**Machine learning-based DNA methylation score for fetal exposure to maternal smoking: development and validation in samples collected from adolescents and adults**

Sebastian Rauschert1, Phillip E Melton2,3,4, Anni Heiskala5, Ville Karhunen6, Graham Burdge7, Jeffrey M Craig8,9, Keith M Godfrey10, Karen Lillycrop11, Trevor A Mori12, Lawrence J Beilin12, Wendy H Oddy4, Craig Pennell13, Marjo-Riitta Järvelin5,6,14, Sylvain Sebert5,15, Rae-Chi Huang1

1 Telethon Kids Institute, University of Western Australia, Nedlands, Perth, Western Australia, Australia

2 Centre for Genetic Origins of Health and Disease, University of Western Australia, Perth, Australia

3 School of Pharmacy and Biomedical Sciences, Faculty of Health Sciences, Curtin University, Perth, Australia

4 Menzies Institute for Medical Research, University of Tasmania, Hobart, Tasmania, Australia

5 Center for Life Course Health Research, University of Oulu, Oulu, Finland

6 Department of Epidemiology and Biostatistics, MRC-PHE Centre for Environment and Health, Imperial College London, London, United Kingdom

7 Institute of Developmental Sciences, University of Southampton, Faculty of Medicine, Southampton, United Kingdom

8 Centre for Molecular and Medical Research, School of Medicine, Deakin University, Geelong, Victoria, Australia

9 Epigenetics, MCRI, Royal Children’s Hospital, Flemington Road, Parkville, Victoria, Australia

10 MRC Lifecourse Epidemiology Unit and NIHR Southampton Biomedical Research Centre, University of Southampton and University Hospital Southampton NHS Foundation Trust, Southampton, UK

11 Biological Sciences, Faculty of Natural and Environmental Sciences, University of Southampton, Southampton, Hampshire, UK

12 Medical School, Royal Perth Hospital Unit, University of Western Australia, Perth, Western Australia

13 University of Newcastle, Newcastle, NSW, Australia

14 Unit of Primary Care, Oulu University Hospital, Oulu, Finland

15 Department of Metabolism, Digestion and Reproduction, Genomic Medicine, Imperial College London, London, United Kingdom

Corresponding author:

Dr Sebastian Rauschert

[Sebastian.Rauschert@telethonkids.org.au](mailto:Sebastian.Rauschert@telethonkids.org.au)

+61863191589

Northern Entrance, Perth Children's Hospital

15 Hospital Ave

Nedlands WA 6009

Competing Financial Interests:

KMG has received reimbursement for speaking at conferences sponsored by companies selling nutritional products and is part of an academic consortium that has received research funding from Abbott Nutrition, Nestec and Danone. The other authors declare they have no actual or potential competing financial interests.

**Abstract**

**Background**

Fetal exposure to maternal smoking during pregnancy is associated with the development of non-communicable diseases in the offspring. It may induce such long-term effects through persistent changes in the DNA-methylome, which therefore holds the potential to be used as a biomarker of this early life exposure. With reducing costs for measuring DNA-methylation, we aimed to develop a DNA-methylation score that can be used on adolescent DNA-methylation data and thereby generate a score for in utero smoke exposure.

**Methods**

We used machine learning methods to create a score reflecting exposure to maternal smoking during pregnancy. This score is based on peripheral blood measurements of DNA methylation (Illumina’s Infinium HumanMethylation450 BeadChip).

The score was developed and tested in the Raine Study with data from 995 Caucasian 17y-old participants using 10-fold cross-validation. The score was further tested and validated in independent data from the Northern Finland Birth Cohort 1986 (NFBC1986) (16y-olds) and 1966 (NFBC1966) (31y-olds). Further, three previously proposed DNA methylation scores were applied for comparison. The final score was developed with 204 CpGs using *elastic net regression.*

**Results**

Sensitivity and specificity values for the best performing previously developed classifier (‘Reese Score’) were 88% and 72% for Raine, 87% and 61% for NFBC1986 and 72% and 70% for NFBC1966, respectively; corresponding figures utilizing the elastic net regression approach were 91% and 76% (Raine), 87% and 75% (NFBC1986) as well as 72% and 78% for NFBC1966.

**Conclusion**

We have developed a DNA methylation score for exposure to maternal smoking during pregnancy, outperforming the three previously developed scores. One possible application of the current score could be for model adjustment purposes or to assess its association with distal health outcomes where part of the effect can be attributed to maternal smoking. Further, it may provide a biomarker for fetal exposure to maternal smoking.

**1. Introduction**

Fetal exposure to maternal smoking during pregnancy increases the risk that the offspring will develop non-communicable diseases (NCDs) (Agrawal et al. 2010; Bhattacharya et al. 2014; DiFranza et al. 2004; Hofhuis et al. 2003; Oken et al. 2005; Wakschlag et al. 2002; Wiklund et al. 2019). On average 6% of the global female population are still smokers, although with a high degree of variability across countries (e.g. due to differences in the social and educational contexts, legislations and cultural factors), according to the WHO report on global tobacco epidemic 2017 (WHO 2017). A 2017 paper describing the smoking rates in Australia and Finland amongst other countries reported the smoking rates amongst young pregnant women are stagnant, despite the overall decrease in smoking rates (Reitan and Callinan 2017).

A meta-analysis by Oken *et al.* including 14 studies showed that offspring of mothers who smoked during gestation had a pooled adjusted odds ratio of 1.50 ( 95% CI: 1.36, 1.65) for the development of obesity (Oken et al. 2007). Timmermans et al. examined the association between maternal smoking during pregnancy and lower birthweight, and higher weight gain and childhood overweight in the offspring (Timmermans et al. 2014). They showed that exposure to maternal smoking during pregnancy associated with an adjusted odds ratio of 3.72 (95% CI 1.33–10.4) for the offspring being in the 85th BMI percentile (Timmermans et al. 2014).

The mechanisms through which maternal smoking may influence the health of the offspring have been suggested to involve the altered epigenetic regulation of genes. Epigenetics is the general term for changes to the DNA that are heritable through cell division and that relate to gene accessibility rather than DNA sequence changes (Goldberg et al. 2007). There are many different epigenetic mechanisms that can affect or alter gene accessibility, such as chromatin structural changes, histone modification or DNA methylation (Goldberg et al. 2007). Many studies have shown that maternal smoking during pregnancy is associated with highly reproducible and specific changes in differentially methylated cytosine-phosphate-guanine (CpG) base pairs in newborns (Joubert et al. 2016), children (Rzehak et al. 2016), young adults (Lee et al. 2015) and middle aged adults (Sun et al. 2013). In a meta-analysis with combined sample size of 6,685 newborns and 3,187 older children, 2,965 (FDR corrected, 568 after Bonferroni correction) differentially methylated CpGs in the offspring were associated with maternal smoking during pregnancy (Joubert et al. 2016). This included CpGs within *AHRR* (aryl-hydrocarbon receptor repressor), *MYO1G (Myosin 1G)*, *CYP1A1* (Cytochrome P450 Family 1 Subfamily A Member 1), *GFI1* (Growth Factor Independent 1 Transcriptional Repressor) and *CNTNAP2* (Contactin-associated protein-like 2) (Rotroff et al. 2016; Rzehak et al. 2016; Tehranifar et al. 2018).These genes are associated with cancer development, detoxification of xenobiotics (*AHRR)* (Esser 2012) and adult BMI (*GFI1*) (Parmar et al. 2018) and suggest a possible epigenetic mechanism linking fetal exposure to maternal smoking during pregnancy with diseases in the offspring. Critically, Joubert et al. showed that the same CpGs were associated with fetal smoke exposure in cord blood, as well as in whole blood from five year olds (Rzehak et al. 2016), and our own research suggests that fetal smoke exposure may induce persistent changes to the DNA methylome still detectable in middle age (Wiklund et al. 2019).

