

Article

Use of Epigenetic Markers to Predict Age and Smoking Status in an Italian Population Sample

Domenico Colloca ¹, Matteo Manfredini ², Fabiano Gentile ¹, Alberto Marino ¹, Maria Carla Gerra ^{2,*} and Cristina Dallabona ^{2,*}

¹ Reparto Carabinieri Investigazioni Scientifiche di Parma—Sezione di Biologia, 43125 Parma, Italy

² Department of Chemistry, Life Sciences and Environmental Sustainability, University of Parma, Parco Area delle Scienze 11/A, 43124 Parma, Italy

* Correspondence: mariacarla.gerra@unipr.it (M.C.G.); cristina.dallabona@unipr.it (C.D.)

Abstract

Background/Objectives: DNA profiling in forensic investigation typically compares genetic profiles, usually derived from the analysis of STR markers. However, this method has limitations when there is no biological reference sample or match in the DNA database. The aim of the current study is thus to replicate, in an Italian cohort, epigenetic markers previously identified in the literature for distinguishing tobacco smokers from non-smokers or estimating chronological age, so as to help narrow down the pool of suspects. **Methods:** DNA methylation at four CpG dinucleotides located around the cg05575921 site of the *AHRR* gene, widely associated with tobacco consumption, was measured. Additionally, five CpG dinucleotides in the *ELOVL2*, *FHL2*, *KLF14*, *TRIM59*, and *C1orf132* genes were examined for chronological age estimation in buccal swab samples of 102 volunteers through pyrosequencing. **Results:** A multiple linear regression model for estimating chronological age shows that *ELOVL2*-C7, *C1orf132*-C1, and *TRIM59*-C7 have a significant effect on age. In this model, the prediction error increases with age. Two logistic regression models were used for determining smoker/non-smoker status, proving that two CpG sites significantly influence the odds of being classified as a smoker. When ex-smokers are included in the non-smoking group, the model correctly classifies the two conditions in about 80% of cases. **Conclusions:** The results demonstrate that the models generated from pyrosequencing data are useful for identifying tobacco smokers and estimating an individual's chronological age, particularly for younger subjects. Further studies are needed to develop models with higher predictive accuracy and to integrate these tools into regular forensic practice.

Keywords: DNA methylation; forensic investigations; smoking status; chronological age



Received: 24 December 2025

Revised: 29 January 2026

Accepted: 17 February 2026

Published: 19 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Multiple studies have demonstrated the significant role of environmental factors and behavior in altering epigenetic modifications without altering the DNA sequence. Diet, stress, lifestyle, drug use, and smoking have been found to influence the epigenome [1]. This could have great potential in forensic contexts.

Epigenetic modifications include DNA methylation, histone modifications, chromatin remodeling, and non-coding RNA [2]. In particular, DNA methylation (5mC), due to its stability and relative ease of measurement for monitoring exposure to various environments and risk factors, is considered a reliable biomarker [3]. It consists of the addition of a methyl

group (-CH₃) to cytosine residues at CpG dinucleotides by DNA methyltransferase (DNMT) enzymes [4–6]. CpG dinucleotides are not uniformly distributed throughout the genome, and regions rich in CpG dinucleotides are called CpG islands (CGIs). CGIs are often located in gene promoters and are typically unmethylated. Conversely, CGIs found in intra- or inter-genic regions may or may not be methylated [7].

Among the environmental factors affecting DNA methylation, smoking is considered one of the most impactful due to the diverse mechanisms it triggers [8]. Many studies have explored the relationship between smoking behaviors and DNA methylation changes. Epigenome-wide analyses have identified and confirmed specific sets of CpG sites in the *AHRR*, *CACNA1D*, *ARRB1*, *MYO1G*, *GFI1*, *F2RL3*, and *PDZD2* genes associated with cigarette smoking using monozygotic twin pairs [9] or large cohorts [8,10]. Methylation patterns at CpG sites of the *AHRR* gene have even been highlighted in cord blood samples from newborns of smoking mothers [11]. The important locus that emerges encodes the aryl hydrocarbon receptor repressor.

Moreover, DNA methylation has been more extensively explored in forensic settings for individuals' age prediction. One of the first studies investigated 34 pairs of identical twins and identified 88 sites showing a strong correlation between methylation and age. This analysis was also replicated in a sample of 31 men and 29 women ranging from 18 to 70 years old [12]. From these initial experiments, the construction of predictive models with significant age methylation correlations began. Various epigenome-wide association studies (EWASs) identified DNA methylation markers associated with epigenetic age; Florath and coworkers (2014) identified 155 significantly age-associated CpG sites mapping to 100 genes [13]. A biomarker based on 391 CpGs was developed to measure the age of different human tissue types accurately [14]. Based on these and other CpG sites, several models for predicting epigenetic age have been built, using, in particular, multiple linear regression models [15,16], quantile regression models, gradient boosting regression, ordinary least squares regression, and others [17].

For a more detailed overview of age-related DNA methylation changes and smoking-associated methylation signatures, comprehensive reviews are available, such as Horvath et al. (2018) and Joehanes et al. (2016) [18,19].

Although the epigenome-wide approach seems to be the best method to discover potential associated CpG sites, replication and validation studies of the results obtained through genome-wide approaches are fundamental. Moreover, in the forensic field, it would not be feasible to analyze a large number of sites, since the process almost always starts with trace evidence and minimal amounts of starting material. Another important aspect to consider is that, since epigenetic modifications are tissue-specific, the validity of the identified markers should be verified in the different tissues that can be detected. However, Jung and coworkers (2019) showed a correlation between specific DNA methylation sites and age in three sample types (blood, saliva, and buccal swab samples), suggesting multi-tissue age-predictive markers [20]. Interestingly, previously reported smoking-specific CpG sites close to cg05575921 (the *AHRR* gene) were investigated by Alghanim and coworkers (2018), and the CpG sites were found to be significantly hypomethylated in blood and saliva of current smokers [21]. As for age prediction, validation studies on candidate genes have confirmed some previously identified markers. In particular, five significantly correlated CpG sites in the *ELOVL2*, *C1orf132*, *TRIM59*, *KLF14*, and *FHL2* genes emerged [16].

