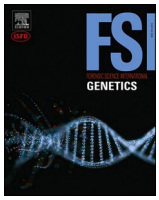
Forensic Science International: Genetics 70 (2024) 103022

Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsigen



# Inference of tobacco and alcohol consumption habits from DNA methylation analysis of blood

A. Ambroa-Conde [a], M.A. Casares de Cal [b], A. Gómez-Tato [b], O. Robinson [c], A. Mosquera-Miguel [a], M. de la Puente [a], J. Ruiz-Ramírez [a], C. Phillips [a], M.V. Lareu [a], A. Freire-Aradas [a,*]

[a] Forensic Genetics Unit, Institute of Forensic Sciences, Universidade de Santiago de Compostela, Spain
[b] CITMAga (Center for Mathematical Research and Technology of Galicia), University of Santiago de Compostela, Spain
[c] MRC Centre for Environment and Health, Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK

ARTICLE INFO

*Keywords:*
DNA methylation
Logistic regression
Quantile regression
Blood
Tobacco
Alcohol
Age estimation

ABSTRACT

DNA methylation has become a biomarker of great interest in the forensic and clinical fields. In criminal investigations, the study of this epigenetic marker has allowed the development of DNA intelligence tools providing information that can be useful for investigators, such as age prediction. Following a similar trend, when the origin of a sample in a criminal scenario is unknown, the inference of an individual's lifestyle such as tobacco use and alcohol consumption could provide relevant information to help in the identification of DNA donors at the crime scene. At the same time, in the clinical domain, prediction of these trends of consumption could allow the identification of people at risk or better identification of the causes of different pathologies. In the present study, DNA methylation data from the UK AIRWAVE study was used to build two binomial logistic models for the inference of smoking and drinking status. A total of 348 individuals (116 non-smokers, 116 former smokers and 116 smokers) plus a total of 237 individuals (79 non-drinkers, 79 moderate drinkers and 79 drinkers) were used for development of tobacco and alcohol consumption prediction models, respectively. The tobacco prediction model was composed of two CpGs (cg05575921 in *AHRR* and cg01940273) and the alcohol prediction model three CpGs (cg06690548 in *SLC7A11*, cg0886875 and cg21294714 in *MIR4435–2HG*), providing correct classifications of 86.49% and 74.26%, respectively. Validation of the models was performed using leave-one-out cross-validation. Additionally, two independent testing sets were also assessed for tobacco and alcohol consumption. Considering that the consumption of these substances could underlie accelerated epigenetic ageing patterns, the effect of these lifestyles on the prediction of age was evaluated. To do that, a quantile regression model based on previous studies was generated, and the potential effect of tobacco and alcohol consumption with the epigenetic age was assessed. The Wilcoxon test was used to evaluate the residuals generated by the model and no significant differences were observed between the categories analyzed.

## 1. Introduction

DNA methylation has become a biomarker of interest in the forensic field. It has been studied for individual age estimation [1,2], tissue determination [3] and differentiation of monozygotic twins [4]. Additionally, since this marker undergoes changes caused by exogenous agents [5] and medical disorders [6], its use has been proposed for the study of environmental factors and diseases in the clinical domain [7]. DNA methylation has also been correlated with lifestyle factors in both clinical [8–10] and forensic [11,12] fields. In relation to the clinical context, development of indices based on epigenetic markers as risk indicators in health disorders related to tobacco and alcohol could be of great interest. In the case of tobacco, an index based on buccal cells has been developed allowing discrimination between normal and cancerous cells [13,14]. For alcohol, a correlation between methylation and consumption disorders has so far been demonstrated [15]. Prediction of these trends of consumption could allow the identification of people at risk or a better identification of the causes of different pathologies. In forensic DNA analysis, the prediction of lifestyles and environmental exposures would allow a better characterization of unknown

* Corresponding author.
*E-mail address:* ana.freire@usc.es (A. Freire-Aradas).

https://doi.org/10.1016/j.fsigen.2024.103022
Received 22 August 2023; Received in revised form 22 December 2023; Accepted 25 January 2024
Available online 28 January 2024

perpetrators from a biological trace. Therefore, the inference of tobacco and alcohol consumption from DNA could be used as a DNA-intelligence tool for police investigations helping to reduce the number of suspects when a DNA sample cannot be matched to any reference sample or profile stored in DNA databases [12].

In recent years, several articles have been published assessing DNA methylation differences between tobacco and alcohol users and non-users. In the case of tobacco, a tendency towards hypomethylation has been observed in smokers compared to non-smokers [16,17]. However, in the case of alcohol, although clear differences have been observed in the methylation patterns of drinkers and non-drinkers [15], no clear methylation trend has been observed. Different studies have identified general tendencies for drinkers to hypermethylate [18,19], and non-drinkers to hypomethylate [20–22] or show both [23]. It has been shown that these differentially methylated patterns between tobacco and alcohol users and non-users can be reversed at some positions [20, 24,25]. For alcohol consumption, a partial recovery of methylation levels was observed after several weeks of abstinence [20]. For tobacco, on the other hand, a more in-depth study identified reversible and irreversible positions over time after cessation of smoking [25–27]. For some positions, the methylation values of former smokers, recovered to levels observed in non-smokers within 0 to 35 years [25–27]. In contrast, other CpG positions maintained methylation levels of smokers even 35 years after cessation [25].

As a result of the discovery articles published to date, many DNA methylation markers (CpG positions) correlated with the consumption of both substances have been identified, and prediction models of smoking and alcohol status built accordingly. In the case of tobacco, the models generated have been developed mainly with blood samples, using different population groups, technologies and statistical models [28–32]. In the model generated by Elliott et al. [28] the differences between two populations were directly assessed (Europe vs South Asia). It was observed that for some CpG positions, the methylation values differed greatly between the groups analyzed, but it was possible to generate a Random Forest model with sensitivity and specificity values of more than 80% for both populations for predicting current active smokers. Furthermore, additional logistic regression models were used to infer years since smoking cessation, number of cigarette per day, as well as years as a smoker [30,33]. Finally, using a single marker, Philibert et al. [31] generated prediction models for blood and saliva while assessing sex and age as covariables.

For the prediction of drinking status, a smaller number of models have been developed to date [32,34–36], all of them blood-based and mainly for European populations. It is worth mentioning the model developed by Liu et al. composed of 144 CpGs obtained Area Under the Curve (AUC) values for heavy drinkers vs non-drinkers of 0.80 to 0.99 in four replication cohorts [34]. For this model, subsequent studies evaluated a possible overestimation of the results by overfitting, obtaining AUC values between 0.50 and 0.75 [36–38]. Further evaluation of this model is necessary to avoid the potential overestimation observed. In addition, two other alcohol prediction models were developed obtaining AUC values of 0.73 [35] and 0.74 [32] for the classification of light to moderate vs heavy drinker, and non-drinker vs heavy drinker, respectively.

The study of such lifestyle factors in the forensic field is of interest not only for the inference of consumption by itself, but also to evaluate their effect on the individual age prediction models developed so far. It has been shown that the consumption of alcohol and tobacco may be correlated with age acceleration, as they are contributing factors in age-related diseases [39]. Age acceleration caused by these lifestyle factors was initially assessed and some correlation was observed, but further research is needed [39–41].

In the present study, DNA methylation data from the UK AIRWAVE study [42] were used to generate binomial logistic regression models for the classification of smoking and drinking alcohol status. The database used was generated from the Airwave Health Monitoring Study [43], that has been recruiting participants among UK police officers since 2004. These models were generated using 348 individuals and 237 individuals for tobacco and alcohol consumption, respectively. In the case of tobacco, a final model comprising 2-CpGs was selected that addressed non-smokers + former smokers vs active smokers. For alcohol, however, non-drinker and moderate drinker were grouped together to be compared with heavy drinkers, selecting a final model comprising 3-CpGs. Finally, age prediction models were assessed showing they were not affected by an individual's smoking or drinking status.

## 2. Material and methods

### 2.1. DNA methylation data and sample classification

DNA methylation data was accessed from the UK AIRWAVE study [42]. A total of 1115 blood samples (452 females and 663 males) were evaluated, which had been analyzed with the Infinium MethylationEPIC BeadChip array, composed of 853,307 CpGs. The age range of the samples analyzed was 19 to 65 years old (standard deviation: ± 13.52 years). For these samples, data related to their lifestyle were available in the form of a questionnaire. Questions relating to tobacco and alcohol consumption were used to classify the samples into different categories.

For smoking status classification, the questions evaluated were: whether or not the volunteer is a smoker, and if currently they do not smoke but have smoked five or more cigarettes a day in the past. Taking into account the first question, individuals who answered "NO" were classified as non-smokers and those who answered "YES" were classified as smokers. The individuals that had answered "YES" to the second question were classified as former smokers. After performing this classification, the individuals were grouped as following: 116 smokers, 728 non-smokers and 271 former smokers (detailed information can be found in Table 1). For model building, the number of individuals in each group was matched, taking as reference the category with the lowest number of classified individuals (116 smokers).

**Table 1**
Summary of the tobacco and alcohol group classification for the 1115 blood samples.

| Lifestyle | Group | Sample size | Gender | Age (years old) |
|---|---|---|---|---|
| Tobacco | Non-smoker | 728 | 281 women<br>447 men | 19-65 |
| | Smoker | 116 | 45 women<br>71 men | 20-65 |
| | Former smoker | 271 | 126 women<br>145 men | 22-64 |
| Alcohol | Non-drinker | 79 | 44 women<br>35 men | 21-65 |
| | Moderate drinker | 956 | 364 women<br>592 men | 19-65 |
| | Heavy drinker | 79 | 44 women<br>35 men | 20-64 |

For alcohol classification, five questions referring to the consumption of different alcoholic beverages for one week were considered. For this purpose, the total amount of alcoholic units consumed by each participant was assessed, classifying them into three groups based on this value. For a correct classification, the gender of the participants was considered. Therefore, the classification was performed according to the following conditions: values equal to 0 as non-drinker; values $> 0$ and $\leq$ 14 units per week for women as moderate drinkers; values $> 0$ and $\leq 21$ units for men as moderate drinkers; values $> 14$ units for women as heavy drinkers; and values $> 21$ units for men as heavy drinkers. From this classification, the individuals were grouped as following: 79 non-drinkers, 956 moderate drinkers and 80 heavy drinkers (detailed information is shown in Table 1). For model building, the number of individuals of the groups was matched, with reference to the category with the lowest number of classified individuals (79 non-drinkers and heavy drinkers).

