nature aging

Analysis

https://doi.org/10.1038/s43587-023-00391-4

Development and validation of DNA methylation scores in two European cohorts augment 10-year risk prediction of type 2 diabetes

In the format provided by the authors and unedited

Supplementary Note

MethylPipeR and MethylPipeR-UI

The analysis pipeline was implemented via a new R package, *MethylPipeR,* along with accompanying user interface, for systematic and reproducible development of complex trait and incident disease predictors. *MethylPipeR* provides functionality for tasks such as model fitting, prediction and performance evaluation as well as automatic logging of experiments and trained models. This is complemented by *MethylPipeR-UI* which provides an interface to the R package functionality while removing the need to write scripts. While *MethylPipeR* was applied to incident T2D prediction with DNA methylation in our experiments, the package is designed for generalised development of predictive models and is applicable to a wide range of omics data and target traits. *MethylPipeR* and *MethlyPipeR-UI* are publicly available at https://github.com/marioni-group/MethylPipeR and https://github.com/marioni-group/MethylPipeR-UI respectively. **Extended Data Figure 3** shows an example from the *MethylPipeR-UI* interface including functionality such as data upload, specification of model options and visualisation of model diagnostics.

Methylation Risk Scores and Incremental Modelling

Composite protein EpiScore models were fit to the training set using 109 protein EpiScores as features [1]. This trained model was used to create a composite protein EpiScore for each individual in the test set. Similarly, direct predictors were fit to the training set using 200,000 CpGs as features and used to create a direct EpiScore for each test set individual.

The following Cox PH models were then fit to the test set to assess the difference in metrics such as the area under the curve (AUC) and area under the precision-recall curve (PRAUC): a risk factors only model; risk factors + protein EpiScore; risk factors + direct EpiScore and a full model (risk factors + protein EpiScore + direct EpiScore). To obtain 10-year survival probabilities in the test set, the cumulative baseline hazard at t=10 was calculated for each model using the Breslow estimator [2].

Calculating Predictions

To obtain predictions from the incremental models, the following should be applied.

Direct EpiScore

For the Cox lasso model, the direct EpiScore can be obtained by summing over the CpG values weighted by their corresponding coefficients (for all non-zero coefficients), as given in **Supplementary Table 13**.

$$Direct\ EpiScore = \beta_1 CpG_1 + \beta_2 CpG_2 + \cdots + \beta_p CpG_p$$

For sBART and RSF, the direct EpiScore requires use of the model object to obtain predictions on new data. These can be found in https://github.com/marioni-group/ and utilised with *MethylPipeR*'s predict function.

Composite Protein EpiScore

The composite protein EpiScore can be obtained by summing over the protein EpiScores weighted by their corresponding coefficients (for all non-zero coefficients), as given in **Supplementary Table 13**.

$$Composite\ Protein\ EpiScore = \beta_1 PES_1 + \beta_2 PES_2 + \cdots + \beta_q PES_q$$

where q is the number of protein EpiScores with a non-zero coefficient.

Each protein EpiScore is calculated as the sum over CpG values weighted by their corresponding coefficients (for all non-zero coefficients); the coefficients for each of these can be found in the supplementary materials in Gadd et al. [1]

Incremental Cox PH

The linear predictor in an incremental Cox PH model can be calculated by multiplying the relevant EpiScores and risk factors by their corresponding coefficient, given in **Supplementary Table 13**. The cumulative baseline hazard is calculated using the basehaz.gbm function in the gbm R package version 2.1.8 [3]. The 10-year survival probabilities can then be calculated as $S(t) = exp[-(\Lambda_0(t))]^{exp(l)}$ at $t = 10$, where $\Lambda_0(t)$ is the cumulative baseline hazard and l is the linear predictor. The 10-year onset probability is therefore $1 - S(10)$.

Penalised Cox Proportional-hazards Regression

Since the number of features (200,000) was much greater than the number of individuals in the training set (=9,835 after data preprocessing), a regularisation method was required to reduce overfitting of the Cox PH models.

Cox PH models with lasso penalisation was fit to the training set using glmnet (version 4.1-1) [4, 5] via *MethylPipeR*.

Hyperparameter optimisation and cross validation (CV) were used to select the λ that minimised the partial-likelihood for each Cox PH model corresponding to the minimum partial-likelihood. The training set was divided equally into nine partitions. For each pre-selected value of λ, nine models were fit, each using eight of the partitions as the training set and the ninth for prediction.

The mean partial likelihood over the nine models was then calculated. The model using the λ that minimised this was chosen to evaluate on the test set. Nine-fold CV was used to balance the advantages provided by using a greater number of folds with the limitations of the number of cases in each fold as well as required computation time. In addition, individuals belonging to the same family were assigned to the same fold and each contained a similar number of cases to avoid folds with too few cases.

Random Forest

Random forest [6] is an ensemble machine learning model that estimates a function by averaging the output from a set of independently trained decision trees. During model fitting, each tree is built using a different subset of the variables and the training set to prevent individual trees from overfitting to the whole dataset. In addition, random survival forests adapts the original method to incorporate right-censored time-to-event data [7].

The hyperparameters corresponding to the number of trees (ntrees), the number of variables considered at each tree split (mtry) and the minimum terminal node size (nodesize) were selected using a grid-search CV method. **Supplementary Table 14** shows the set of values that were considered for each hyperparameter. The R package randomForestSRC (version 2.11.0) [8] was used for fitting random survival forests via *MethylPipeR*.

