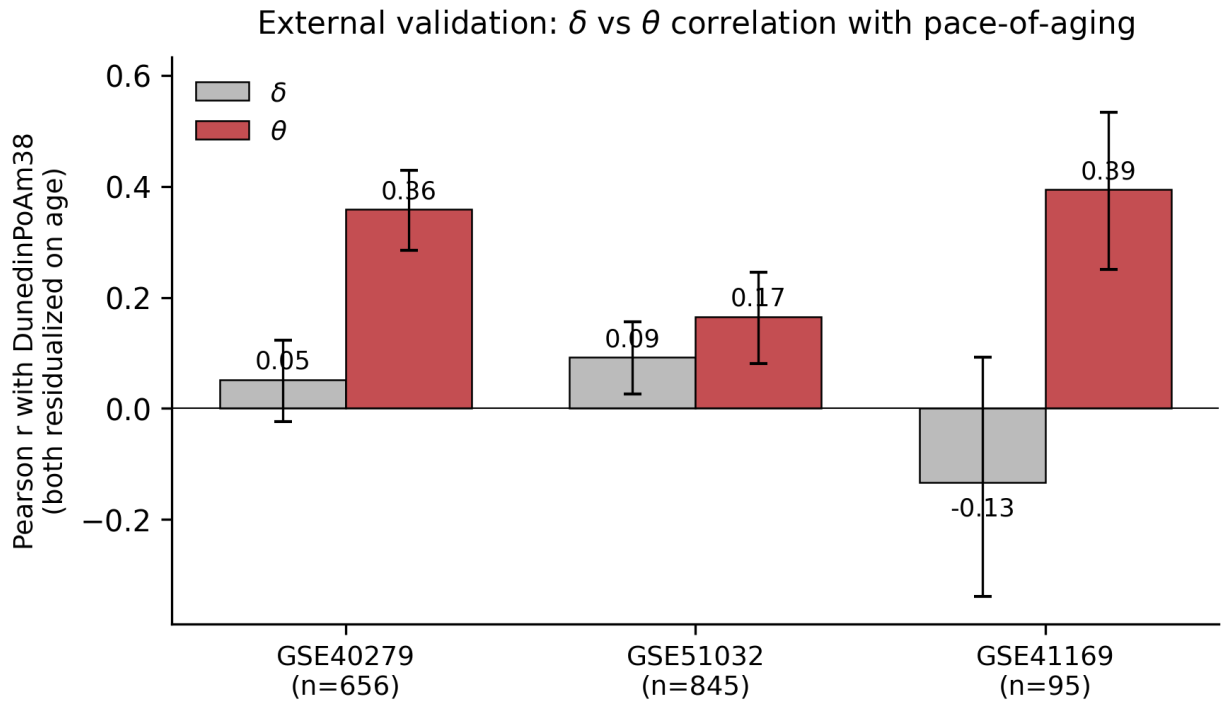


Figure 5. External verification: bootstrap correlations of (δ, θ) with DunedinPoAm38 across the three cohorts. θ correlates positively with PoAm in all three cohorts (r = 0.17–0.39); δ does not. DunedinPoAm38 is not part of the six-clock factor model.

4. Discussion

4.1 Two separable dimensions, three components per person

Our results demonstrate that inter-clock discordance in epigenetic age acceleration reflects two separable biological dimensions. The first, epigenomic erosion (δ), captures stochastic methylation drift and is measured primarily by first-generation clocks (Horvath, PCHorvath). The second, specific biological aging (θ), captures targeted methylation changes associated with mortality-related phenotypes and is measured by second-generation clocks (GrimAge, PhenoAge). This is a structural statement about the clocks: it tells us what each clock predominantly measures, not what any individual patient's biological age "truly" is.

This finding inverts how the discordance between clocks should be read. In the introduction, we presented it as a problem, different clocks disagree about who is accelerated, and the field has no clean answer to which one is right. The factor decomposition recasts the same observation as a property of the panel: the disagreement is exactly what the analysis needs to operate on. If all six clocks measured the same thing with independent noise, the correlation matrix would have a single dominant eigenvalue, the factor model would extract one dimension, and there would be no clock-specific structure to interpret. The fact that the clocks disagree in a structured way, first-generation clocks agreeing with each other but not with mortality-trained clocks, is what makes it possible to separate erosion from pathway aging in the first place. Discordance is not noise to be

averaged away; it is the empirical signature of multiple underlying processes being weighted differently by different clocks, and the factor decomposition is what reads that signature.