In light of these observations, this study aims to generate an age/smoking-predictive model in the Italian population for potential forensic applications using DNA methylation analysis. Specifically, the first objective of this study is to assess whether analyzing DNA methylation in buccal swabs of Italian population samples at previously identified CpG sites indicative of tobacco smoking [21] can narrow down potential suspects by iden-

tifying whether they are smokers or non-smokers. The second goal is to estimate the chronological age of an unknown individual by measuring the percentage of methylation at specific CpG sites, which are known chronological age markers [16]. To assess how habitual tobacco smoking affects the prediction of chronological age, smoking status will also be considered within a specific age range, and the associated error rate will be determined. Finally, considering the available techniques, DNA methylation analysis using bisulfite conversion followed by sequencing is a specific and fast technique, measuring the 5mC level of each CpG dinucleotide with a high degree of accuracy. It is considered the gold standard for 5mC analysis. This study aims to investigate whether the evidence obtained through the simple and rapid technique of bisulfite conversion, followed by quantitative pyrosequencing, can provide reliable and informative results that may be useful for forensic applications. The decision to use pyrosequencing is based on its ability to deliver results quickly, which is crucial in forensic analysis, where timely data is essential for investigations. Moreover, the technique is cost-effective, has simple procedures, and requires user-friendly equipment. In the future, the data collected may be validated by comparison with trace DNA.

2. Materials and Methods

2.1. Participants

One hundred two (102) unrelated Caucasian subjects (73.4% males, aged 20–72 years, mean age 44.5 ± 13.1 years) were recruited for this study. This study was performed in line with the principles of the Declaration of Helsinki, and written informed consent was obtained from all participants included in the study. Approval was granted by the Local Ethics Committee (Comitato Etico Area Vasta Emilia Nord, Italy) (PROT. n. 20181, 12 May 2023), and the approved procedures and protocols were strictly followed. The subjects were not paid for their participation and were accepted as volunteers. The sample in the present study included cigarette smokers and non-smokers aged 18 years or older. Smokers were defined as people smoking at least one cigarette per day (88.5% of the participants smoked ≥ 5 cigarettes/day). Exclusion criteria were being under 18 years old and exhibiting overt pathologies at the time of sampling. All the participants were interviewed about their demographic data.

2.2. Sample Collection, DNA Extraction and DNA Quantification

Two buccal swabs were collected from each participant and immediately subjected to the pretreatment of samples using the Investigator Lyse&Spin Basket Kit (Qiagen) and labeled with a unique QR code. Subsequently, an automated purification procedure with the EZ1 Advanced XL (Qiagen Inc., Valencia, CA, USA) was used for DNA extraction. The Applied Biosystems Quantifiler[®] Trio DNA Quantification Kit (Life Technologies[™], Carlsbad, CA, USA) was applied to accurately estimate the quantity of DNA, even when the sample concentration was as low as 5 pg/ μ L or less [22]. The samples were then stored at -20 °C until use.

2.3. DNA Methylation Analysis in Nine CpG Sites at Six Genomic Loci for Age and Smoking Status Prediction

The EpiTect Fast DNA Bisulfite Conversion Kit (Qiagen) was applied to convert 102 DNA samples (200 ng) with sodium bisulfite. The procedure includes sample preparation, bisulfite-mediated conversion of unmethylated cytosines into uracil, purification of the converted single-stranded DNA using MinElute DNA spin columns, and DNA desulfonation, followed by washes. Finally, the converted DNA samples were eluted in 15 μ L.

After the bisulfite conversion phase, the samples were amplified using the PyroMark PCR kit (Qiagen), which allowed highly specific amplification of the DNA regions of interest: the AgePlex markers (BioVectis, Warsaw, Poland) (CpG sites located in the *ELOVL2*, *C1orf132*, *TRIM59*, *KLF14*, *FHL2* genes) were amplified to identify chronological age [16]. Four consecutive CpG sites located in the area around the cg05575921 site of the aryl hydrocarbon receptor repressor (*AHRR*) gene (CpG1, Chr5:373,476; CpG2, Chr5:373,490; CpG3, Chr5:373,494 and CpG4, Chr5:373,529) were considered for smoking status prediction [21]. Candidate CpG sites and genome localization are reported in Table 1.

Table 1. Candidate CpG sites and genome localization.

CpG Site	Gene	Genome Location (GRCh38)	Sequence to Analyze	Authors (PMID)
CpG1 CpG2 CpG3 CpG4	<i>AHRR</i>	Chr5:373,476 Chr5:373,490 Chr5:373,494 Chr5:373,529	TTGYGGGATTAGTAGG TYGGGYGGTGGTTGGGAGG TAGGGGATAGGTTGGTTT TTGYGGTTTT GGAGGTATAGAGTTTTTTT	Alghanim et al. 2018 [21] (29956353)
C7	<i>ELOVL2</i>	Chr6: 11044661	CCRTAAACRTTAAACC RCCRCRCRAAACCRCAC	
C1	<i>C1orf132</i>	Chr1: 207823681	AAATCTACRCA AACRACRATAAATAATCC	Zbieć-Piekarska et al., 2015 [16] (26026729)
C7	<i>TRIM59</i>	Chr3: 160450199	GGTTTGGYGYGGGAYGAGG YGAAGYGTYGGT GGTYGAYGGTTT	
C1	<i>KLF14</i>	Chr7: 130734355	TTGAGGAATTATTTTTTATTT TYGYGTTT- TTTTTTT TGTYGGYGAGTT	
C2	<i>FHL2</i>	Chr2: 105399288	AGGTAATGGTAATAGAG AGTTATYGGGAGY GTYGTTTTYGGYGTG GGTTTTYGGGYG YGAGTTTYGGAYGA GGTTTGGG	

Y—location of CpGs (pyrimidine C or T); R—location of CpGs (purine A or G).

Age prediction analysis was performed using PCR amplification of AgePlex methylation markers (BioVectis) [16]. Each of the five markers (*ELOVL2*, *C1orf132*, *TRIM59*, *KLF14*, *FHL2*) used by BioVectis to estimate chronological age must be amplified in separate amplification reactions. Therefore, five different PCR reactions were performed on each sample to determine chronological age. PCR reactions were carried out in a total volume of 25 µL, according to the manufacturer's instructions.

