Contents lists available at ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com

Research paper

Evaluation of three statistical prediction models for forensic age prediction based on DNA methylation

Inge Smeers[a, 1], Ronny Decorte[a, b], Wim Van de Voorde[a, b], Bram Bekaert[a, b, *]

[a] Forensic Biomedical Sciences, Department of Imaging & Pathology, KU Leuven - University of Leuven, Leuven, Belgium
[b] Department of Forensic Medicine, Laboratory of Forensic Genetics and Molecular Archaeology, University Hospitals Leuven, Leuven, Belgium

ARTICLE INFO

Keywords:
Forensic age prediction
DNA methylation
Statistical regression modelling

1. Introduction

In a forensic context, age prediction is a useful tool in the identification of victims as well as unknown donors of crime scene traces. DNA methylation has recently emerged as an interesting biomarker of chronological age. This epigenetic phenomenon, in which the addition of a methyl group to cytosine forms 5-methylcytosine, occurs mainly at CpG dinucleotides [1]. Biological ageing is reflected in the epigenome and is characterised by a global hypomethylation [2], but certain regions in CpG islands undergo hypermethylation [3]. Throughout the genome, many CpG sites have been uncovered of which the methylation status is highly correlated with chronological age. These CpGs have been successfully used as markers for forensic age prediction. However, in previous studies we and others [4–7] have pointed out that the mean average deviation (MAD) between predicted age and actual age becomes larger as age increases. The prediction models used in those studies were built using on ordinary least squares regression, a linear regression method which is based on several assumptions, one of which being that the variance remains constant across the data [8]. This assumption was not fulfilled in the data of our previous study [4], which was heteroscedastic, meaning that there was non-constant variance. This heteroscedasticity should be taken into account when developing an age prediction model based on methylation data, such that the increasing prediction error can be reflected in the resulting prediction intervals.

To this end, Freire-Aradas et al. [5] have proposed a quantile regression model, in which an X quantile regression line will be drawn so that 100*X% of the data points are below that line. They predicted chronological age based on the median (0.50 quantile) and used the 0.10 and 0.90 quantiles for the limits of their prediction intervals. An alternative method for dealing with heteroscedasticity is weighted least squares (WLS) regression, in which every datapoint receives a weight based on the expected variance of that point. Based on the expectation that the model will fit the data better in areas of lower variance, the weights are inversely proportional to the variance. Hence datapoints with a low expected variance will receive a higher weight on the fit of the model [8].

The scope of the current study is to develop these three linear regression models based on the same dataset so that their performances can be compared. Besides assessing the accuracy in terms of the MAD, the ability to produce appropriate prediction intervals will also be evaluated. When providing an age estimation in support of a police investigation, it is more convenient to estimate prediction intervals rather than just reporting an average error. Since the prediction error increases with age, the average is only accurate for people in the middle of the age spectrum of the study population, whereas prediction intervals could change in size with increasing age to properly reflect this growing error. In addition, intervals are more straightforward to interpret and they avoid tunnel vision from the investigators, by providing a range wherein the age of the victim or suspect is situated with a large amount of certainty. In this study, 95% prediction intervals will be used to compare the success rate of the models, whereby a prediction is

[*] Corresponding author at: UZ Leuven, Campus Sint-Rafaël, Kapucijnenvoer 33, 3000 Leuven, Belgium.
Email address: bram.bekaert@uzleuven.be (B. Bekaert)

[1] Present address: Department of Human Genetics, Laboratory of Reproductive Genomics, KU Leuven - University of Leuven, Leuven, Belgium.

https://doi.org/10.1016/j.fsigen.2018.02.008
Received 5 September 2017; Received in revised form 27 December 2017; Accepted 5 February 2018
Available online xxx
1872-4973/ © 2017.

considered to be correct if the actual age falls between the limits of the predicted interval. The dataset will be that of our previous study [4].

2. Results

Methylation data was obtained through pyrosequencing of bisulphite converted DNA of 206 blood samples, for a total of 16 CpGs in 4 genes (*ELOVL2*, *EDARADD*, *PDE4C* and *ASPA*) (Supplementary table S1). The most highly correlated CpG of every gene was selected and included in the dataset: CpG1 for *EDARADD*, *PDE4C* and *ASPA*, and CpG6 for *ELOVL2* [4]. These data were submitted to regression modelling using three different methods (OLS, WLS and quantile regression), the results of which are discussed below.