### 2.2. Statistical analysis

The selection of candidate CpG sites to infer the studied lifestyles was based on the AUC (Area under the ROC Curve), and on the percentage of correct classifications (%CC), both obtained from binomial logistic regression analysis. Firstly, AUC was calculated for all the 853,307 CpGs included in the Infinium MethylationEPIC BeadChip array using the pROC R package [44], and those presenting values equal or higher to 0.7 were retained. For these retained CpGs and taking into account the maximum number of CpGs recommended to be included in logistic regression models without overfitting – $p + 1 \leq min\ (n0, n1, n2)/10$ *parameters* [45] – %CC was calculated and those markers depicting values equal or higher to 70% were selected. Statistical significance was set at p-value $\leq 0.05$ (E-2).

For the lifestyle prediction models two different statistical approaches were used, binomial and multinomial logistic models, developed using the nnet package [46] for multinomial regression. The corresponding predictive accuracy for the logistic regression models was measured with the following performance metrics: sensitivity, specificity, AUC and percentage of correct classifications (%CC). For the evaluation of these parameters in binomial logistic models, it should be considered that smoker and heavy drinker groups were set as class 1, so a better prediction of belonging to this group produces a higher sensitivity. For the other group assessed in each model, class 0, the specificity will be the parameter related to its classification. Principal Component Analysis (PCA) was carried out using the factoextra R package [47]. Cross-validation of the developed logistic regression models was performed with a leave-one-out cross validation using the pROC R package.

An age prediction model based on multivariate quantile regression was built using the quantreg R package [48] taking as reference a previous model [49]. Cross-validation of the age prediction model was performed with a k-fold cross-validation ($k = 10$) using cvTools R package [50]. For the age prediction model, the median absolute error (MAE) was used to measure the predictive accuracy. Representations of the DNA methylation values, as well as the predicted vs chronological age were made using the ggplot2 R package [51]. Correlations between DNA methylation levels and chronological age were evaluated using the Spearman correlation test ($r_s$), and the inter-group variability was analyzed using the standard deviation (SD). All statistical analyses were carried out using R software v.4.1.1 [52] with scripts developed in-house.

## 3. Results

### 3.1. Selection of candidate CpGs

For the selection of the markers correlated with tobacco and alcohol, matched samples for age and sex for a total of 116 non-smokers (age range: 19–62, mean: 40.56 years; female/male ratio: 0.55) vs 116 smokers (age range: 20–65, mean: 41.73 years; female/male ratio: 0.63), plus 79 non-drinkers (age range: 21–65, mean: 42.13 years; female/male ratio: 1.26) vs 79 drinkers (age range: 20–64, mean: 42.65 years; female/male ratio: 1.19) were assessed. A total of 853,307 binomial logistic regression models were generated (one per CpG included in the Infinium MethylationEPIC BeadChip array). Of all the evaluated models, those with an AUC value below 0.7 were discarded, keeping a total of 67 tobacco-correlated CpGs and 30 alcohol-correlated CpGs, that were selected for further evaluation.

In order to reduce the number of markers analyzed, the formula $p + 1 \leq min\ (n0, n1, n2)/10$ *parameters* [53] was used in order to define the maximum number of CpGs recommended to be included in logistic regression models without overfitting. Thus, the number of events per variable was evaluated based on the number of individuals, taking into account that a minimum of 10 events per parameter is advisable to avoid overfitting [45,53]. The number of markers that could be used depends on the number of parameters in the model, with binomial models (dichotomous variables) allowing a larger number of markers than multinomial models (polytomous variables). Therefore, with the evaluated groups defined as *n0*, *n1* and *n2* using the previous formula [45], for tobacco, with a maximum of 116 individuals in one of the evaluated groups (smokers), the generated models should not contain more than 4 and 10 CpGs for multinomial and binomial logistic models, respectively. For alcohol, both non-drinker and heavy drinker groups contained an equal number of 79 individuals, reducing the number of recommended markers to 3 and 6 for multinomial and binomial models, respectively. Since many CpGs presented very similar AUC values and considering that a forward approach was used according to the CpGs sorted first by AUC and then by %CC, it was decided to select as many markers as allowed in the binomial models for each lifestyle. For this final selection, following the order of markers defined by the AUC, percentage of correct classifications was calculated and only those CpGs, up to the maximum CpG number, as defined above, with a value equal to or higher than 70% of correct classifications were selected.

Based on these criteria and avoiding those CpGs that had more than 50 'not analyzed' (NA), a total of 10 tobacco-correlated and 5 alcohol-correlated CpGs were selected. Figs. 1 and 2 show the corresponding boxplots for the DNA methylation values per group for tobacco and alcohol, respectively. Although selection of candidate CpGs was performed using the extreme groups for both cases, boxplots have been arranged to show the DNA methylation levels for the three groups per lifestyle. Table 2 shows the selected markers for both lifestyles ordered by the criteria defined above. As it can be observed, all the selected markers (10 and 5 CpGs for tobacco and alcohol, respectively) were statistically significant (p-values range between E-15 to E-5), although showing the tobacco-correlated CpGs higher correlation values than the corresponding CpGs for alcohol (mean: E-12 versus E-6, respectively).

Potential correlation with age for the 15 selected CpG sites was evaluated for each category inside each lifestyle, with all markers showing $r_s$ values below or close to |0.50| (Supplementary Table S1). When assessing the correlation with age, generally higher $r_s$ values were observed for the smoker group (mean: |0.27|) compared to non-smokers (mean: |0.12|) and former smokers (mean: |0.14|). In the case of alcohol, the moderate and heavy drinker groups showed similar mean $r_s$ values (mean of |0.26| and |0.29|, respectively), with non-drinkers giving lower values for the majority of markers (mean: |0.08|). It is noteworthy that differences were observed in some specific markers, e.g., *SLC7A11* and *MIR4435–2HG*, with mainly one category standing out from the rest (−0.46 for heavy drinkers and −0.34 for moderate drinkers, respectively). In addition, evaluation of the dispersion of DNA methylation values of the groups, indicated similar trends observed between the groups of the studied lifestyles (tobacco presenting mean values for non-smokers of 0.04, former smokers: 0.05 and smokers: 0.06; alcohol presenting for all groups a mean value of 0.04).

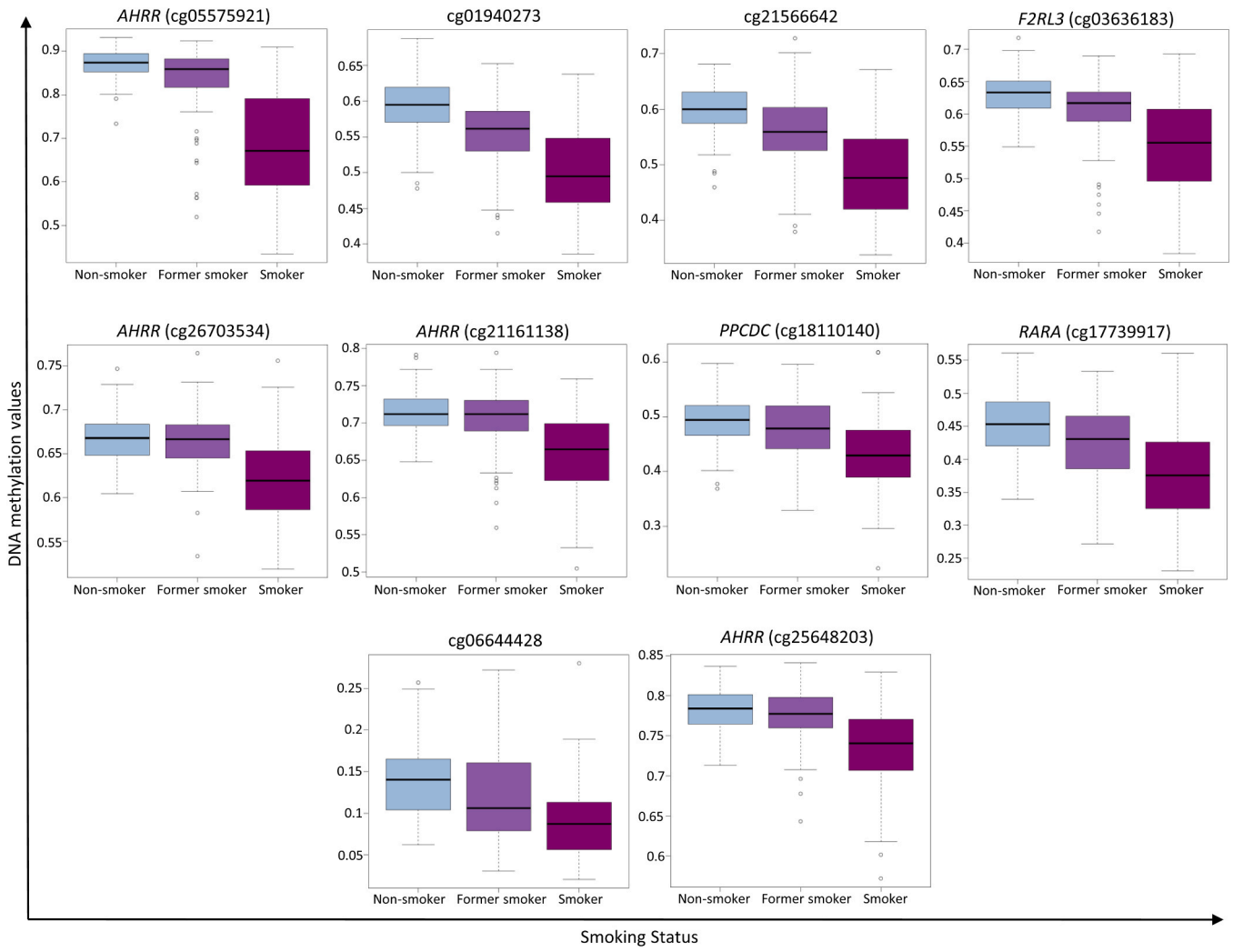



**Fig. 1.** Boxplots representing the DNA methylation values for the 10 CpGs correlated with tobacco consumption presenting an AUC and %CC higher than 0.7 and 70%, respectively, for N = 116 non-smoker, N = 116 former smoker and N = 116 smoker individuals. Markers are sorted in descending order by AUC and %CC.