The amplification program consisted of an initial step to activate the DNA polymerase at 95 °C for 15 min, followed by 45 cycles comprising denaturation (94 °C for 30 s), annealing (56 °C for 30 s) and extension (72 °C for 30 s), and a final extension of 72 °C for 10 min. The Qiagen PyroMark PCR kit was also used to amplify the smoking status prediction sites; the amplification conditions were similar to the ones used for age estimation but using the primers reported by Alghanim and coworkers (2018) [21].

To exclude the formation of nonspecific bands or primer dimers, samples were subjected to 1.5% TBE agarose gel electrophoresis before pyrosequencing. Negative PCR controls were included in each PCR amplification. Pyrosequencing was performed using the PyroMark Q48 Autoprep instrument [23], following the manufacturer's instructions. The generated pyrograms were automatically analyzed using PyroMark Q48 Autoprep

Software v.2.4.2—Build 3 (Qiagen, Hilden, Germany). The results obtained using Pyromark Q48 Autoprep pyrosequencing are expressed as methylation percentages and described in the Results Section.

2.4. Statistical Analysis

A multiple linear regression model (defined as the Italian model) was estimated to analyze the relationship between the five AgePlex markers (ELOVL2-C7, C1orf132-C1, TRIM59-C7, KLF14-C1, FHL2-C2) and chronological age. Sex was introduced as a covariate. The models were estimated using a robust variance estimator to correct standard errors for issues of heteroscedasticity and clustering in the data [24]. Furthermore, multicollinearity issues were checked using the Variance Inflation Factor (VIF). Goodness-of-fit and accuracy were assessed to ensure reliable age estimation, including the calculation of the adjusted R^2 to determine the proportion of variation in the dependent variable associated with the explanatory variables once controlled for the number of predictors, and two different metrics to assess the performance of regression models, the Root Mean Square Error (RMSE) and the Mean Absolute Difference (MAD). RMSE, which is the most common metric for such analyses, represents the average magnitude of the prediction error of a regression model and it is calculated as

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N}}$$

where y_i is the observed value, $\hat{y}_i$ the predicted value and N the total number of observations. It is measured in the same units of the response variable. The squaring process penalizes outliers, making RMSE more useful when the analysis considers larger errors as particularly undesirable.

MAD is calculated as

$$MAD = \frac{\sum_{i=1}^N |y_i - \hat{y}_i|}{N}$$

It is more intuitive to interpret than RMSE, as it is the average of the absolute differences between observed and predicted values. It is also less sensitive to outliers, so it does not disproportionately penalize larger errors. A group of participants who had smoked in the past but no longer do (87.5% of ex-smokers had quit more than 10 years earlier) took part in the study. Ex-smokers were analyzed either as smokers or as non-smokers; thus, two models were developed based on different classification criteria. The first model (Model 1) defines smokers as subjects who reported smoking at least one cigarette per day (51.0%), while non-smokers were those who reported never having smoked or who had quit smoking (49.0%). In the second model (Model 2), ex-smokers were included in the smoking group (58.8%), which was contrasted with the never-smoking group (41.2%). In both models, age and sex of individuals were included as control variables.

As logistic models are estimated using an iterative procedure based on the maximum likelihood function, the choice of the best model among the two above described was made by means of the AIC and BIC criteria, which are among the best statistical tools for model selection in case of non-nested logistic models. The lower the value, the better the model fit.

Cross-validation techniques were ultimately applied to ensure evaluation of model performance, specifically of how effectively the results would generalize to an independent dataset [25–27]. For the linear regression model, the k-Fold and the Leave-One-Out (LOO) methods were used to estimate the prediction error of the model. Both are indicated for small sample sizes, and they differ in the number of training sets extracted from the sample. In the k-fold method, the sample is first divided into k subsamples (folds) of equal size,

after which the model is trained on $k - 1$ folds and tested on the remaining fold. This procedure is repeated k times, with each fold serving as the testing set only once. The model's performance is ultimately assessed by averaging the MSE values across all k iterations. The Leave-One-Out (LOO) technique can be viewed as an extreme generalization of the k -fold procedure where $k = n$. In other words, in LOO, the model is trained on $n - 1$ observations and tested on the single observation that is not included in the training set. This procedure is then repeated n times and the model's performance is assessed just like in the k -fold technique.

The k -fold technique was also applied on logistic models, for which it provides a cross-validation of the Area Under the Curve (AUC) estimates [28].

3. Results

3.1. Chronological Age Estimation

Overall, the model demonstrates a significant increase in model fit compared to the model with the sole intercept ($F = 37.5$, $p < 0.001$), and the residuals are normally distributed according to the Shapiro–Wilk test for normality ($W = 0.991$, $p = 0.736$). However, the model fails to meet the homoscedasticity assumption, as confirmed by the Breusch–Pagan test ($\chi^2 = 11.32$, $p = 0.001$). As a result, a robust standard error estimator was employed. Additionally, the VIF statistic does not indicate the presence of multicollinearity issues (mean VIF = 1.54). Table 2 reports the multiple linear regression model of age on AgePlex markers (ELOVL2-C7, C1orf132-C1, TRIM59-C7, KLF14-C1, FHL2-C2).

Table 2. Multiple linear regression of age on Ageplex markers (ELOVL2-C7, C1orf132-C1, TRIM59-C7, KLF14-C1, FHL2-C2) ^a. Regression coefficients (β) and standardized coefficients (β_{STAN}).

Marker	β	β_{STAN}	Robust SE	t	p-Value	95% CI
Sex (ref. F)	−0.340	−0.011	1.539	−0.22	0.825	−3.395–2.714
ELOVL2-C7	0.339	0.191	0.112	3.04	0.003	0.117–0.561
C1orf132-C1	−0.412	−0.279	0.111	−3.72	<0.001	−0.632–−0.192
TRIM59-C7	0.829	0.565	0.123	6.75	<0.001	0.585–1.073
KLF14-C1	0.363	0.117	0.231	1.57	0.119	−0.095–0.820
FHL2-C2	0.208	0.110	0.123	1.70	0.092	−0.035–0.452
N	102					
F (6, 95)	37.51					
R ² adjusted	0.695			$p < 0.001$		
RMSE	7.400					
MAD	5.661					

^a RSE = robust standard error; RMSE = Root Mean Square Error; MAD = Mean Average Deviation. Intercept (constant) not shown.