2.1. Quadratic relationships

Evidence of non-linearity of varying degrees was observed in the relationships between methylation measurements and age. Fig. 1 shows the residuals plots for linear models predicting age in function of the markers. A parabolic shape indicates that a quadratic regression model, where the response variable is modelled by the squared predictor variable, is more appropriate than a linear one. The decision of whether a quadratic term should be included based on these plots is rather subjective. Since there were only four CpGs to consider as predictor variables, we opted to include the quadratic term of every CpG in the final dataset.

2.2. Age prediction models

A stepwise variable selection was conducted to select the best possible combination of predictors, being that which explains most of the variability in chronological age without overfitting the data. The selected group of predictor variables included the methylation values of *ASPA* CpG1, the squared methylation values of *ELOVL2* CpG6, and both the methylation values and the squared values of *EDARADD* CpG1 and *PDE4C* CpG1. The resulting models predicted chronological age according to a regression formula of the form $y = \alpha + \beta_1{*}x_1^2 + \beta_2{*}x_2 + \beta_3{*}x_2^2 + \beta_4{*}x_3 + \beta_5{*}x_4 + \beta_6{*}x_4^2$, where x_1 is the methylation level (in percentage) of *ELOVL2* CpG6, x_2 is the methylation level of *EDARADD* CpG1, x_3 is the methylation level of *ASPA* CpG1 and x_4 is the methylation level of *PDE4C* CpG1. The intercepts (α) and the coefficients (β) are provided in Table 1. In the quantile regression model, the limits of the 95% prediction intervals are calculated according to two separate formulas for the 0.025 and 0.975 quantiles. To illustrate the increasing error with age, an error plot was made based on the OLS regression model and this is shown in Fig. 2. Error plots of the WLS and quantile regression model are provided in Supplemental Fig. 1. A Shapiro-Wilk test was performed on the residuals of each model and they were found to be normally distributed ($p = 0.5085$ for OLS, $p = 0.1189$ for WLS and $p = 0.132$ for quantile regression). Normal QQ plots of these residuals are included in Supplemental Fig. 2.

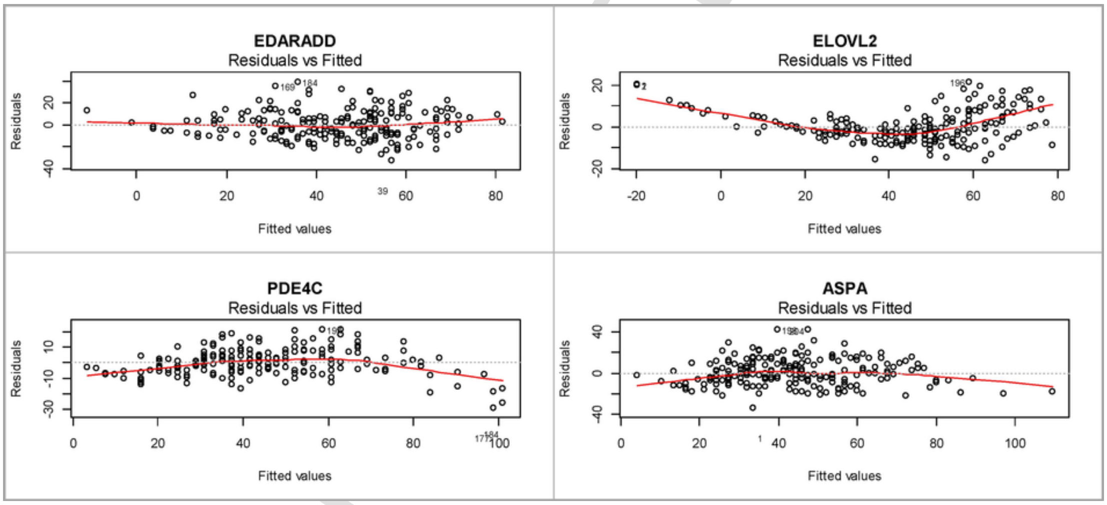

Fig. 1. Residuals plot for each predictor. The residuals were obtained from linear regression models fitting age in function of the predictor variable in question. For every age estimate (fitted value) on the x-axis, the deviation with the actual age (residual) is given on the y-axis. The numbers in the plot indicate, by sample number, data points that are flagged by R as possible outliers. *Note to the editor: This is a 2-column fitting image.*