The structural statement has a direct consequence at the individual level. For each person, the factor model yields three sets of quantities: (i) an erosion score δ_i, (ii) a pathway-aging score θ_i, and (iii) six clock-specific residuals ε_{ik}. This is the practically useful output of the framework. The two shared scores integrate information across all six clocks and place the individual on the two-dimensional latent aging plane; the six residuals flag any clock that sees something in this person not predicted by the shared dimensions. Section 3.5 illustrates three contrasting profiles — one with strong opposing shared scores (δ high, θ low), one with the same shared scores but negligible residuals, and one with negligible shared scores but two large residuals — making three points concrete: that the shared scores alone determine the structural part of the multi-clock signal; that the apparent discordance across clocks comes from the loading matrix rather than from any property of the person; and that the residual vector ε is not a passive remainder but can carry the dominant signal in its own right.

This three-component structure is also why multi-clock panels are not interchangeable with single clocks even when the panel is reduced to a single composite. A single number — Horvath, GrimAge, or a panel average — projects the three-component profile onto one axis chosen by the clock's training procedure, and information is lost in directions orthogonal to that axis. The decomposition recovers the full information that the panel was capable of providing all along, and presents it in a coordinate system (δ, θ, ε) that is biologically rather than algorithmically chosen.

The core message of this paper is correspondingly twofold. At the population level: all these clocks have a lot in common (70% of the total variance in the GSE40279 correlation matrix is captured by two shared dimensions, in the eigenvalue sense made precise in Section 3.2, and the loading pattern replicates across three independent cohorts), and they also have things that are not in common (GrimAge devotes 71% of its variance to a private signal, Horvathv1 essentially none). At the individual level: this shared-and-private structure is exactly what enables a differentiated per-person interpretation, where two interpretable shared scores summarise the common signal and the six residuals carry the clock-specific information.

4.2 Why factor analysis rather than PCA

PCA yields qualitatively similar loading patterns (rank correlation > 0.95) but cannot provide the clock-specific uniquenesses ψ_k that distinguish shared from private structure, nor formal factor-score precision estimates via the Guttman reliability matrix. PCA components by construction span the full residual variance; the factor model explicitly separates shared (h_k^2) from clock-specific (ψ_k) variation. For the question this paper asks — what do the clocks have in common, and what is private to each — the factor-analytic distinction is essential.

4.3 Methodological considerations

The ML factor analysis approach with post-estimation Procrustes rotation has several methodological advantages. No distributional priors are needed, the factor structure is entirely data-driven. The Tong stochasticity coefficients enter solely through the rotation target, which is an interpretive choice analogous to choosing between Varimax and Promax rotation in classical factor analysis. Computation requires seconds rather than minutes, and the Guttman reliability formula provides exact (not sampled) standard errors for factor scores.

Future applications should incorporate Horvath's logarithmic age transformation or use non-linear regression for residualization; complementary approaches such as conformalized quantile regression (Li et al., 2025) offer adaptive individual-level prediction intervals that capture heteroscedasticity directly

Several limitations remain. First, the two-factor model leaves 45–73% of variance unexplained for several clocks, suggesting additional latent dimensions or, equivalently, that the panel does not span the full biology these clocks could in principle access. Second, without cell-type deconvolution, the specific factor may partly reflect compositional shifts rather than intracellular aging. Third, without mortality follow-up, the critical question, which component predicts lifespan, remains unanswered. Fourth, GSE40279 served as the training set for the Hannum clock, potentially inflating its loading in this cohort.

4.4 Beyond the structural decomposition

This paper has restricted itself to a structural question: how do six clocks decompose into shared and clock-specific variation? Two related questions are deliberately left to companion papers. (i) How should the resulting factor scores be reported to clinicians and patients, with appropriate individual-level uncertainty and population context? This is developed in [Uhlig et al., 2026b]. (ii) What is the full measurement-uncertainty budget for an epigenetic clock, including the between-laboratory component that the present within-platform analysis cannot quantify? This is the subject of Section 4.5 and of [Uhlig et al., 2026b].

4.5 Toward proficiency testing and validation of epigenetic clocks

An important limitation is that all six clocks are computed from the same Illumina 450K array, so any clock-specific residual variability captured in ψ_k reflects only within-platform, within-laboratory variation (probe-level noise, positional effects, normalization artifacts). In the language of ISO 5725, this is the repeatability component; the full reproducibility across laboratories and platforms is not represented. A companion paper [Uhlig et al., 2026b] develops the complete measurement uncertainty budget, proposes a proficiency testing scheme under ISO/IEC 17043, and specifies requirements for the between-platform and between-laboratory component.

The critical next step is the design and execution of interlaboratory comparison studies in which identical DNA reference samples are processed independently by multiple laboratories using the same or different clock computation pipelines. Such studies serve a dual purpose. First, as proficiency tests (PT) under ISO/IEC 17043, they enable laboratories to demonstrate competence

in epigenetic age estimation and to identify systematic biases in their workflows. Second, they provide the empirical basis for estimating method performance parameters, repeatability and reproducibility standard deviations, under ISO 5725. Without these quantities, a formal measurement uncertainty budget for an epigenetic clock cannot be constructed: the present analysis captures inter-clock disagreement, but not the technical variability introduced by laboratory processing, array batch effects, and computational pipeline choices.