### 3.2. *Development of a prediction model for tobacco consumption*

A multinomial logistic regression model was explored to differentiate three categories for smoking status (N = 116 smokers, N = 116 former smokers and N = 116 non-smokers). For this evaluation, several combinations of the selected CpG sites were tested. Considering the maximum number of markers recommended for the tobacco multinomial models (4 CpGs), four combinations of markers were evaluated following a forward approach, i.e. the addition of one marker per subsequent model. The addition of CpG sites stopped when no additional improvement was discernible from the model. The corresponding percentage of correct classifications is shown in Table 3. Following the order established for the selected CpGs (Table 2), models were generated including one by one, the CpGs presenting the highest AUC values in descending order. As shown in Table 3, a slight increase in the global correct classification rate was observed, as the number of markers in the models increased (from 58.91% to 66.09%, for the 1-CpG through to the 4-CpGs model, respectively). In detailed assessments of the specific categories, we observed that the extreme groups gave higher classification rates (mean: 73.28% and 70.48% for non-smokers and smokers, respectively) compared with former smokers (mean: 46.55%).

Considering the results achieved, a multinomial logistic model does not readily provide correctly classified consumption habits for the three groups under study. Therefore, binomial logistic regression was subsequently explored. To perform these analyses while keeping all the samples from the three categories represented, it was decided to group two categories into one. For this purpose, the classification table (Table 3) of the most accurate multinomial model (4-CpGs) was used to assess the classification trend of the intermediate group (former smokers). In the multinomial model, 52.59% of former smokers were correctly classified, with most of the remaining individuals (36.21%) being classified as non-smokers. Consequently, between these two groups, 88.80% of the former smokers were present, with a greater tendency to be classified as non-smokers than smokers (36.21% and 11.21%, respectively). Therefore, for the creation of the binomial logistic models, it was decided to combine non-smoker and former smoker as a single category. Additionally, in order to avoid losing informativeness, all the samples from both groups were retained. Therefore, binomial logistic models were generated by comparing the non-smokers + former smokers group of 232 individuals with 116 smokers. Following the same approach as the one used for multinomial models, up to three models were assessed for this analysis. The corresponding performance metrics are shown in Table 4. In this case, since no improvement in the classifications was achieved by the third CpG included in the model, the models tested stopped at three CpG sites, and a 2-CpG model of *AHRR* (cg05575921) and cg01940273 was selected (Fig. 3A), providing 86.49% of correct classifications. The additional performance metrics included an AUC level of 0.87, while specificity (0.90) was higher than sensitivity (0.79), indicating a better classification of the non-smokers + former smokers group vs smokers (%CC: 90.09% vs 79.31%, respectively).

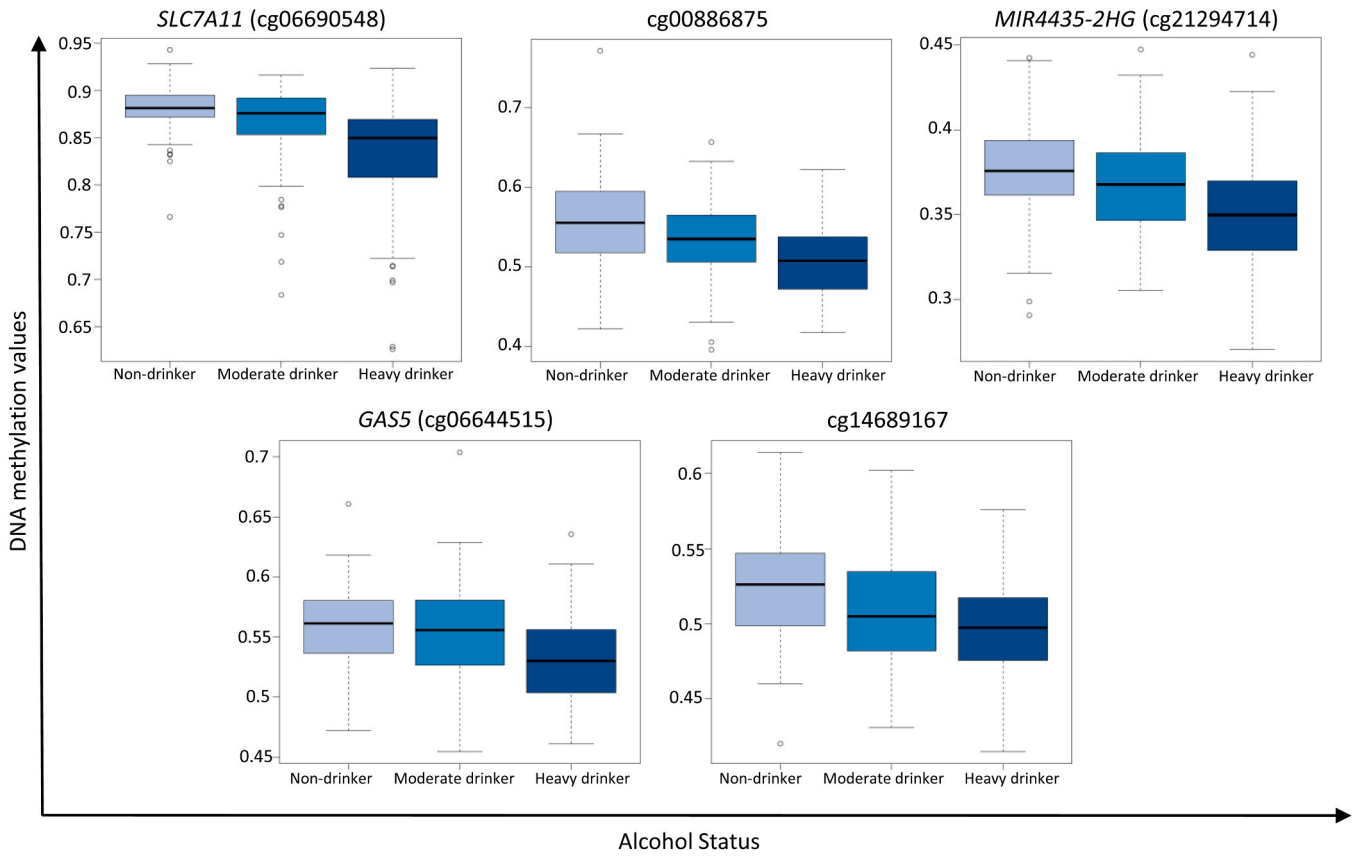



**Fig. 2.** Boxplots representing the DNA methylation values for the 5 CpGs correlated with alcohol consumption presenting an AUC and %CC higher than 0.7 and 70%, respectively, for N = 79 non-drinker, N = 79 moderate drinker and N = 79 drinker individuals. Markers are sorted in descending order by AUC and %CC.

**Table 2**
Preliminary selection of 10 smoking and 5 drinking related CpGs showing a value higher than 0.7 for AUC, percentage of correct classifications (%CC) and statistical significance (p-value), based on DNA methylation data from the DPUK platform.

| Lifestyle | Gene | CpG_ID | GRCh38 chromosome position | AUC | %CC | P-value |
|---|---|---|---|---|---|---|
| Tobacco | *AHRR* | cg05575921 | chr5:373263 | 0.90 | 88.79% | 1.77E-11 |
| | none | cg01940273 | chr2:232420224 | 0.89 | 86.21% | 3.73E-15 |
| | none | cg21566642 | chr2:232419951 | 0.89 | 83.19% | 9.57E-15 |
| | *F2RL3* | cg03636183 | chr19:16889774 | 0.86 | 81.47% | 1.71E-12 |
| | *AHRR* | cg26703534 | chr5:377243 | 0.82 | 76.29% | 1.27E-12 |
| | *AHRR* | cg21161138 | chr5:399245 | 0.81 | 77.59% | 3.99E-12 |
| | *PPCDC* | cg18110140 | chr15:75058039 | 0.81 | 73.71% | 1.30E-11 |
| | *RARA* | cg17739917 | chr17:40321320 | 0.80 | 76.72% | 3.12E-12 |
| | none | cg06644428 | chr2:232419402 | 0.80 | 72.84% | 1.92E-11 |
| | *AHRR* | cg25648203 | chr5:395329 | 0.79 | 73.28% | 5.67E-11 |
| Alcohol | *SLC7A11* | cg06690548 | chr4:138241654 | 0.82 | 77.85% | 3.44E-7 |
| | none | cg00886875 | chr3:106635478 | 0.76 | 70.89% | 4.84E-7 |
| | *MIR4435-2HG* | cg21294714 | chr2:111429551 | 0.74 | 72.15% | 4.63E-6 |
| | *GAS5* | cg06644515 | chr1:173865693 | 0.73 | 70.89% | 5.64E-6 |
| | none | cg14689167 | chr10:4267023 | 0.71 | 71.52% | 1.92E-5 |

Considering the reduced number of samples, a leave-one-out cross-validation was performed to validate the selected model. The cross-validation showed the same values for sensitivity, specificity and %CC as the model (0.79, 0.90 and 86.49%, respectively), with only a slight difference in the AUC value obtained (0.86).

