

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

Received: 19 November 2021

Accepted: 27 February 2023

Published online: 6 April 2023

 Check for updates

Yipeng Cheng ^{1,2}, Danni A. Gadd ¹, Christian Gieger ^{3,4,5}, Karla Monterrubio-Gómez ⁶, Yufei Zhang¹, Imrich Berta ¹, Michael J. Stam⁷, Natalia Szlachetka⁷, Evgenii Lobzaev ⁷, Nicola Wrobel⁸, Lee Murphy ⁸, Archie Campbell ¹, Cliff Nangle ¹, Rosie M. Walker^{1,9}, Chloe Fawns-Ritchie^{1,10}, Annette Peters ^{4,5,11}, Wolfgang Rathmann^{5,12}, David J. Porteous ¹, Kathryn L. Evans¹, Andrew M. McIntosh ¹³, Timothy I. Cannings ¹⁴, Melanie Waldenberger ^{3,4}, Andrea Ganna², Daniel L. McCartney ¹, Catalina A. Vallejos ^{6,15}  & Riccardo E. Marioni ¹ 

Type 2 diabetes mellitus (T2D) presents a major health and economic burden that could be alleviated with improved early prediction and intervention. While standard risk factors have shown good predictive performance, we show that the use of blood-based DNA methylation information leads to a significant improvement in the prediction of 10-year T2D incidence risk. Previous studies have been largely constrained by linear assumptions, the use of cytosine–guanine pairs one-at-a-time and binary outcomes. We present a flexible approach (via an R package, MethylPipeR) based on a range of linear and tree-ensemble models that incorporate time-to-event data for prediction. Using the Generation Scotland cohort (training set $n_{\text{cases}} = 374$, $n_{\text{controls}} = 9,461$; test set $n_{\text{cases}} = 252$, $n_{\text{controls}} = 4,526$) our best-performing model (area under the receiver operating characteristic curve (AUC) = 0.872, area under the precision–recall curve (PRAUC) = 0.302) showed notable improvement in 10-year onset prediction beyond standard risk factors (AUC = 0.839, precision–recall AUC = 0.227). Replication was observed in the German-based KORA study ($n = 1,451$, $n_{\text{cases}} = 142$, $P = 1.6 \times 10^{-5}$).

Diabetes mellitus is one of the most prevalent diseases in the world and a leading cause of mortality. Around half a billion people live with diabetes worldwide, with type 2 diabetes (T2D) making up about 90% of these cases¹. Individuals with diabetes can suffer from debilitating complications including nerve damage, kidney disease and blindness².

The disease also increases the future risk of dementia and cardiovascular disease³, with recent studies highlighting obesity and T2D as risk factors for coronavirus disease 2019 (COVID-19) disease severity and intensive care unit admission⁴. Furthermore, the risk of complications increases over time and is exacerbated if blood glucose levels are poorly


A full list of affiliations appears at the end of the paper.  e-mail: catalina.vallejos@ed.ac.uk; riccardo.marioni@ed.ac.uk

Table 1 | Summary information for the Generation Scotland training and test sets

	Training		Test	
	Cases	Controls	Cases	Controls
<i>n</i>	374	9,461	252	4,526
TTE (years to onset or censoring)	5.7 (3.4)	11.1 (1.8)	5.9 (3.4)	11.3 (1.7)
Age (onset or censoring)	61.2 (10.7)	58.1 (14.6)	60.4 (9.4)	59.2 (13.9)
Sex (male)	184 (49.2)	3,903 (41.3)	133 (52.8)	1,681 (37.1)
BMI (kg m ⁻²)	31.7 (5.7)	26.3 (4.8)	32.2 (6.2)	26.5 (5.0)
Self-reported parent or sibling diabetes	137 (36.6)	1,553 (16.4)	105 (41.7)	858 (19.0)
Self-reported hypertension	117 (31.3)	1,022 (10.8)	90 (35.7)	575 (12.7)

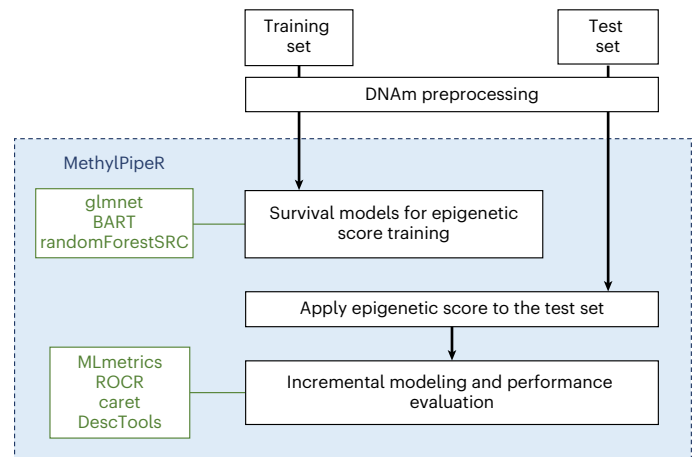
Summary information is shown as the mean (s.d.) or *n* (%).

managed. Despite developments in the way T2D can be managed for patients, these treatments are reactive, focusing on patients that have already been diagnosed. Early intervention with metformin or lifestyle changes have been shown to delay the onset of T2D, although they do not reduce the risk of all-cause mortality⁵.

Beyond public health costs, T2D also presents a substantial financial burden to the National Health Service (NHS), with estimated annual spending of £10 billion on diabetes in the UK. Around 80% of these costs are for the treatment of complications, many of which are preventable with early intervention⁶.

While the mechanisms of insulin resistance in T2D are well known, the interactions between genetic and environmental factors that increase T2D susceptibility are less understood. Previous T2D risk prediction models have used a range of health risk factors⁷. However, these have not used epigenetic information. Epigenetics is the study of heritable changes to DNA that do not modify its nucleotide sequence. A commonly studied form of this is DNA methylation (DNAm), whereby methyl groups are attached to the DNA molecule, most commonly to the five-carbon on a cytosine in a cytosine–guanine pair (CpG). Due to its involvement with gene expression and gene environment interactions, DNAm can provide dynamic predictive information for disease risk for an individual. For example, epigenetic scores built via penalized regression models have been used to show that weighted linear CpG predictors can explain a substantial proportion of phenotypic variance (R^2) of modifiable health factors including body mass index (BMI) (12.5%), high-density lipoprotein cholesterol (15.6%) and smoking status (60.9%)⁸. Blood-based DNAm is of particular interest in predictive modeling and biomarker development because of its comparatively noninvasive sampling procedure. Epigenetic scores have also shown the ability to explain that up to 58% of variance in plasma protein levels are associated with several incident diseases including T2D and several comorbidities⁹. Epigenome-wide association studies (EWAS) have identified a number of CpG sites significantly associated with T2D^{10–14} and related risk factors such as cardiovascular disease¹⁵ and obesity^{16,17}. While these provide some predictive performance for T2D prevalence, incident T2D has been less well studied. One such EWAS with 563 cases and 701 controls identified 18 CpGs associated with incident T2D but did not consider any prediction models¹⁰. Given that preventive lifestyle changes have been shown to effectively reduce T2D onset¹⁸, prediction of T2D incidence years ahead of time would be greatly beneficial in stratifying populations so those at high risk can be monitored and treated with early interventions.