Machine learning is a sub-field of artificial intelligence focusing on pattern detection. Machine learning methods can be divided into supervised and unsupervised methods. In supervised methods, labels are known, and the model tries to fit the data according to the label. In unsupervised methods, the algorithm tries to find clustering of similar data points. In both approaches, the aim is to create a model – with minimal assumptions on the data-generating process – that is generalizable to an external dataset. Machine learning has proven to be useful in classification problems in medical research and diagnosis, especially in cancer and image classification (Capper et al. 2018; Díaz-Uriarte and Alvarez de Andrés 2006; Quraishi et al. 2015; Schmidhuber 2015; Yoo et al. 2014).

Successful examples of implementation of machine learning in epigenetics are Houseman’s cell counts (E. A. Houseman et al. 2012) and Horvarth’s epigenetic age acceleration (S. Horvath 2013). Both are widely adopted in the field and are based on the elastic net regression approach (Zou and Hastie 2005).

In light of this, we applied machine learning methods to develop a DNA methylation score in adolescents and adults as a proxy for fetal exposure to maternal smoking. Similar to that described by Reese et al. (Reese et al. 2017), we aimed to generate a score that could be applied to studies using HumanMethylation450 and EPIC BeadChip DNA methylation data. Compared to the score by Reese et al, we have extended the DNA methylation score to older ages, including adolescence and adulthood. In data sets without information on maternal smoking during pregnancy, establishing and validating the score would enable its implementation in adjusting epigenome-wide DNA methylation association studies for this important early life exposure. It may also serve as covariate to any model in more conventional epidemiological studies to adjust for possible confounding by maternal smoking in the absence of the measure. With reducing costs for DNA methylation arrays, the availability of such a DNA methylation score would be a valuable tool for epidemiological studies in disease pathways.

**2. Methods**

***2.1 Studies***

*(i) The Raine Study*

The study design and initial characteristics of the Raine Study have been previously described (Newnham et al. 1993). From 1989 to 1991, a total of 2900 pregnant women were enrolled. This included multiparous pregnancies. Recruitment took place at King Edward Memorial Hospital and surrounding private hospitals. The 2868 live births have been followed up at 1, 2, 3, 5, 8, 10, 14 and 17 years during which anthropometric (e.g. height, weight, skinfolds), clinical and biochemical data have been collected. Ethics approval for conducting the epigenetics analysis at the 17-year follow-up was given by the Human Ethics Committee of the University of Western Australia. Institutional ethics approval has been obtained through the University of Western Australia (approval numbers: RA/4/1/6613, 1214-EP, RA-4-1-2646). Informed and written consent was provided by the participants and their parents or legal guardians. The present analyses included 995 participants that were of Caucasian ethnicity.

*(ii) The Northern Finland Birth Cohort 1986 (NFBC1986)*

The Northern Finland Birth Cohort 1986 consists of 99% of all children who were born in the provinces of Oulu and Lapland in Northern Finland between 1 July 1985 and 30 June 1986 (Järvelin et al. 1993). There were 9,432 live-born individuals that entered the study. At the age of 16, those living still in Finland (n=9,215) were invited to participate in a follow-up study including a clinical examination. 7344 participants attended the study in the year 2001/2002, of which 5654 completed the postal questionnaire, the clinical examination, provided a blood sample and gave written informed consent (parents and children). Approval for the studies was granted by the ethics committee of the Northern Ostrobothnia Hospital District in Oulu, Finland in accordance with the declaration of Helsinki.

*(iii) The Northern Finland Birth Cohort (NFBC1966)*

The Northern Finland Birth Cohort 1966 is a prospective follow-up study of children from the two northernmost provinces of Finland (Rantakallio 1988). Of all women in this region with expected delivery dates in 1966, 96% were recruited through maternity health Centres (12,058 live births). All individuals still living in northern Finland or the Helsinki area (n = 8,463) were contacted and invited for clinical examination when they turned 31 years of age (Jarvelin et al. 2004). A total of 6007 participants attended the clinical examination. DNA was extracted from blood samples given at the clinical examination (5,753 samples available). The samples were selected to resemble the original study cohort (Jarvelin et al. 2004). An informed consent for the use of the data including DNA was obtained from all subjects and approval for the study was granted by the ethics committee of the Northern Ostrobothnia Hospital District in Oulu, Finland in accordance with the declaration of Helsinki.

*DNA methylation profiling*

*The Raine Study*

DNA methylation was measured in peripheral whole blood samples from participants at age 17 years using the Illumina HumanMethylation450 BeadChip. Venous blood samples were taken by phlebotomists after an overnight fast. Samples were stored at -80 degrees until analysis. Processing of the Illumina Infinium HumanMethylation450 BeadChips was carried out by the Centre for Molecular Medicine and Therapeutics (CMMT) <http://www.cmmt.ubc.ca>. We excluded three samples as outliers and one sample for biological sex inconsistency, as this might be indicative of a sample mix-up. Outliers were defined by the R packages *shinyMethyl* (Fortin and Hansen 2015) and *MethylAid* (Van Iterson et al. 2014) as samples that did not cluster together with the rest. Annotation of the CpG to the nearest gene was performed using Illumina’s genome coordinates (GRCh37/hg19). DNA methylation data of 996 Caucasian study participants on475429 probes were available for analysis.

*NFBC1986*

DNA was extracted from all 5654 blood samples at the 16-year follow-up. DNA methylation was recorded on Illumina HumanMethlation450 array for 546 randomly selected subjects at the Department of Genomics Imperial College London (London, UK). Of those, 24 technical replicates were excluded. A total of 18 samples did not reach a call rate of >95% applying a detection P-value filter of 10e-16. We excluded 7 samples with biological sex inconsistency, no sample was outlying from the overall data structure (1st principle component (PC) score of the DNA methylation values outside mean +/- 4 standard deviations (SD)). DNA methylation data of 517 Caucasian samples with 466290 autosomal probes (call rate filter 95%) each were available for this analysis.

*NFBC1966*

DNA methylation at 31 years was measured for 807 randomly selected Caucasian subjects that attended the clinical examination and completed the questionnaire at both 31 and 46 years. For this, the Illumina HumanMethlation450 array was utilized at the Department of Genomics Imperial College London (London, UK). For DNA methylation marker calling we used a detection p-value threshold of <10-16 A call rate filter of 95% was applied to all autosomal Illumina probes yielding 459378 probes for association testing. Due to low marker call rate (<95%), 67 samples were excluded. Seven samples were excluded for biological sex inconsistency; and one sample for globally outlying DNA methylation values (1st PC score of the DNA methylation values outside mean +/- 4SD).

For all studies, we used the raw methylation betas without plate normalization, as normalization methods might introduce bias into the model for the machine learning approach, by changing the variance and residual structure of the DNA methylation data. This might also improve the application of the score to different data sets (Reese et al. 2017).

*Smoking Variables*

*The Raine Study*

Mothers reported smoking behaviour in questionnaires administered at the 18th and 34th week of gestation. Maternal smoking during pregnancy was coded as ‘yes’ vs ‘no’ in regard to smoking during pregnancy, based on a combination of the categorical variables for the number of cigarettes smoked daily at 18 and 34 weeks of gestation. In a previous epigenome wide association study on the Raine Study data set, we did not find any differences between the CpGs associated with in utero smoke exposure at 18 and at 34 weeks (Rauschert et al. 2019). To not sacrifice sample size, we decided to use the combined timepoints as any smoking, as described above.

We also present data on the number of adolescents that ever smoked. Smoking behaviour of the adolescents at 16 years of age was self-reported in a confidential online questionnaire, and we re-coded the variable asking for cigarette consumption over the life-time to “(any) smoking” versus “no smoking”.

Supplement Figure S1 showing the distribution of exposure to maternal smoking across the plates for measurement of DNA methylation indicates no sign of a potential batch induced bias.

*NFBC1986 and NFBC1966*

Information on maternal smoking was self-reported in questionnaires by mothers during pregnancy. The questions asked: Did you smoke before pregnancy? yes, no; Did you smoke when pregnancy was discovered? yes, no; number of cigarettes after the 2nd gestational month: none, <10, 10 or more; mother’s smoking after the 2nd gestational month. This information was re-coded to a binary variable indicating any smoking during pregnancy as opposed to no smoking during pregnancy to harmonize the data with the Raine study variable.

The adolescents own smoking in NFBC1986 and adult smoking in NFBC1966 were also assessed using questionnaire data. We coded them into ever-smoker, which includes occasional/former smoker, and never smoker.