The methodological foundation for such PT schemes already exists. A collaboration between the U.S. Food and Drug Administration and QuoData, documented in several peer-reviewed publications since 2020, has developed and applied ISO 13528 consensus estimation and hierarchical Bayesian methodology to proficiency testing programs in food safety and veterinary diagnostics. This established framework transfers directly to the epigenetic clock context: robust consensus values for age-acceleration residuals can be estimated from multi-laboratory measurements on shared reference samples, and the resulting repeatability and reproducibility variances can subsequently be integrated into the factor-analytic structure presented here as additional variance components.

Nevertheless, the covariance structure is remarkably consistent across the three cohorts (GSE40279, GSE51032, GSE41169), despite differences in population, laboratory, and sample handling, suggesting that between-laboratory effects do not dominate the factor structure at the level of the latent dimensions identified here. Strictly speaking, however, the variability patterns described in this paper apply only within the specific cohort and laboratory context. When a different laboratory, platform (e.g., EPIC v2; Garma & Quintela-Fandino, 2024), or pipeline is used, additional variability must be expected. The design of a proficiency testing scheme for epigenetic clocks, including reference material requirements, accreditation pathway, and the complete uncertainty budget combining within-laboratory and between-laboratory components, is presented in the companion paper [Uhlig et al., 2026].

The framework also serves as the basis for validating new clocks: a new clock can be projected into the existing factor space to determine its loading on erosion and specific aging, its uniqueness, and whether it accesses a novel third dimension. The critical gatekeeper is the technical repeatability standard deviation, without it, the decomposition cannot distinguish signal from noise. We advocate that any new clock report replicate-based repeatability data as a minimum requirement [Uhlig et al., 2026b].

5. Conclusions

Multi-clock epigenetic age acceleration can be decomposed into two latent dimensions — epigenomic erosion and specific biological aging — via maximum likelihood factor analysis with theory-guided Procrustes rotation. In the primary analysis (GSE40279, n = 656), the decomposition reveals a consistent clock ranking by erosion content, with first-generation clocks capturing stochastic drift and mortality-trained clocks capturing specific pathology. The ranking replicates across three independent cohorts.

The framework operates at two levels. At the population level, the loadings characterise what each clock predominantly measures, identifying a substantial shared component (70% of total variance in two dimensions) and a clock-specific component that varies enormously across the panel, from essentially zero for Horvathv1 to 70% for GrimAgeV1. At the individual level, the same decomposition yields for each person a three-component profile (δ_i, θ_i, ε_{i1}, …, ε_{i6}) that resolves multi-clock discordance into two shared scores plus six clock-specific residuals. The per-person profile is the practically useful output of the framework: it transforms six discordant numbers into a differentiated interpretation that single-clock and panel-mean approaches alike cannot deliver. As a methodological validation, the decomposition recovers the previously established slower epigenetic aging in Hispanic vs. Caucasian individuals (Horvath et al., 2016; Faul et al., 2023) and localises it specifically to the erosion dimension.

By treating clock discordance as structured information rather than noise, the approach provides a principled basis for multi-clock panel design and a foundation on which a quality-assured epigenetic diagnostics framework can be built, including the formal measurement uncertainty budget that is the subject of [Uhlig et al., 2026b].

Natural extensions include a Bayesian scoring variant yielding person-specific posterior distributions for δ_i, θ_i, and the residual vector, and the integration of technical replicates from interlaboratory comparison studies (see Section 4.5) to characterise the between-laboratory variance component absent from the present analysis. New clocks should report replicate-based repeatability data as a minimum validation requirement [Uhlig et al., 2026b].

Code and Data Availability

All three GEO datasets used in this study are publicly available: GSE40279 (Hannum et al., 2013), GSE51032 (EPIC-Italy), and GSE41169 (Dutch Blood). Clock computations were performed using the Biolearn library (Ying et al., 2024).