Currently, most studies generating DNAm predictors consider marginal CpG effects or assume only linear additive effects between CpGs. The use of predictive models that can incorporate both interaction and nonlinear effects could capture more complex relationships between variables, resulting in greater prediction accuracy. Therefore, our study aimed to evaluate both the additional predictive benefit that

**Fig. 1 | The prediction pipeline and functionality provided in MethylPipeR.**

MethylPipeR is an R package designed to facilitate reproducible prediction pipelines using DNAm or other types of high-dimensional omics data. The green text boxes indicate functionality incorporated from external R packages. The blue area indicates the functionality included within MethylPipeR.

DNAm can provide for 10-year T2D risk and the applicability of linear and tree-ensemble survival models.

In this study, we use one of the world's largest studies with paired genome-wide DNAm and data linkage to electronic health records, Generation Scotland ($n = 14,613$, $n = 626$ incident T2D cases over 15 years of follow-up), to develop and validate epigenetic scores for T2D. We show the added contribution of these epigenetic scores to prediction over and above standard risk factors, for example, age, sex and BMI and externally validate these results in the KORA S4 cohort.

Results

After preprocessing, the mean time-to-onset of T2D was 5.7 and 5.9 years for the training ($n = 374$ cases) and test ($n = 252$ cases) sets, respectively. Mean age at onset was also similar between the training and test set at 61.2 and 60.4 years and the mean BMI for cases (at baseline) was 31.7 and 32.2 kg m⁻². The full set of cohort summary details for cases and controls in both sets are shown in Table 1. The preprocessing steps and sample sizes at each step are shown in Extended Data Fig. 1. The machine learning prediction pipeline of the MethylPipeR package is shown in Fig. 1.

Null model for the incremental modeling approach

A Cox proportional-hazards model in the test set with age, sex, BMI, self-reported hypertension and family history of diabetes as predictors yielded good classification metrics: area under the receiver operating

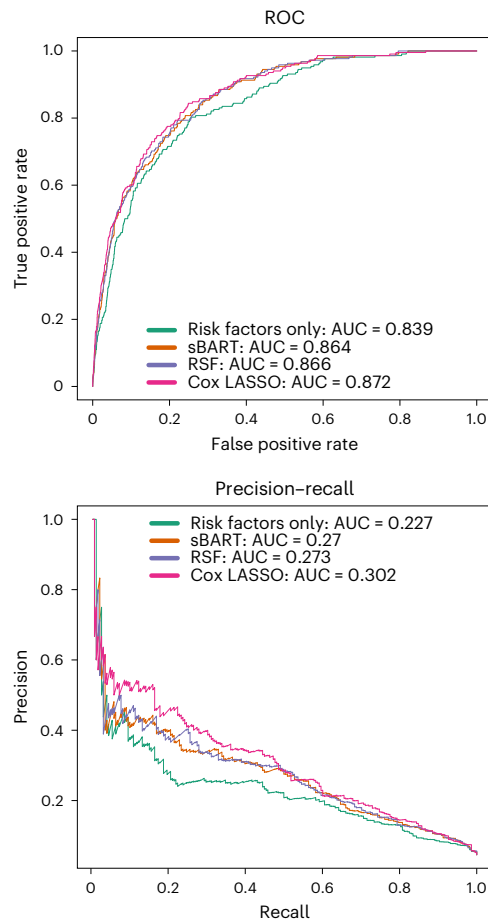


Fig. 2 | ROC and precision–recall curves for the full models. Full models include the risk factors, composite protein epigenetic score and either the Cox proportional-hazards LASSO, RSF or sBART direct epigenetic score.

characteristic curve (AUC) = 0.839, area under the precision-recall curve (PRAUC) = 0.227.

Incremental model using the direct epigenetic score

In the risk factors plus direct epigenetic score test set model, Cox proportional-hazards with least shrinkage and selection operator (LASSO) performed the best, showing an AUC and PRAUC of 0.870 and 0.299, respectively ($P = 3.6 \times 10^{-27}$ for the epigenetic score coefficient). This corresponds to an increase of 3.1% and 7.2% over the standard risk factor model.

Overall, the tree-ensemble models used for the direct epigenetic score resulted in lower performance compared to the Cox proportional-hazards LASSO. AUC values for random survival forests (RSF) and survival Bayesian additive regression trees (sBART) were 0.852 and 0.840 and the PRAUC values were 0.247 and 0.230, respectively (Supplementary Table 1). *P* values for the epigenetic score coefficients and receiver operating characteristic (ROC) curves for all models are given in Supplementary Table 2.

Incremental model using the composite protein epigenetic score

The composite protein epigenetic score (with 109 possible input protein epigenetic score features) derived using a Cox proportional-hazards LASSO model showed good performance with AUC and PRAUC of 0.864 and 0.270, respectively (epigenetic score coefficient $P = 1.61 \times 10^{-18}$). The increase in PRAUC was smaller for the composite protein epigenetic score compared to the direct epigenetic score but still a notable improvement over using risk factors only.

Incremental model using composite protein plus direct epigenetic scores

The full model (risk factors plus composite protein epigenetic score plus direct epigenetic score) with a Cox proportional-hazards LASSO direct epigenetic score gave an AUC and PRAUC of 0.872 and 0.302, respectively. The full models with RSF and sBART-derived direct epigenetic scores showed AUCs of 0.866 and 0.864, respectively. The corresponding PRAUC values were 0.273 and 0.270. Therefore, the best overall model used direct and composite protein epigenetic scores from Cox proportional-hazards LASSO models. The ROC and precision–recall curves for the full models and risk factor-only model are shown in Fig. 2. We also examined if our findings were robust to potential lag effects in T2D diagnosis¹⁹. Increases to both the AUC and PRAUC were observed when adding the epigenetic scores to a risk factor-only model after excluding cases diagnosed within the first 2 years of follow-up (Supplementary Table 3).

Binary classification metrics and model calibration

Supplementary Table 4 shows how confusion matrix metrics vary for the null (risk factor-only) model and the Cox proportional-hazards LASSO model across a range of probability classification thresholds. As expected, as the classification probability threshold is increased, sensitivity and negative predictive value decrease while specificity increases. The effects of these differences on the number of true positives and false negatives are illustrated in Fig. 3. The two models also show differences in their calibration plots (Extended Data Fig. 2). In addition, the difference in the number of correctly classified individuals between the two models are given. These were calculated assuming, arbitrarily, a 10-year incidence rate of 33%, for example, in a scenario where high-risk individuals have been selected for screening.

Selected CpGs

The Cox proportional-hazards LASSO model assigned nonzero coefficients to 145 CpGs (Supplementary Table 5). After filtering the EWAS Catalog by *P* value ($P < 3.6 \times 10^{-8}$)²⁰ and sample size ($n > 1,000$), 119 (82%) of the model-selected CpGs were present. These CpGs corresponded to 742 entries and showed epigenome-wide associations with traits including serum high-density lipoprotein cholesterol, serum triglycerides, smoking, C-reactive protein, BMI and age (Supplementary Table 6).

Selected protein epigenetic scores

The composite protein epigenetic score Cox proportional-hazards LASSO model assigned nonzero coefficients to 46 protein epigenetic scores. Details on the corresponding proteins and genes are given in Supplementary Table 7. Out of the selected protein epigenetic scores, 21 previously showed associations with incident T2D⁹.

Validation in the KORA S4 cohort

Prediction of incident diabetes in the KORA S4 cohort using the Cox proportional-hazards LASSO model showed good replication of direct epigenetic score performance ($P = 1.6 \times 10^{-5}$) with increases of 1.6% in absolute terms above the null model values for both AUC and PRAUC. Further details are provided in Supplementary Table 8.

