Devising reliable and accurate epigenetic predictors: 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;

Abstract

Illumina DNA methylation arrays are frequently used for the discovery of methylation signatures associated with aging and disease. 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. Here, we test the performance of different versions of age and cancer prediction models trained either directly on the original features (CpGs) or derived principal components (PCs). Utilizing PCA for dimension reduction consistently led to small improvements in the reliability of the age prediction models, measured in terms of the repeatability of technical replication. However, this improvement came at the cost of a notable reduction in their predictive accuracy. Moreover, by modeling prediction performance as a function of the training set size, we show that the PC-based models need far larger training set sizes to be similarly accurate as CpG-based models. Dimension reduction by PCA also resulted in a markedly lower predictive accuracy when replacing simple penalized regression models by weighted ensembles of deep-learning models for cancer prediction.

Keywords

Epigenetic predictors, PC clocks, DNA methylation, aging, cancer, machine learning, dimension reduction, PCA.

Main text

The methylome is a flexible layer of the genome that plays a crucial role in the regulation of gene expression and chromatin structure. The strong correlation of aberrant DNA methylation patterns with various disease phenotypes can be used for training diagnostic tools with machine learning. For example, we previously developed the WID-BC [1], WID-OC [2], WID-EC [3] signatures that predict breast, ovarian and

endometrial cancer from cervical smear samples analyzed using the Illumina Infinium-Methylation EPIC BeadChip v1.0, which covers approx. 850,000 CpG sites across the human genome. Because the methylome also changes with age in a predictable manner, another popular use for Illumina methylation array data is training age predictors, also called epigenetic clocks. Initially clocks were trained to predict only chronological age, i.e., the time from birth to a specific date [4–6]. The observation that deviations in actual versus predicted age correlate with certain phenotypic traits ranging from gender to mortality risk, sparked new DNA methylation-based investigations into biological aging, which is a measure of the age-related health status of an individual. Improved biological clock versions are now trained directly incorporating biomarkers for general health and physiological decline associated with age [7].

Two major obstacles present when training DNA methylation-based trait-prediction models. First, the microarray data are highly dimensional, with typically 100- or 1000-fold more CpGs than samples assessed. Secondly, in most large-scale DNA methylation-based studies, each sample is processed and measured only a single time. Because each measurement carries a certain degree of uncertainty which cannot be estimated without replication, continuous effort goes to understanding and reducing various sources of variability. Various data-driven pre-processing pipelines have been developed that remove noise, e.g. [8–10], and recently, Illumina released the MethylationEPIC v2.0 BeadChip, claiming improved precision by removing about 10% probes on the EPIC v1.0 that were deemed unreliable, at the same time adding probes to a total of 900K CpG sites to the array to better cover genomic regions relevant for clinical research [11].

Recently, Higgins-Chen *et al* [12] proposed that a simple dimension reduction done by training penalized regression models on extracted principal components rather than on the original features directly, results in more reliable age clocks because the PCA

feature space incorporates information from all of the original features. The performance of a predictive model designed to estimate continuous variables, such as age, may be judged both on its accuracy (the closeness of the predicted with the true values) and on its reliability or precision (the closeness of two repeated predicted values), using external validation sets to test the prediction outcomes the model was trained for. To first verify that we could reproduce the PCA approach and age predictions from [12], we retrained PC and PC proxy versions of the Hannum clock (trained to predict chronological age and Hannum age, respectively) on the original subset of 78,464 CpGs in a HumanMethylation450 dataset derived from 656 peripheral blood samples "Hannum" [4] (for details on the methods and all datasets analyzed in this study see Additional file 1 and Additional file 2), and compared age predictions on a blood-derived data set comprising 36 duplicate measurements "450K_BloodRep" (GSE55763 [13]; Additional file 3, panels (a-b)). Reliability as estimated by the Intraclass Correlation Coefficient (ICC) improved for the predictions of the PC versus the original Hannum clock version, but age predictions from the PC clock were significantly less accurate, with a 19% lower root-mean-squared-error (RMSE; Extended data item 2, panel (c)) and a significantly lower regression slope for predicted versus actual age (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 clock; Extended data item 2, panels (d-e)). This loss in accuracy was not reported by [12]. Training a second PC clock on a larger subset of 473,034 CpGs shared between the "Hannum" training set and a "450K_BloodFull" external validation set (GSE55763 [14], n = 2639 blood samples) aggrevated the loss of accuracy, with a regression slope of only 0.59 (95% CI 0.57 - 0.60) for the Hannum PC2 clock (Fig. 1).

