Supplementary Information for: "A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking."

Supplementary Results

GrimAge reliability is related to its two-step calculation

GrimAge is a unique case because it is calculated from chronological age, sex, and 7 DNAm-based components in a two-step process[1]. Inclusion of age and sex bolsters reliability, because age and sex are the same for technical replicates but different between samples. To isolate the variation attributed to DNAm, we re-calculated GrimAge setting age to 50 and sex to female for all samples (GrimAge50F ICC 0.963, GrimAge50F acceleration ICC 0.959). GrimAge50F reliability was still significantly higher than most GrimAge components (Fig. 1g, Supplementary Tables 3-4), demonstrating that combining multiple epigenetic biomarkers can bolster reliability.

Superior reliability of PC clocks is not related to use of new or substitute datasets, or number of CpGs or samples

For PCHorvath1, PCHorvath2, and PCPhenoAge, we utilized new and/or substitute datasets for training (Supplementary Table 6). This was because our PC clocks could not use 27K array data, or because some original training data was not available. To ensure the superior reliability of PC clocks was not solely a result of using new training datasets, we retrained Horvath1, Horvath2, and PhenoAge in new training data using traditional clock methods (applying elastic net regression directly on CpGs). In the case of PCPhenoAge, we also trained a PC clock using only the original data. To ensure maximum comparability with the PC clocks, we also used the same starting set of 78,464 CpGs for these new CpG-based clocks. This analysis confirmed that PC clock reliability improvements were primarily due to the PC clock methodology (Extended Data Fig. 5a-d).

We had added methylation and phenotypic data (N = 3593) from the Health and Retirement Study (HRS)[2] to the original InCHIANTI training data (N = 912), as it became available after the original development of PhenoAge. Indeed, using this larger sample size to train a CpG version of PhenoAge was sufficient to raise ICC of age acceleration from 0.76 to 0.91, and to enhance mortality prediction (Extended Data Fig. 5e). However, PCPhenoAge showed the same very high reliability whether it was trained in InCHIANTI or the combined dataset, suggesting CpG clock reliability is far more sensitive to training sample size than PC clock reliability. Thus, the improvement in reliability for PCPhenoAge was not due to increased sample size.

We had pre-selected only CpGs that were present across all our training and test data sets, and this resulting set of 78,464 CpGs was larger than the ~21K CpGs used to train the original Horvath1 and PhenoAge. However, using 78,464 CpGs to train a CpG version of PhenoAge in InCHIANTI did not improve the ICC (Extended Data Fig. 5e). Conversely, using the original ~21K CpGs to train PCPhenoAge still led to very high reliability. Thus, PCPhenoAge's superior reliability was not a result of considering more CpGs during training.

Investigating alternative methods to training reliable clocks

Traditional clocks utilize a subset of CpGs that provide information about a larger network of multicollinear CpGs but retain noise from individual CpGs. PC clocks show enhanced reliability due to two properties: 1) PCs incorporates information from many intercorrelated CpGs which dilutes noise, and 2) PCA tends to ignore noise which is not correlated between CpGs. To determine if one of these factors is more important, we investigated alternative methods to training reliable clocks that only share one of these properties. For this purpose, we trained variants of Hannum and PhenoAge to investigate how these methods affect age and mortality clocks respectively.

To test if diluting noise across many CpGs is sufficient for high reliability, we trained CpG-based clocks by ridge regression which retains all CpGs but shrinks coefficients of correlated CpGs toward zero. Ridge regression may improve reliability if the sample size is low but has no effect at higher sample sizes (Extended Data Fig. 5e-f). In all cases, the reliability of ridge regression clocks remains far below that of PC clocks trained from the same data. Mortality prediction is also not improved. This suggests that simply diluting noise by using many CpGs in the final model is not sufficient to improve reliability, and that the noise filtering properties of PCA are also required.

Our proposed PC clock training procedure first defines PCs in an unsupervised manner, followed by supervised selection of PCs using elastic net regression. This results in relevant information being spread out across many PCs. An alternative is to try to capture this information in fewer PCs by using supervised PCA,[3] a previously described method which pre-selects CpGs correlated with the outcome of interest (e.g. age or phenotypic age) and then performs PCA only on those CpGs. However, supervised PCA does not improve in reliability over traditional CpG clocks (Extended Data Fig. 5e-f), likely because this method can substantially limit how many CpGs are included. Supervised PCA PhenoAge predictors incorporated 50-500 CpGs, substantially less than the 8,000-50,000 CpGs that are needed for high reliability (Fig. 4). Less restrictive filtering of CpGs may lead to better reliability of supervised PCA predictors, but this approach would be similar to the PC clocks method. Mortality prediction is also not improved by supervised PCA. Thus, many CpGs and PCs are needed to construct high reliability clocks.