Epigenetic score prediction of COVID-19 outcomes

In all models, incident T2D was predictive of hospitalization with COVID-19 infection. However, neither the composite protein nor the direct epigenetic score were predictive of the same outcome (Supplementary Table 9). Additionally, neither the (direct or protein-based) epigenetic scores nor incident T2D were predictive of ongoing symptomatic COVID-19 after COVID-19 infection.

Discussion

Using a large cohort with genome-wide epigenetic data and health records linkage to longitudinal primary and secondary care data, we

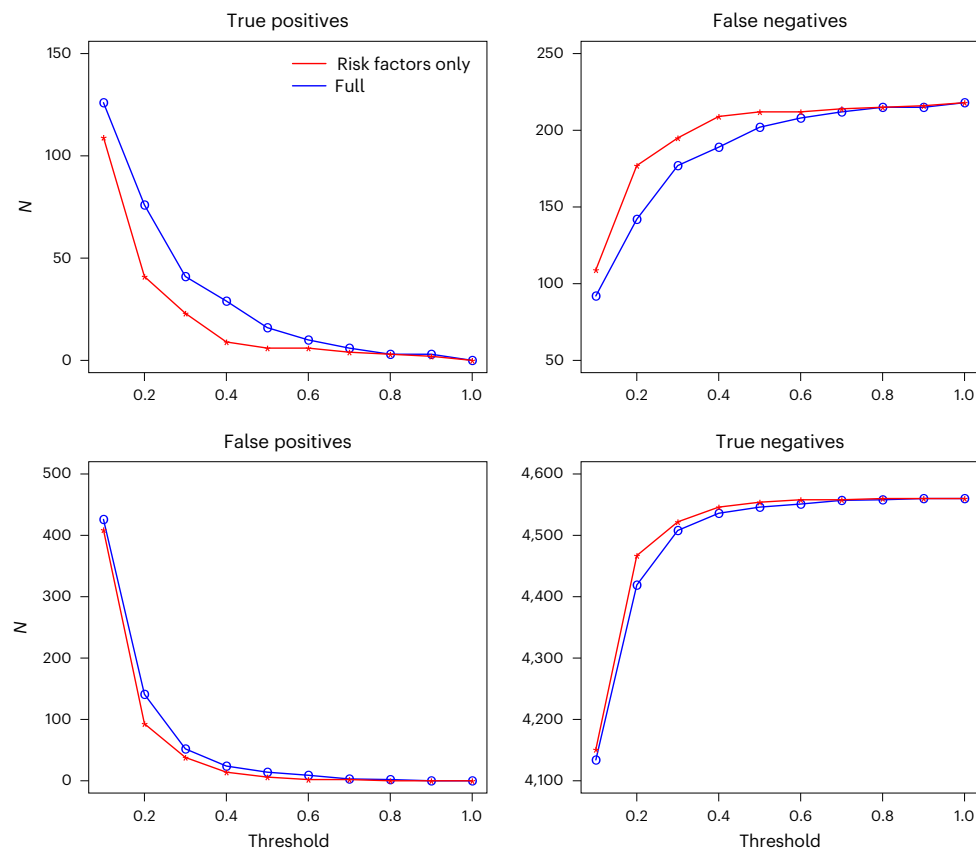


Fig. 3 | Confusion matrix plot of true positives/false negatives and false positives/true negatives in the Generation Scotland test dataset. In each panel, the full model is shown in blue and the risk factors only model is shown in red. (The full model uses direct and composite protein epigenetic scores from the Cox proportional-hazards LASSO model.)

showed that DNAm-based predictors augment standard risk factors in the prediction of incident T2D. The best model with traditional risk factors yielded an AUC of 0.839 compared to 0.872 when DNAm was also considered and the PRAUC increased from 0.227 to 0.305. Using several linear and nonlinear survival models, we showed that overall, the Cox proportional-hazards LASSO model produced the most predictive direct epigenetic score. A composite protein epigenetic score also notably increased predictive performance. The direct epigenetic score also showed good external validation performance in the KORA S4 cohort. Beyond the T2D analysis presented in this study, we developed the MethylPipeR R package, along with a user interface version MethylPipeR-UI (shown in Extended Data Fig. 3), to facilitate reproducible machine learning time-to-event (TTE) and binary prediction using DNAm or other types of high-dimensional omics data.

Determining a ‘best’ model is complicated and depends on the trade-off that a user wishes to make. In this study, we optimized AUC and PRAUC but binary classification metrics vary by method or classification threshold. When using classifiers in clinical settings, decisions need to be made about the number of patients that can be recommended for intervention and the acceptable proportion of false positives and false negatives. We showed an increase in the correct identification of positives and negatives at varying probability thresholds when adding direct and composite epigenetic scores above standard risk factors. For instance, given an (arbitrary) incidence rate of 33% (commonly used as a cutoff for high risk of T2D)²¹ over 10 years in a sample of 10,000 individuals, our best model would correctly classify an additional 449 individuals compared to the risk factor-only model at a threshold of 0.2 (Supplementary Table 4). Given the costs of treating T2D-related complications, our study gives evidence for possible benefits of epigenetic scores on public health that could also alleviate the

financial burden to the NHS. In addition, an assessment of calibration is also critical^{22,23}. Investigation of these related criteria could assist in deciding an optimal threshold given clinical constraints and provide a more comprehensive assessment of model predictions than AUCs or metrics at the commonly used threshold of 0.5.

Several CpGs from the direct epigenetic score were previously identified as epigenome-wide significant correlates of traits commonly linked to T2D^{14,17,24–28}. Future work could investigate the overlap between these and TTE EWAS studies. Further studies could also include DNAm predictors for traditional risk factors that are included in the null model, such as BMI⁶.

Limitations include the relatively small number of disease cases in the dataset, the limited hyperparameter optimization performed for sBART and the relatively simple variable preselection method for tree-ensemble methods. Given the lower performance of these methods compared to the best models in this study, there is potential for additional improvement in predictive performance with further investigation of more advanced feature (pre)selection. This is particularly important when we consider that the preselection step used linear models before the nonlinear model fitting. Model fitting and preselection were also performed using the same training set, which may have introduced issues associated to post-selection inference^{29,30}. In addition, factors such as overfitting, related individuals in the test set and batch effects between the three rounds of DNAm data processing may all have had an effect on the test set AUC. Future studies may also take into account multimorbidities because the presence of competing risks can lead to bias in onset predictions³¹. Finally, a small proportion of the linkage codes used to define diabetes included broad terms that were nonspecific to T2D; however, the late age of onset in these individuals meant that there was a high likelihood that they had

developed T2D. Epigenetic scores for T2D-associated proteins have also been shown to replicate incident T2D–protein associations in this sample⁹, suggesting that the case definitions we used captured biological signals relevant to T2D.

There are many strengths to our study. First, the models used captured relationships between CpGs and TTE information, which is not possible using traditional EWAS methods. Second, data linkage to health care measures provided comprehensive T2D incidence data in a very large cohort study, that is, Generation Scotland. Validation performance in the KORA S4 cohort also strengthened evidence for the applicability of the models to other populations. Finally, the R package MethylPipeR encourages reproducibility and allows others to develop similar predictors on new data with minimal setup.