Next, we investigated how the observed tradeoff between accuracy and reliability of age prediction models underlying clocks directly depends on the additional PCA step and the size of the training set. Increasing, randomly fixed sample subsets of the

"450K_BloodFull" (n = 2639 blood samples) and a EPIC v1.0 cervical smear sample data set "3CDisc" (n = 1647 samples) were used to train Elastic net regression models in the original or the PCA-derived feature space, recording for each training round accuracy and reliability of age predictions in external validation sets of corresponding sample and data types (Fig. 2). 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 much lower predictive accuracy when trained on PCs instead of CpGs for both the blood (Fig. 2a) and cervical smear DNA methylation data (Fig. 2c). Whereas the clocks trained on the original feature space neared the modeled maximum accuracy for the full training sets (Ymax = 0.90 ± 0.05 for blood and Ymax = 0.94 ± 0.04 for cervical smears in the modeled fits), hypothetically the PC clocks would reach a comparable accuracy of 0.9 only for approximately 11,707 blood or 17,000 cervical smear samples (calculated from the modeled equations). In contrast, differences in reliability were minimal between PC and non-PC clocks, mostly showing an excellent ICC well above 0.9 [15] (Fig. 2b/d). Only for the cervical smear data, few individual clocks showed random drops in ICC, something the PC-trained clocks did not suffer from (Fig. 2d). Thus, Fig. 2 clearly demonstrates that training penalized regression models on the PCA-derived feature space can come with a considerable loss in accuracy and that the accuracy of PC clocks stronlgy depends on the sample size of the training set.

The variability in age-related health outcomes can be interpreted as an interindividual difference in biological age relative to chronological age [7]. The deviation between chronological age and predicted age using epigenetic biomarkers has previously been used as an estimate of biological age, and has been shown to predict multiple age-related health outcomes, such as mortality. Interestingly, Higgins-Chen *et al* [12] report a stronger association of some versions of their PC clocks compared to the corresponding non-PC clocks. For instance, PCHorvath1 and PCPhenoAge exhibits a stronger

association with mortality than Horvath1 and PhenoAge, respectively. Another recent DNA methylation-based study also found a stronger correlation between predicted and actual telomer length when transforming the data with PCA before training [16]. However, in agreement with our models (Fig. 2), 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. As the association of a clock with chronological age increases, the association with health outcomes shrinks, and a trained epigenetic biomarker then only informs on chronological age. Many recent epigenetic biomarkers are now training to predict biological outcomes rather than chronological age alone to overcome this limitation [7]. Although Higgins-Chen et al. *et al* [12] demonstrate an increased performance for predicting age-related outcomes for many of their PC clocks, our data provide additional context and demonstrate that the loss of accuracy that comes by training penalized regression models on PCA-transformed DNA methylation data needs to be considered for the relevant outcome, providing a practical example based on cancer detection below (Additional file 4).

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 [5], age-matched with breast (BC) [1], ovarian (OC) [2] and endometrial cancer (EC) [3] cases. Simple penalized regression models were replaced by ensembles of rich deep-learning base models trained with the AutoGluon framework [17] 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 [1], AUC WID-EC = 0.92, 95% CI: 0.88-0.97 [3], AUC WID-OC = 0.76, 95% CI: 0.68–0.84 [2]). 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 (Additional file 4).