***2.2 Analysis, model training and model selection***

***Overview***

A flow chart of the distinct modelling steps, including a brief description and what data was utilized is provided in Figure 1. The overall aim was to identify the best performing algorithm to identify study participants as being exposed to maternal smoking during pregnancy based on DNA methylation data. Performance in this context is defined as the model accuracy when compared to the known information of exposure to maternal smoking during pregnancy. To measure accuracy, we focused on Cohen’s kappa to identify the best model in this study. First, we split the Raine study data into training and test set; then we applied 11 different machine learning algorithms to the training data to make a pre-selection of best performing algorithms. The four best performing models were taken forward for further refinement of the model parameters. Finally, the best performing model was selected based on Cohen’s kappa.

***Machine learning models***

To check for the performance of different algorithms as defined in the previous paragraph, we derived scores for the exposure to maternal smoking during pregnancy from several different models after training with default settings for the model parameters in the statistical packages. The exact R code used for this can be found in Supplement 1. This was carried out to provide an overview of which methods to focus on for further modelling. There is no standard way of selecting machine learning algorithms. Some models are more suited for specific tasks than others and a good way of starting is to systematically test different algorithms on the data set and pre-select those with the best initial performance for further training. It is advisable to select a variety of different algorithms, such as tree-, regression and clustering based methods, as this allows for the testing of linear and non-linear associations in the data.

All statistical and predictive modelling was conducted using R 3.5.1 and the *caret* package (Kuhn 2008). The primary models evaluated in this study were gradient boosting machine (Friedman 2001) using the *gbm* package (Ridgeway et al. 2013), elastic net regression (Zou and Hastie 2005) using the *glmnet* package (Hastie and Qian 2014), random forest (Breiman 2001) using the *randomForest* package (Liaw and Wiener 2002), and support vector machine (Cortes and Vapnik 1995) using *e1071* (Meyer and Wien 2015). In addition, we evaluated C5.0 (Pandya and Pandya 2015), Classification with Bagging (Breiman 1996), linear discriminant analysis (Duda et al. 2012), k-nearest neighbour (Altman 1992), naive Bayes classifier (Rish 2001), logistic regression, and classification and regression trees (Breiman 2017), which were applied by setting the caret variable “*method*” to *C5.0*, *treebag*, *lda*, *knn*, *nb*, *glm*, and *rpart*, respectively. All of the evaluated models other than k-nearest neighbour are supervised machine learning models.

***Variable Pre-selection***

Overfitting in the variable selection when using the training data for that step was accounted for by selecting CpGs for the modelling process from the meta-analysis of Joubert et al. (Joubert et al. 2016). Joubert et al. used both FDR and the Bonferroni correction to define significant CpGs for their study. The table we selected the CpGs from is Supplement 4, Excel Table S3 in Joubert et al., with the column titled “Meta-Analysis of sustained smoking and newborn methylation adjusted for cell type”. We decided to consider an arbitrary p-value <0.00001 for CpGs to be included in the modelling, which is in-between the FDR and Bonferroni threshold. We acknowledge that linear models can detect some relevant associations in the data but believe restricting the selection only to Bonferroni p-values limits the possibility to identify non-linear associations using machine learning modelling. Including all ~450k variables would technically be possible, but this would be computationally very expensive in terms of time and resources. Hence, our pre-selection process resulted in the inclusion of 1511 CpGs.

As the aim of a predictive model is to be as parsimonious as possible, we excluded the highly correlated variables for the elastic net regression model, only retaining the CpGs that retain most of the information based on correlation structures of the data. This was done by examining the pairwise correlation structures of the CpGs in the Raine study data, before splitting it into training and test set. Given two CpGs were correlated with an R-squared >0.75, we removed the CpG with the largest mean absolute correlation and thereby reducing multicollinearity issues. For this we used the R function *findCorrelation* (Kuhn 2008). In total, 267 CpGs were removed from the initial set of 1511 CpGs, as they had correlation coefficients >0.75 with at least one other CpG, leaving 1,244 CpGs for analysis (Excel Table S1). The tree based and support vector models are not as vulnerable to correlated data as the linear regression-based model, hence all 1511 CpGs were used for those.

***Fitting Method***

We created the smoking score based on the study by Reese et al. using the exact coefficients and CpGs they identified with their LASSO approach (Reese et al. 2017). Briefly, to retrieve the score, one needs to multiply the CpG methylation values with the respective coefficient provided by Reese et al. in their supplement, Tables S1 and then add up the results from all 28 CpGs. The R code used to calculate the Reese score exemplified in the Raine study, is provided in Supplement 1.

Richmond et al. describe two different scores for exposure to maternal smoking during pregnancy (Richmond et al. 2018). One being created based on 568 CpGs from cord blood methylation data and a second score, using 19 CpGs from adult methylation data from the Avon Longitudinal Study of Parents and Children (ALSPAC) (Joubert et al. 2016; Richmond et al. 2018). Richmond et al. describe the calculation of the score as multiplying the model coefficients from the Joubert et al. meta-analysis of an EWAS for maternal smoking during pregnancy with the CpG methylation betas (Joubert et al. 2016). For that, the Supplement 4, Excel table S3 of the Richmond et al. publication is required. Hence, we identified the CpGs required for the score creation as per Richmond et al. in the Raine study, NFBC1966 and NFBC1986 and multiplied the individual participants DNA methylation values with the respective coefficient from the Joubert et al. study. For the 568 CpG score, the column titled “Meta-Analysis of sustained smoking and newborn methylation adjusted for cell type” is used, and for the 19 CpG score the column “Meta-Analysis of sustained smoking and methylation in older children”.

The random forest algorithm proposed by Breiman (Breiman 2001) is a decision tree-based algorithm. Rather than a single decision tree, this algorithm uses an ensemble approach. Every tree is created by only using a bootstrapped sample of the entire data. A second step of randomness is added for each split by only selecting a random subset of all predictive variables. This means, random forest implements both bagging (a method to combine multiple unstable learners, like decision trees, to gain more stable predictions) and random variable selection to build the trees, which leads to low correlation between the trees in the forest. For the purpose of this study, we set the *tuneLength* variable in the *caret* model to 20, which means a maximum of 20 different settings for the random forest parameters are evaluated. The following settings for the parameter *mtry* (number of input variables at each split) were tested: 2, 3, 5, 7, 10, 14, 19, 27, 38, 53, 74, 102, 142, 198, 275, 382, 530, 736, 1023. For random forest the *caret* package defaults the number of trees to 500, as the algorithm has been shown to plateau in its performance around this value.

The gradient boosting machine algorithm is also a tree-based method (Friedman 2002). Gradient boosting machines grow trees sequential and try to improve upon those trees that show weak predictions making the method useful in cases of imbalanced data, like in this study. The parameters that can be tuned in a gradient boosting machine are minimum observations per node, number of trees and interaction depth (the number of splits to be performed on a tree (starting from a single node)). The following values and all their combinations were tested: all values from 1 to 20 for interaction depth, the minimum number of observations per node were kept at the *caret* default of 10, and the number of trees was tested in steps of 50 from 50 to 1000.

The Support vector machine (SVM) algorithm uses a subset of data points as so-called support vectors (Cortes and Vapnik 1995). In a two dimensional case, support vectors are those data points that are closest to the line indicating the greatest separation between two classes. For the linear version of this model, the parameter *C* can be tuned. Also known as *Cost*, this parameter determines the possible misclassifications that are allowed. Simply speaking, it imposes an error penalty to the model. That means, the higher the value of *C*, in theory, the less likely it should be that the SVM algorithm will misclassify the data. In our tuning step, with *tuneLength* set to 20, the following values of *C* were tested: 0.00, 0.01, 0.05, 0.10, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00 and 5.00.

The code and the final models to use the three models above to create the score are available in Supplement 1 and require the *caret* package function *predict*.