Bayesian Additive Regression Trees

Bayesian Additive Regression Trees (BART) [9] is a nonparametric method that estimates a function as a sum over a set of regression trees. BART incorporates the ability to model both additive and interaction effects and has shown high predictive performance in comparison with similar methods. To reduce overfitting and model uncertainty in parameters and predictions, BART uses prior distributions over tree-related parameters. Posterior estimates are obtained in a Bayesian framework through Markov Chain Monte Carlo (MCMC).

A variant of BART for survival analysis [10] was used for 10-year onset prediction. 1,000 posterior samples for model parameters were kept after 500 burn-in samples. These were used to generate 10-year survival probabilities on the test set. This resulted in 1,000 survival probabilities for each individual in the test set, the mean of which was used as their survival prediction. MCMC convergence was assessed using Geweke's diagnostic (gewekediag function in the BART R package). This was calculated on the BART output (yhat.test with t<=10 [10]) for the 1,000 posterior samples using 30 randomly chosen individuals in the test set. Most Z-values (77%) fell within the interval [-1.96, 1.96] suggesting convergence.

Due to the computation time requirements of MCMC sampling and the apparent robustness of BART to hyperparameter misspecification [9], the sBART model was run with hyperparameters set to default. This was performed using the R package BART (version 2.9) [11] via *MethylPipeR*.

Validation of Best Performing Model in KORA S4

The present analyses are based on a subsample of the participants of the KORA S4 study. KORA (Cooperative Health Research in the Region of Augsburg) is a research platform performing population-based surveys and subsequent follow-ups in the region of Augsburg in Southern

Germany [12]. Participants were of German nationality, aged between 25-74 years, 50% female, and sampled from the population registers in the study area where main place of residence was: Ausburg city town, county Ausburg or county Aichach-Friedberg. Each participant completed a health questionnaire, providing details on health status and medication. Blood samples were also taken for assaying of omics data. KORA S4 participants were recruited between 25/10/1999-28/04/2001. This study used a subsample of the 1,451 participants of the KORA S4 study with DNAm and incident T2D data available and no prevalent diabetes at baseline.

The best performing model selected for the Generation Scotland cohort (Cox PH lasso) was used for prediction of incident T2D in the KORA S4 cohort.

For diabetes morbidity, the data are limited to a follow-up of 10 years - starting from KORA S4 recruitment. For incident T2D all prevalent diabetics as well all other diabetes types except T2D cases are excluded. Age, BMI, hypertension, sex as well as self-reported family (mother or father) history of T2D were taken at the baseline of KORA S4. BMI was calculated as the individual's weight in kg divided by the square of their height in metres.

EpiScore Prediction of COVID-19 Outcomes

Self-reported COVID-19 phenotypes were available in a subset of individuals from the Generation Scotland DNA methylation sample who had also participated in the CovidLife surveys (n=2,399) [13]. Ongoing symptomatic COVID-19 phenotypes were ascertained from CovidLife survey 3, (n=1,802 Generation Scotland participants with DNAm profiled), where participants were asked about the total overall time they experienced symptoms in their first/only episode of illness, as well as the whole of their COVID-19 illness. Ongoing symptomatic COVID-19 was defined here

as symptoms persisting for at least 4 weeks from infection and is correct as of February 2021, when the survey 3 was administered. Hospitalisation information, derived from the Scottish Morbidity Records (SMR01), was used to obtain COVID-19 hospital admissions using ICD10 codes U07.1 (lab-confirmed COVID-19 diagnosis), and U07.2 (clinically-diagnosed COVID-19). Hospitalisation data is correct as of February 2021. Logistic regression was used to assess the predictive performance of the T2D EpiScores in relation to ongoing symptomatic COVID-19 and severe COVID-19 (i.e. hospitalisation), adjusting for the standard risk factors and incident T2D before a positive COVID-19 test.

1. Gadd, D.A., et al., *Epigenetic scores for the circulating proteome as tools for disease prediction.* Elife, 2022. **11**: p. e71802.
2. Lin, D., *On the Breslow estimator.* Lifetime data analysis, 2007. **13**(4): p. 471-480.
3. Ridgeway, G., *Generalized Boosted Models: A guide to the gbm package.* Update, 2007. **1**(1): p. 2007.
4. Friedman, J., T. Hastie, and R. Tibshirani, *Regularization paths for generalized linear models via coordinate descent.* Journal of statistical software, 2010. **33**(1): p. 1.
5. Simon, N., et al., *Regularization paths for Cox's proportional hazards model via coordinate descent.* Journal of statistical software, 2011. **39**(5): p. 1.
6. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.
7. Ishwaran, H., et al., *Random survival forests.* Annals of Applied Statistics, 2008. **2**(3): p. 841-860.
8. Ishwaran, H. and U. Kogalur, *Fast unified random forests for survival, regression, and classification (RF-SRC).* R package version, 2019. **2**(1).
9. Chipman, H.A., E.I. George, and R.E. McCulloch, *BART: Bayesian additive regression trees.* The Annals of Applied Statistics, 2010. **4**(1): p. 266-298.
10. Sparapani, R.A., et al., *Nonparametric survival analysis using Bayesian additive regression trees (BART).* Statistics in medicine, 2016. **35**(16): p. 2741-2753.
11. Sparapani, R., C. Spanbauer, and R. McCulloch, *Nonparametric machine learning and efficient computation with bayesian additive regression trees: the BART R package.* Journal of Statistical Software, 2021. **97**(1): p. 1-66.
12. Holle, R., et al., *KORA-a research platform for population based health research.* Das Gesundheitswesen, 2005. **67**(S 01): p. 19-25.

13. Fawns-Ritchie, C., et al., *CovidLife: a resource to understand mental health, well-being and behaviour during the COVID-19 pandemic in the UK.* Wellcome Open Research, 2021. **6**(176): p. 176.