In conclusion, we have demonstrated the potential for DNAm data to provide notable improvement in predictive performance for incident T2D, compared to traditional risk factors (age, sex, BMI, hypertension and family history). We evaluated different models with a systematic approach and presented a framework with the ability to generalize to other traits and datasets for training and testing predictors in future studies.

Methods

Statistics and reproducibility

To enhance reproducibility, the analysis pipeline and results presented in this study have been reported using the Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD) checklist³² (Supplementary File 1). Sample sizes for training, testing and validation of the statistical models were determined by the cohort sizes. Exclusions were made based on DNAm quality control and missingness in health data. Randomization and blinding were not applicable during data collection because the study used observational data from population-based cohorts. Statistical tests for model coefficients assumed a normal distribution but this was not formally tested.

Generation Scotland

Blood-based DNAm and linked health data were obtained from Generation Scotland³³, a family-structured, population-based cohort. The cohort consists of 23,960 volunteers across Scotland aged 18–99 years at recruitment (between 2006 and 2011), of whom over 18,000 currently have genome-wide DNAm data available (Illumina EPIC array). In DNAm quality control, CpG sites were filtered by removing those with a low bead count in 5% or more of the samples or a high detection *P* value (>0.05) in more than 5% of samples. Probes on the X and Y chromosomes were also removed. Samples were filtered by removing those with a mismatch between predicted and recorded sex or 1% or more of CpGs with a detection *P* > 0.05. Missing CpG values were mean-imputed. To enable the predictors to be applied to existing cohort studies with older Illumina array data, CpGs were filtered to the intersection of the 450k and EPIC array sites (*n* = 453,093 CpGs).

This study considered DNAm data from three large subsets of the Generation Scotland cohort, with 5,087 (set 1), 4,450 (set 2) and 8,877 (set 3) individuals. Processing took place in 2017, 2019 and 2021, respectively. Set 1 and set 3 included related individuals within and between each set while all individuals in set 2 were unrelated to each other and to individuals in set 1 (genetic relationship matrix threshold < 0.05). In our experiments, the training set consisted of sets 2 and 3 combined; set 1 was used as the test set. To avoid the presence of families with individuals across both training and test sets, any individuals in the training set from the same family as an individual in the test set were excluded from the analysis (*n*_{excluded} = 3,138).

Participant health measures including age, BMI, sex, self-reported hypertension and family (parent or sibling) history of T2D were taken at baseline (DNAm sampling) via questionnaire. BMI was calculated as the individual's weight in kilograms divided by the square of their height in meters. Missing values in the set 1 health measures were treated as

missing completely at random and the corresponding individuals were excluded (*n*_{set1} = 99). This was not performed in sets 2 and 3 because the health measures were used for incremental modeling (set 1 only).

Disease cases were ascertained through data linkage to NHS Scotland health records consisting of hospital (International Statistical Classification of Diseases and Related Health Problems, 10th Revision (ICD-10) codes) and GP records (Read2 codes). Prevalent cases were identified from a baseline questionnaire (self-reported) or from ICD-10/Read2 codes dated before baseline and removed from the dataset. Type 1 and juvenile cases were treated as control observations. A total of 757 incident cases were observed over the follow-up period (from the recruitment date to January 2022); after preprocessing, 626 cases remained. Mean time to T2D onset was 5.9, 5.4 and 6.0 years for sets 1, 2 and 3, respectively, with ranges of 0.2–14.8 (set 1), 0.2–14.8 (set 2) and 0.1–14.8 (set 3) years. In GP record-derived cases, 81% of cases had a C10F 'type 2 diabetes mellitus' code; 12% had a C10 'diabetes mellitus' code and 4% had a C109 'non-insulin dependent diabetes mellitus' code. The full list of included and excluded terms are given in Supplementary Table 10.

Composite protein epigenetic score

A composite protein epigenetic score model for incident T2D was trained using a set of 109 protein epigenetic scores previously shown to associate with a range of incident diseases⁹. For each protein, the epigenetic score was calculated for each individual in the training and test sets. A Cox proportional-hazards LASSO model was fitted to the training set with the 109 protein epigenetic scores (scaled within set to a mean of 0 and variance of 1) as features. The linear predictor from the Cox proportional-hazards LASSO model was then used as the composite protein epigenetic score in the incremental test set model.

Direct epigenetic score

The direct epigenetic score Cox proportional-hazards LASSO model for incident T2D was fitted to the DNAm data in the training set. Due to memory limitations in the model fitting R package (glmnet³⁴), the CpGs were filtered to the 200,000 sites with the highest variance. The linear predictor from the Cox proportional-hazards LASSO model was then used as the direct epigenetic score in the incremental test set modeling. For the tree-ensemble models, the Cox proportional-hazards LASSO-selected CpGs were used as input and the 10-year onset risk was subsequently used as the direct epigenetic score.

Outcome definition for the 10-year onset prediction

The link to NHS Scotland health records provided dates for disease diagnoses from which age at onset was calculated. Along with age at baseline (DNAm sampling), these were used to calculate the TTE, measured in years, for each individual. For incident T2D cases and controls, TTE was defined as the time from baseline to disease onset and censoring, respectively. Controls were censored at the latest date of available GP records (September 2020). In addition, controls who died during the follow-up period were censored at time of death.

While all models were trained as survival models, our primary prediction outcome was incident T2D diagnosis within 10 years. Therefore, predictions on the test set were calculated using the 10-year onset probability (one minus survival probability). When calculating the binary outcome metrics, cases with TTE > 10 were treated as controls. These metrics included confusion matrices, AUC and PRAUC. After preprocessing and case thresholding (TTE > 10), there were 218 cases and 4,560 controls in the test set.

The numbers of individuals, cases and controls after each preprocessing step are shown in Extended Data Fig. 1.

Incremental modeling

Composite protein epigenetic scores and direct epigenetic scores were generated in the training dataset using different machine learning

methods with the MethylPipeR package (Fig. 1), before being applied to the test set using an incremental modeling approach (further detail in the Supplementary Note). In the test set, a (null) risk factor-only model was defined via a Cox proportional-hazards model for T2D with age, sex, BMI, self-reported hypertension and self-reported family (sibling or parental) history of diabetes as predictors. A multitude of risk factors have been used in previous T2D risk prediction tools, most of which include the set that we have used in this study. While additional risk factors, such as waist:hip ratio, may also be relevant⁷, we selected the null model covariates based on those used in the Diabetes UK: Type 2 Diabetes Know Your Risk tool (<https://riskscore.diabetes.org.uk/start>) to compare our results to an existing widely used tool. This was with the exception of ethnicity because of the relative homogeneity of the Generation Scotland cohort. These also closely match the top risk factors identified in a systematic review of previous T2D risk predictors (Fig. 2 in Collins et al.⁷).

Penalized regression predictors

Because the number of CpGs ($n_{\text{CpG}} = 200,000$) was much greater than the number of rows in the training set ($n = 9,835$ after preprocessing), a regularization method was required to reduce overfitting of the Cox proportional-hazards regression models.

Penalized regression models reduce overfitting by applying a regularization penalty in the model fitting process. This forces regression parameters to remain small or possibly to shrink them to zero. The latter allows the method to be used for variable selection by keeping only the variables with resulting nonzero coefficients.

LASSO penalization was fitted to the training set DNAm and protein epigenetic scores using glmnet^{34,35} via MethylPipeR with the best shrinkage parameter (λ) chosen by nine-fold cross-validation.