The fourth model tested was elastic net regression (Zou and Hastie 2005). This model is a logistic regression-based model, that allows to specify two parameters: lambda and alpha. The model does not only fit the data, like a logistic regression model, but it also performs variable selection. For this, the penalty parameter lambda can be tuned, which, based on its size, will penalise uninformative variables more. The alpha parameter can be set to 0, 1 or any integer in-between, where 1 means LASSO regression, like in the Reese et al. study. There, the model strictly drops uninformative or correlated variables. The setting 0 for alpha does not perform variable selection, but rather calculates weights for all variables, based on their importance for the classification. Elastic net regression keeps the alpha value between 0 and 1, which is a mix of both options described above; it will perform feature selection, but in the case of correlated variables, that are both potentially meaningful for the classification, it will not randomly select one of the two, like LASSO.

With *tuneLength* of 20, the following parameter settings and all their combinations were tested: alpha of 0.10, 0.1473684, 0.1947368, 0.2421053, 0.2894737, 0.3368421, 0.3842105, 0.4315789, 0.4789474, 0.5263158, 0.5736842, 0.6210526, 0.6684211, 0.7157895, 0.7631579, 0.8105263, 0.8578947, 0.9052632, 0.9526316, and 1.0; lambda of 0.003785885, 0.004714173, 0.005870074, 0.007309400, 0.009101643, 0.011333339, 0.014112241, 0.017572522, 0.021881253, 0.027246473, 0.033927228, 0.042246086, 0.052604703, 0.065503224, 0.081564423, 0.101563783, 0.126466926, 0.157476249, 0.196088967 and 0.244169411.

For the final, best performing model using elastic net regression, the coefficients for the scoring are provided in Excel Tables S2, with instructions as to how to apply the score in Supplement 1. Briefly, the model can both produce a probability score (a value between 0 and 1, with 0 meaning not exposed and 1 meaning exposed) and a binary class (with a cut-off of 0.5; values above this fall into the “exposed” class, while values below fall into the “not exposed class”). To generate the DNA-methylation risk score (ranging from 0 to 1), the steps are as follow: (1) multiply the CpG beta values by their respective coefficients generated by elastic net regression and (2) sum these across the 204 CpGs with the provided coefficients (Excel Table S2).

We include a guide for easy application of this score (Supplement 1). This is taken from our R package, that is developed on github, so anyone can apply the score via the R programming language (<https://github.com/Hobbeist/DNAsmokeR)>.

Of note, elastic net regression is the only machine learning model used in this study that not only fits a predictive model, but also performs variable selection. This is why the creation of the score requires only 204 CpGs for elastic net regression, whereas all input CpGs are required for score creation in the other methods.

All models were trained using an 80% training and model-fitting subset and a 20% test subsample of the Raine Study data (Figure 1). The 80% training and model-fitting sample was determined using stratified randomized selection to retain the ratio between smoke exposure groups. For the training step in the 80% subset, we applied 10-fold cross-validation with 5 repeats. Therefore, the Raine model-fitting dataset was randomly sampled 5 times into 10 groups, and for each sampling, 10 models were fitted (each time the model was fitted excluding one group and predicted values were estimated for the remaining group). The average Cohen’s Kappa results of those 50 modelling steps were used for comparing and selecting the best model. All CpG values were centred and scaled for the modelling. The R code for model training and testing of the above four machine learning models can be found in Supplement 1.

***Imbalanced data problem***

Classification algorithms try to reduce the overall error rate in classification; highly imbalanced data sets, where the minority class is very small compared to the majority class, tend to show good prediction accuracy but an overrepresentation of classification into the majority class. This also means, that prediction accuracy is not a feasible measure for overall classification quality. We used the following three approaches to address this: The data set was split into training and test data, stratified by exposure to smoking, meaning the ratio of smoke exposed to not exposed was the same in the training and test sets. To overcome the imbalance problem further, we applied a synthetic minority oversampling technique (Chawla et al. 2002), which outperforms oversampling the minority class (smoke exposed in our example) or under-sampling the majority class (not exposed). In this approach, new, synthetic minority instances are created between existing data points, based on k-nearest neighbours.

We trained all our models on the Kappa () metric by Cohen (Cohen 1960; Viera and Garrett 2005). This metric compares the observed prediction accuracy with the expected prediction accuracy under random-guess circumstances:

= ,

where is the observed prediction accuracy and is the expected prediction accuracy. For values between 0 and 1, 1 indicates a perfect prediction. We also report sensitivity, specificity and area under the curve (AUC), as AUC has been established as a model comparison measure in machine learning (Held et al. 2016; Jin and Ling 2005).

Further, we report the brier score as another quality measure, which is calculated by:

,

where *N* is the number of forecasts, *fi* is the score for participant *i* and *oi* is the observed class: 0 for not exposed and 1 for exposed. The brier score is a good quality measure for a probability score (Rufibach 2010). In terms of interpretation, values close to 0 indicate very good predictive power, whereas values close to 1 indicate bad performance.

***Criteria for model selection***

For the selection of the best model, the quality measure of interest was Cohen’s kappa (prediction accuracy and Cohen’s kappa for all models are reported in Supplement Figure S2). The top four algorithms were elastic net regression, gradient boosting machine, support vector machine and random forest. Hence, we decided to train those four models further and compare them based on Cohen’s kappa, sensitivity, specificity and AUC. The model quality measures sensitivity, specificity, accuracy as well as AUC were derived from the ROC-curve, utilizing the point that minimizes the distance from the ROC curve to the top left corner, using the R packages *pROC*, *caret* and *plotROC*. The *caret* package was utilized to calculate the kappa statistic.

***Final model exploration***

*Model quality measures and variable importance*

For each machine learning model, we determined the CpGs that contributed the most towards the classification based on the absolute value of their estimated coefficients and compared the top 20 CpGs from each model to identify CpGs that were common across multiple models. To facilitate comparisons across models, we derived a measure of relative importance for each CpG by scaling the absolute value of each coefficient to the CpG coefficient with the largest value for each model.

This was done as these variables might hold insights into potential biological associations between maternal smoking during pregnancy and DNA methylation in the offspring.

*Significance test for ROC differences*

The elastic net regression based smoking score, the Reese and the Richmond score are all continuous values that we inspect via ROC curves. To identify whether the observed difference in the ROC-curves and area under the ROC curves is statistically significant, we used DeLongs test. This test is used to check whether ROC curves are uncorrelated and is implemented in the R function *roc.test* (package *pROC*) (DeLong et al. 1988).

*EPIC array sensitivity analysis*

With the availability of the newer BeadChip array “EPIC” from Illumina, we also analysed the performance of the score when only using the subset of CpGs in the elastic net model that are in both the 450k and EPIC array. In total, there are 23 elastic net score CpGs missing compared to 450k, totalling to 181 CpGs. We performed this sensitivity analysis only for the final best performing model, elastic net.

**3. Results**

Participants’ characteristics (Table 1) showed no significant differences in the comparison of age, sex, adolescent smoking and exposure to maternal smoking between the training (n=797) and test sets (n=198) of the Raine study. Maternal smoking rates were similar in the training (ratio of exposed to not-exposed: 0.42 (237/560)) and test sets (0.42 (59/139)), as expected due to the stratified split of the Raine study into training and test sets. In both NFBC1986 and 1966, approximately 20% of study participants were exposed to maternal smoking during pregnancy. The NFBC cohorts had smoking rates of 34.3% at the 16 year (NFBC1986) and 52% at the 31-year follow-ups (NFBC1966). This is higher than in the Raine study, where the proportion of smokers at the 17-year follow-up was 21.2% in the testing and 24% in the training set.

*Machine Learning Models: Quality Measures*

Taking into account that some CpGs available in the Raine study were not available in the NFBC studies, due to post-processing and outlier exclusion, and due to the exclusion of highly correlated CpGs to create a sparse model, as well as variable selection via elastic net regression, the final number of CpGs to create the DNA-methylation score was 204 for the elastic net (Excel Table S2).

*Scores based on gradient boosting machine, random forest and support vector machine*

For the gradient boosting machine approach, the cross-validation step resulted in a final model with 1000 trees, an interaction depth of 6 and a minimum number of observations randomly selected per tree of 10.

The final support vector machine algorithm is a model with a penalisation parameter C of 0.75. And finally, for the random forest: default number of trees (500) and *mtry* of 198 variables.