As no further data for smokers was available, the discarded individuals from the non-smokers and former smokers group were used as a testing set to assess the accuracy of the model for this category. Thus, tobacco consumption status was predicted for a total of 606 non-smokers and 151 former smokers, obtaining a 93.39% of correctly classified non-smokers and former smokers vs only 6.61% classified as smokers.

### 3.3. Development of a prediction model for alcohol consumption

To explore multinomial logistic regression for alcohol consumption, a similar strategy was used to that outlined above for tobacco. DNA methylation values available for 79 non-drinkers, 79 moderate drinkers and 79 heavy drinkers were used to generate a drinking status model. Models were generated following a forward approach, adding in descending order the CpGs listed in Table 2. The maximum number of variables (3 CpGs) considering that groups had 79 individuals was taken into account and subsequently, a maximum of three models were tested. The corresponding performance metrics are detailed in Table 3. The multinomial models generated for the prediction of drinking status

**Table 3**
Summary of the accuracy of the evaluated multinomial logistic regression models for tobacco (N = 116 non-smoker, N = 116 former smoker and N = 116 smoker), and alcohol status prediction (N = 79 non-drinker, N = 79 moderate drinker and N = 79 heavy drinker).

| | Model | CpGs | Correct Classifications | Classification Table | | | |
|---|---|---|---|---|---|---|---|
| Tobacco | 1-CpG model | *AHRR* (cg05575921) | 58.91% | **Predicted\Real** | **Non-smoker** | **Former smoker** | **Smoker** |
| | | | | **Non-smoker** | 74.14% | 52.59% | 12.93% |
| | | | | **Former smoker** | 25.00% | 32.76% | 17.24% |
| | | | | **Smoker** | 0.86% | 14.66% | 69.83% |
| | 2-CpGs model | *AHRR* (cg05575921), cg01940273 | 64.37% | **Predicted\Real** | **Non-smoker** | **Former smoker** | **Smoker** |
| | | | | **Non-smoker** | 72.41% | 37.07% | 11.21% |
| | | | | **Former smoker** | 25.86% | 50.86% | 18.97% |
| | | | | **Smoker** | 17.24% | 12.07% | 69.83% |
| | 3-CpGs model | *AHRR* (cg05575921), cg01940273, cg21566642 | 64.37% | **Predicted\Real** | **Non-smoker** | **Former smoker** | **Smoker** |
| | | | | **Non-smoker** | 73.28% | 37.07% | 11.21% |
| | | | | **Former smoker** | 25.86% | 50.00% | 18.97% |
| | | | | **Smoker** | 0.86% | 12.93% | 69.83% |
| | 4-CpGs model | *AHRR* (cg05575921), cg01940273, cg21566642, *F2RL3* | 66.09% | **Predicted\Real** | **Non-smoker** | **Former smoker** | **Smoker** |
| | | | | **Non-smoker** | 73.28% | 36.21% | 10.34% |
| | | | | **Former smoker** | 25.86% | 52.59% | 17.24% |
| | | | | **Smoker** | 0.86% | 11.21% | 72.41% |
| Alcohol | 1-CpG model | *SLC7A11* | 48.52% | **Predicted\Real** | **Non-drinker** | **Moderate drinker** | **Heavy drinker** |
| | | | | **Non-drinker** | 70.89% | 56.96% | 21.52% |
| | | | | **Moderate drinker** | 21.52% | 20.25% | 24.05% |
| | | | | **Heavy drinker** | 7.59% | 22.78% | 54.43% |
| | 2-CpGs model | *SLC7A11*, cg00886875 | 50.21% | **Predicted\Real** | **Non-drinker** | **Moderate drinker** | **Heavy drinker** |
| | | | | **Non-drinker** | 65.82% | 43.04% | 17.73% |
| | | | | **Moderate drinker** | 22.78% | 26.58% | 24.05% |
| | | | | **Heavy drinker** | 11.39% | 30.38% | 58.23% |
| | 3-CpGs model | *SLC7A11*, cg00886875, *MIR4435-2HG* | 52.74% | **Predicted\Real** | **Non-drinker** | **Moderate drinker** | **Heavy drinker** |
| | | | | **Non-drinker** | 63.29% | 49.37% | 12.66% |
| | | | | **Moderate drinker** | 25.32% | 27.85% | 21.52% |
| | | | | **Heavy drinker** | 11.39% | 22.78% | 65.82% |

showed low percentages of correct classifications (mean: 50.49%). The highest %CC was obtained for the 3-CpGs model (52.74%), with a large proportion of the misclassified individuals been related to the moderate drinker category (> 70%).

The results obtained for multinomial logistic regression do not allow an adequate classification of the three categories under study. Therefore, binomial logistic regression was then evaluated. As in the case of tobacco, it was decided to combine two of the categories, while retaining all the samples from the three groups represented. When evaluating the classification table of the most accurate multinomial model (3-CpGs), the classification trend of the intermediate group was used (moderate drinkers). It was observed that 49.37% of the moderate drinkers were classified as non-drinkers, with the combination of these categories accounting for the majority of the observed moderate drinkers (77.22%). Therefore, it was decided, considering the trend of the intermediate group, to combine non-drinkers and moderate drinkers into a single category. Additionally, to avoid exclusion of individuals used in marker selection, we grouped non-drinkers and moderate drinkers adding all individuals of each group. Thus, binomial logistic regression models were built for 158 individuals of the non-drinkers + moderate drinkers group against 79 heavy drinkers. The forward approach was used for model building and the corresponding performance metrics evaluated (Table 4). In this case, a 3-CpG model composed of *SLC7A11* (cg06690548), cg0886875 and *MIR4435–2HG* (cg21294714) was selected (Fig. 3B), providing 74.26% correct classifications. Additional performance metrics comprised an AUC level of 0.80 and sensitivity (0.81) was higher than specificity (0.71), reflecting a better classification of the heavy drinker group vs non-drinkers + moderate drinkers (%CC: 81.01% vs 70.89%, respectively).

To evaluate the accuracy of the selected model, a leave-one-out cross-validation was carried out showing similar values to those obtained in the selected model (AUC: 0.79, sensitivity: 0.80, specificity: 0.70 and %CC: 72.57%).

In the case of alcohol, the remaining individuals in the available dataset were moderate drinkers (N = 858), with any non-drinkers and/or heavy drinkers available for testing. It was decided to use these individuals as a testing set to assess the accuracy of the model for such a heterogeneous group as moderate drinkers. Of the 858 individuals available, 852 (99.30%) were correctly classified as non-drinkers + moderate drinkers, and only 6 individuals (0.7%) were classified as heavy drinkers.

### 3.4. Tobacco and alcohol effects on age prediction

To evaluate the effects of the lifestyle factors assessed in this study on the prediction of epigenetic age, a quantile regression model was developed. This model was performed using the DNA methylation data from the same sample set used for building the tobacco and alcohol prediction models, except for three samples which presented missing data for the CpGs under study in this section (N = 1112, 19 to 65 years old, analyzed with Infinium MethylationEPIC BeadChip array). Based on the markers employed in a previous publication [49], five out of the seven reported CpGs were used for this analysis: *ELOVL2* (cg21572722), *ASPA* (cg02228185), *FHL2* (cg06639320), *CCDC102B* (cg19283806)

**Table 4**
Summary of the accuracy of the evaluated binomial logistic regression models for tobacco (N = 232 non-smoker + former smoker vs N = 116 smoker), and alcohol status prediction (N = 158 non-drinker + moderate drinker vs N = 79 heavy drinker). The selected final models are highlighted in bold.

| | Model | CpGs | AUC | Sensitivity | Specificity | Correct Classifications | Classification Table | | |
|---|---|---|---|---|---|---|---|---|---|
| Tobacco | 1-CpG model | AHRR (cg05575921) | 0.87 | 0.81 | 0.87 | 85.06% | **Predicted\Real** | **Non-smoker + Former smoker** | **Smoker** |
| | | | | | | | **Non-smoker + Former smoker** | 87.07% | 18.97% |
| | | | | | | | **Smoker** | 12.93% | 81.03% |
| | **2-CpGs model** | **AHRR (cg05575921), cg01940273** | **0.87** | **0.79** | **0.90** | **86.49%** | **Predicted\Real** | **Non-smoker + Former smoker** | **Smoker** |
| | | | | | | | **Non-smoker + Former smoker** | 90.09% | 20.69% |
| | | | | | | | **Smoker** | 9.91% | 79.31% |
| | 3-CpGs model | AHRR (cg05575921), cg01940273, cg21566642 | 0.87 | 0.79 | 0.90 | 86.21% | **Predicted\Real** | **Non-smoker + Former smoker** | **Smoker** |
| | | | | | | | **Non-smoker + Former smoker** | 89.66% | 20.69% |
| | | | | | | | **Smoker** | 10.34% | 79.31% |
| Alcohol | 1-CpG model | SLC7A11 | 0.77 | 0.70 | 0.77 | 74.26% | **Predicted\Real** | **Non-drinker + Moderate drinker** | **Heavy drinker** |
| | | | | | | | **Non-drinker + Moderate drinker** | 76.58% | 30.38% |
| | | | | | | | **Heavy drinker** | 23.42% | 69.62% |
| | 2-CpGs model | SLC7A11, cg00886875 | 0.78 | 0.75 | 0.70 | 71.31% | **Predicted\Real** | **Non-drinker + Moderate drinker** | **Heavy drinker** |
| | | | | | | | **Non-drinker + Moderate drinker** | 69.62% | 25.32% |
| | | | | | | | **Heavy drinker** | 30.38% | 74.68% |
| | **3-CpGs model** | **SLC7A11, cg00886875, MIR4435-2HG** | **0.80** | **0.81** | **0.71** | **74.26%** | **Predicted\Real** | **Non-drinker + Moderate drinker** | **Heavy drinker** |
| | | | | | | | **Non-drinker + Moderate drinker** | 70.89% | 18.99% |
| | | | | | | | **Heavy drinker** | 29.11% | 81.01% |
| | 4-CpGs model | SLC7A11, cg00886875, MIR4435-2HG, *GAS5* | 0.80 | 0.81 | 0.70 | 73.84% | **Predicted\Real** | **Non-drinker + Moderate drinker** | **Heavy drinker** |
| | | | | | | | **Non-drinker + Moderate drinker** | 70.25% | 18.99% |
| | | | | | | | **Heavy drinker** | 29.75% | 81.01% |

and cg07082267. The scatter plots of the corresponding age-correlated CpGs representing the DNA methylation values against the chronological age are shown in Supplementary Fig. S1. Quantile regression was performed using the five selected CpGs in order to generate the age prediction model, giving a median absolute error (MAE) of ± 2.8 years. To evaluate the accuracy of the model, a k-fold cross-validation was carried out, providing a MAE of ± 2.79 years.