We also experimented with introducing a penalty factor for each CpG inversely proportional to ICC into elastic net regression, utilizing using M-values, or winsorizing beta-values (Supplementary Table 8). However, these did not yield better reliability than the CpG filtering approach.

Low-variance PCs capture heterogeneity in aging

We noted that adding progressively more PCs, in order from highest- to lowest-variance, for consideration (but not necessarily inclusion) in elastic net regression in training data (HRS and InCHIANTI) led to PCPhenoAge models with improved mortality prediction in independent test data (FHS) (Fig. 4). This improvement is clear up to at least PC1000 with only minor gains beyond PC1000. Meanwhile, reliability and age prediction were maintained. This suggested that low-variance PCs contain some consistent mortality signals across multiple datasets, even if they may typically be discarded by scree plots or random matrix theory methods. We performed PCA on a random noise matrix of equal size to the original PCPhenoAge training data upon which PCA was performed (4505x78464). For each column of the matrix, 4505 random draws from a normal distribution were used to populate the matrix, using a normal distribution with mean and standard deviation calculated from the CpGs of the PCPhenoAge data matrix. This approximates a random structure while accounting for the fact that the CpGs themselves tend to be approximated by normal distributions of varying mean and variance. We find that in such a matrix, the mean and maximum eigenvalues of the principal components are 0.050 (corresponding to PC636) and 0.199 (corresponding to PC42) respectively (Extended Data Fig. 7a). We also show that the variance explained by the components of PCPhenoAge outpaces that of the random matrix PCA up to PC126. We note these methods are meant for dimensionality reduction and it is still possible that lower-variance PCs contain important signals.

We tested if lower-variance PCs continue to introduce information from new CpGs. Due to the procedures of SVD/PCA, all CpGs used as inputs will have some weight across all PCs, but we determined which CpGs have a weight in the PC greater than expected to identify “driver” CpGs for each PC. We then determined, for each CpG, the first (highest-variance) PC where that CpG is identified as a driver CpG. This revealed that lower variance PCs throughout the entire range of PCs continue to contribute unique CpGs that are not significantly represented by high-variance PCs (Extended Data Fig. 7b-c). Interestingly, a spike in unique CpGs occurs after PC4037, the last PC incorporated into the PCPhenoAge model. These CpGs may primarily constitute noise and suggests that elastic net regression can efficiently exclude them from the PCPhenoAge model.

We also sought to determine if lower-variance PCs contain useful information for prediction. Phenotypic age[4], which PCPhenoAge is trained to predict, is a composite measure of chronological age and 9 clinical biomarkers for physiological dysregulation and age-related disease. Each biomarker is dysregulated in a different, limited subset of participants – for example, creatinine is above normal limits for 13.4% of HRS participants, C-reactive protein is elevated in 9.1%, and both are elevated in 2.8%. There may be even more heterogeneity within disease groups. We hypothesized that low-variance PCs, while not capturing much variance across the entire cohort, may capture variance relevant to these subsets.

To test this hypothesis, we first categorized PCs according to the following criteria (Extended Data Fig. 7d): High-variance PCs were defined as PCs 1-126 where eigenvalues are larger than that of a randomized matrix. Medium-variance PCs were defined as the next-highest PCs up to PC1000 (each explaining less than 0.047% of the total variance), as we did not find that adding PCs beyond PC1000 significantly contributed to mortality prediction (Fig. 4). Low-variance PCs were defined as the remaining PCs (each explaining less than 0.016% of the total variance).

We then calculated univariate associations between PCs and PhenoAge biomarkers in the training data (most biomarkers are not available in test data). Many medium- and low-variance PCs show associations with at least one biomarker (Extended Data Fig. 7e). The selected PhenoAge PCs are enriched in associations with at least one biomarker (50%) compared to the unselected PCs (25%) (Extended Data Fig. 7f-g). Interestingly, chronological age is mostly associated with top PCs – consistent with the chronological age predictors such as PCHorvath1 utilizing primarily the top 100-150 PCs. The notable biomarker that did not show any associations with low-variance PCs was lymphocyte percentage, demonstrating that low-variance PCs do not simply show spurious associations with any given variable.

The relevant PC PhenoAge signal appears spread out across many PCs. Hence the advantage of elastic net regression: it can combine PCs (each with small signal) while using cross-validation to discard other PCs and prevent overfitting. Thus, we did not adjust the univariate associations for multiple testing, as the PCs are meant to be combined in a predictor. Instead, we split the PCPhenoAge summary score into the signal stemming from high-variance, medium-variance, and low-variance PCs. We found that in both training data and independent test data, many of the biomarkers and diseases, as well as mortality, show associations with the PCPhenoAge signal stemming from low- and medium-variance PCs (Extended Data Fig. 7h-j) in multivariate models.