*Final DNA-methylation score*

In the Raine data set, gradient boosting machine outperformed elastic net, random forest, and support vector machine scores on every measure except sensitivity, which was the same for gradient boosting machine and elastic net regression (Table 2). However, in the NFBC data sets, elastic net outperformed gradient boosting machine, random forest, and support vector machine scores for all quality measures except sensitivity (higher for gradient boosting machine than elastic net in all datasets, and for support vector machine in NFBC 1966). Based on this assessment, we concluded that elastic net regression, with an alpha of 0.1 and a lambda of 0.1264669, had the best overall performance of the 4 machine learning methods evaluated.

*Previous “gold standard” scores*

*Reese et al. score*

The model evaluation metrics for the Reese score can be found in Table 2. For all metrics, including our key metric Cohen’s kappa, the elastic net regression-based score outperformed the Reese score in the Raine study. However, when comparing the ROC curves (Figure 2), the elastic net-based curve did not significantly differ from the Reese score curve for the Raine study based on the DeLong test (Table 3). In NFBC1986, the sensitivity is the same between the elastic net and the Reese score, however all other model measures, including Cohen’s kappa are better in the elastic net score. The DeLong test, as indicated in Table 3, shows a significant difference between the elastic net and the Reese score, with the elastic net based score outperforming the Reese score. And lastly, the same is true for NFBC1966. There is no difference in sensitivity, but all other measures show that the elastic net score outperforms the Reese score.

*Richmond et al. scores*

For the Raine study, the elastic net score outperforms the Richmond score with 568 and 19 CpGs in all model metrics. Further, as can be seen in Table 2, the ROC curves are significantly different between the elastic net and the Richmond based scores. The same is true for both NFBC studies, although the specificity is slightly better for the Richmond score with 568 CpGs in NFBC1986. Further, the Reese score outperforms both Richmond scores in all studies, except for the specificity in NFBC1986. Between the two Richmond scores, the 568 CpG score performs better than the 19 CpG score in all studies.

*Variable importance*

Seven CpGs were included in the top 20 CpGs for each of the four machine learning models, including the top four CpGs from the elastic net model: cg14179389 (*GFI1*, with the highest coefficient for the elastic net model), cg25949550 (*CNTNAP2*, 94% importance relative to cg14179389), cg22132788 (*MYO1G*, 80% relative importance), and cg11207515 (also in *CNTNAP2*, 66% relative importance) (Excel Table S2). The remaining CpGs included in the top 20 CpGs for all machine learning models were cg13570656 (*CYP1A1*), cg17924476 (*AHRR*), and cg08474748 (*ANKRD31*) (Excel Tables S3–S6).

For the gradient boosting machine algorithm, the first four CpGs were also frequently amongst the top CpGs in epigenome wide association studies: cg22132788 (*MYO1G*, 100% importance), cg14179389 (*GFI1*, 99.99% importance), cg25949550 (*CNTNAP2*, 80.9% importance) and cg17924476 (*AHRR*, 20.8% importance).

The support vector machine algorithm identified CpGs in associated with *MYO1G*, *CNTNAP2*, *GFI1*, *CYP1A1*, *AHRR* and *FTO* amongst the top 10 most important variables in developing the score. And lastly, the random forest algorithm chose CpGs as the top 10 most important variables that are all frequently associated with maternal smoking during pregnancy. Those CpGs were associated with the genes *MYO1G*, *CNTNAP2*, *GFI1*, *AHRR*, *CYP1A1* and *FTO*.

*EPIC array sensitivity analysis*

The results for applying the elastic net score to those CpGs available in EPIC data (with 23 score CpGs missing compared to 450k, totalling to 181 CpGs for this analysis) can be found in Supplement 1, Table S1. For all metrics, the elastic net score based on the 450k version of Illumina’s BeadChip array outperforms the elastic net score using only the 181 CpGs also available on the EPIC array. The overall performance, however, is still good, with Cohen’s kappa values all exceeding 0,3, with the best value in the Raine study, being 0.65.

**4. Discussion**

In this study, we have developed a DNA-methylation score for exposure to maternal smoking during pregnancy that outperforms an existing composite score (Reese et al. 2017), utilizing DNA methylation probes (CpGs) measured in peripheral blood at the 17 year follow up of the Raine study. We believe that with reducing costs for measuring DNA methylation, such a DNA-methylation score could be a valuable contribution to epidemiological studies and clinical diagnostics.

To identify the most promising machine learning algorithms for creating the score, a range of models were chosen that performed well in similar tasks and reflect a broad range of approaches, from linear to non-linear. We decided to further test the performance of elastic net regression, gradient boosting, support vector and random forest algorithms, as they showed the best performance with respect to Cohen’s kappa in the algorithm selection step and are algorithms previously utilized in other epigenetic predictive modelling problems (Capper et al. 2018; S. Horvath 2013; Eugene Andres Houseman et al. 2012).

Our study aimed to establish a score utilizing only variables created by the Illumina HumanMethylation450 BeadChip, as this makes the score independent of any other variables that studies otherwise would need to have collected. Compared to more classical statistical approaches, the single aim of creating a score is that it is as accurate as possible in differentiating individuals exposed to, from individuals not exposed to in utero smoke, with as little input as necessary.

Further, we purposely utilized raw DNA methylation betas, to avoid skewing the models based on normalization methods, which alter the residual and variance structure of the data. This also avoids the need for studies that aim to apply this score to normalize their data in a specific way, making it simple to apply. The main reason for doing so is, that there is no gold standard of correcting the probes from the BeadChip and the methods all perform slightly differently (Marabita et al. 2013). Ideally, the same normalization method would be applied to the training data as well as any data that applies the score. Using the raw values overcomes the issue initially and, as shown in this study, performs very well when applied to other studies.

In the Raine study and NFBC1986 our score performs moderately better than the Reese et al. (Reese et al. 2017) score created from cord blood. Reese et al. applied the LASSO penalized model rather than an elastic net approach to derive their score. Elastic nets have been mathematically shown to outperform LASSO regression when the number of variables is much larger than the number of cases (Zou and Hastie 2005). LASSO regression only selects at most as many variables as there are cases, which might not be feasible in the case of smaller sample sizes.

Nevertheless, the score by Reese et al. performs surprisingly well in the NFBC whole blood DNA samples collected at 16 and 31 years of age, with sensitivity consistently in excess of 70%, despite being derived based on cord blood DNA methylation. As the cord blood measurement is closer to the exposure to maternal gestational smoking, the Reese et al. model might pick up stronger associations that potentially decrease over time, as DNA methylation has been shown to change with age (Steve Horvath 2013).

The good predictive capability of our score applied to both the Raine test set and the validation study NFBC1986 (both 16/17 years of age) suggests that methylation could follow similar patterns across exposed and non-exposed individuals in the same age group and still holds some structural similarities when applied to a different age group, as in the NFBC1966.

Richmond et al. examined the relationship between maternal smoking and DNA methylation applying two different scores utilizing 19 and 568 CpGs, and reported AUCs of 0.69 and 0.72, respectively, for their population of 656 women with measurements at two timepoints and in 230 men (Richmond et al. 2018). The two Richmond scores both underperformed the Reese et al. score and our own score when applied to the Raine data set and the two NFBC data sets. The CpGs were selected based on Bonferroni significance in association with maternal smoking, excluding the possibility that non-significant associations might still be contributing to the differences in association with maternal smoke exposure, by, for example, multivariate effects, that a linear regression model itself is not able to assess.

All four models trained in our study, independent of their performance, selected cg22132788 (*MYO1G*), cg25949550 (*CNTNAP2*), cg14179389 (*GFI1*), cg11207515 (*CNTNAP2*), cg13570656 (*CYP1A1*), cg17924476 (*AHRR*) and cg08474748 (*ANKRD31*) amongst the top 20 most influential variables for the DNA-methylation score. These data concur with findings of several studies investigating DNA methylation in different age groups, showing that the same CpGs are differentially associated with exposure (Joubert et al. 2014; Richmond et al. 2015; Rzehak et al. 2016).

