Devising reliable and accurate epigenetic clocks: choosing the optimal computational solution

Charlotte D. Vavourakis[1,2], Chiara M. Herzog[1,2*], Martin Widschwendter[1,2,3,4*]

[1]European Translational Oncology Prevention and Screening Institute, University of Innsbruck, Milser Str. 10, Hall in Tirol, 6060, Austria.
[2]Institute for Biomedical Ageing Research, University of Innsbruck, Rennweg 10, Innsbruck, 6020, Austria.
[3]Department of Women's Cancer, UCL EGA Institute for Women's Health, University College London, Medical School Building, Room 340, 74 Huntley Street, London, WC1E 6AU, UK.
[4]Department of Women's and Children's Health, Karolinska Institutet,, Tomtebodavägen 18A,, Solna, 171 77, Sweden.

*Corresponding author(s). E-mail(s): chiara.herzog@uibk.ac.at; martin.widschwendter@uibk.ac.at;
Contributing authors: charlotte.vavourakis@uibk.ac.at;

ARISING FROM A.T. Higgins-Chen et al. *Nature Aging* https://doi.org/10.1038/s43587-022-00248-2 (2022) [1]

Summary

Illumina DNA methylation arrays are frequently used for the discovery of methylation signatures associated with aging and disease [2]. With the latest InfiniumMethylation EPIC BeadChips v1.0 and v2.0 covering 850-900K CpG sites across the human

genome, one of the major hurdles to overcome when training trait prediction models is the high dimensionality of the data, with the number of features (target CpGs) greatly exceeding the typical number of samples assessed. In addition, most large-scale DNA methylation-based studies do not include replicate measurements for a given sample, making it impossible to estimate the degree of measurement uncertainty. Hence, the robustness of the assay and reliability of the prediction models are critical to ensure potential clinical utility. Recently, Higgins-Chen *et al* [1] proposed that training models on principal components (PCs) calculated from the data rather than directly on a selection of original features, results in more reliable age predictions. We found that dimension reduction with PCA consistently improves the reliability of age prediction models but may compromise their predictive accuracy. Our model projections further suggest that the balance in the tradeoff between accuracy and precision we observed, may tip over in favor of PC-based models with increasing training sample numbers.

Main

We argue that the PC approach proposed by [1] does not necessarily lead to increased performance compared to the original age prediction models, because model performance should be judged based on both accuracy (the closeness of the predicted with the true values) and reliability or precision (the closeness of two repeated predicted values). To demonstrate our argument on data presented in [1], we retrained a PC version of the Hannum clock for chronological age prediction on the original 450K training dataset GSE40279 derived from whole blood samples [3]. After verifying we could reproduce the age predictions from [1] (Extended data item 2), performance of the Hannum original and PC clock versions was compared on a 450K test set derived from 36 duplicate measurements on whole blood samples (GSE55763 [4], Fig. 1). Although reliability as estimated by the Intraclass Correlation Coefficient (ICC) improved for the predictions of the PC versus the original Hannum clock version, we found that the

age predictions were less accurate for the former, with a 19% lower root-mean-squared-error (RMSE; Fig. 1a) and a significantly lower regression slope for predicted versus actual age (Fig. 1b, slope = 0.661, 95% CI: 0.556-0.767 for the Hannum PC clock and 0.855, 95% CI: 0.771-0.940 for the Hannum original clock predictions), which was not reported by [1].

The authors did report a stronger association of the PCPhenoAge with mortality than PhenoAge. In addition, one recent DNA methylation-based study found a stronger correlation between predicted and actual telomer length when transforming the data with PCA before training [5]. Based on the underlying reasoning that individuals show a different rate of health decline with age, the deviations between actual and predicted age for the PC version of the Hannum clock could be interpreted as differences between chronological and biological age. However, Zhang *et al* [6] previously demonstrated that highly accurate chronological age predictions from blood samples are feasible when the sample set size used for training is large enough, with error rates compared to chronological age approximating zero. Moreover, because the Hannum clock is trained on data labeled with chronological age and not on biomarkers for biological age, the accuracy of the prediction model underlying this clock can be objectively evaluated comparing predicted versus chronological age. Possible correlations of age predictions with age acceleration or mortality would be an interesting, but unintended side effect.