Thus, low-variance PCs likely capture mortality and morbidity risk. Aging increases heterogeneity across a population - different people age to varying degrees in different physiological systems, get different diseases, and get different treatments. Thus, any given mortality-related signature may only be present in a small subset of the population and are best captured by low-variance PCs. Our method utilizing elastic net regression can effectively identify useful combinations of low-variance PCs to improve mortality predictions, while filtering out other PCs to prevent overfitting and maintain high reliability.

Additional longitudinal data reveals effects of cell composition shifts

We examined two short-term longitudinal data sets for stress and schizophrenia which are associated with altered epigenetic aging[5,6]. First, we replicated the increased stability of PC clock trajectories in short-term longitudinal data in a cohort of 13 schizophrenia patients treated with clozapine, measured at 2-3 time points over 1 year including just prior to clozapine initiation (Extended Data Fig. 8c-d). DNAmTL increased rapidly during this period (p = 0.0077, 129 bp/year), but PCDNAmTL did not (p = 0.371, 20 bp/year) (Supplementary Table 12). DNAmTL's increase was likely due to a combination of noise and small sample size. Thus, the PC clocks may be useful in avoiding false positives in small pilot studies of interventions targeting epigenetic age.

We also examined 132 combat-exposed military personnel (baseline age range 18-54) from the Prospective Research in Stress-related Military Operations (PRISMO) study[7] with 2-3 time points and up to 500 days follow-up. Again, longitudinal changes in PC clocks were far more intercorrelated than CpG clocks (Extended Data Fig. 9a), consistent with reduced noise. Interestingly, the PC clocks continued to show substantial fluctuations, possibly reflecting relevant biological variance (Extended Data Fig. 9b). We noted that longitudinal changes in both CpG and PC clocks were strongly correlated with shifts in DNAm-estimated cell proportions for granulocytes and lymphocytes (Extended Data Fig. 9a), more strongly than in the SATSA dataset (Fig. 7g). Correcting for the within-individual longitudinal change in 5 cell types most associated with epigenetic age resulted in greatly improved stability for PC clocks but not CpG clocks (Extended Data Fig. 9b, Supplementary Table 11)

Parameters from PRISMO were used to model trials to protect younger adults from pathological aging under stressful conditions (Extended Data Fig. 9c, Supplementary Table 13). Consistent with our findings of cell composition shifts, power only nominally improved (and worsened in the case of Horvath2) using the PC clocks alone. When PC clocks were adjusted for cell composition, the required sample size was substantially reduced (approximately Horvath1 3-fold reduction; Horvath2 1.5; Hannum 3; PhenoAge 6; DNAmTL 5; GrimAge 2.5).

These results suggest short-term changes in epigenetic clocks may be affected by biological factors such cell composition shifts, but this phenomenon can only be corrected for after technical noise is minimized. Cell composition shifts may be magnified in the PRISMO dataset because of the exposure to stress and cortisol.

1. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany. NY).* **11**, 303–327 (2019).
2. Crimmins, E. M., Thyagarajan, B., Levine, M. E., Weir, D. R. & Faul, J. Associations of Age, Sex, Race/Ethnicity, and Education With 13 Epigenetic Clocks in a Nationally Representative U.S. Sample: The Health and Retirement Study. *Journals Gerontol. Ser. A* **76**, 1117–1123 (2021).
3. Bair, E., Hastie, T., Paul, D. & Tibshirani, R. Prediction by supervised principal components. *J. Am. Stat. Assoc.* **101**, 119–137 (2006).
4. Levine, M. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany. NY).* **10**, 276162 (2018).
5. Higgins-Chen, A. T., Thrush, K. L. & Levine, M. E. Aging biomarkers and the brain. *Semin. Cell Dev. Biol.* **116**, 180–193 (2021).
6. Higgins-Chen, A. T., Boks, M. P., Vinkers, C. H., Kahn, R. S. & Levine, M. E. Schizophrenia and Epigenetic Aging Biomarkers: Increased Mortality, Reduced Cancer Risk, and Unique Clozapine Effects. *Biol. Psychiatry* (2020) doi:10.1016/j.biopsych.2020.01.025.
7. van der Wal, S. J. *et al.* Associations between the development of PTSD symptoms and longitudinal changes in the DNA methylome of deployed military servicemen: A comparison with polygenic risk scores. *Compr. Psychoneuroendocrinology* **4**, 100018 (2020).