Table 1
R output for the three regression models. Intercepts (α) and coefficients (β) of the regression formulas are given for every model, along with the residual standard error (RSE) and adjusted R^2 value for the least squares models. *For quantile regression, an RSE and adjusted R^2 were calculated manually. It should be noted that since R^2 is a least squares concept, the adjusted R^2 for quantile regression is an analogous R^2, calculated using the median age rather than the mean, as described in the methods.

Model	OLS	WLS	Quantile regression		
			0.5 quantile	0.025 quantile	0.975 quantile
α	28.482919	25.6875818	20.92851	34.42774	47.35468
β_1	0.005849	0.0067625	0.00621	0.00654	0.00720
β_2	−0.890577	−0.6723617	−0.072758	−1.31748	−0.78918
β_3	0.006739	0.0046379	0.00542	0.01082	0.00606
β_4	−0.178642	−0.1354516	−0.12192	−0.17347	−0.39766
β_5	1.501235	0.9016329	1.30874	1.32473	1.29014
β_6	−0.012502	−0.0038442	−0.00905	−0.01558	−0.01204
RSE	4.664	1.698	4.702*	n/a	n/a
Adj. R^2	0.9534	0.9701	0.9527*	n/a	n/a

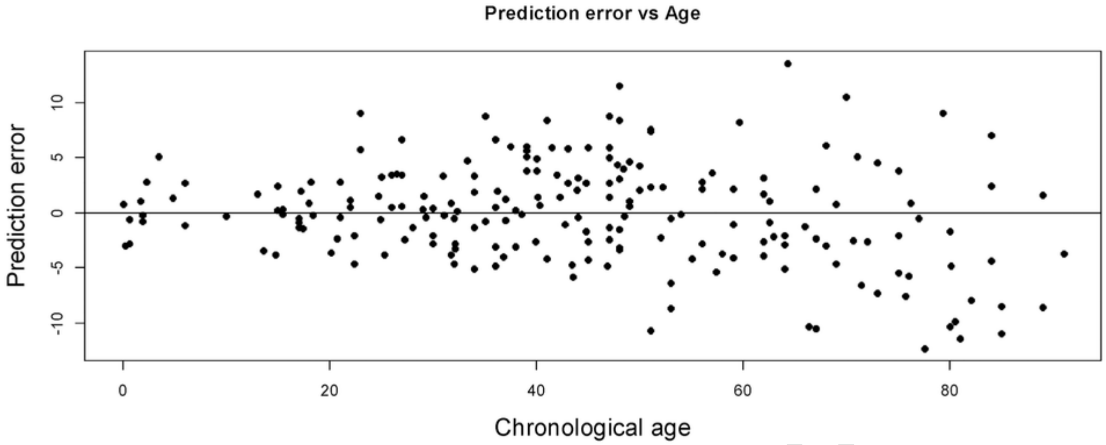

Fig. 2. Prediction error plotted against chronological age for the prediction obtained from the ordinary least squares regression model. *Note to the editor: This is a single-column fitting image.*

2.3. Validation using a training and test set

The root-mean-square error (RMSE) was used for the validation of the models since it can be easily calculated based on the predictions made for new data, rather than the residual standard error (RSE) which requires a number of degrees of freedom. Since the errors are squared, larger errors have a disproportionately larger influence on the mean. Therefore, the variance between errors is also taken into account, which is not the case when calculating the MAD. Table 2 shows the results of the model validation when randomly splitting the data into a training set (n = 137, average age 44.52 years) and a test set (n = 69, average age 42.77 years). A comparison of the RMSEs before and after validation indicates that there is no overfitting of the data. In fact, for every method the model fit is slightly better when applying the training model to the test set, compared to when a model is fitted on and then applied to the same data.