To further demonstrate how dimension reduction by PCA may affect prediction model accuracy in the context of a directly-linked clinical outcome, we compared different versions of a cancer prediction model trained on an EPIC v1.0 dataset derived from 1647 cervical smear samples from healthy women [7], age-matched with breast (BC) [8], ovarian (OC) [9] and endometrial cancer (EC) [10] cases. Simple penalized regression models were replaced by ensembles of rich deep-learning base models trained with the AutoGluon framework [11] to allow for multi-class predictions. The accuracy

of the classifier trained on the original feature space was comparable to those of individual cancer prediction models trained with simple penalized regression models (AUC WID-BC = 0.81, 95% CI: 0.76–0.86 [8], AUC WID-EC = 0.92, 95% CI: 0.88-0.97 [10], AUC WID-OC = 0.76, 95% CI: 0.68–0.84 [9]). Meanwhile, the PC version of the classifier suffered from a lower accuracy, with areas under the receiver operator curve (AUC) dropping from 0.82 to 0.66 for predicting BC, from 0.93 to 0.89 for predicting EC, and from 0.70 to 0.55 for predicting OC, respectively (Extended data item 3).

Next, we investigated how the observed tradeoff between accuracy and precision of age prediction models underlying clocks depends on feature engineering and the size of the training set. An increasing, randomly fixed sample subset of the EPIC v1.0 cervical smear sample data set (n = 1647 samples) was used in separate model training rounds, recording each time accuracy of age predictions in an external validation set (cervical smears, n = 449) and precision in a repeatability dataset (n = 4 samples, 4 x measured each). The same experiment was repeated three times and power-law curves were fitted to the model performance estimates. Over the complete size range of training subsets, the Elastic net regression models had a lower predictive accuracy when trained on PCs instead of CpGs (Fig. 2a). Whereas the models trained on the original feature space were already close to the modeled maximum accuracy (Ymax = 0.85 ± 0.05), the PC-based models were expected to reach a comparable accuracy of 0.9 for approximately 12,640 samples. On the other hand, PC-based models reached maximal reliability already for small training sizes and did not suffer from random drops in ICC as the CpG-based models did (Fig. 2b).

Some important considerations should be noted when comparing the performance of PC-based and CpG-based age prediction models as a function of the size of the training set, as we have done here. Firstly, when combining methylation data from an increasing number of samples, the chances that a given target CpG successfully yields

a methylation readout in all samples decreases, and the resulting elimination of features that automatically occurs when combining methylation data from larger sample sets was not accounted for in the above model projections. It may positively affect performance when unreliable features are eliminated, but negatively affect performance when important features are lost. Secondly, for the CpG-based models the number of input features were fixed, hence for any n samples the same number of features (CpGs) are evaluated during training. In contrast, by design each PC-based model is trained on n-1 PCs and there is a direct dependence of the number of input features (PCs) on the sample size of the training set. Thirdly, the modeled Ymax values exceed 1 for the PC-approach. The modeled curves describe the recorded accuracies well in the tested sample set size range, but the extrapolated path is nonetheless hard to predict. Hypothetically, increasing the number of samples and hence the number of PCs could lead to over-fitting, which can be overcome again either by increasing training size even more or by increasing model complexity [12]. Finally, increasing model complexity by using state-of-the-art deep-learning approaches instead of simple penalized regression models such as Elastic Net further opens the potential for training multimodal trait prediction models that may evaluate methylation data, various omics data, biomarker data, and medical images at once. The dependence of the richer models on the sample size of the training set for the direct versus the PCA approach remains to be investigated.

In summary, feature engineering methods such as data dimension reduction with PCA or the use of deep learning algorithms alone are no silver bullets that can overcome the issues that come with high dimensional data, and the optimal machine learning approach depends on the available data and intended purpose of the prediction model. Our recommendations for future studies are to incorporate technical replicates where possible so that both reliability and accuracy of prediction models

can be carefully evaluated, and to consider various approaches before settling on a final machine learning strategy.

Data availability statement

The 450K microarray data analyzed here is publicly available in NCBI's Gene Expression Omnibus and are accessible through GEO Series Accession Numbers GSE40279 and GSE55763. The EPIC microarray data is available under restricted access through the European Genome-phenome Archive (EGA), which is hosted by the EBI and the CRG, under the accession numbers EGAS00001007184, EGAS00001005055, EGAS00001005045, EGAS00001005033.

Availability of computer code

All of the analyses were performed with publicly available R packages (R version 4.2.0) and Python libraries (Python version 3.8, supplied within the Docker Hub digest ace9e4217b43 for the framework AutoGluon-Tabular) using custom scripts that are made publicly available on https://github.com/ChVav/EPIC_PCsignatures. Individual package version numbers are given in the Extended Methods section provided as Supplementary Information.