The set of explanatory variables accounts for about 70% of overall variability in age. More specifically, the results prove that only ELOVL2-C7, C1orf132-C1, and TRIM59-C7 have a significant effect on age. ELOVL2-C7 ($\beta = 0.339$) and TRIM59-C7 ($\beta = 0.829$) are positively associated, while C1orf132-C1 ($\beta = -0.412$) is negatively associated with age. The analysis of standardized coefficients reveals that TRIM59-C7 has the greatest impact on age estimation ($\beta_{\text{STAN}} = 0.565$). The two metrics of model performance and accuracy show that predictions have an RMSE of 7.4 years and a MAD of 5.7 years. Sex does not show statistical significance at $p < 0.05$, which is why sex-stratified models were not tested.

Once predicted values are disaggregated by chronological age group (<40, 40–54, 55+ years) (Table 3), the heterogeneity in MAD values becomes apparent. MAD values range from 4–5 years up to 54 years of age, but they increase significantly and become

more substantial from 55 years onwards, with a value of MAD between the predicted and observed values of 8.4 years. Therefore, predictions become less accurate as age increases, although overall model significance (F test) for the 55+ age group is slightly over the statistical threshold, possibly due to the lower number of subjects in that category. The performance of the model was then evaluated using cross-validation techniques. Table 3 shows, among other indicators, the MAD values obtained through the LOO procedure. The overall MAD is 6.1, only slightly higher than the non-validated value of 5.7. Moreover, as in the non-validated model, the 55+ age group shows lower prediction accuracy. The validated R^2 is 0.663.

Table 3. Age-stratified models. Overall statistical significance (F test) and Mean Absolute Difference by age group. Raw (MAD) and cross-validated (LOO) values.

Age Group	N	Mean Absolute Deviation			Overall Test	
		Raw	95% CI	Validated	F	p-Value
<40 years	41	5.4	4.0–6.7	5.7	9.99	<0.001
40–54 years	38	4.3	2.9–5.5	4.6	8.69	<0.001
55+ years	23	8.4	6.8–10.2	9.3	2.56	0.062
Total	102	5.7	4.8–6.5	6.1	37.51	<0.001

Similar indications emerge from the k-fold cross-validation. We used five folds, and the mean RMSE values obtained from various repetitions of the procedure ranged between 7.762 and 7.972 (Table 4), again only slightly higher than the RMSE value of the non-validated model (7.400).

Table 4. Cross-validation: RMSE values for different replications of the k-fold technique.

Replications	RMSE	95% CI
1	7.972	-
5	7.869	7.788–7.950
10	7.762	7.666–7.857
15	7.795	7.679–7.887
20	7.807	7.746–7.868

After validation of the accuracy of our model, we conducted a comparison with an identical model estimated by Zbiec-Piekarska and coworkers (2015) for a Polish sample of 300 individuals. This approach allows the model to also be viewed as further validation of the Polish model. We found a consistent direction of the effects of each marker on age; although all the coefficients were statistically significant in the Polish model, only three showed significance in the Italian model (Table 5). (The coefficients reported in Table 5 for the Italian sample differ slightly from those in Table 3 because they were estimated without controlling for individuals' sex, ensuring comparability with the Polish model.) This discrepancy has implications for the fit and performance of the two models. The Polish model presents a better fit to the observed data, as evidenced by its higher adjusted R^2 (0.941 vs. 0.669) and lower overall MAD value (3.9 years vs. 6.0 years), indicating better performance and accuracy.

However, these differences, either in magnitude or statistical significance, may be influenced by the different sample sizes or structural characteristics of the samples. Therefore, a more formal comparison of the two models involves fitting the Italian sample with the Polish regression model, determining the predicted values, and comparing them with

those resulting from our model. According to the literature, this comparison is best carried out using predicted age values as the independent X variable and observed age values as the dependent Y variable [29]. Thus, we constructed Figure 1 as the Observed–Predicted (OP) plot of both the Polish and the Italian regression models, including the two regression lines and their respective 95% confidence intervals. This methodology clearly allows the regression line for the Italian model to have no bias from the 1:1 line, a perfect bisector with a slope of 1. In contrast, the corresponding slope from the Polish model is 1.179. A Wald test with the null hypothesis $H_0: \beta = 1$ was applied, and the result supports the rejection of such a hypothesis ($F = 4.51$; $p\text{-value} = 0.036$). Consequently, the Polish model, once applied on our data, presents a statistically significantly higher slope than the Italian one, suggesting that chronological age increases more rapidly with increasing methylation.

Table 5. Regression coefficients from the Italian (IT) and Polish (PL) model.

Marker	IT	PL
ELOVL2-C7	0.338	0.465
C1orf132-C1	−0.411	−0.355
TRIM59-C7	0.827	0.306
KLF14-C1	0.360	0.833
FHL2-C2	0.210	0.237