References

Belsky, D. W., Caspi, A., Arseneault, L., Baccarelli, A., Corcoran, D. L., Gao, X., et al. (2020). Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. eLife, 9, e54870. https://doi.org/10.7554/eLife.54870

Belsky, D. W., Caspi, A., Corcoran, D. L., Sugden, K., Poulton, R., Arseneault, L., et al. (2022). DunedinPACE, a DNA methylation biomarker of the pace of aging. eLife, 11, e73420. https://doi.org/10.7554/eLife.73420

Sehgal, R., Borrus, D. S., Gonzalez, J. T., Markov, Y., ADNI, & Higgins-Chen, A. T. (2025). Biological versus technical reliability of epigenetic clocks and implications for disease prognosis and intervention response. bioRxiv 2025.10.13.682176. https://doi.org/10.1101/2025.10.13.682176

Faul, J. D., Kim, J. K., Levine, M. E., Thyagarajan, B., Weir, D. R., & Crimmins, E. M. (2023). Epigenetic-based age acceleration in a representative sample of older Americans: Associations with aging-related morbidity and mortality. *Proceedings of the National Academy of Sciences*, 120(9), e2215840120. https://doi.org/10.1073/pnas.2215840120

Garma, L. D., & Quintela-Fandino, M. (2024). Applicability of epigenetic age models to next-generation methylation arrays. Genome Medicine 16, 116. https://doi.org/10.1186/s13073-024-01387-4

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., & Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell, 49*(2), 359–367. https://doi.org/10.1016/j.molcel.2012.10.016

Higgins-Chen, A. T., Thrush, K. L., Wang, Y., Minteer, C. J., Kuo, P.-L., Wang, M., Niimi, P., Sturm, G., Lin, J., Moore, A. Z., Bandinelli, S., Väisänen, S. B., Perola, M., Salomaa, V., Kähönen, M., Lehtimäki, T., Lind, L., Ingelsson, E., Jylhävä, J., Levine, M. E., & Ferrucci, L. (2022). A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking. *Nature Aging, 2*(7), 644–661. https://doi.org/10.1038/s43587-022-00248-2

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology, 14*(10), R115. https://doi.org/10.1186/gb-2013-14-10-r115

Horvath, S., Gurven, M., Levine, M. E., Trumble, B. C., Kaplan, H., Allayee, H., Ritz, B. R., Chen, B., Lu, A. T., Rickabaugh, T. M., Jamieson, B. D., Sun, D., Li, S., Chen, W., Quintana-Murci, L., Fagny, M., Kobor, M. S., Tsao, P. S., Reiner, A. P., Assimes, T. L., & Li, Y. (2016). An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biology, 17*, 171. https://doi.org/10.1186/s13059-016-1030-0

Kriukov, D., Kuzmina, E., Efimov, E., Dylov, D. V., & Khrameeva, E. E. (2024). Epistemic uncertainty challenges aging clock reliability in predicting rejuvenation effects. *Aging Cell, 23*(11), e14283. https://doi.org/10.1111/acel.14283

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

Li, Y., Goodrich, J. M., Peterson, K. E., Song, P. X.-K., & Luo, L. (2025). Uncertainty quantification in epigenetic clocks via conformalized quantile regression. *Genetic Epidemiology, 49*(4), e70008. https://doi.org/10.1002/gepi.70008

Uhlig, S., Uhlig, C., Hettwer, K., Nichani, K., Simon, K. (2026). Toward proficiency testing of epigenetic clocks: a metrological framework for multi-clock quality assurance under ISO 5725 / ISO 17043. *Publication in preparation*.

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., & Horvath, S. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY), 11*(2), 303–327. https://doi.org/10.18632/aging.101684

Tong, H., Dwaraka, V. B., Chen, Q., Luo, Q., Lasky-Su, J. A., Smith, R., & Teschendorff, A. E. (2024). Quantifying the stochastic component of epigenetic aging. *Nature Aging, 4*(6), 886–901. https://doi.org/10.1038/s43587-024-00600-8

Ying, K., Paulson, S., Eames, A., Tyshkovskiy, A., Li, S., Eynon, N., Jacques, M., Grolaux, R., Seale, K., Jacques, E., Goeminne, L. J. E., Cipriano, A., Perez-Guevara, M., Emamifar, M., Casas Martínez, M., Kwon, D., Kosheleva, A., Snyder, M., Gobel, D., Herzog, C., McCartney, D. L., Marioni, R. E., Lasky-Su, J., Poganik, J. R., Moqri, M., & Gladyshev, V. N. (2025). A unified framework for systematic curation and evaluation of aging biomarkers. *Nature Aging, 5*(11), 2323–2339. https://doi.org/10.1038/s43587-025-00987-y

Zhang, Q., Vallerga, C. L., Walker, R. M., Lin, T., Henders, A. K., Montgomery, G. W., He, J., Fan, D., Fowdar, J., Kennedy, M., Pitcher, T., Pearson, J., Scott, R. J., Holliday, E. G., Attia, J. R., Mather, K. A., Sachdev, P. S., Ames, D., Wright, M. J., Trollor, J. N., Zheng, F., Liu, L., Reppermund, S., Schofield, P. R., Kwok, J. B. J., Brodaty, H., Debette, S., & McRae, A. F. (2019). Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. Genome Medicine, 11, 54. https://doi.org/10.1186/s13073-019-0667-1