2.4. Comparison of performances

Model performance was compared in terms of the MAD and of the ability to correctly predict a subject's age within the limits of the given prediction interval. The MAD was used as a measure of accuracy so that the results may easily be compared to other literature on this subject, where this is the most often used metric. It is also more straightforward to interpret than the RMSE as a depiction of prediction error since it describes the average deviation alone, without the added implication of taking variance among errors into account. The deviations are consistent (within 0.06 years) across all model types, as shown in Table 3. The average width of a prediction interval is remarkably smaller in quantile regression compared to OLS and WLS, which results in a success rate below 95%. The samples which were incorrectly predicted are indicated by red crosses in Fig. 3.

The major difference between the models, however, was the gradual increase in prediction interval range with increasing age in WLS and quantile regression, whereas in OLS regression the intervals remained constant across all ages. This difference is illustrated by the graphs displayed in Fig. 3.

Table 2
Validation results for the three regression models. RMSE = root-mean-square error.

Model	RMSE	Test RMSE
Ordinary least squares	4.58	4.29
Weighted least squares	4.67	4.44
Quantile regression	4.64	4.46

Table 3
Comparison of the three regression models. MAD = mean absolute deviation.

Model	MAD	Average prediction interval range	Correct predictions	% Correct
Ordinary least squares	3.21	19.89 years	67	97.10%
Weighted least squares	3.20	18.59 years	66	95.65%
Quantile regression	3.26	16.02 years	63	91.30%

3. Discussion

The OLS, WLS and quantile regression models showed an equally strong performance in terms of absolute prediction errors. This indicates that putting a lower weight on datapoints with a higher expected variance, does not heavily alter the overall fit of the prediction model. In other words, the datapoints with high variance did not have much effect the slope of the regression line in OLS regression. It also shows that in this data there is no difference in predicting the median age (as quantile regression does), in comparison to predicting the average age. In the OLS and WLS models, where the average age is predicted, the residuals are normally distributed around the regression line, which explains why the predicted average and median age are quite similar. If the residuals were to fall mostly underneath the regression line due to high outliers, for instance, predicting the median age would shift the regression line downwards to compensate for this as the assumption of normality is dropped [12].

Although the accuracies in terms of absolute errors are similar across all three models, there is a notable difference in how the prediction interval limits progress with increasing age. The intervals clearly adapt to the increasing prediction error by becoming increasingly wider in WLS and quantile regression, whereas their limits run parallel to the regression line in the OLS model. In WLS regression, predictions are obtained in roughly the same fashion as in OLS regression, but the addition of weights according to the expected variance allows for prediction intervals to correctly reflect the accuracy of the predicted ages. However, this did result in a slight drop in success rate, with three samples being incorrectly predicted, compared to two with the OLS