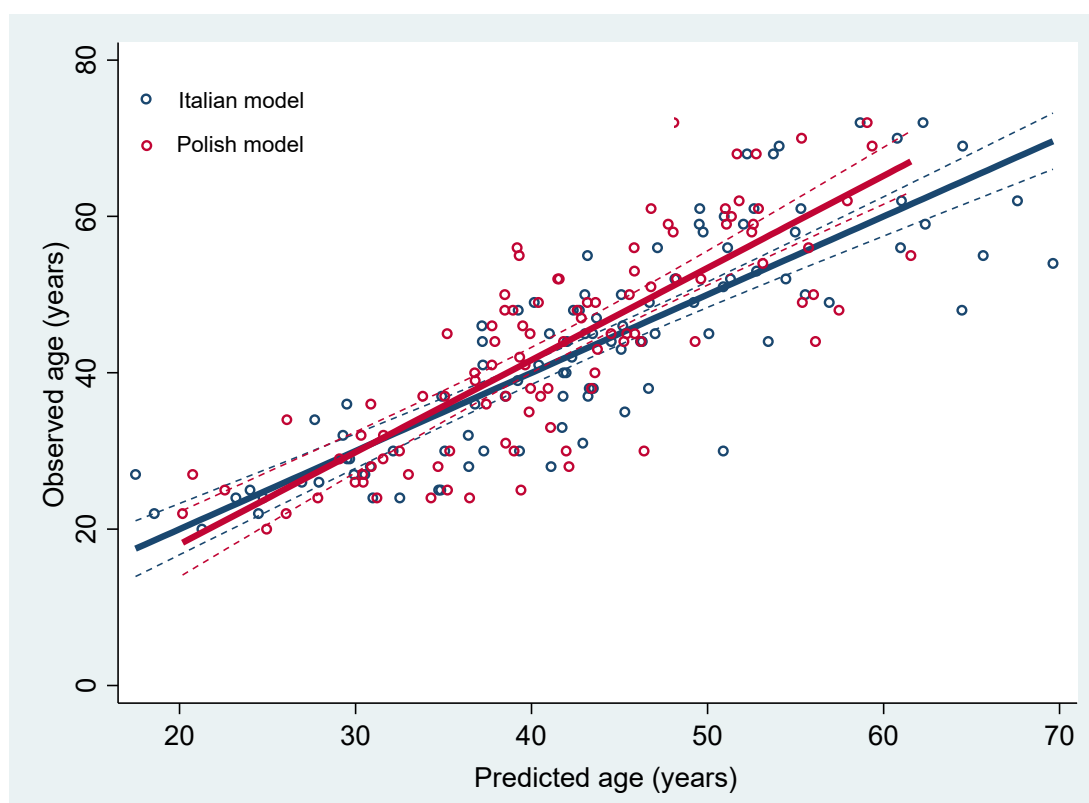


Figure 1. Observed versus predicted age scatter plot for the Italian cohort analyzed in this study ($n = 102$; age range 20–72 years). Predicted ages are derived from either the Italian regression model (blue) or the Polish regression model (red). Solid lines represent the fitted regression relationships, with shaded bands indicating the corresponding 95% confidence intervals. The dash lines are the confidence intervals, defined as shaded bands in the text.

3.2. Smoking Status Prediction

Two logistic regression models were estimated (Table 6) to investigate the odds of smoking status based on four consecutive CpG sites in the *AHRR* gene region outlined in the study by Alghanim et al. (2018) and previously associated with tobacco exposure [21], specifically Chr5:373,476 (CpG1), Chr5:373,490 (CpG2), Chr5:373,494 (CpG3), and Chr5:373,529 (CpG4). Sex and age were included as covariates as well. As described in the Statistical Analysis Section, the two models involve distinct characterizations of smoking habits, namely the classification of ex-smokers as either smokers or non-smokers.

Table 6. Logistic regression models of smoking status on four CpG sites ^a.

Marker	OR	Robust SE	Z	p-Value	95% CI
<i>Model 1</i>					
Age	0.927	0.023	−3.04	0.002	0.884–0.974
Sex (ref. F)	0.397	0.248	−1.48	0.140	0.116–1.353
CpG1	0.820	0.050	−3.26	0.001	0.728–0.924
CpG2	0.926	0.056	−1.28	0.200	0.823–1.042
CpG3	1.092	0.039	2.46	0.014	1.018–1.172
CpG4	1.064	0.064	1.03	0.304	0.945–1.197
N	102				
Log-likelihood	−45.3				
Wald chi (6)	24.95	<i>p</i> < 0.001			
AIC—BIC	104.5	122.9			
Pseudo R ²	0.360				
<i>Model 2</i>					
Age	0.940	0.021	−2.75	0.006	0.899–0.982
Sex (ref. F)	0.814	0.466	−0.36	0.719	0.264–2.503
CpG1	0.863	0.040	−3.21	0.001	0.788–0.944
CpG2	0.903	0.049	−1.89	0.058	0.812–1.004
CpG3	1.092	0.036	2.68	0.007	1.024–1.164
CpG4	1.069	0.053	1.34	0.179	0.970–1.179
N	102				
Log-likelihood	−50.5				
Wald chi (6)	25.11	<i>p</i> < 0.001			
AIC—BIC	114.9	133.3			
Pseudo R ²	0.270				

^a OR = odds ratio; RSE = robust standard error. Constant not shown.

Consistently, the results across both models indicate that CpG1 and CpG3 are the only statistically significant markers influencing the odds of being classified as a smoker, along with age. The odds of being identified as a smoker decrease by 6–7% for each unit increase in age. CpG1 demonstrates a negative association with the response variable, with the odds decreasing by 14–18% (according to the model) for each unit increase in methylation percentage. Conversely, CpG3 shows a 9% increase in odds in both models. Based on AIC and BIC statistics, Model 1 performs better than Model 2.

The same conclusion can be drawn by looking at the two classification tables of predicted observations, based on a threshold of $p = 0.5$. In other words, an observation is classified as “being a smoker” if its predicted probability $p \geq 0.5$. Table 7 shows that Model 1 performs better than Model 2 in classifying observations, with accuracy rates of 79.4% and 73.5%, respectively. While both models share similar levels of sensitivity (the percentage of correct classification of smokers out of true smokers in the sample), ranging from 77% to 79%, they differ in specificity (the percentage of correct classification of non-smokers out of true non-smokers in the sample), with Model 1 achieving 80% and Model 2 achieving 69%.

This distinction is also evident when looking at the ROC curves (Figure 2), where AUC is larger for Model 1 (0.874) compared to Model 2 (0.829).

Table 7. Classification table of predicted observations from the two estimated logistic models (The use of different cutoffs results in lower overall classification performance. In Model 1, a higher cutoff (0.75) correctly classifies 70.6% of the total observations but substantially reduces sensitivity (50%). In contrast, a lower cutoff (0.50) leads to an overall accuracy of 74.5%, with a marked decrease in specificity (54.0%)).