Tree-ensemble models

Tree ensembles are nonparametric models capable of estimating complex functions using a set of decision trees. Two tree-ensemble approaches were used: RSF³⁶ and sBART³⁷. RSF³⁸ 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 from the training set to prevent individual trees from overfitting to the whole dataset. sBART is a nonparametric method that estimates a function as a sum over a set of regression trees. sBART incorporates the ability to model both additive and interaction effects and has shown high predictive performance compared with similar methods^{37,39}.

RSF and sBART were fitted to the training set using the R packages randomForestSRC (v.2.11.0)⁴⁰ and BART (v.2.9)⁴¹, respectively via MethylPipeR. Details on hyperparameter selection are given in the Supplementary Note.

Because of computational limitations and probable overfitting in using the tree-ensemble models on all CpGs in the dataset, variable preselection was based on the coefficients in the penalized Cox proportional-hazards models. Each tree-ensemble model was evaluated with the features corresponding to nonzero coefficients from the Cox proportional-hazards LASSO model.

Evaluating predictive performance

Survival models can be used to predict the risk of incident T2D for an arbitrary prediction period. In this study, we focused on classification performance for the binary outcome defined by a 10-year T2D incidence. Incidence probabilities were calculated as one minus 10-year survival probabilities and the binary outcomes were calculated by truncating the observed TTE at 10 years (see 'Outcome definition for the 10-year onset prediction' section of this article).

The AUC and PRAUC were calculated as measures of predictive performance because they do not require the choice of a fixed discrimination threshold. PRAUC is more informative in situations where there

is a class imbalance in the test set⁴². Additionally, binary classification metrics consisting of sensitivity (recall), specificity, positive predictive value (precision) and negative predictive value were calculated. These metrics require selection of a discrimination threshold to assign positive and negative class predictions. We evaluated their behavior across a range of discrimination thresholds, between 0 and 1 in increments of 0.1.

Differences in correctly classified individuals between the risk factor-only and Cox proportional-hazards LASSO models were calculated assuming, arbitrarily, a 10-year incidence rate of 33%, for example, in a scenario where high-risk individuals have been selected for screening in a population of 10,000. The numbers of true positives and true negatives were calculated as follows:

$TP = \text{sensitivity} \times N_{\text{actual positives}}$ and $TN = \text{specificity} \times N_{\text{actual negatives}}$ respectively, where $N_{\text{actual positives}} = 3,300$ and $N_{\text{actual negatives}} = 7,700$. The difference between the two was then taken at each classification threshold.

Model calibration was examined by comparing predicted probabilities with actual case or control proportions⁴³.

Selected CpG comparison with the EWAS Catalog

The CpG sites selected by the Cox proportional-hazards LASSO model were queried in the EWAS Catalog⁴⁴ to identify traits that have previously been linked to these sites. The EWAS Catalog is a database that allows users to search EWAS results from the existing literature. We performed a tissue-agnostic query using the selected CpGs and filtered results to those with an epigenome-wide significance threshold of $P < 3.6 \times 10^{-8}$ (ref. 20) in studies with a sample size greater than 1,000. Almost all (739 out of 742; 99.6%) of the results after filtering were from blood-based studies. The remaining results were from saliva and prefrontal cortex-based studies.

Validation in the KORA S4 cohort

The Cox proportional-hazards LASSO model using the direct epigenetic score was applied to the KORA S4 cohort⁴⁵. This cohort consisted of 1,451 individuals in southern Germany, aged 25–74 years. Cohort summary details are shown in Supplementary Table 11. Individuals with missing values in the health measures were removed from the dataset. Missing CpG values in the DNAm data were mean-imputed. Like the approach in the Generation Scotland test set, an epigenetic score was computed for each individual in the KORA S4 dataset. Evaluation was then performed using an incremental modeling approach. Additional cohort and methods details such as the outcome definition, follow-up period and preprocessing numbers are provided in the Supplementary Note.

Epigenetic score prediction of COVID-19 outcomes

The subset of the Generation Scotland cohort with reported COVID-19 infection (clinically diagnosed or positive test from linked test data), who had also participated in the CovidLife study⁴⁶ were used to predict ongoing symptomatic COVID-19 and hospitalization from COVID-19 ($n = 703$). Ongoing symptomatic COVID-19 cases were defined as individuals with self-reported symptoms lasting 4 weeks or longer⁴⁷. Hospitalization cases were defined as hospital admissions with accompanying ICD-10 codes U07.1 (confirmed COVID-19 test) and U07.2 (clinically diagnosed), derived from the Scottish Morbidity Records (SMR01). Details of the method and summary statistics are shown in Supplementary Note and Supplementary Table 12.

Ethics approval and consent to participate

All components of Generation Scotland received ethical approval from the NHS Tayside Committee on Medical Research Ethics (research ethics committee (REC) ref. no. 05/S1401/89). Generation Scotland has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC ref. no. 20-ES-0021), providing generic ethical approval for a wide range of uses within medical research.

Written, informed consent was provided by Generation Scotland participants.

The KORA S4 studies were approved by the ethics committee of the Bavarian Medical Association (no. 99186) and were conducted according to the principles expressed in the Declaration of Helsinki (World Medical Association Declaration of Helsinki 2008). All study participants gave their written informed consent.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

According to the terms of consent for Generation Scotland participants, access to data must be reviewed by the Generation Scotland Access Committee. Applications should be made to access@generationscotland.org. The informed consent given by the KORA S4 study participants does not cover data posting in public databases. However, data are available upon request from the KORA Project Application Self-Service Tool (<https://epi.helmholtz-muenchen.de/>). Data requests can be submitted online and are subject to approval by the KORA board.

Code availability

Analysis scripts for this study are available at <https://github.com/marioni-group/episcoroes-diabetes-prediction> and <https://doi.org/10.5281/zenodo.7628959>. MethyPipeR v.0.1.0 is available at <https://github.com/marioni-group/MethyPipeR> and <https://doi.org/10.5281/zenodo.7628816>. MethyPipeR-UI is available at <https://github.com/marioni-group/MethyPipeR-UI> and <https://doi.org/10.5281/zenodo.7635952>.