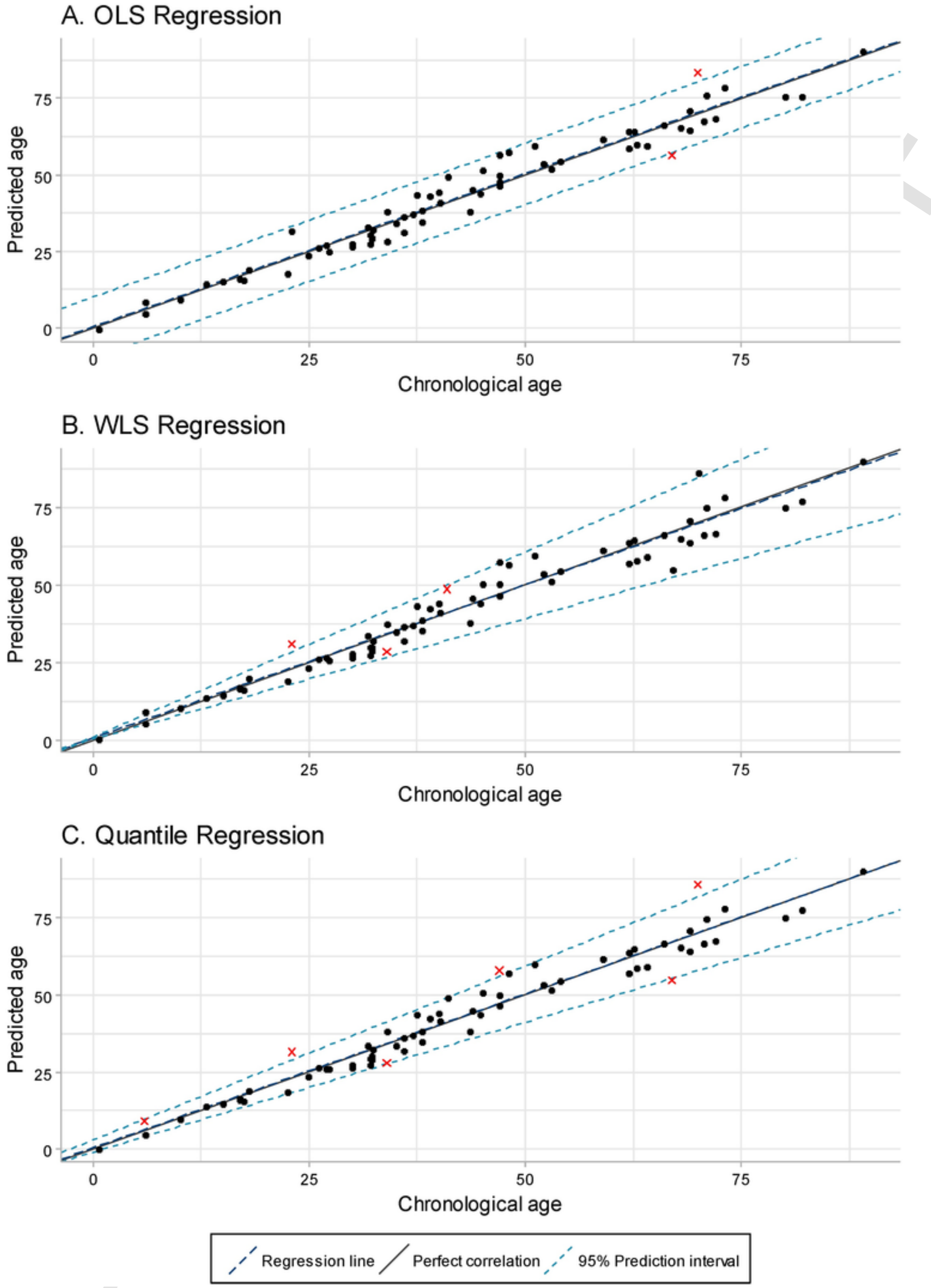

Fig. 3. Predictions for the test set with each of the three regression models, plotted against the actual chronological age. The red crosses are samples for which the obtained prediction interval did not include the actual age of that individual. (A) In ordinary least squares (OLS) regression, the prediction interval range remains constant across age. (B) In weighted least squares (WLS) regression, the intervals become gradually larger as the prediction error increases with age. (C) In quantile regression, the intervals also become gradually larger and they are not symmetrical around the predictions allowing a non-normal distribution of the variance. *Note to the editor: This is a 2-column fitting image.* (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

model. Nonetheless, a correct prediction was still obtained for 95.65% of the samples, an acceptable success rate considering 95% prediction intervals were used. While predictions obtained through OLS regression failed for two individuals above the average age of 42 years, the three individuals who received incorrect predictions from the WLS model were all younger than average (Fig. 3). This suggests that widening the prediction intervals for older people and thereby improving the success rate at that end of the spectrum, comes with a sacrifice at the younger end of the spectrum, where intervals are narrower and the success rate consequently drops.

In the quantile regression model, the prediction intervals show the same trend, but overall the age ranges are smaller, and this results in a notably lower success rate. In contrast to the WLS model, there is no shift in the age for which predictions are incorrect, but rather an addition of younger individuals with failed predictions compared to those in OLS regression (Fig. 3). In the same way that the median is less affected by outliers than the average, quantiles as interval limits are less affected by outliers as well, which can explain why the quantile prediction intervals are smaller than those of OLS and WLS regression in the current study. However, this also implies that outliers are less likely to be correctly predicted, hence the lower success rate observed in the quantile model.

Furthermore, in OLS and WLS regression, the variance is assumed to be normally distributed around the regression line and therefore the prediction intervals are calculated to be symmetrical around the predicted age. The quantile regression approach is less strict, dropping this assumption and allowing the prediction intervals to be asymmetrical around the predictions. This could be particularly advantageous in cases where the variance is heavily skewed in regards to the regression line.