In order to check if the studied lifestyles could have an influence on age prediction, the samples used for building the age prediction model were grouped by categories in the same way as in the final logistic regression models, and errors per cluster were calculated. For the samples categorized by tobacco intake, the predicted errors obtained for the two groups of interest were evaluated, resulting in a MAE of ± 2.78 years for the non-smokers + former smokers category and a MAE of ± 3.12 years for the smokers category (Fig. 4A). To assess whether there were significant differences between the residuals of the evaluated groups (Fig. 4B), a Wilcoxon test was applied and a p-value of 0.78 obtained. Hence, no significant differences in age predictions were observed for non-smokers or former smokers and smokers.

Regarding the alcohol intake, the predicted errors obtained for the groups of interest were also evaluated, yielding a MAE of ± 2.82 years for the non-drinker + moderate drinker category and a MAE of ± 2.59 years for the heavy drinker group (Fig. 4C). To determine if there were significant differences between the residuals of the groups studied (Fig. 4D), a Wilcoxon test was carried out and a p-value of 0.10 obtained. Therefore, no significant differences were observed between the predicted ages of non-drinkers or moderate drinkers and heavy drinkers.

## 4. Discussion

DNA methylation has become an increasingly studied biomarker of interest in forensic genetics. In recent years, the different applications of this epigenetic marker have been evaluated, highlighting the prediction of age [1,2], tissue identification [3], and the study of lifestyles [11,12]. Tobacco and alcohol consumption have created the most interest, and several discovery articles have been published identifying markers that show differences between consumers and non-consumers [21,22,26,27]. As these behaviors are related to disease risk/status, the methylation levels in certain genes correlated with medical conditions such as cancer have been evaluated, and many show differences between smokers or drinkers, and non-consumers [8,9]. At the same time, associations have been observed between dependency and DNA methylation, therefore, the generation of predisposition indices could be useful for preventative diagnostics [13,14]. From these studies, tools of forensic interest have been generated to identify whether a person is a smoker or non-smoker and a drinker or non-drinker [31,32,36] - information that could be

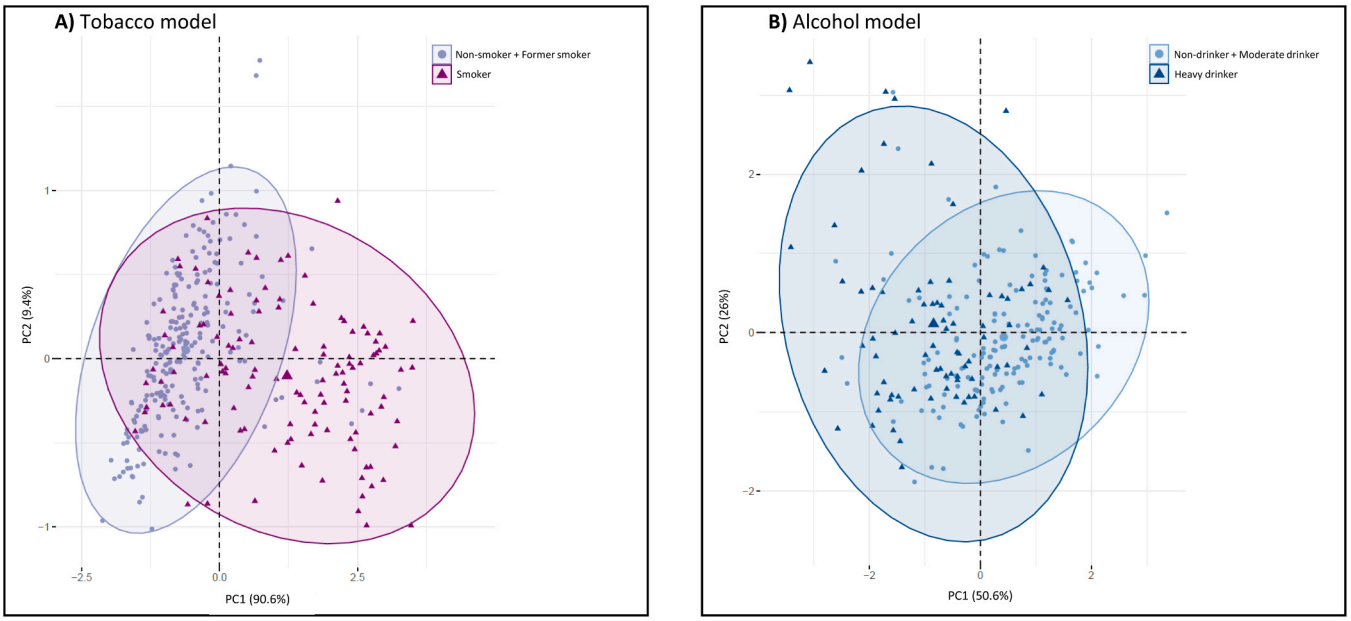



**Fig. 3.** PCA representation of the selected models for each lifestyle. **A)** Tobacco binomial logistic regression model composed of *AHRR* and cg01940273 representing 232 individuals classified as non-smoker + former smoker as well as 116 smokers. **B)** Alcohol binomial logistic regression model composed of *SLC7A11,* cg0886875 and *MIR4435–2HG* representing 158 individuals classified as non-drinker + moderate drinker as well as 79 heavy drinkers.

relevant in criminal investigations. Finally, the relationship between these lifestyle habits and age acceleration has been evaluated, causing epigenetic modifications of biological age to differ more from the chronological age than under normal conditions [39–41]. Considering both clinical and forensic applications, markers correlated with smoking and drinking status were selected to develop predictive models for these lifestyles.

In the present study, a total of 15 markers were selected, 10 CpGs for smoking and 5 CpGs for alcohol consumption. Regarding the underlying DNA methylation data, it is important to mention that inter-individual variation within each smoking or alcohol category was observed, with the most marked variation found in consumers, i.e. former and current smokers, as well as moderate and heavy drinkers. Similar findings have been previously reported by Vidaki et al. [54], suggesting a potential consumer-behaviour effect in terms of intensity and duration.

To identify the most accurate consumption status prediction models among the 10 tobacco-related CpGs and the 5 alcohol-related CpGs selected in our study, a forward approach was explored using logistic regression. Considering the three categories defined for each lifestyle (non-consumer, intermediate consumer, consumer), multinomial and binomial models were evaluated to represent the different consumption states analyzed. As Maas et al. [36] mentioned in reference to overfitting, for all the models generated in our study, the recommendations for the amount of predictors per number of participants were taken into account. To avoid overfitting, the generation of the models was limited to a maximum number of CpGs depending on whether the analysis was multinomial or binomial [45,53]. Firstly, the multinomial logistic regression models were evaluated, presenting percentages of correct classifications of 66.09% for the 4-CpG tobacco model and 52.74% for the 3-CpG alcohol model. As observed in other published models, the intermediate groups show lower accuracies than the extreme ones. In the case of tobacco consumption, Alghanim et al. [29] developed a 4-CpG multinomial logistic regression model that gave high percentages of correct classifications for non-smokers (84.9%) and smokers (90%), although these percentages were reduced to 66.7% for former smokers. At the same time, the model of 13-CpGs published by Maas et al. [30] shows sensitivity and specificity values (sensitivity: 0.78 for non-smokers, 0.65 for former smokers and 0.67 for smokers; specificity: 0.75 for non-smokers, 0.77 for former smokers and 0.99 for smokers) similar to the 4-CpG multinomial model of our study (sensitivity: 0.73 for non-smokers, 0.53 for former smokers and 0.72 for smokers; specificity: 0.77 for non-smokers, 0.78 for former smokers and 0.94 for smokers). The sensitivity for former smokers, which evaluates the degree of correct prediction of this category, presents values lower than 0.7 in both the Maas et al. model (0.65) and from our study (0.53). For the alcohol model developed at the present study, the percentages of correct classifications were lower (63.29% for non-drinkers, 27.85% for moderate drinkers and 65.82% for heavy drinkers), with the intermediate group being the most challenging to be classified (0.28 sensitivity).

Considering the difficulty observed in predicting the three groups separately, grouping of the intermediate category with one of the extreme groups was evaluated. For tobacco, grouping of non-smokers and former smokers was previously considered by the models of Elliott et al. [28] and Maas et al. [30], obtaining high values for the statistical parameters evaluated in those models. Moreover, in the 1-CpG binomial logistic regression models developed by Alghanim et al. [29], a reduction in AUC was observed for non-smokers vs former smokers compared to smokers vs former smokers in blood (mean AUC 0.73 and 0.91, respectively) and in saliva (mean AUC 0.69 and 0.81, respectively). Thus, former smokers were differentiated better from smokers than from non-smokers. In the case of drinking status, Maas et al. [36] evaluated different clustering of categories in the models analyzed, with the highest AUCs obtained for the models comparing heavy drinkers vs non-drinkers and light drinkers (AUC > 0.7 for the different CpG combinations evaluated). Evaluation of the grouping of consumers (heavy + moderate) vs non-consumers was evaluated by Chamberlain et al. [32], obtaining in their model an AUC of 0.64. With all this, it could be concluded that, if a multinomial model for these lifestyles does not correctly predict the three categories, it would be advisable to evaluate the grouping of the intermediate category with the non-consumers.