References

- Saeedi, P. et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res. Clin. Pract.* **157**, 107843 (2019).
- Gregg, E. W., Sattar, N. & Ali, M. K. The changing face of diabetes complications. *Lancet Diabetes Endocrinol.* **4**, 537–547 (2016).
- Biessels, G. J. & Despa, F. Cognitive decline and dementia in diabetes mellitus: mechanisms and clinical implications. *Nat. Rev. Endocrinol.* **14**, 591–604 (2018).
- McGurnaghan, S. J. et al. Risks of and risk factors for COVID-19 disease in people with diabetes: a cohort study of the total population of Scotland. *Lancet Diabetes Endocrinol.* **9**, 82–93 (2021).
- Lee, C. G. et al. Effect of metformin and lifestyle interventions on mortality in the Diabetes Prevention Program and Diabetes Prevention Program Outcomes Study. *Diabetes Care* **44**, 2775–2782 (2021).
- Keng, M. J. et al. Impact of achieving primary care targets in type 2 diabetes on health outcomes and healthcare costs. *Diabetes Obes. Metab.* **21**, 2405–2412 (2019).
- Collins, G. S., Mallett, S., Omar, O. & Yu, L.-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC Med.* **9**, 103 (2011).
- McCartney, D. L. et al. Epigenetic prediction of complex traits and death. *Genome Biol.* **19**, 136 (2018).
- Gadd, D. A. et al. Epigenetic scores for the circulating proteome as tools for disease prediction. *eLife* **11**, e71802 (2022).
- Cardona, A. et al. Epigenome-wide association study of incident type 2 diabetes in a British population: EPIC-Norfolk study. *Diabetes* **68**, 2315–2326 (2019).
- Meeks, K. A. C. et al. Epigenome-wide association study in whole blood on type 2 diabetes among sub-Saharan African individuals: findings from the RODAM study. *Int. J. Epidemiol.* **48**, 58–70 (2019).
- Walaszczyk, E. et al. DNA methylation markers associated with type 2 diabetes, fasting glucose and HbA_{1c} levels: a systematic review and replication in a case-control sample of the Lifelines study. *Diabetologia* **61**, 354–368 (2018).
- Al Muftah, W. A. et al. Epigenetic associations of type 2 diabetes and BMI in an Arab population. *Clin. Epigenetics* **8**, 13 (2016).
- Chambers, J. C. et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. *Lancet Diabetes Endocrinol.* **3**, 526–534 (2015).
- Nakatochi, M. et al. Epigenome-wide association of myocardial infarction with DNA methylation sites at loci related to cardiovascular disease. *Clin. Epigenetics* **9**, 54 (2017).
- Wang, X. et al. An epigenome-wide study of obesity in African American youth and young adults: novel findings, replication in neutrophils, and relationship with gene expression. *Clin. Epigenetics* **10**, 3 (2018).
- Wahl, S. et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
- Haw, J. S. et al. Long-term sustainability of diabetes prevention approaches: a systematic review and meta-analysis of randomized clinical trials. *JAMA Intern. Med.* **177**, 1808–1817 (2017).
- Samuels, T. A., Cohen, D., Brancati, F. L., Coresh, J. & Kao, W. H. Delayed diagnosis of incident type 2 diabetes mellitus in the ARIC study. *Am. J. Manag. Care* **12**, 717–724 (2006).
- Saffari, A. et al. Estimation of a significance threshold for epigenome-wide association studies. *Genet. Epidemiol.* **42**, 20–33 (2018).
- Ekoe, J.-M., Goldenberg, R. & Katz, P. Screening for diabetes in adults. *Can. J. Diabetes* **42**, S16–S19 (2018).
- Van Calster, B. et al. Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 230 (2019).
- Van Calster, B. et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J. Clin. Epidemiol.* **74**, 167–176 (2016).
- Demerath, E. W. et al. Epigenome-wide association study (EWAS) of BMI, BMI change and waist circumference in African American adults identifies multiple replicated loci. *Hum. Mol. Genet.* **24**, 4464–4479 (2015).
- Mendelson, M. M. et al. Association of body mass index with DNA methylation and gene expression in blood cells and relations to cardiometabolic disease: a Mendelian randomization approach. *PLoS Med.* **14**, e1002215 (2017).
- Sayols-Baixeras, S. et al. Identification and validation of seven new loci showing differential DNA methylation related to serum lipid profile: an epigenome-wide approach. The REGICOR study. *Hum. Mol. Genet.* **25**, 4556–4565 (2016).
- Braun, K. V. E. et al. Epigenome-wide association study (EWAS) on lipids: the Rotterdam Study. *Clin. Epigenetics* **9**, 15 (2017).
- Kriebel, J. et al. Association between DNA methylation in whole blood and measures of glucose metabolism: KORA F4 study. *PLoS ONE* **11**, e0152314 (2016).
- Lee, J. D., Sun, D. L., Sun, Y. & Taylor, J. E. Exact post-selection inference, with application to the lasso. *Ann. Stat.* **44**, 907–927 (2016).
- Taylor, J. & Tibshirani, R. Post-selection inference for ℓ_1 -penalized likelihood models. *Can. J. Stat.* **46**, 41–61 (2018).
- Austin, P. C., Lee, D. S. & Fine, J. P. Introduction to the analysis of survival data in the presence of competing risks. *Circulation* **133**, 601–609 (2016).
- Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br. J. Surg.* **102**, 148–158 (2015).

33. Smith, B. H. et al. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS: SFHS). The study, its participants and their potential for genetic research on health and illness. *Int. J. Epidemiol.* **42**, 689–700 (2013).
 34. Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for Cox's proportional hazards model via coordinate descent. *J. Stat. Softw.* **39**, 1–13 (2011).
 35. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).
 36. Ishwaran, H., Kogalur, U. B., Blackstone, E. H. & Lauer, M. S. Random survival forests. *Ann. Appl. Stat.* **2**, 841–860 (2008).
 37. Sparapani, R. A., Logan, B. R., McCulloch, R. E. & Laud, P. W. Nonparametric survival analysis using Bayesian additive regression trees (BART). *Stat. Med.* **35**, 2741–2753 (2016).
 38. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
 39. Chipman, H. A., George, E. I. & McCulloch, R. E. BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4**, 266–298 (2010).
 40. Ishwaran, H. & Kogalur, U. Fast unified random forests for survival, regression, and classification (RF-SRC). R package version 2.11.0 (2021).
 41. Sparapani, R., Spanbauer, C. & McCulloch, R. Nonparametric machine learning and efficient computation with Bayesian additive regression trees: the BART R package. *J. Stat. Softw.* **97**, 1–66. (2021).
 42. Saito, T. & Rehmsmeier, M. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **10**, e0118432 (2015).
 43. De Cock, B., Nieboer, D., Van Calster, B., Steyerberg, E. W. & Vergouwe, Y. The CalibrationCurves package: validating predicted probabilities against binary events. R package version 0.1.5. <https://github.com/BavoDC/CalibrationCurves> (2023).
 44. Battram, T. et al. The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res.* **7**, 41 (2022).
 45. Wichmann, H.-E., Gieger, C. & Illig, T. KORA-gen—resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen* **67**, S26–S30 (2005).
 46. 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 Res.* **6**, 176 (2021).
 47. Shah, W., Hillman, T., Playford, E. D. & Hishmeh, L. Managing the long term effects of covid-19: summary of NICE, SIGN, and RCGP rapid guideline. *BMJ* **372**, n136 (2021).
- (awardee: H. C. Whalley). Y.C. is supported by the University of Edinburgh and University of Helsinki joint PhD program in Human Genomics. D.A.G. is supported by funding from the Wellcome Trust 4-year PhD in Translational Neuroscience—training the next generation of basic neuroscientists to embrace clinical research (no. 108890/Z/15/Z). C.A.V. is a Chancellor's Fellow funded by the University of Edinburgh. D.L.M. and R.E.M. are supported by an Alzheimer's Research UK major project grant no. ARUK-PG2017B-10. R.E.M. is supported by an Alzheimer's Society major project grant no. AS-PG-19b-010. M.J.S., N.S. and E.L. are supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. Recruitment to the CovidLife study was facilitated by the Scottish Health Research Register (SHARE) and Biobank. SHARE is supported by NHS Research Scotland, the universities of Scotland and the Chief Scientist Office of the Scottish Government. The KORA S4 study was initiated and financed by the Helmholtz Zentrum München—German Research Center for Environmental Health, which is funded by the German Federal Ministry of Education and Research and by the State of Bavaria. Furthermore, the KORA research has been supported by the Munich Center of Health Sciences, Ludwig-Maximilians-Universität München as part of LMUinnovativ and is supported by the German Centre for Cardiovascular Research. The KORA S4 study is funded by the Bavarian State Ministry of Health and Care through the research project DigiMed Bayern (www.digimed-bayern.de).