The final best-performing overall score of this study, the elastic net, utilizes the *FTO* gene-associated cg00253658 for classification. The associated CpG is the 20th most important CpG based on the variable importance. This gene has previously been shown to associate with the development of obesity (Frayling et al. 2007). There is evidence, however, that SNPs in the *FTO* gene might rather affect expression of the *IRX3* gene, which is related to obesity (Smemo et al. 2014). Further, methylation in the *AHRR* gene, with its associates CpG cg17924476 ranked 6th in the elastic net model, was associated with the development of eczema in boys and girls in a previous study (Mukherjee et al. 2016). Eczema in young children is a possible pre-cursor of asthma and allergies, highlighting the potential association between exposure to maternal smoking during pregnancy and the development of asthma and allergies later in life (Almqvist et al. 2007).

This raises the possibility of using this score as a risk score for phenotypes associated with in utero smoke exposure such as obesity and allergic disease (Agrawal et al. 2010; DiFranza et al. 2004; Oken et al. 2005) in future studies that may give insights into pathways affected in fetal programming associated with maternal smoking.

*Strengths and Limitations*

DNA methylation is strongly associated with exposure to maternal smoking during pregnancy, as several studies have shown (Joubert et al. 2016; Rzehak et al. 2016). Hence, it is a good starting point to test which machine learning algorithms have the potential to be utilized as predictive models in the future.

With availability of DNA methylation data as measured by the Infinium HumanMethylation450 BeadChip and for all chosen models, the DNA methylation variables chosen for the classification, including the parameters for each model are easily accessible and stated in this study. As interpretability is important for reassurance when using in the context of clinical practice, we decided to approach the modelling with this in mind.

All maternal smoking variables were assessed via questionnaires rather than by the more objective measurement of cotinine, which is a limitation of this study.

Further, the score was developed utilizing whole blood DNA methylation, which might not be the optimal sample type for specific DNA methylation and also might be affected by differences in white blood cell counts. We did not adjust for blood cell counts in our models, but our methylation score seemed to perform well despite this.

The performance of the score for correctly classifying exposure to maternal smoking during pregnancy might be influenced by current or recent smoking by adolescent or adult offspring, which might be more frequent in offspring exposed to maternal smoking during pregnancy. About half of the 31-year old NFBC1966 study participants and over a third of the 16-year old NFBC 1986 participants reported ever smoking. The proportion was lower in the Raine study (21% in the training dataset) but information on smoking was missing for almost one-quarter of the participants.

We used data from Caucasian study participants in training, testing and validation data, hence its performance needs to be confirmed in other ethnic groups. By utilizing three different studies with two different age groups from culturally different countries (Australia and Finland), however, we were able to assess whether the models were overfit to the training study data or generalisable. The match between the Raine Study and NFBC1986 in terms of sex and follow-up is by chance, as both studies were independently developed, and data collection performed independently.

The model was created using data from the 450 version of the Illumina HumanMethylation series and not the newer EPIC version. The score still performed well when based only on the 181 CpGs available in the EPIC array using the coefficients in Excel Table S2. Further, Investigators whose data are missing other CpGs can also derive a score based on available CpGs, though the score’s performance might differ and the performance will be somewhat uncertain. We did not have pyro-sequenced data available to test how the smoking score would compare to the score derived from the 450 BeadChip array.

Further, although utilizing raw DNA-methylation data in our study showed very good results, future studies should systematically evaluate the effect of available normalization methods on the DNA-methylation score.

**Conclusion**

Our study shows that DNA methylation in late adolescence and early adulthood can be utilized to establish a score for the exposure to maternal smoking during pregnancy. The score was validated externally in study populations from Finland and Australia.

We have evaluated the different machine learning approaches by utilizing imbalanced data specific measures (Cohen’s kappa) as well as established comparative measures such as AUC, as in similar studies (Held et al. 2016; Jin and Ling 2005).

Our findings suggest that the score can be utilized by studies that have Illumina HumanMethylation450 or EPIC data available. As maternal smoking during pregnancy is one of the most well-established early life variables to be strongly associated with DNA methylation later in life, this score allows for studies that do not have information on maternal smoking behaviour during pregnancy to account for its variance. For future studies, it might be interesting to test the interaction of this score with other risk factors related to maternal smoking during pregnancy.

The score combines information on DNA methylation and early life exposure, and potentially a means to examine associations between this score and health outcomes such as cardiometabolic or respiratory diseases.

Acknowledgements

We acknowledge the Raine Study participants and their families, The Raine Study Team for cohort co-ordination and data collection, The NHMRC for their long-term contribution to funding the study over the last 25 years, The Telethon Kids Institute for long term support of the Raine Study. We also acknowledge The University of Western Australia, Curtin University, Women and Infants Research Foundation, Edith Cowan University, Murdoch University, The University of Notre Dame Australia and the Raine Medical Research Foundation for providing funding for Core Management of the Raine Study.

This work was supported by resources provided by The Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia.

We further thank Professor Paula Rantakallio (launch of NFBC1966 and initial data collection). We gratefully acknowledge the contributions of the participants in the Northern Finland Birth Cohort 1966 study and the Northern Finland Birth Cohort 1986. We also thank all the field workers and laboratory personnel for their efforts.

Funding statements

SR received funding for this work from the European Union’s Horizon 2020 project LifeCycle, as an awardee of the LifeCycle fellowship 2018.

The DNA methylation work was supported by NHMRC grant 1059711. RCH and TAM are supported by NHMRC Fellowships (grant number 1053384 and 1042255, respectively).

KMG is supported by the UK Medical Research Council (MC\_UU\_12011/4), the National Institute for Health Research (as an NIHR Senior Investigator (NF-SI-0515-10042) and through the NIHR Southampton Biomedical Research Centre) and the European Union's Erasmus+ Capacity-Building ENeASEA Project and Seventh Framework Programme (FP7/2007-2013), projects EarlyNutrition and ODIN under grant agreement numbers 289346 and 613977.

NFBC1966 received financial support from University of Oulu Grant no. 65354, Oulu University Hospital Grant no. 2/97, 8/97, Ministry of Health and Social Affairs Grant no. 23/251/97, 160/97, 190/97, National Institute for Health and Welfare, Helsinki Grant no. 54121, Regional Institute of Occupational Health, Oulu, Finland Grant no. 50621, 54231. NFBC1986 received financial support from EU QLG1-CT-2000-01643 (EUROBLCS) Grant no. E51560, NorFA Grant no. 731, 20056, 30167, USA / NIHH 2000 G DF682 Grant no. 50945. SS, AH, VK and MRJ received support by H2020-633595 DynaHEALTH, H2020 733206 LifeCycle, the academy of Finland EGEA-project (285547), the Biocenter Oulu, H2020-733206 LifeCycle, H2020-824989 EUCANConnect, EU-H2020 (grant no. 82576), EU-H2020 EarlyCause (grant no. 848158), EU-H2020 LongITools (grant no. 873749), EU H2020-MSCA-ITN-2016 CAPICE Marie Sklodowska-Curie grant (grant no. 721567) and the Medical Research Council, UK (grants no. MR/M013138/1, MRC/BBSRC MR/S03658X/1 (the EU JPI HDHL)) . VK is funded by the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant (721567).

References

Agrawal A, Scherrer JF, Grant JD, Sartor CE, Pergadia ML, Duncan AE, et al. 2010. The effects of maternal smoking during pregnancy on offspring outcomes. Preventive medicine 50:13-18.

Almqvist C, Li Q, Britton WJ, Kemp AS, Xuan W, Tovey ER, et al. 2007. Early predictors for developing allergic disease and asthma: Examining separate steps in the ‘allergic march’. Clinical & Experimental Allergy 37:1296-1302.

Altman NS. 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician 46:175-185.

Bhattacharya S, Beasley M, Pang D, Macfarlane GJ. 2014. Maternal and perinatal risk factors for childhood cancer: Record linkage study. BMJ open 4:e003656.

Breiman L. 1996. Bagging predictors. Machine learning 24:123-140.

Breiman L. 2001. Random forests. Machine Learning 45:5-32.

Breiman L. 2017. Classification and regression trees:Routledge.

Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, et al. 2018. DNA methylation-based classification of central nervous system tumours. Nature 555:469-474.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. 2002. Smote: Synthetic minority over-sampling technique. Journal of artificial intelligence research 16:321-357.