While in this paper the focus was on statistical linear regression models, it should be noted that recent advances in bioinformatics have brought forth some machine learning methods which can also prove useful for forensic age prediction modelling. The major advantage of machine learning is that there are a lot less assumptions to be fulfilled, allowing more flexible and complex relationships between variables in the model, although one should be cautious of overfitting in small datasets [9]. Machine learning can be used to narrow down large pools of predictor variables to the most significant ones [13], but also in datasets with only a small selected set of markers it can be used for prediction modelling. Proposed methods in forensic age prediction based on DNA methylation include support vector regression [9,13], artificial neural networks [10,14] and random forest regression [11].

4. Conclusion

When providing an age estimate to police officers based on DNA methylation data, this estimate should be accompanied by a prediction interval that appropriately reflects the variance in the prediction model. The purpose of this study was to evaluate three types of linear regression models to determine which would be most appropriate to yield not only a high prediction accuracy, but also sufficiently representative prediction intervals in heteroscedastic data. When encountering heteroscedasticity, quantile regression and WLS regression are two viable alternatives for the mostly used OLS regression. They both allow prediction intervals to become wider with increasing age as the prediction error increases. In addition, quantile regression does not assume a normal distribution of the variance either, allowing asymmetry of the prediction intervals in case the variance is skewed to one side. In the current dataset, weighted regression seems to be the most appropriate alternative since it yielded a higher success rate than quantile regression. However, based on the findings presented in this paper, we argue that the choice of which type of regression model to use in future studies should always depend on the characteristics of the data at hand.

5. Materials and methods

The dataset, which was adopted from our previous study, was based on blood samples from 206 individuals between 0 and 91 years old. DNA was extracted, bisulphite converted and pyrosequencing was performed to determine methylation status as described in our previous paper [4]. Methylation levels were measured in four genes (*EDARADD*, *PDE4C*, *ELOVL2* and *ASPA*), including only the most highly correlated CpG of each gene in the final dataset (Supplementary file S2). These were CpG6 of *ELOVL2* and CpG1 of *EDARADD*, *PDE4C* and *ASPA*. Linearity of the relationship between methylation status and age was evaluated for every CpG by modelling age in function of each individual marker and looking at the residuals plots. The squared values of the methylation measurements were consequently included to account for quadrilinear relationships with age, leading to a total of eight prediction variables.

The most suitable combination of variables was selected through a stepwise selection, which uses the Akaike information criterion (AIC) and confirmed by comparison of the Bayesian information criterion (BIC) and goodness of fit (R^2). The selected variables were used to fit three models through OLS, WLS and quantile regression. Weights for the WLS model were determined through an auxiliary model which predicts the expected variance in function of the predictors, based on the relationship between the residuals of the OLS model and the predictor variables. For the quantile regression model an analogous R^2 was calculated using the median age rather than the mean, resulting in the following formula: $Adj.R^2 = 1 - \frac{\sum_i e_i^2/df_e}{\sum_i (y_i - \tilde{y})^2/df_t}$. The corresponding 95% prediction intervals of the OLS and WLS predictions were calculated by the predict function in R. In the quantile model, age was predicted as the median (.5 quantile) and the lower and upper limits of the 95% prediction interval were modelled as the 0.025 and 0.975 quantiles, respectively. In this way, 95% of the data points supposedly fall between these two limits. A Shapiro-Wilk test was performed on the residuals of every model to check whether they are normally distributed, along with a visual inspection by means of a normal QQ plot.

For validation purposes the dataset was randomly split (ratio 2:1) into a training set (n = 137) and a test set (n = 69) and the root-mean-square error (RMSE) of the original model was compared to that of the training model when applied to the test set. The performance of the prediction methods was assessed in the same training and test set based on their MAD and ability to predict age correctly within an appropriate 95% prediction interval.

Data analysis and prediction modelling were performed in R for Windows 3.4.0 with RStudio v1.0.143 using the leaps v3.0, mgcv v1.8-17, quantreg v5.33 and caTools v1.17.1 packages. The R script is available in Supplementary file S3.

Acknowledgement

The authors would like to thank Professor Anne-Marie De Meyer of the KU Leuven Department of Mathematics and Akshat Dwivedi for their advice.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fsigen.2018.02.008.

References

[1] P.J. Russell, iGenetics: A Molecular Approach, 3rd ed.Pearson Benjamin Cummings, San Fransisco, 2010.