Although the grouping of two categories (non-consumers and intermediate consumption) could be considered a disadvantage, in a forensic context, the value of a test could lie in identifying the group that is less frequent in the general population. Bearing this in mind, according to Eurostat data, only 19.7% of the EU population smokes daily and only 8.4% drinks daily. In the case of alcohol, almost one in five individuals are considered heavy drinkers, showing differences related either to the gender (26.3% for men and 11.4% for women), and to the country of

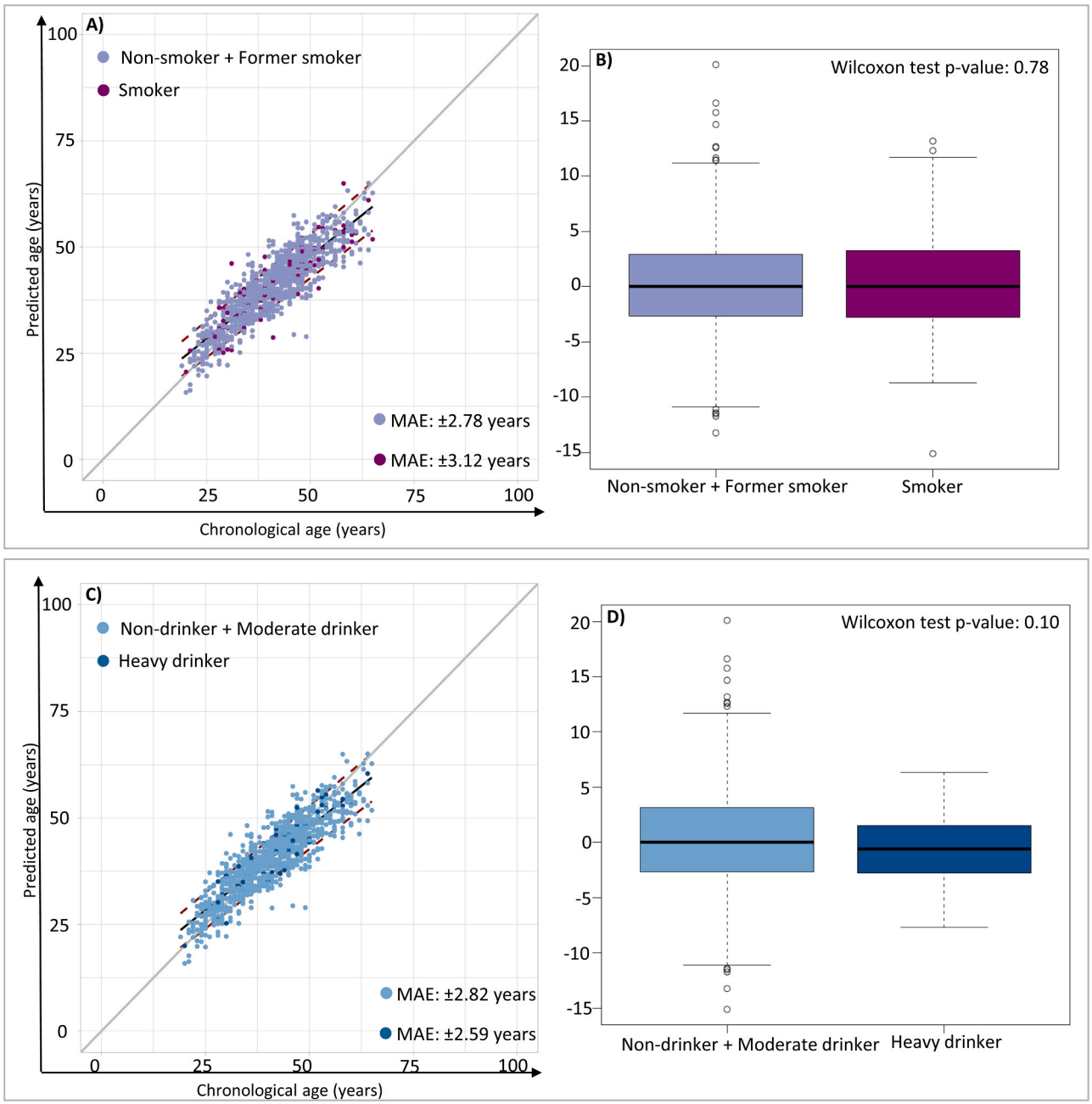



**Fig. 4.** Graphical representations derived from the age prediction quantile regression model generated with 1112 blood samples previously used for building lifestyle prediction models. **A)** and **C**) Predicted age vs chronological age of the sample set grouped by tobacco and alcohol categories, respectively. The black diagonal line represents the 0.5 quantile and the discontinuous dark red lines the corresponding 0.1 and 0.9 quantiles. The gray line represents perfect correlation. **B)** and **D)** Residuals obtained in the model generated for the defined categories for tobacco and alcohol consumption, respectively.

origin, showing Denmark and Romania the highest percentage of prevalence of heavy episodic drinking at least once a month among alcohol drinkers (37.8% and 35.0%, respectively). Therefore, correctly identifying individuals who might fall into these categories would further reduce the number of suspects in a criminal case.

For tobacco consumption, a 2-CpG binomial logistic regression model confronting non-smokers + former smokers vs smokers composed of *AHRR* (cg05575921) and cg01940273 gave performance metrics of: AUC: 0.87, sensitivity: 0.79, specificity: 0.90 and %CC: 86.49%. To validate the model using an independent testing set, since a larger pool of samples was not available, a complete external validation was not performed, evaluating only individuals classified as non-smoker (N = 606) and former smoker (N = 151). This allowed evaluation of the efficiency of the model for the pooled group, with 93.39% of the validation set samples correctly assigned. Although these results demonstrate that a correct classification of the non-smokers + former smokers was obtained, further analysis of the models would be necessary to fully evaluate a larger validation set covering all three evaluated categories.

Most published models make a partial evaluation of the categories, evaluating two of three possible groups. Only those models built with the extreme categories (non-smokers vs smokers) have given AUC values close to or above 0.90 [32,55–57]. However, biased results could be generated, since former smokers are not represented.

A very comprehensive model is described by Maas et al. [30], reporting a binomial logistic model (grouping non-smoker and former smoker in one category) giving an AUC of 0.90 and a multinomial model

with AUCs of 0.84, 0.77 and 0.93 for non-smoker, former smokers and smokers, respectively. Our study has certain similarities with Maas et al., but the different maker selection criteria and differences in model construction (backward and forward approaches) led to the generation of very different prediction models. The model developed in our study, built with 2 CpGs, that are among the 13 selected by Maas et al., gave similar results with a lower number of markers (AUC: 0.87). It is worth mentioning that there is a large difference in the sensitivity of both models, with that of Maas et al. predicting smokers with a value of 0.59 compared to our model with 0.79. A possible justification for the high AUC and correct classifications obtained by Maas et al. despite the low sensitivity could be a large imbalance in the number of individuals among the groups in the model. The Maas et al. multinomial model presented difficulties to predict former smokers, obtaining sensitivities of 0.65 for the model and 0.39 for the validation set. The prediction of the intermediate category has been more challenging, with worse predictions being observed in the models that evaluate it independently (Shenker et al. [58] AUC: 0.82, sensitivity: 0.69 and specificity: 0.71; Maas et al. [30] AUC: 0.77, sensitivity: 0.65 and specificity: 0.77). Considering the reversibility of DNA methylation for some of the markers correlated with tobacco and the effect on this reversibility of time since cessation of smoking and intensity of consumption demonstrated by McCartney et al. [33], the group of former smokers is a difficult category to classify consistently. Moreover, as shown in Fig. 3A, and in the study of McCartney et al., it is difficult to obtain a complete separation of the categories analyzed, as there are several factors that can modify the methylation patterns associated with this lifestyle. More accurate prediction of the intermediate group is likely to require development of: a larger number of markers; specific markers for former smokers (e.g., positions that exhibit irreversible methylation patterns over time); or complementing the generated models with consumption cessation time models.

The selected tobacco markers for the logistic regression model for predicting smoking status have been previously reported. The *AHRR* gene is a protein-coding gene related to cell growth and differentiation. The CpG position selected in this study, cg05575921, has been reported on multiple occasions as one of the markers that shows the highest correlation with tobacco consumption [25–27,59] and is present in almost all smoking status prediction models published to date [28, 30–32,55–57,60]. Some of these obtained individual AUCs for this marker of 0.88 [30] and 0.99 [31], similar results to our 1-CpG model generated with this marker (non-smokers vs smokers of 0.90, and non-smokers + former smokers vs smokers of 0.87). The *AHRR* marker is of interest for not presenting differences between European and South Asian populations [28], for recovering methylation values to a non-smoking state after 5 years of consumption cessation [17], for being correlated with age acceleration [9] and for being correlated with different mortality factors [14] - representing as it does a biomarker of lung cancer [10]. Position cg01940273 has also been previously reported to correlate with smoking status [25–27]. This position is present in the 13 CpG model of Maas et al. [30] presenting individual AUCs of 0.89, similar to those of our 1-CpG model for non-smokers vs smokers (0.89). Maas's study also evaluated the time since smoking cessation, a characteristic with which this marker had been related in other publications [14]. Finally, it has been observed that cg01940273 is related to breast cancer risk [22,23].