Cohen J. 1960. A coefficient of agreement for nominal scales. Educational and psychological measurement 20:37-46.

Cortes C, Vapnik V. 1995. Support-vector networks. Machine Learning 20:273-297.

DeLong ER, DeLong DM, Clarke-Pearson DL. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics 44:837-845.

Díaz-Uriarte R, Alvarez de Andrés S. 2006. Gene selection and classification of microarray data using random forest. BMC bioinformatics 7:3.

DiFranza JR, Aligne CA, Weitzman M. 2004. Prenatal and postnatal environmental tobacco smoke exposure and children’s health. Pediatrics 113:1007-1015.

Duda RO, Hart PE, Stork DG. 2012. Pattern classification:John Wiley & Sons.

Esser C. 2012. Biology and function of the aryl hydrocarbon receptor: Report of an international and interdisciplinary conference. Archives of toxicology 86:1323-1329.

Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, Lindgren CM, et al. 2007. A common variant in the fto gene is associated with body mass index and predisposes to childhood and adult obesity. Science (New York, NY) 316:889-894.

Friedman JH. 2001. Greedy function approximation: A gradient boosting machine. Annals of statistics:1189-1232.

Friedman JH. 2002. Stochastic gradient boosting. Computational statistics & data analysis 38:367-378.

Goldberg AD, Allis CD, Bernstein E. 2007. Epigenetics: A landscape takes shape. Cell 128:635-638.

Hastie T, Qian J. 2014. Glmnet vignette. Retrieve from <http://www> web stanford edu/~ hastie/Papers/Glmnet\_Vignette pdf Accessed September 20:2016.

Held E, Cape J, Tintle N. 2016. Comparing machine learning and logistic regression methods for predicting hypertension using a combination of gene expression and next-generation sequencing data. BMC Proc 10:141-145.

Hofhuis W, de Jongste JC, Merkus PJFM. 2003. Adverse health effects of prenatal and postnatal tobacco smoke exposure on children. Archives of Disease in Childhood 88:1086-1090.

Horvath S. 2013. DNA methylation age of human tissues and cell types. Genome biology 14:3156.

Horvath S. 2013. DNA methylation age of human tissues and cell types. Genome biology 14:R115.

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics 13:86.

Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. 2012. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics 13:86.

Jarvelin MR, Sovio U, King V, Lauren L, Xu B, McCarthy MI, et al. 2004. Early life factors and blood pressure at age 31 years in the 1966 northern finland birth cohort. Hypertension (Dallas, Tex : 1979) 44:838-846.

Järvelin MR, Hartikainen‐Sorri AL, Rantakallio P. 1993. Labour induction policy in hospitals of different levels of specialisation. BJOG: An International Journal of Obstetrics & Gynaecology 100:310-315.

Jin H, Ling CX. 2005. Using auc and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering 17:299-310.

Joubert BR, Haberg SE, Bell DA, Nilsen RM, Vollset SE, Midttun O, et al. 2014. Maternal smoking and DNA methylation in newborns: In utero effect or epigenetic inheritance? Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology 23:1007-1017.

Joubert BR, Felix JF, Yousefi P, Bakulski KM, Just AC, Breton C, et al. 2016. DNA methylation in newborns and maternal smoking in pregnancy: Genome-wide consortium meta-analysis. American journal of human genetics 98:680-696.

Kuhn M. 2008. Building predictive models in r using the caret package. Journal of statistical software 28:1-26.

Lee KWK, Richmond R, Hu P, French L, Shin J, Bourdon C, et al. 2015. Prenatal exposure to maternal cigarette smoking and DNA methylation: Epigenome-wide association in a discovery sample of adolescents and replication in an independent cohort at birth through 17 years of age. Environmental health perspectives 123:193-199.

Liaw A, Wiener M. 2002. Classification and regression by randomforest. R news 2:18-22.

Marabita F, Almgren M, Lindholm ME, Ruhrmann S, Fagerström-Billai F, Jagodic M, et al. 2013. An evaluation of analysis pipelines for DNA methylation profiling using the illumina humanmethylation450 beadchip platform. Epigenetics 8:333-346.

Meyer D, Wien FT. 2015. Support vector machines. The Interface to libsvm in package e1071:28.

Mukherjee N, Patil V, Chen S, Zhang H, Arshad SH, Holloway JW, et al. 2016. Interaction of <em>ahrr-</em>methylation and gestational smoking influences adolescent eczema, but not asthma. European Respiratory Journal 48:PA4594.

Oken E, Huh SY, Taveras EM, Rich-Edwards JW, Gillman MW. 2005. Associations of maternal prenatal smoking with child adiposity and blood pressure. Obesity research 13:2021-2028.

Oken E, Levitan EB, Gillman MW. 2007. Maternal smoking during pregnancy and child overweight: Systematic review and meta-analysis. International Journal Of Obesity 32:201.

Pandya R, Pandya J. 2015. C5. 0 algorithm to improved decision tree with feature selection and reduced error pruning. International Journal of Computer Applications 117:18-21.

Parmar P, Lowry E, Cugliari G, Suderman M, Wilson R, Karhunen V, et al. 2018. Association of maternal prenatal smoking gfi1-locus and cardio-metabolic phenotypes in 18,212 adults. EBioMedicine 38:206-216.

Quraishi BM, Zhang H, Everson TM, Ray M, Lockett GA, Holloway JW, et al. 2015. Identifying cpg sites associated with eczema via random forest screening of epigenome-scale DNA methylation. Clinical epigenetics 7:68.

Rantakallio P. 1988. The longitudinal study of the northern finland birth cohort of 1966. Paediatric and perinatal epidemiology 2:59-88.

Rauschert S, Melton PE, Burdge GC, Craig JM, Godfrey KM, Holbrook JD, et al. 2019. Maternal smoking during pregnancy induces persistent epigenetic changes into adolescence, independent of postnatal smoke exposure and is associated with cardiometabolic risk. Frontiers in genetics 10:770.

Reese SE, Zhao S, Wu MC, Joubert BR, Parr CL, Haberg SE, et al. 2017. DNA methylation score as a biomarker in newborns for sustained maternal smoking during pregnancy. Environmental health perspectives 125:760-766.

Reitan T, Callinan S. 2017. Changes in smoking rates among pregnant women and the general female population in australia, finland, norway, and sweden. Nicotine & tobacco research 19:282-289.

Richmond RC, Simpkin AJ, Woodward G, Gaunt TR, Lyttleton O, McArdle WL, et al. 2015. Prenatal exposure to maternal smoking and offspring DNA methylation across the lifecourse: Findings from the avon longitudinal study of parents and children (alspac). Human molecular genetics 24:2201-2217.

Richmond RC, Suderman M, Langdon R, Relton CL, Davey Smith G. 2018. DNA methylation as a marker for prenatal smoke exposure in adults. International journal of epidemiology.

Ridgeway G, Southworth MH, RUnit S. 2013. Package ‘gbm’. Viitattu 10:40.

Rish I. An empirical study of the naive bayes classifier. In: Proceedings of the IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, Vol. 3IBM New York, 41-46.

Rotroff DM, Joubert BR, Marvel SW, Haberg SE, Wu MC, Nilsen RM, et al. 2016. Maternal smoking impacts key biological pathways in newborns through epigenetic modification in utero. BMC genomics 17:976.

Rufibach K. 2010. Use of brier score to assess binary predictions. Journal of Clinical Epidemiology 63:938-939.

Rzehak P, Saffery R, Reischl E, Covic M, Wahl S, Grote V, et al. 2016. Maternal smoking during pregnancy and DNA-methylation in children at age 5.5 years: Epigenome-wide-analysis in the european childhood obesity project (chop)-study. PloS one 11:e0155554.

Schmidhuber J. 2015. Deep learning in neural networks: An overview. Neural networks 61:85-117.

Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, et al. 2014. Obesity-associated variants within fto form long-range functional connections with irx3. Nature 507:371-375.

Sun YV, Smith AK, Conneely KN, Chang Q, Li W, Lazarus A, et al. 2013. Epigenomic association analysis identifies smoking-related DNA methylation sites in african americans. Human genetics 132:1027-1037.