Author contributions

Y.C., D.A.G. and C.G. performed the data analysis. Y.Z., I.B., M.J.S., N.S. and E.L. conducted the preliminary analyses. Generation Scotland cohort: N.W. and L.M. were responsible for the data collection. A.C., C.N., R.M.W., K.L.E. and D.L.M. prepared the data. D.J.P. and A.M.M. were responsible for data collection and cohort management. KORA S4 cohort: A.P., W.R. and M.W. were responsible for data collection and preparation. CovidLife: C.F.-R. performed the data collection and preparation. K.M.-G. and T.I.C. provided input on statistical modeling. A.G., C.A.V. and R.E.M. were responsible for study design and methodology. All authors contributed to drafting the paper and the figures.

Competing interests

R.E.M. has received a speaker fee from Illumina and is an advisor to the Epigenetic Clock Development Foundation. A.M.M. has previously received speaker fees from Janssen and Illumina and research funding from The Sackler Trust. L.M. has received payment from Illumina for presentations and consultancy. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at <https://doi.org/10.1038/s43587-023-00391-4>.

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s43587-023-00391-4>.

Correspondence and requests for materials should be addressed to Catalina A. Vallejós or Riccardo E. Marioni.

Peer review information *Nature Aging* thanks Srikanth Bellary and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Acknowledgements

This research was funded in whole, or in part, by the Wellcome Trust (nos. 104036/Z/14/Z, 108890/Z/15/Z and 216767/Z/19/Z). For the purpose of open access, we have applied a CC BY public copyright licence to any author accepted manuscript version arising from this submission. Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates (no. CZD/16/6) and the Scottish Funding Council (no. HR03006) and is currently supported by the Wellcome Trust (no. 216767/Z/19/Z). DNAm profiling of the Generation Scotland samples was carried out by the Genetics Core Laboratory at the Edinburgh Clinical Research Facility and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award 'Stratifying Resilience and Depression Longitudinally' (ref. no. 104036/Z/14/Z)). The DNAm data assayed for Generation Scotland was partially funded by a 2018 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (ref. no. 27404; awardee: D. M. Howard) and by a JMARS SIM fellowship from the Royal College of Physicians of Edinburgh

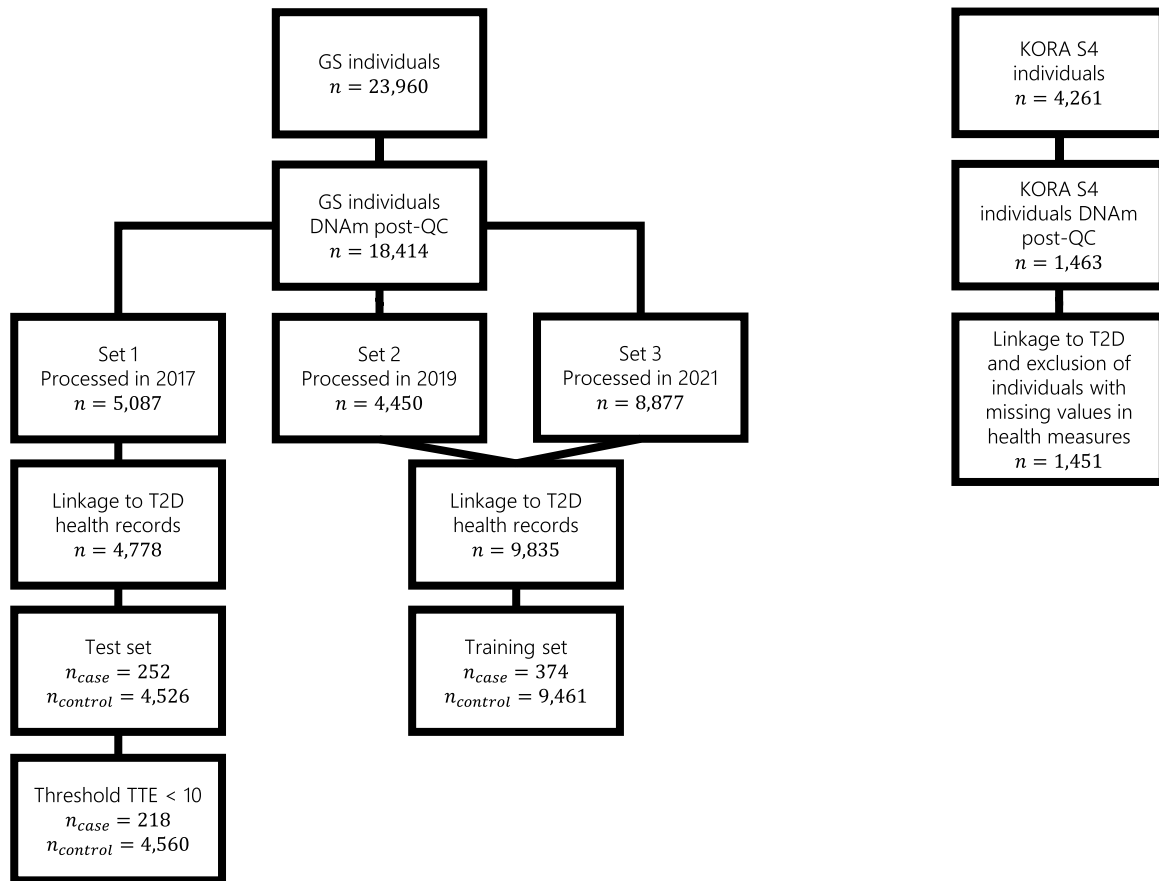
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with