Some important technical 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 overfitting, which can be overcome again either by increasing training size even more or by increasing model complexity [18]. 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.

Finally, to further showcase how dimension reduction by PCA may affect prediction model accuracy in the context of a directly-linked clinical outcome beyond age

and richer machine learning models, we compared different versions of a cancer prediction model trained on the EPIC v1.0 dataset "3CDisc" derived from 1647 cervical smear samples from healthy women [5], age-matched with breast (BC) [1], ovarian (OC) [2] and endometrial cancer (EC) [3] cases. Simple penalized regression models were replaced by ensembles of rich deep-learning base models trained with the AutoGluon framework [17] to allow for multi-class predictions. The accuracy of the classifier trained on the original feature space was comparable to those of previously published individual cancer prediction models trained with simple penalized regression models (AUC WID-BC = 0.81, 95% CI: 0.76–0.86 [1], AUC WID-EC = 0.92, 95% CI: 0.88-0.97 [3], AUC WID-OC = 0.76, 95% CI: 0.68–0.84 [2]). 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 (Additional file 4).

In summary, feature engineering methods such as data dimension reduction with PCA or the use of deep learning algorithms 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.

List of abbreviations

AUC: Area Under the Curve

BC: breast cancer

CI: confidence interval

CpG: Cytosine-phosphate-Guanine

EC: endometrial cancer

OC: ovarian cancer

PC: principal component

PCA: principal component analysis

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The 450K microarray datasets analyzed during the current study are available in the NCBI's Gene Expression Omnibus and are accessible through GEO Series Accession Numbers GSE40279 and GSE55763. The EPIC microarray datasets analyzed during the current study are 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. 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 detailed methods section provided in Additional file 1.

Competing interests

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.

Authors' contributions

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.

Additional files

- Additional file 1 (.pdf): extended methods section.
- Additional file 2 (.pdf): datasets analyzed in this study.
- Additional file 3 (.pdf): reproducing clock calculations presented by [12].
- Additional file 4 (.pdf): Effect of PCA transformation (feature extraction) on the performance of cancer predition models.

Figure legends

Figure 1. Accuracy of the original and a PC version of the Hannum clock, the latter trained as described by [12] on a 450K training set derived from blood (n = 656

samples). Linear regressions of predicted age versus chronological age for the Hannum **(a)** and Hannum_PC2 clock **(b)** were calculated on an independent 450K test dataset (n=2639 blood samples).

Figure 2. Performance of Elastic net regression models trained as a function of training size directly on the same set of CpGs as input features (Elnet, green shades and point shapes) or on n-1 derived PCs (PC, purple shades and triangle shapes). Performance paramaters of each clock from three independent experiments (points) and modeled fits (lines; Y-axis) are plotted as a function of n samples in the training set (X-axis), with n max = 2639 for the blood-derived **(a-b)** and n max = 1647 for the cervical-smear derived DNA methylation dataset **(c-d)**. **(a)** Predictive accuracy of the age clocks in blood (n = 656 samples). Elnet fit: $Y = 0.94(\pm 0.04) - 4.5(\pm 0.6)X^{-0.63\pm 0.04}$; PC fit: $Y = 6.94(\pm 0.02) - 8(\pm 4)X^{-0.03\pm 0.02}$; $\pm SE$. **(b)** Test-retest reliability of the corresponding clocks (n = 36 samples, 2 x measured each). Elnet fit: $Y = 0.99(\pm 0.08) - 1.5(\pm 0.5)X^{-0.92\pm 0.08}$; PC fit: $Y = 1.0(\pm 0.2) - 7(\pm 4)X^{-1.1\pm 0.2}$; $\pm SE$. **(c)** Predictive accuracy of the age clocks in cervical smears (n = 449 samples). Elnet fit: $Y = 0.90(\pm 0.05) - 9(\pm 2)X^{-0.79\pm 0.05}$; PC fit: $Y = 1.07(\pm 0.03) - 2.6(\pm 0.2)X^{-0.28\pm 0.03}$; $\pm SE$. **(d)** Test-retest reliability of the corresponding clocks (n = 4 samples, 4 x measured each). Elnet fit: $Y = 1.0(\pm 0.3) - 2(\pm 2)X^{-1.0\pm 0.3}$; PC fit: $Y = 1.0(\pm 0.2) - 0.2(\pm 0.2)X^{-0.7\pm 0.2}$; $\pm SE$.