<i>Model 1</i>	True		Total
Classified	Smoker	Non-Smoker	
Smoker	41	10	51
Non-smoker	11	40	51
Total	52	50	102
Correctly classified	79.4%		
Sensitivity	78.9%		
Specificity	80.0%		
AUC area	0.874		
<i>Model 2</i>	Smoker	Non-smoker	Total
Smoker	46	13	59
Non-smoker	14	29	43
Total	60	42	102
Correctly classified	73.5%		
Sensitivity	76.7%		
Specificity	69.0%		
AUC area	0.829		

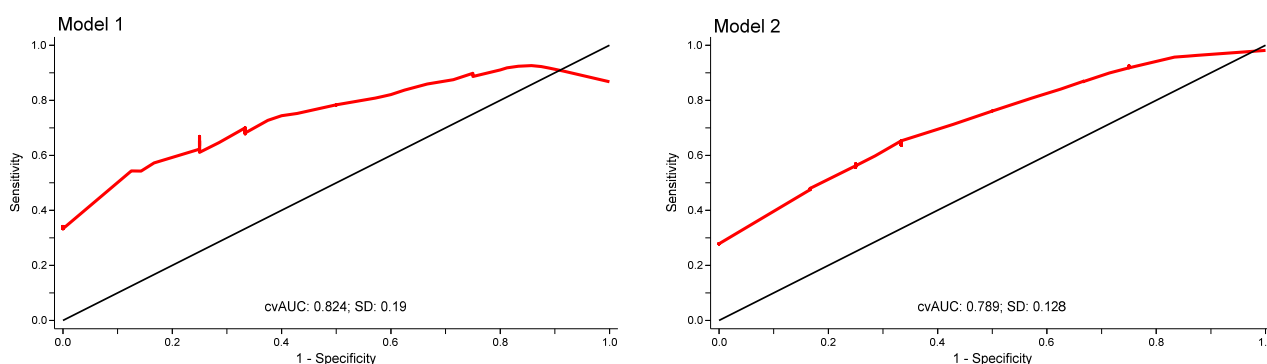


Figure 2. Cross-validated ROC curves for predicted smoking status based on a threshold of $p = 0.5$. Values derived from the 5 markers used in this study ($n = 102$). Model 1, which compares current smokers vs. non-smokers (defined as never-smokers + ex-smokers), outperforms Model 2, which compares current + ex-smokers vs. never-smokers. The corresponding areas under the curve (AUCs) are 0.824 for Model 1 and 0.789 for Model 2.

The better performance of Model 1 over Model 2 is further confirmed after cross-validation of the respective AUC values and areas (Figure 2): Model 1 shows, in fact, a value of 82.4% (95% CI: 72.1–89.4%) against a value of 78.9% (95% CI: 64.5–84.5%) for Model 2. Overall, these statistical indicators suggest that including ex-smokers in the non-smoking group enhances the model's ability not only to correctly classify observations but also to accurately identify non-smoking individuals.

4. Discussion

In the present study, forensic epigenetic markers for predicting age and smoking status in the Italian population were replicated in an Italian cohort. Although genome-wide DNA methylation arrays are considered highly effective tools for identifying age- or smoking-associated CpG markers during the discovery phase, this study emphasizes the importance of targeted sequencing to validate previously identified associations. Replicating the observed methylation markers allows us to confirm the reliability of the selected CpG sites and to evaluate their predictive performance in independent datasets. Moreover, analyzing a limited number of markers addresses the need to work with a small quantity of biological material that often characterizes samples collected from crime scenes. For instance, the selected marker set may not provide precise and accurate data if the analyzed fluid is seminal fluid [30]. A preliminary evaluation of the nature of the biological fluid is always essential to ensure the appropriate selection of markers.

Here, ELOVL2-C7, C1orf132-C1, and TRIM59-C7 significantly influenced age estimation, with TRIM59-C7 having the most substantial effect. It is noteworthy that the accuracy of predictions decreases with age: although this higher error in the older group may be due to the smaller sample size in the over-55 age group, studies based on DNA methylation for age estimation suggest that the underlying reason could be increased exposure to environmental factors over a longer period in the oldest individuals [31]. In addition, the Italian model, developed based on an Italian population sample, was also used as a further validation of the Polish model proposed by Zbiec-Piekarska and colleagues (2015) [16]. This analysis revealed that the coefficients are substantially consistent in direction and significance, comparing the two models, although only three were significant in the Italian model. The differences observed may be due either to the distinct populations considered or to different sampling methods. While promising results have been obtained, validation with additional models, populations and forensic casework conditions such as mixed samples, trace-level DNA, and degraded templates, is essential to ensure that these markers are reliably applicable in forensic contexts. Previous studies analyzing DNA methylation differences for age estimation among other populations have reported mixed results [2,26–28]. Therefore, further research including different populations is recommended to achieve standardization.

Concerning smoking status, two CpG sites emerged as statistically significant markers affecting the odds of being classified as a smoker. Notably, our results provide an important indication, as the model in which ex-smokers are classified as non-smokers demonstrates better predictive ability. It is thus essential to accurately identify non-smoking individuals and validate this classification in other studies and populations. Our findings confirm that two of the four CpG sites analyzed in the present study, identified as markers by Alghanim (2018) [21] and located near the cg05575921 marker site, are reliable markers for distinguishing between the two statuses. Two of these markers significantly influence the odds of smoking status, but with opposite effects: methylation of CpG3 is associated with a higher probability of being classified as a smoker, whereas an increase in methylation of CpG1 is linked to a reduced probability of being classified as a smoker. Interestingly, these markers are located in a gene encoding a repressor of the aryl hydrocarbon receptor pathway, which regulates the expression of xenobiotic-metabolizing enzymes involved in the response to environmental pollutants, such as polycyclic aromatic hydrocarbons found in cigarette smoke [32]. Consequently, changes in *AHRR* DNA methylation induced by cigarette smoke, along with the resulting alteration in its expression, may impair the body's ability to metabolize pollutants [33].

Although sex-specific DNA methylation differences can influence forensic predictions, especially age estimation, because methylation patterns may vary between males and

females, sex did not show statistical significance in our cohort. This is consistent with previous analyses showing that sex-specific models did not outperform unisex models based on the tested markers [34].