Competing interests statement

C.H. and M.W. are shareholders of Sola Diagnostics GmbH, a company that holds patents for epigenetic signatures for cancer prevention and aging research.

Author contribution statement

The work reported in the paper has been performed by the authors, unless clearly specified in the text. Methodology, Software, Formal Analysis, Validation, Visualization, Writing - original draft preparation: C.D.V.; Writing - Review & Editing: C.D.V., C.H., M.W.; Conceptualization, Data curation, Supervision and Project Administration: C.D.V., C.H.; Funding Acquisition: M.W.

Acknowledgments

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 874662 (HEAP); The Eve Appeal (https://eveappeal.org.uk/); and the Land Tirol. We would like to cordially thank Dr. Sara Sleigh for critically revising this manuscript.

Figure legends

Figure 1. Performance of the original and PC versions of the Hannum clock, the latter trained as described by [1] on the original 450K training set derived from blood (n = 656 samples). **(a)** Precision (ICC; error bars represent 95% confidence intervals) and accuracy (RMSE) of the linear regression between predicted age versus chronological age as shown in **(b-c)** for the Hannum and Hannum_PC clocks using an independent test data set (n = 36 samples, 2 x measured each).

Figure 2. Performance of Elastic net regression models trained in combination with different feature engineering approaches as a function of training size. Age prediction models were trained directly on the same set of 632,385 CpGs as input features (Elnet, green shades and point shapes) or on n-1 derived PCs (PC, purple shades and triangle shapes, n max = 1647 samples). Recorded outcomes for the performance of

each model (points) and modeled fits (lines; Y-axis) from three experiments are plotted as a function of n samples in the training set (X-axis). **(a)** Predictive accuracy of the models estimated in an independent test set (n = 449 samples) as the slope of the regression curve predicted versus actual chronological age. Elnet fit: $Y = 0.89(\pm 0.05) - 17(\pm 3)X^{-0.94\pm 0.05}$; PC fit: $Y = 1.37(\pm 0.03) - 2.13(\pm 0.06)X^{-0.16\pm 0.03}$; $\pm SE$. **(b)** Test-retest reliability of the models given by the intraclass correlation coefficient (ICC) calculated from repeated measurements of cervical smears within the Repeatability set (n = 4 samples, 4 x measured each). Elnet fit: $Y = 1.0(\pm 0.7) - 40(\pm 90)X^{-1.8\pm 0.7}$; PC fit: $Y = 1.0(\pm 0.2) - 0.7(\pm 0.4)X^{-1.1\pm 0.2}$; $\pm SE$.

References

[1] Higgins-Chen, A. T. *et al.* A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking. *Nature Aging* **2**, 644–661 (2022).

[2] Noguera-Castells, A., García-Prieto, C. A., Álvarez-Errico, D. & Esteller, M. Validation of the new epic dna methylation microarray (900k epic v2) for high-throughput profiling of the human dna methylome. *Epigenetics* **18**, 2185742 (2023).

[3] Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell* **49**, 359–367 (2013).

[4] Lehne, B. *et al.* A coherent approach for analysis of the illumina humanmethylation450 beadchip improves data quality and performance in epigenome-wide association studies. *Genome biology* **16**, 1–12 (2015).

[5] Doherty, T. *et al.* A comparison of feature selection methodologies and learning algorithms in the development of a dna methylation-based telomere length

estimator. *BMC bioinformatics* **24**, 178 (2023).

[6] Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome medicine* **11**, 1–11 (2019).

[7] Barrett, J. E. *et al.* Susceptibility to hormone-mediated cancer is reflected by different tick rates of the epithelial and general epigenetic clock. *Genome Biology* **23**, 52 (2022).

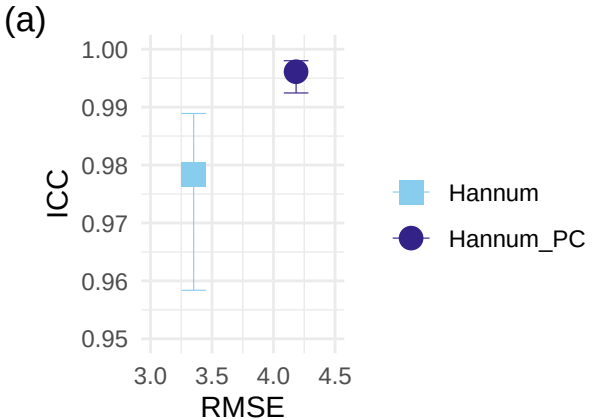
[8] Barrett, J. E. *et al.* The wid-bc-index identifies women with primary poor prognostic breast cancer based on dna methylation in cervical samples. *Nature Communications* **13**, 449 (2022).