Tehranifar P, Wu HC, McDonald JA, Jasmine F, Santella RM, Gurvich I, et al. 2018. Maternal cigarette smoking during pregnancy and offspring DNA methylation in midlife. Epigenetics 13:129-134.

Timmermans SH, Mommers M, Gubbels JS, Kremers SPJ, Stafleu A, Stehouwer CDA, et al. 2014. Maternal smoking during pregnancy and childhood overweight and fat distribution: The koala birth cohort study. Pediatric Obesity 9:e14-e25.

Viera AJ, Garrett JM. 2005. Understanding interobserver agreement: The kappa statistic. Fam Med 37:360-363.

Wakschlag LS, Pickett KE, Edwin Cook J, Benowitz NL, Leventhal BL. 2002. Maternal smoking during pregnancy and severe antisocial behavior in offspring: A review. American Journal of Public Health 92:966-974.

WHO. 2017. Who report on the global tobacco epidemic, 2017: Monitoring tobacco use and prevention policies. Geneva: World Health Organization.

Wiklund P, Karhunen V, Richmond RC, Parmar P, Rodriguez A, De Silva M, et al. 2019. DNA methylation links prenatal smoking exposure to later life health outcomes in offspring. Clinical epigenetics 11:97.

Yoo C, Ramirez L, Liuzzi J. 2014. Big data analysis using modern statistical and machine learning methods in medicine. International Neurourology Journal 18:50-57.

Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67:301-320.

**Tables**

**Table 1.** Characteristics for the Raine study training and test data subset and the Northern Finland Birth Cohort 1986 and 66. ***p*** is the p-value for the t-test and chi2 test between the Raine Study training and test set. ***SD***: standard deviation.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **Raine Study: Testing** | **Raine Study Training** | ***p*** | **NFBC1986** | **NFBC1966** |
| **n** | 198 | 797 |  | 478 | 602 |
| **Age in years (mean (SD*a*))** | 17.20 (0.49) | 17.27 (0.61) | 0.132*b* | 16.06 (0.36) | 31.01 (0.34) |
| **Sex (%)** |  |  |  |  |  |
| Male | 99 (50.0) | 402 (50.4) | 0.975*c* | 221 (46.2) | 261 (43.4) |
| Female | 99 (50.0) | 395 (49.6) |  | 257 (53.8) | 341 (56.6) |
| **Adolescent Smoking (%)d** |  |  | 0.401*e* |  |  |
| Non-smoker | 108 (54.5) | 392 (49.2) |  | 288 (60.3) | 283 (47.0) |
| Ever-Smoker | 42 (21.2) | 191 (24.0) |  | 164 (34.3) | 313 (52.0) |
| Missing | 48 (24.2) | 214 (26.9) |  | 26 (5.4) | 6 (1.0) |
| **Maternal smoking during pregnancy (%)f** |  |  | 1*c* |  |  |
| Exposed | 59 (29.8) | 237 (29.7) |  | 95 (19.9) | 130 (21.6) |
| Not-Exposed | 139 (70.2) | 560 (70.3) |  | 383 (78.7) | 472 (78.4) |

a SD: Standard Deviation

*b*: Wilcoxon Mann-Whitney U-test;

*c*: Chi-square test

d adolescent smoking status was defined as ever smoked during the lifetime versus never smoked as based on questionnaires.

*e* Wilcoxon Mann-Whitney U-test.

f maternal smoking was defined as any smoking during pregnancy.

**Table 2.** Model quality measures (Sensitivity, Specificity, Cohen’s kappa, Accuracy, Area under the Receiver Operator Curve (AUC) curve and brier score) for the elastic net machine learning model, Reese et al. cord blood, Richmond et al. 568 CpG, Richmond et al. 19 CpG score the gradient boosting machine, random forest and support vector machine models that were amongst the four best performing models in our analysis. Results provided int his table are based on the Raine Study test data (n=198), NFBC1986 (n=478) and NFBC1966 (n=602).

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Sensitivity** | **Specificity** | **Cohen’s kappa** | **Accuracy** | **AUC** | **Brier Score** | **# CpGs required** |
| **Raine Study test data set** |  |  |  |  |  |  |  |
| Elastic Net Score | 0.91 | 0.76 | 0.68 | 0.83 | 0.87 | 0.13 | 204 |
| Gradient Boosting Machine | 0.91 | 0.82 | 0.72 | 0.88 | 0.88 | 0.1 | 1511 |
| Random Forest | 0.87 | 0.73 | 0.58 | 0.83 | 0.83 | 0.17 | 1511 |
| Support Vector Machine | 0.87 | 0.73 | 0.6 | 0.83 | 0.85 | 0.13 | 1511 |
| Reese Score | 0.88 | 0.72 | 0.6 | 0.83 | 0.85 | 0.21 | 28 |
| Richmond Score 568 CpGs | 0.7 | 0.68 | 0.34 | 0.69 | 0.72 | 0.22 | 568 |
| Richmond Score 19 CpGs | 0.79 | 0.58 | 0.37 | 0.72 | 0.73 | 0.22 | 19 |
| **NFBC1986** |  |  |  |  |  |  |  |
| Elastic Net Score | 0.87 | 0.75 | 0.56 | 0.84 | 0.85 | 0.13 | 204 |
| Gradient Boosting Machine | 0.95 | 0.29 | 0.19 | 0.54 | 0.74 | 0.39 | 1511 |
| Random Forest | 0.79 | 0.16 | 0.06 | 0.64 | 0.54 | 0.24 | 1511 |
| Support Vector Machine | 0.87 | 0.44 | 0.33 | 0.77 | 0.79 | 0.16 | 1511 |
| Reese Score | 0.87 | 0.61 | 0.46 | 0.82 | 0.8 | 0.18 | 28 |
| Richmond Score 568 CpGs | 0.65 | 0.76 | 0.34 | 0.74 | 0.71 | 0.22 | 568 |
| Richmond Score 19 CpGs | 0.65 | 0.77 | 0.31 | 0.68 | 0.73 | 0.22 | 19 |
| **NFBC1966** |  |  |  |  |  |  |  |
| Elastic Net Score | 0.72 | 0.78 | 0.39 | 0.73 | 0.8 | 0.19 | 204 |
| Gradient Boosting Machine | 0.88 | 0.26 | 0.1 | 0.45 | 0.68 | 0.48 | 1511 |
| Random Forest | 0.77 | 0.18 | 0.05 | 0.64 | 0.48 | 0.24 | 1511 |
| Support Vector Machine | 0.88 | 0.45 | 0.33 | 0.76 | 0.75 | 0.2 | 1511 |
| Reese Score | 0.72 | 0.7 | 0.32 | 0.71 | 0.73 | 0.18 | 28 |
| Richmond Score 568 CpGs | 0.66 | 0.63 | 0.22 | 0.69 | 0.72 | 0.22 | 568 |
| Richmond Score 19 CpGs | 0.61 | 0.72 | 0.23 | 0.63 | 0.73 | 0.22 | 19 |

Table 3. DeLong test for significant difference between all ROC curves in Figure 2.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **DeLong-Test p-values** | | |
| **Test** | **Raine Study** | **NFBC1986** | **NFBC1966** |
| **Elastic net vs Reese score** | 0.49 | 0.12 | 0.03 |
| **Elastic net vs Richmond young** | 0.00058 | 0.008 | 0.006 |
| **Elastic net vs Richmond old** | 0.01 | 0.04 | 0.004 |
| **Reese vs Richmond young** | 1.57E-06 | 0.23 | 0.43 |
| **Reese vs Richmond old** | 0.002 | 0.67 | 0.47 |
| **Richmond young vs Richmond old** | 0.001 | 0.41 | 0.91 |

**Figure 1 :** Flow chart for the modelling steps. This includes details of the steps undertaken in the training, testing and validation phase, as well as the data used per step.

**Figure 2.** Receiver operator curve (ROC) for the four different model scores tested: elastic net regression, Reese et al. methylation score, Richmond et al. 568 CpG and 19 CpG scores. Area under the curve provided for every score, applied to the Raine Study test set, NFBC1986 and NFBC1966. ***AUC***: Area under the ROC