Some limitations and perspectives should be considered. This study was conducted on the Italian population. Thus, it was not possible to evaluate the extent to which possible ancestral influences affect the markers analyzed, limiting the generalizability of the results to other populations or contexts. Most proposed models are based on European or Asian populations and testing them on samples from diverse biogeographical origins would help determine whether the predictive model is universally applicable. Additionally, it should be noted that no studies that evaluate whether passive smoking has similar effects at the epigenetic level, which might require differentiating active smoking from passive smoke exposure, have been identified. Another important aspect to highlight is that some DNA methylation sites exhibit differential methylation across different tissues [35], and most prediction models developed so far, including this study, are tissue-specific. This indicates that results obtained from buccal swabs cannot be directly extrapolated to blood, semen, or mixed biological samples. Therefore, the tissue of origin in the sample must be known. Studies such as the one by Jung and coworkers (2018), which proposed a multi-tissue age prediction test (blood, saliva, and buccal swabs) including five CpG sites [20], emphasize the challenge of identifying accurate universal markers. Further experiments and technological advancements may facilitate the development of models with markers applicable regardless of the biological origin of the sample. Although this was not included in this study, it could potentially influence the DNA methylation level at the various CpG sites analyzed. Finally, it should also be noted that the analyzed cohort has a relatively small sample size, particularly within the older age group, and therefore these findings could be confirmed through validation in larger samples.

Overall, the findings support the potential of targeted DNA methylation analysis as a complementary approach in forensic investigations, especially for age and lifestyle profiling.

5. Conclusions

The models proposed in this study are useful tools for estimating chronological age and distinguishing between smoker and non-smoker status based on DNA methylation markers. Future studies should test these models under conditions that better reflect operational forensic scenarios, such as degraded and trace-level DNA and mixed samples, to verify their applicability in forensic caseworks. The influence of additional lifestyle and environmental factors, such as alcohol consumption, physical inactivity, dietary habits, and medical conditions, on methylation at CpG sites associated with age prediction and smoking status should also be explored. These variables may contribute to discrepancies between predicted epigenetic age and the actual chronological age. Including these variables could enhance the models' accuracy and support the identification of more robust and widely applicable epigenetic markers for forensic investigations.

Author Contributions: Conceptualization, D.C., M.M., F.G., A.M., M.C.G. and C.D.; methodology, D.C., M.M., M.C.G. and C.D.; formal analysis, D.C. and M.M.; investigation, D.C., M.M., F.G., M.C.G. and C.D.; data curation, D.C. and M.M.; writing—original draft preparation, M.C.G. and M.M.; writing—review and editing, D.C., M.M., F.G., A.M., M.C.G. and C.D.; visualization, M.M.; supervision, F.G., A.M., M.C.G. and C.D.; project administration, M.C.G. and C.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Local Ethics Committee (Comitato Etico Area Vasta Emilia Nord, Italy) (PROT. n. 20181, 12 May 2023).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: Maria Carla Gerra is supported by a fixed-term research assistant contract by the COMP-HUB Initiative, funded by the ‘Departments of Excellence’ program of the Italian Ministry for Education, University and Research (MIUR, 2018–2022).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Tiffon, C. The Impact of Nutrition and Environmental Epigenetics on Human Health and Disease. *Int. J. Mol. Sci.* **2018**, *19*, 3425. [\[CrossRef\]](#)
2. Vidaki, A.; Kayser, M. From forensic epigenetics to forensic epigenomics: Broadening DNA investigative intelligence. *Genome Biol.* **2017**, *18*, 238. [\[CrossRef\]](#) [\[PubMed\]](#)
3. Heikkinen, A.; Bollepalli, S.; Ollikainen, M. The potential of DNA methylation as a biomarker for obesity and smoking. *J. Intern. Med.* **2022**, *292*, 390–408. [\[CrossRef\]](#) [\[PubMed\]](#)
4. Bird, A.; Taggart, M.; Frommer, M.; Miller, O.J.; Macleod, D. A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA. *Cell* **1985**, *40*, 91–99. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Varriale, A.; Bernardi, G. Distribution of DNA methylation, CpGs, and CpG islands in human isochores. *Genomics* **2010**, *95*, 25–28. [\[CrossRef\]](#)
6. Cain, J.A.; Montibus, B.; Oakey, R.J. Intragenic CpG Islands and Their Impact on Gene Regulation. *Front. Cell Dev. Biol.* **2022**, *10*, 832348. [\[CrossRef\]](#)
7. Ioshikhes, I.P.; Zhang, M.Q. Large-scale human promoter mapping using CpG islands. *Nat. Genet.* **2000**, *26*, 61–63. [\[CrossRef\]](#)
8. Maas, S.C.E.; Vidaki, A.; Wilson, R.; Teumer, A.; Liu, F.; van Meurs, J.B.J.; Uitterlinden, A.G.; Boomsma, D.I.; de Geus, E.J.C.; Willemsen, G.; et al. Validated inference of smoking habits from blood with a finite DNA methylation marker set. *Eur. J. Epidemiol.* **2019**, *34*, 1055–1074. [\[CrossRef\]](#)
9. van Dongen, J.; Willemsen, G.; BIOS Consortium; de Geus, E.J.C.; Boomsma, D.I.; Neale, M.C. Effects of smoking on genome-wide DNA methylation profiles: A study of discordant and concordant monozygotic twin pairs. *eLife* **2023**, *12*, e83286. [\[CrossRef\]](#)
10. Breitling, L.P.; Yang, R.; Korn, B.; Burwinkel, B.; Brenner, H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. *Am. J. Hum. Genet.* **2011**, *88*, 450–457. [\[CrossRef\]](#)
11. Joubert, B.R.; Håberg, S.E.; Nilsen, R.M.; Wang, X.; Vollset, S.E.; Murphy, S.K.; Huang, Z.; Hoyo, C.; Midttun, Ø.; Cupul-Uicab, L.A.; et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environ. Health Perspect.* **2012**, *120*, 1425–1431. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Bocklandt, S.; Lin, W.; Sehl, M.E.; Sánchez, F.J.; Sinsheimer, J.S.; Horvath, S.; Vilain, E. Epigenetic predictor of age. *PLoS ONE* **2011**, *6*, e14821. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Florath, I.; Butterbach, K.; Müller, H.; Bewerunge-Hudler, M.; Brenner, H. Cross-sectional and longitudinal changes in DNA methylation with age: An epigenome-wide analysis revealing over 60 novel age-associated CpG sites. *Hum. Mol. Genet.* **2014**, *23*, 1186–1201. [\[CrossRef\]](#) [\[PubMed\]](#)
14. Horvath, S.; Oshima, J.; Martin, G.M.; Lu, A.T.; Quach, A.; Cohen, H.; Felton, S.; Matsuyama, M.; Lowe, D.; Wilson, J.G.; et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging* **2018**, *10*, 1758–1775. [\[CrossRef\]](#)
15. Park, J.-L.; Kim, J.H.; Seo, E.; Bae, D.H.; Kim, S.-Y.; Lee, H.-C.; Woo, K.M.; Kim, Y.S. Identification and evaluation of age-correlated DNA methylation markers for forensic use. *Forensic Sci. Int. Genet.* **2016**, *23*, 64–70. [\[CrossRef\]](#)
16. Zbieć-Piekarska, R.; Spólnicka, M.; Kupiec, T.; Parys-Proszek, A.; Makowska, Ż.; Pałeczka, A.; Kucharczyk, K.; Płoski, R.; Branicki, W. Development of a forensically useful age prediction method based on DNA methylation analysis. *Forensic Sci. Int. Genet.* **2015**, *17*, 173–179. [\[CrossRef\]](#)
17. Ghemrawi, M.; Tejero, N.F.; Duncan, G.; McCord, B. Pyrosequencing: Current forensic methodology and future applications—a review. *Electrophoresis* **2023**, *44*, 298–312. [\[CrossRef\]](#)
18. Joehanes, R.; Just, A.C.; Marioni, R.E.; Pilling, L.C.; Reynolds, L.M.; Mandaviya, P.R.; Guan, W.; Xu, T.; Elks, C.E.; Aslibekyan, S.; et al. Epigenetic Signatures of Cigarette Smoking. *Circ. Cardiovasc. Genet.* **2016**, *9*, 436–447. [\[CrossRef\]](#)