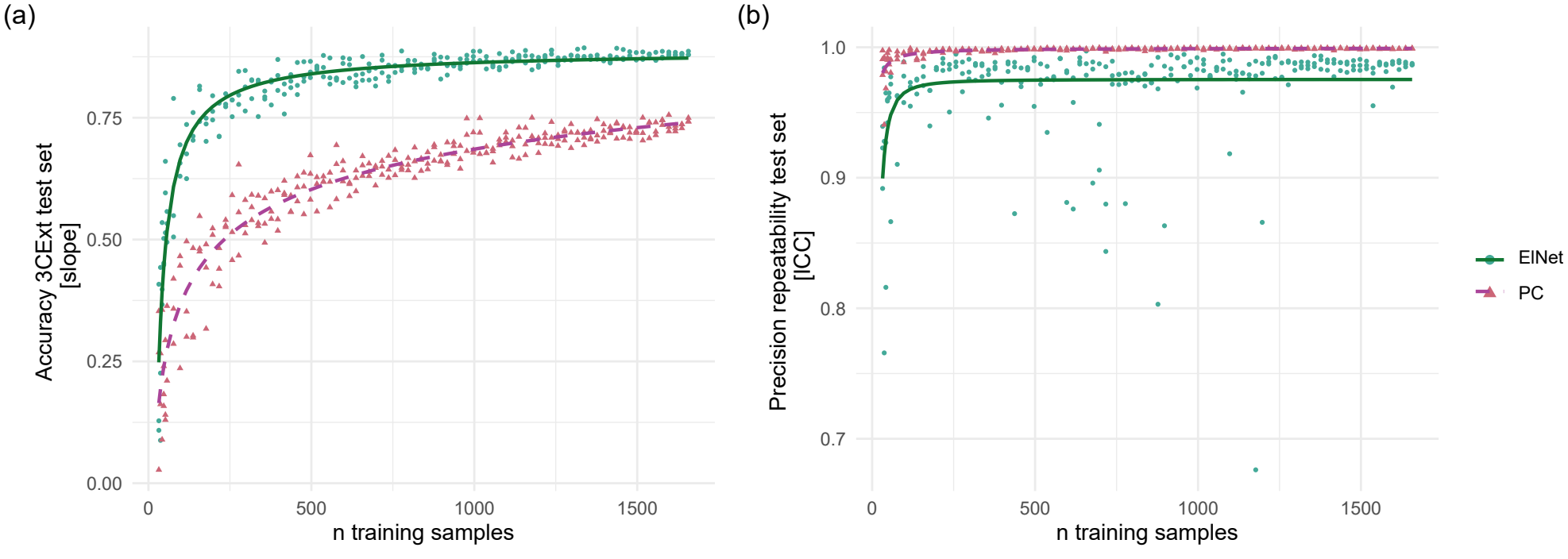
[9] Barrett, J. E. *et al.* The dna methylome of cervical cells can predict the presence of ovarian cancer. *Nature Communications* **13**, 448 (2022).

[10] Barrett, J. E. *et al.* The wid-ec test for the detection and risk prediction of endometrial cancer. *International Journal of Cancer* (2022).

[11] Erickson, N. *et al.* Autogluon-tabular: Robust and accurate automl for structured data. *arXiv preprint arXiv:2003.06505* (2020).

[12] Belkin, M., Hsu, D., Ma, S. & Mandal, S. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences* **116**, 15849–15854 (2019).


(a)
1.00
0.99
0.98
0.97
0.96
0.95
ICC
3.0
3.5
4.0
4.5
RMSE
Hannum
Hannum_PC
(b)
Hannum predicted age
80
70
60
50
40
R2 = 0.853
y = 0.86x + 16.58
40
50
60
70
80
Age
(c)
Hannum_PC predicted age
80
70
60
50
40
R2 = 0.69
y = 0.66x + 21.9
40
50
60
70
80
Age


(a)
Accuracy 3CExt test set
[slope]
0.00
0.25
0.50
0.75
0
500
1000
1500
n training samples
(b)
Precision repeatability test set
[ICC]
0.7
0.8
0.9
1.0
0
500
1000
1500
n training samples
ElNet
PC