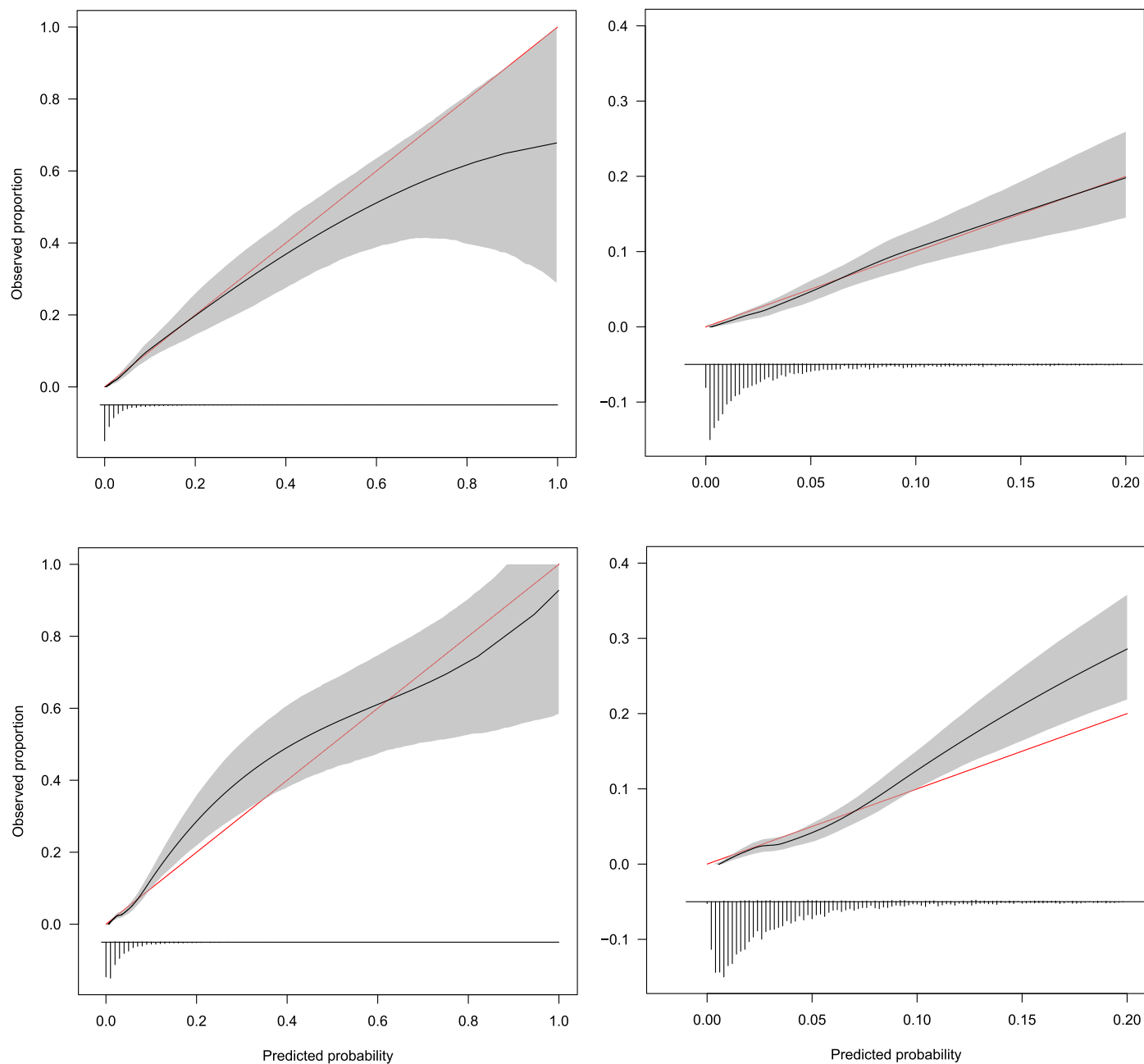
the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2023

¹Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. ²Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland. ³Research Unit Molecular Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. ⁴Institute of Epidemiology, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. ⁵German Center for Diabetes Research, München-Neuherberg, Germany. ⁶MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. ⁷School of Informatics, University of Edinburgh, Edinburgh, UK. ⁸Edinburgh Clinical Research Facility, University of Edinburgh, Western General Hospital, Edinburgh, UK. ⁹Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh BioQuarter, Edinburgh, UK. ¹⁰Department of Psychology, University of Edinburgh, Edinburgh, UK. ¹¹German Centre for Cardiovascular Research, Partner Site Munich Heart Alliance, München, Germany. ¹²Institute for Biometrics and Epidemiology, German Diabetes Center, Leibniz Institute for Diabetes Research at Heinrich Heine University, Düsseldorf, Germany. ¹³Division of Psychiatry, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK. ¹⁴School of Mathematics, University of Edinburgh, Edinburgh, UK. ¹⁵The Alan Turing Institute, London, UK.
✉ e-mail: catalina.vallejos@ed.ac.uk; riccardo.marioni@ed.ac.uk



Extended Data Fig. 1 | Preprocessing steps for Generation Scotland and KORA S4. The number of individuals/cases and controls in are given after each step.



Extended Data Fig. 2 | Calibration plots for incremental models in Generation Scotland. Plots are shown for the full model (risk factors + composite protein epigenetic score + Cox PH lasso direct epigenetic score) (top-left) and the risk factors only model (bottom-left). The black line shows the loess calibration regression curve. The grey area shows 95% confidence intervals calculated from 2000 bootstrap samples. The ideal calibration line (observed

= predicted) is shown in red. The histogram shows the distribution of predicted probabilities. The wider confidence intervals at higher predicted probabilities are due to the small number of predictions in those ranges. Most predictions are low in the probability range, emphasised in the zoomed-in plots (top-right and bottom-right).



Extended Data Fig. 3 | An example from the *MethylPipeR-UI* Shiny app. The left hand panel provides functionality for uploading data and specifying pipeline parameters. The right hand tabs show output such as model diagnostics, performance metrics and console output.

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
- A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.
- A description of all covariates tested
- A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
- A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
- For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
Give P values as exact values whenever suitable.
- For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
- For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
- Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection

Data collection was performed by Generation Scotland (GS) and Cooperative Health Research in the Region of Augsburg (KORA). For further information please contact access@generationscotland.org (GS) and kora-studienzentrum@helmholtz-muenchen.de (KORA). Data cannot be publicly disclosed due to them containing information that could compromise participant consent and confidentiality.

Data analysis

All analysis code was written by the authors and is publicly available on GitHub in the following repositories:

<https://github.com/marioni-group/episcopes-diabetes-prediction>
<https://github.com/marioni-group/MethylPipeR>
<https://github.com/marioni-group/MethylPipeR-UI>

R packages:

- BART (version 2.9)
- gbm (version 2.1.8)
- glmnet (version 4.1-1)
- randomForestSRC (version 2.11.0)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our [policy](#)

According to the terms of consent for GS participants, access to data must be reviewed by the GS Access Committee. Applications should be made to access@generationscotland.org.

Applications for access to KORA should be made using the KORA.PASST system (<http://epi.helmholtz-muenchen.de/>).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

- Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used 9,537 participants from GS for whom we had DNA methylation data. For model validation, we used data from 1,451 participants in the KORA S4 study with DNA methylation and incident T2D data available. Sample sizes for training, testing and validation of the statistical models were determined by the cohort sizes
Data exclusions	Exclusions were made based on DNAm quality control and missingness in health data (outlined under the Generation Scotland section in Methods)
Replication	Replication of model performance was performed through validation in an external cohort, KORA.
Randomization	This study used observational data from a population-based cohort
Blinding	This study used observational data from a population-based cohort

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Involved in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input checked="" type="checkbox"/>	<input type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	<p>Generation Scotland: the Scottish Family Health Study (GS) is a large, family-structured, population-based cohort study of > 24,000 individuals from across Scotland. Of these individuals, 9,537 had DNA methylation measures available. Full summary statistics are shown in Table 1 and Supplementary Table 4.</p> <p>KORA is a research platform performing population-based surveys and subsequent follow-ups in the region of Augsburg in Southern Germany. This study used a subsample of the 1,451 participants of the KORA S4 study with DNA methylation and incident T2D data available. Summary statistics are provided in Supplementary Table 2.</p>
----------------------------	--

Recruitment

Recruitment for Generation Scotland: the Scottish Family Health Study (GS) took place between 2006 and 2011 with a clinical visit where detailed health, cognitive, and lifestyle information was collected along with biological samples (blood, urine, saliva).

Recruitment for KORA S4 took place between 1999 and 2001. Each participant completed a health questionnaire, providing details on health status and medication. Blood samples were also taken for assaying of omics data.

Ethics oversight

All components of Generation Scotland received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). Generation Scotland has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 20-ES-0021), providing generic ethical approval for a wide range of uses within medical research.

The KORA studies were approved by the Ethics Committee of the Bavarian Medical Association (Bayerische Landesärztekammer; S4: #99186) and were conducted according to the principles expressed in the Declaration of Helsinki. All study participants gave their written informed consent.

Note that full information on the approval of the study protocol must also be provided in the manuscript.