19. Horvath, S.; Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **2018**, *19*, 371–384. [CrossRef]
20. Jung, S.-E.; Lim, S.M.; Hong, S.R.; Lee, E.H.; Shin, K.-J.; Lee, H.Y. DNA methylation of the ELOVL2, FHL2, KLF14, C1orf132/MIR29B2C, and TRIM59 genes for age prediction from blood, saliva, and buccal swab samples. *Forensic Sci. Int. Genet.* **2019**, *38*, 1–8. [CrossRef]
21. Alghanim, H.; Wu, W.; McCord, B. DNA methylation assay based on pyrosequencing for determination of smoking status. *Electrophoresis* **2018**, *39*, 2806–2814. [CrossRef] [PubMed]
22. Knight, K.L.; Mauriello, A.; Williams, G. Quantitation of DNA Using the Applied Biosystems Quantifiler® Trio DNA Quantification Kit. *Methods Mol. Biol.* **2023**, *2685*, 175–188. [CrossRef] [PubMed]
23. Kreutz, M.; Schock, G.; Kaiser, J.; Hochstein, N.; Peist, R. PyroMark® Instruments, Chemistry, and Software for Pyrosequencing® Analysis. *Methods Mol. Biol.* **2015**, *1315*, 17–27. [CrossRef] [PubMed]
24. White, H. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica* **1980**, *48*, 817–838. [CrossRef]
25. Stone, M. Cross-Validatory Choice and Assessment of Statistical Predictions. *J. R. Stat. Soc. Ser. B (Methodol.)* **1976**, *36*, 111–133. Available online: <https://api.semanticscholar.org/CorpusID:62698647> (accessed on 22 December 2025). [CrossRef]
26. Bates, S.; Hastie, T.; Tibshirani, R. Cross-validation: What does it estimate and how well does it do it? *J. Am. Stat. Assoc.* **2024**, *119*, 1434–1445. [CrossRef]
27. Hastie, T.; Tibshirani, R.; Friedman, J.H.; Friedman, J.H. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. In *Springer Series in Statistics*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2009.
28. Luque-Fernandez, M.A.; Redondo-Sánchez, D.; Maringe, C. cvauroc: Command to compute cross-validated area under the curve for ROC analysis after predictive modeling for binary outcomes. *Stata J. Promot. Commun. Stat. Stata* **2019**, *19*, 615–625. [CrossRef]
29. Piñeiro, G.; Perelman, S.; Guerschman, J.P.; Paruelo, J.M. How to evaluate models: Observed vs. predicted or predicted vs. observed? *Ecol. Modell.* **2008**, *216*, 316–322. [CrossRef]
30. Heidegger, A.; Pisarek, A.; de la Puente, M.; Niederstätter, H.; Pośpiech, E.; Woźniak, A.; Schury, N.; Unterländer, M.; Sidstedt, M.; Junker, K.; et al. Development and inter-laboratory validation of the VISAGE enhanced tool for age estimation from semen using quantitative DNA methylation analysis. *Forensic Sci. Int. Genet.* **2022**, *56*, 102596. [CrossRef]
31. Onofri, M.; Delicati, A.; Marcante, B.; Carlini, L.; Alessandrini, F.; Tozzo, P.; Carnevali, E. Forensic Age Estimation through a DNA Methylation-Based Age Prediction Model in the Italian Population: A Pilot Study. *Int. J. Mol. Sci.* **2023**, *24*, 5381. [CrossRef]
32. Opitz, C.A.; Litzenburger, U.M.; Sahm, F.; Ott, M.; Tritschler, I.; Trump, S.; Schumacher, T.; Jestaedt, L.; Schrenk, D.; Weller, M.; et al. An endogenous tumour-promoting ligand of the human aryl hydrocarbon receptor. *Nature* **2011**, *478*, 197–203. [CrossRef]
33. Lee, K.W.K.; Pausova, Z. Cigarette smoking and DNA methylation. *Front. Genet.* **2013**, *4*, 132. [CrossRef]
34. Carlsen, L.; Holländer, O.; Danzer, M.F.; Vennemann, M.; Augustin, C. DNA methylation-based age estimation for adults and minors: Considering sex-specific differences and non-linear correlations. *Int. J. Leg. Med.* **2023**, *137*, 635–643. [CrossRef]
35. Lokk, K.; Modhukur, V.; Rajashekar, B.; Märten, K.; Mägi, R.; Kolde, R.; Koltšina, M.; Nilsson, T.; Vilo, J.; Salumets, A.; et al. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome Biol.* **2014**, *15*, 3248. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.