Evaluating the markers correlated with alcohol consumption status, a 3-CpG prediction model for non-drinkers + moderate drinkers vs heavy drinkers composed of *SLC7A11* (cg06690548), cg0886875 and *MIR4435–2HG* (cg21294714) gave an AUC of 0.80, sensitivity of 0.81, specificity of 0.71 and %CC of 74.26%. Different approaches to clustering the categories in binomial models have been evaluated in the published studies for alcohol. Those models composed only of non-drinkers vs heavy drinkers generally have higher values, presenting AUCs around 0.80 [32,34,36]. Of the published models, only Liu et al. [34] and Maas et al. [36] attempt to assess all possible categories for alcohol consumption. Focusing on Maas's model, AUCs in a range of 0.70–0.75 were obtained for heavy drinkers vs non-drinkers + light drinkers, lower than those obtained with our model with similar grouping (0.80). Considering the classification methods, the model for heavy + at risk drinkers vs non-drinkers + light drinkers could also be compared with our study. The classification of alcohol consumption presents a greater challenge than tobacco consumption, with consistently lower AUC and %CC values, as well as a smaller separation between the defined categories, as can be seen in Fig. 3B. This might be due to a smaller separation in the methylation values of the individuals classified in the different categories evaluated, with the intermediate group overlapping with the two extreme categories to a greater extent (Fig. 2). This could also be observed in the classification tables of the multinomial models generated, with the predictions of the intermediate group more evenly distributed than former smokers in the tobacco models. Therefore, more studies are needed to obtain a better separation of the intermediate group for which the introduction of other variables such as time or intensity of consumption could be useful.

Of the markers correlated with alcohol consumption used in the selected model, *SLC7A11*, has been reported in lifestyle-related discovery studies. For the other selected CpGs, to the best of our knowledge, our study detected these CpG positions to be correlated with alcohol consumption for the first time. The *SLC7A11* gene, is the most reported marker related to alcohol consumption, its correlation has been observed in different discovery studies [21,22] and it is among the selected CpGs for all the previous published drinking status models. Moreover, this marker has been associated with the number of drinks consumed per week, and may be a marker of interest to identify heavy or at-risk drinkers [22]. Both cg21294714 of the *MIR4435–2HG* gene and cg0886875, were identified to be correlated with alcohol consumption in this study. These CpG positions presented individually, for the extreme categories (non-drinkers vs heavy drinkers), gave AUC values of 0.70 and 0.71 respectively.

One limitation in our study is the absence of a complete independent sample set for both smoking and alcohol consumption. Additionally, both CpG selection and model building were performed using identical samples and this could partially bias our findings. However, the resulting selected and most informative CpG sites are consistent with previous studies, which adds value to our results. For those CpG sites reported in the present study as correlated with alcohol for the first time will, additional studies will be necessary for validation purposes.

Different publications have shown that tobacco and alcohol consumption influence age acceleration, causing discordance between chronological and biological age data. When developing age prediction models based on DNA methylation, it is often not possible to check whether the selected markers are influenced by other factors. Indeed, the present study has been developed using samples from the UK AIRWAVE study, where the volunteers were police officers. Since this occupational group is exposed to high levels of stress, a potential cofounding effect cannot be discarded in our analyses. Considering that these lifestyle factors produce global alterations in DNA methylation values, an age prediction model based on quantile regression was generated to evaluate the effect of the consumption of these substances in age prediction. A MAE of ± 2.79 years was obtained. It should be noted that in the study of Freire-Aradas et al. [49], the MAE obtained was ± 3.07. The difference in the evaluated parameter is probably due to the range of age analyzed. While Freire-Aradas's model covers 18 to 104 years, the dataset used in the current study covers 19 to 65 years old. The Wilcoxon test indicated no significant differences (tobacco p-value of 0.78; alcohol p-value of 0.10) between the residuals generated by the model for the groups analyzed in each lifestyle. Although an effect of smoking and alcohol consumption on age prediction was not observed, this only means that the markers included in the age prediction model developed on the present study are not sensitive to these lifestyle factors. This cannot be taken to mean that smoking and alcohol do not have an effect on aging in general.

Additionally, in accordance with Vidaki et al., 2023 [54], signs of association with age were found for the majority of smoking-CpGs depicting methylation reduction through the lifespan, although all $r_s$ values were below or close to |0.50|. From these, the highest values were observed for the smoker group in our study. Similar findings were obtained for the alcohol-CpGs, although in this case, no specific group was detected to present higher correlations in comparison to the others. Although the correlation with age is low in scale, it has been detected and therefore, the age predictive potential of the selected CpGs should be further studied.

## Declaration of Competing Interest

The authors have declared no conflict of interest.

## Acknowledgements

MVL is supported by the Ministerio de Educación, Cultura y Ciencia, Spain (PID2019-107876RB-I00). MdlP is supported by a postdoctoral fellowship awarded by the Gobierno de España: IJC2020-042638-I, funded by MCIN/AEI/10.13039/501100011033 and the European Union "NextGenerationEU/PRTR". J.R. is supported by the "Programa de axudas á etapa predoutoral" funded by the Consellería de Cultura, Educación e Ordenación Universitaria e da Consellería de Economía, Emprego e Industria from Xunta de Galicia, Spain (ED481A-2020/039). The Airwave Health Monitoring Study is funded by the Medical Research Council (MRC), (MR/R023484/1), the National Institute for Health Care Research (NIHR) Health Protection Research Unit in Chemical and Radiation Threats and Hazards (NIHR-200922), the Imperial College Biomedical Research Centre (BRC) 2017–22, and the Imperial College Healthcare NHS Trust. The initial phase of the study, including participant recruitment, was funded by the Home Office (780-TETRA; 2003-18). Views expressed are those of the authors and not necessarily those of the study sponsors. We thank all study participants for their involvement. DPUK provided data access for this project: Elliott, P. (2017). Airwave [Data set]. Dementias Platform UK. https://doi.org/10.48532/002000 through MRC grant ref MR/L023784/2″ (core funding).

## Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.fsigen.2024.103022.