References

[1] 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).

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

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

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

[5] 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).

[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] Moqri, M. *et al.* Biomarkers of aging for the identification and evaluation of longevity interventions. *Cell* **186**, 3758–3775 (2023).

[8] Aryee, M., Jaffe, A. & Corrada-Bravo, H. Ladd-acosta c. feinberg ap. hansen kd. irizarry ra. *Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics* **30**, 1363–9 (2014).

[9] Tian, Y. *et al.* Champ: updated methylation analysis pipeline for illumina beadchips. *Bioinformatics* **33**, 3982–3984 (2017).

[10] Xu, Z., Niu, L. & Taylor, J. A. The enmix dna methylation analysis pipeline for illumina beadchip and comparisons with seven other preprocessing pipelines. *Clinical Epigenetics* **13**, 216 (2021).

[11] 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).

[12] 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).

[13] 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).

[14] Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).

[15] Koo, T. K. & Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* **15**, 155–163 (2016).

[16] 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).

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

[18] 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).

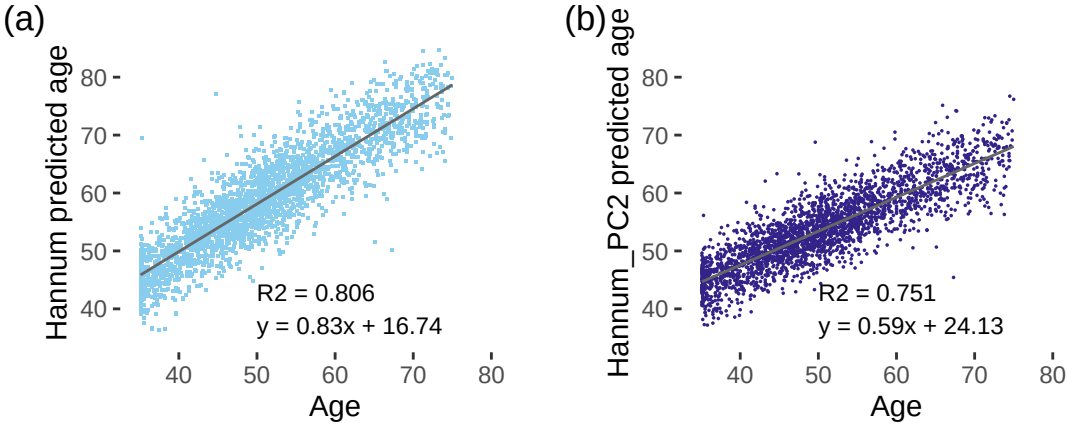
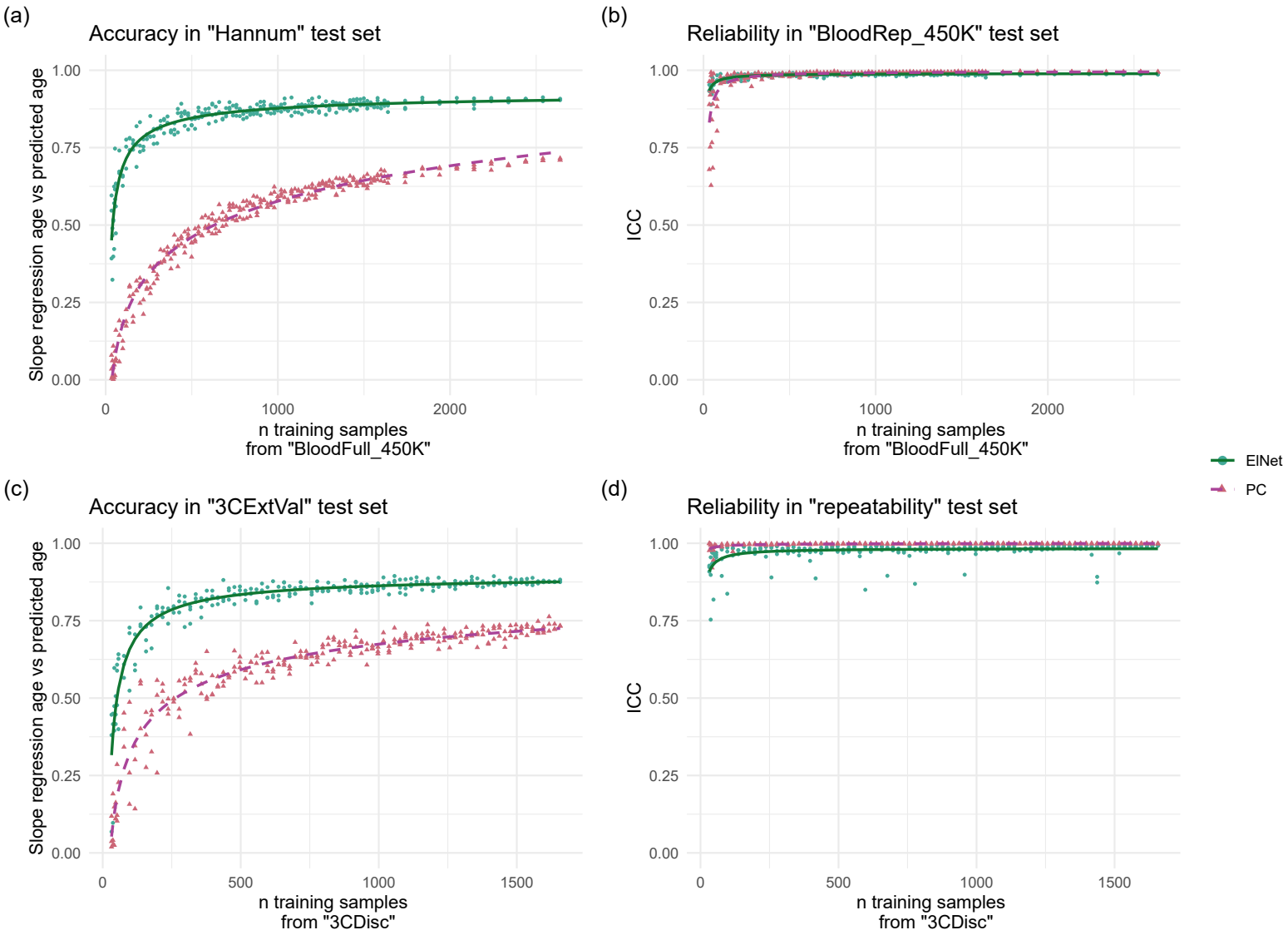

(a)
Hannum predicted age
80
70
60
50
40
R2 = 0.806
y = 0.83x + 16.74
40
50
60
70
80
Age
(b)
Hannum_PC2 predicted age
80
70
60
50
40
R2 = 0.751
y = 0.59x + 24.13
40
50
60
70
80
Age

Figure 2


(a)
Accuracy in "Hannum" test set
Slope regression age vs predicted age
1.00
0.75
0.50
0.25
0.00
0
1000
2000
n training samples
from "BloodFull_450K"
(b)
Reliability in "BloodRep_450K" test set
ICC
1.00
0.75
0.50
0.25
0.00
0
1000
2000
n training samples
from "BloodFull_450K"
ElNet
PC
(c)
Accuracy in "3CExtVal" test set
Slope regression age vs predicted age
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
n training samples
from "3CDisc"
(d)
Reliability in "repeatability" test set
ICC
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
n training samples
from "3CDisc"