## References

[1] K. Schwender, O. Holländer, S. Klopfleisch, M. Eveslage, M.F. Danzer, H. Pfeiffer, et al., Development of two age estimation models for buccal swab samples based on 3 CpG sites analyzed with pyrosequencing and minisequencing, Forensic Sci. Int Genet 53 (2021) 102521.
[2] A. Wózniak, A. Heidegger, D. Piniewska-Róg, E. Pospiech, C. Xavier, A. Pisarek, et al., Development of the VISAGE enhanced tool and statistical models for epigenetic age estimation in blood, buccal cells and bones, Aging (Albany NY) 13 (5) (2021) 6459–6484.
[3] H.Y. Lee, S.E. Jung, E.H. Lee, W.I. Yang, K.J. Shin, DNA methylation profiling for a confirmatory test for blood, saliva, semen, vaginal fluid and menstrual blood, Forensic Sci. Int Genet 24 (2016) 75–82.
[4] A. Vidaki, C. Díez López, E. Carnero-Montoro, A. Ralf, K. Ward, T. Spector, et al., Epigenetic discrimination of identical twins from blood under the forensic scenario, Forensic Sci. Int Genet 31 (2017) 67–80.
[5] A.V. Probst, E. Dunleavy, G. Almouzni, Epigenetic inheritance during the cell cycle, Nat. Rev. Mol. Cell Biol. 10 (3) (2009) 192–206.
[6] C.B. Santos-Rebouças, M.M.G. Pimentel, Implication of abnormal epigenetic patterns for human diseases, Eur. J. Hum. Genet 15 (1) (2007) 10–17.
[7] K.M. Bakulski, M.D. Fallin, Epigenetic epidemiology: promises for public health research, Environ. Mol. Mutagen 55 (3) (2014) 171–183.
[8] M. Varela-Rey, A. Woodhoo, M.L. Martinez-Chantar, J.M. Mato, S.C. Lu, Alcohol, DNA methylation, and cancer, Alcohol Res Curr. Rev. 35 (1) (2012) 25–35.
[9] X. Gao, Y. Zhang, L.P. Breitling, H. Brenner, Tobacco smoking and methylation of genes related to lung cancer development, Oncotarget 7 (37) (2016) 59017–59028.
[10] D. Fragou, E. Pakkidi, M. Aschner, V. Samanidou, L. Kovatsi, Smoking and DNA methylation: Correlation of methylation with smoking behavior and association with diseases and fetus development following prenatal exposure, Food Chem. Toxicol. 129 (2019) 312–327.
[11] H.Y. Lee, S.D. Lee, K.J. Shin, Forensic DNA methylation profiling from evidence material for investigative leads, BMB Rep. 49 (7) (2016) 359–369.
[12] A. Vidaki, M. Kayser, From forensic epigenetics to forensic epigenomics: broadening DNA investigative intelligence, Genome Biol. 18 (1) (2017) 1–13.
[13] A.E. Teschendorff, Z. Yang, A. Wong, C.P. Pipinikas, Y. Jiao, A. Jones, et al., Correlation of smoking-associated DNA methylation changes in buccal cells with DNA methylation changes in epithelial cancer, JAMA Oncol. 1 (4) (2015) 476–485.
[14] Y. Zhang, B. Schöttker, I. Florath, C. Stock, K. Butterbach, B. Holleczek, et al., Smoking-associated DNA methylation biomarkers and their predictive value for all-cause and cardiovascular mortality, Environ. Health Perspect. 124 (1) (2016) 67–74.
[15] R. Zhang, Q. Miao, C. Wang, R. Zhao, W. Li, C.N. Haile, et al., Genome-wide DNA methylation analysis in alcohol dependence, Addict. Biol. 18 (2) (2013) 392–403.
[16] L.P. Breitling, R. Yang, B. Korn, B. Burwinkel, H. Brenner, Tobacco-smoking-related differential DNA methylation: 27K discovery and replication, Am. J. Hum. Genet 88 (4) (2011) 450–457.
[17] S. Zeilinger, B. Kühnel, N. Klopp, H. Baurecht, A. Kleinschmidt, C. Gieger, et al., Tobacco smoking leads to extensive genome-wide changes in DNA methylation, PLoS One 8 (5) (2013) e63812.
[18] D. Bönsch, B. Lenz, U. Reulbach, J. Kornhuber, S. Bleich, Homocysteine associated genomic DNA hypermethylation in patients with chronic alcoholism, J. Neural Transm. 111 (12) (2004) 1611–1616.
[19] Alcohol abuse and cigarette smoking are associated with global DNA hypermethylation: results from the German Investigation on Neurobiology in Alcoholism ( GINA). 2015;49:97–101.
[20] R.A. Philibert, B. Penaluna, T. White, S. Shires, T. Gunter, J. Liesveld, et al., A pilot examination of the genome-wide DNA methylation signatures of subjects entering and exiting short-term alcohol dependence treatment programs, Epigenetics 9 (9) (2014) 1212–1219.
[21] P.A. Dugué, X. Wang, L. Baglietto, R. Wilson, B. Lehne, E. Makalic, et al., Alcohol consumption is associated with widespread changes in blood DNA methylation: Analysis of cross - sectional and longitudinal data, Addict. Biol. 26 (1) (2021) e12855.
[22] L.E. Wilson, Z. Xu, S. Harlid, A.J. White, M.A. Troester, D.P. Sandler, et al., Alcohol and DNA methylation: an epigenome-wide association study in blood and normal breast tissue, Am. J. Epidemiol. 188 (6) (2019) 1055–1065.
[23] R. Zhao, R. Zhang, W. Li, Y. Liao, J. Tang, Q. Miao, et al., Genome-wide DNA methylation patterns in discordant sib pairs with alcohol dependence, Asia -Pac. Psychiatry 5 (1) (2013) 39–50.
[24] L. Tsaprouni, T. Yang, J. Bell, K. Dick, S. Kanoni, J. Nisbet, et al., Cigarette smoking reduces DNA methylation levels at multiple genomic loci but the effect is partially reversible upon cessation, Epigenetics 9 (10) (2014) 1382–1396.
[25] F. Guida, T.M. Sandanger, R. Castagné, G. Campanella, S. Polidoro, D. Palli, et al., Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation, Hum. Mol. Genet 24 (8) (2015) 2349–2359.
[26] S. Ambatipudi, C. Cuenin, H. Hernandez-Vargas, A. Ghantous, F. Le Calvez-Kelm, R. Kaaks, et al., Tobacco smoking-associated genome-wide DNA methylation changes in the EPIC study, Epigenomics 8 (5) (2016) 599–618.
[27] P. Dugué, C. Jung, J.E. Joo, X. Wang, E. Ming, E. Makalic, et al., Smoking and blood DNA methylation: an epigenome-wide association study and assessment of reversibility, Epigenetics 15 (4) (2020) 358–368.
[28] H.R. Elliott, T. Tillin, W.L. McArdle, K. Ho, A. Duggirala, T.M. Frayling, et al., Differences in smoking associated DNA methylation patterns in South Asians and Europeans, Clin. Epigenetics 6 (1) (2014) 4.
[29] H. Alghanim, W. Wu, B. Mccord, DNA methylation assay based on pyrosequencing for determination of smoking status, Electrophoresis 39 (21) (2018) 2806–2814.
[30] S.C.E. Maas, A. Vidaki, R. Wilson, A. Teumer, F. Liu, J.B.J.Van Meurs, Validated inference of smoking habits from blood with a finite DNA methylation marker set, Eur. J. Epidemiol. 34 (11) (2019) 1055–1074.
[31] R. Philibert, J.A. Mills, M. Dogan, S.R.H. Beach, J.D. Long, AHRR methylation predicts smoking status and smoking intensity in both saliva and blood DNA, Am. J. Med Genet Part B Neuropsychiatr. Genet 183 (1) (2020) 51–60.
[32] J.D. Chamberlain, S. Nusslé, L. Chapatte, C. Kinnaer, D. Petrovic, S. Pradervand, et al., Blood DNA methylation signatures of lifestyle exposures: tobacco and alcohol consumption, Clin. Epigenetics 14 (1) (2022) 155.
[33] D.L. McCartney, A.J. Stevenson, R.F. Hillary, R.M. Walker, M.L. Bermingham, S. W. Morris, et al., Epigenetic signatures of starting and stopping smoking, EBioMedicine 37 (2018) 214–220.
[34] C. Liu, R.E. Marioni, A.K. Hedman, L. Pfeiffer, P.C. Tsai, L.M. Reynolds, et al., A DNA methylation biomarker of alcohol consumption, Mol. Psychiatry 23 (2) (2018) 422–433.
[35] D.L. McCartney, R.F. Hillary, A.J. Stevenson, S.J. Ritchie, R.M. Walker, Q. Zhang, et al., Epigenetic prediction of complex traits and death, Genome Biol. 19 (1) (2018) 136.
[36] S.C.E. Maas, A. Vidaki, A. Teumer, R. Costeira, R. Wilson, J. van Dongen, et al., Validating biomarkers and models for epigenetic inference of alcohol consumption from blood, Clin. Epigenetics 13 (1) (2021) 198.
[37] M. Hattab, S. Clark, E. van den Oord, Overstimation of the classification accuracy of a biomarker for assessing heavy alcohol use. Mol. Psychiatry 23 (11) (2018) 2114–2115.
[38] P.D. Yousefi, R. Richmond, R. Langdon, A. Ness, C. Liu, D. Levy, et al., Validation and characterisation of a DNA methylation alcohol biomarker across the life course, Clin. Epigenetics 11 (1) (2019) 163.

[39] X. Gao, Y. Zhang, L.P. Breitling, H. Brenner, Relationship of tobacco smoking and smoking-related DNA methylation with epigenetic age acceleration, Oncotarget 7 (30) (2016) 46878–46889.
[40] M. Stephenson, S. Bollepalli, E. Cazaly, J.E. Salvatore, W.F. Street, Associations of alcohol consumption with epigenome-wide DNA methylation and epigenetic age acceleration: individual-level and co-twin comparison analyses, Alcohol Clin. Exp. Res 45 (2) (2022) 318–328.
[41] J.K. Kresovich, A.M.M. Lopez, E.L. Garval, Z. Xu, A.J. White, P. Dale, et al., Alcohol consumption and methylation-based measures of biological age, J. Gerontol. Ser. A Biol. Sci. Med Sci. 76 (12) (2021) 2107–2111.
[42] Elliott P. Airwave [Data set]. Dementias Platform UK. [Internet]. 2017. Available from: ⟨https://doi.org/10.48532/002000⟩.
[43] P. Elliott, A.C. Vergnaud, D. Singh, D. Neasham, J. Spear, A. Heard, The airwave health monitoring study of police officers and staff in great britain: Rationale, design and methods, Environ. Res 134 (2014) 280–285.
[44] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. Sanchez, et al., pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinforma. 12 (2011) 77.
[45] D.W.J. Hosmer, S. Lemeshow, R.X. Sturdivant. Applied logistic regression, Third edit.,, John Wiley & Sons,, 2013.
[46] W. Venables, B. Ripley, Modern Applied Statistics with S, Springer New York,, Springer US, 2002.
[47] Kassambara A., Mundt F. Factoextra: Extract and visualize the results of multivariate data analyses [Internet]. 2020. Available from: ⟨https://cran.r-project.org/package=factoextra⟩.
[48] Koenker R., Portnoy S., Ng P., Zeileis A., Grosjean P., Ripley B. Package quantreg: Quantile Regression. 2015.
[49] A. Freire-Aradas, C. Phillips, A. Mosquera-Miguel, L. Girón-Santamaría, A. Gómez-Tato, M. Casares De Cal, et al., Development of a methylation marker set for forensic age estimation using analysis of public methylation data and the Agena Bioscience EpiTYPER system, Forensic Sci. Int Genet 24 (2016) 65–74.
[50] Alfons A. Package cvTools: Cross-validation tools for regression models. 2015.
[51] Wickham H., Chang W. Package ggplot2: An implementation of the grammar of graphics. 2015.
[52] R. Team, R. Core, A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2020.
[53] P. Peduzzi, J. Concato, E. Kemper, T.R. Holford, A.R. Feinstein, A simulation study of the number of events per variable in logistic regression analysis, J. Clin. Epidemiol. 49 (12) (1996) 1373–1379.
[54] A. Vidaki, B. Planterose Jiménez, B. Poggiali, V. Kalamara, K.J. van der Gaag, S.C. E. Maas, et al., Targeted DNA methylation analysis and prediction of smoking habits in blood based on massively parallel sequencing, Forensic Sci. Int Genet 65 (2023) 102878.
[55] R. Philibert, A quantitative epigenetic approach for the assessment of cigarette consumption, Front Psychol. 6 (2015) 656.
[56] Y. Zhang, I. Florath, K. Saum, H. Brenner, Self-reported smoking, serum cotinine, and blood DNA methylation, Environ. Res 146 (2016) 395–403.
[57] N. Kondratyev, A. Golov, M. Alfimova, T. Lezheiko, V. Golimbet, Prediction of smoking by multiplex bisulfite PCR with long amplicons considering allele-specific effects on DNA methylation, Clin. Epigenetics 10 (1) (2018) 130.
[58] N.S. Shenker, P.M. Ueland, S. Polidoro, K. Van Veldhoven, F. Ricceri, R. Brown, et al., DNA methylation as a long-term biomarker of exposure to tobacco smoke, Epidemiology 24 (5) (2013) 712–716.
[59] X. Gao, M. Jia, Y. Zhang, L.P. Breitling, H. Brenner, DNA methylation changes of whole blood cells in response to active smoking exposure in adults: A systematic review of DNA methylation studies, Clin. Epigenetics 7 (2015) 113.
[60] N.S. Shenker, S. Polidoro, K. van Veldhoven, C. Sacerdote, F. Ricceri, M.A. Birrel, et al., Epigenome-wide association study in European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking, Hum. Mol. Genet 22 (5) (2013) 843–851.