[2] S.D. van Otterdijk, J.C. Mathers, G. Strathdee, Do age-related changes in DNA methylation play a role in the development of age-related diseases?, Biochem. Soc. Trans. 41 (2013) 803–807.

[3] Å. Johansson, S. Enroth, U. Gyllensten, Continuous aging of the human DNA methylome throughout the human lifespan, PLoS One 8 (2013) https://doi.org/10.1371/journal.pone.0067378.

[4] B. Bekaert, A. Kamalandua, S.C. Zapico, W. Van De Voorde, R. Decorte, Improved age determination of blood and teeth samples using a selected set of DNA methylation markers, Epigenetics 10 (2015) 922–930, https://doi.org/10.1080/15592294.2015.1080413.

[5] A. Freire-Aradas, C. Phillips, A. Mosquera-Miguel, L. Girón-Santamaría, A. Gómez-Tato, M. Casares De Cal, J. Álvarez-Dios, J. Ansede-Bermejo, M. Torres-Español, P.M. Schneider, E. Pospiech, W. Branicki, Á. Carracedo, M.V. Lareu, 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, https://doi.org/10.1016/j.fsigen.2016.06.005.

[6] G. Hannum, J. Guinney, L. Zhao, L. Zhang, G. Hughes, S. Sadda, B. Klotzle, M. Bibikova, J.B. Fan, Y. Gao, R. Deconde, M. Chen, I. Rajapakse, S. Friend, T. Ideker, K. Zhang, Genome-wide methylation profiles reveal quantitative views of human aging rates, Mol. Cell 49 (2013) 359–367, https://doi.org/10.1016/j.molcel.2012.10.016.

[7] R. Zbieć-Piekarska, M. Spólnicka, T. Kupiec, A. Parys-Proszek, Z. Makowska, A. Pałeczka, K. Kucharczyk, R. Płoski, W. Branicki, Development of a forensically use-

ful age prediction method based on DNA methylation analysis, Forensic Sci. Int. Genet. 17 (2015) 173–179, https://doi.org/10.1016/j.fsigen.2015.05.001.

[8] B. Rosner, Fundamentals of Biostatistics, 8th ed., Cengage Learning, Boston, 2016.

[9] C. Xu, H. Qu, G. Wang, B. Xie, Y. Shi, Y. Yang, Z. Zhao, L. Hu, X. Fang, J. Yan, L. Feng, A novel strategy for forensic age prediction by DNA methylation and support vector regression model, Sci. Rep. 5 (2015) https://doi.org/10.1038/srep17788.

[10] A. Vidaki, D. Ballard, A. Aliferi, T.H. Miller, L.P. Barron, D. Syndercombe Court, DNA methylation-based forensic age prediction using artificial neural networks and next generation sequencing, Forensic Sci. Int. Genet. 28 (2017) 225–236, https://doi.org/10.1016/j.fsigen.2017.02.009.

[11] J. Naue, H.C.J. Hoefsloot, O.R.F. Mook, L. Rijlaarsdam-Hoekstra, M.C.H. van der Zwalm, P. Henneman, A.D. Kloosterman, P.J. Verschure, Chronological age prediction based on DNA methylation: massive parallel sequencing and random forest regression, Forensic Sci. Int. Genet. 31 (2017) 19–28, https://doi.org/10.1016/j.fsigen.2017.07.015.

[12] B.S. Cade, B.R. Noon, A gentle introduction to quantile regression for ecologists, Front. Ecol. Environ. 1 (2003) 412–420, https://doi.org/10.2307/3868138.

[13] F. Sarac, H. Seker, A. Bouridane, Exploration of unsupervised feature selection methods to predict chronological age of individuals by utilising CpG dinucleotics from whole blood, Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS (2017) 3652–3655, https://doi.org/10.1109/EMBC.2017.8037649.

[14] M. Spólnicka, E. Pośpiech, B. Pepłońska, R. Zbieć-Piekarska, Makowska, A. Pięta, J. Karłowska-Pik, B. Ziemkiewicz, M. Wężyk, T. Gasperowicz, M. Bednarczuk, M. Barcikowska, C. Żekanowski, R. Płoski, W. Branicki, DNA methylation in ELOVL2 and C1orf132 correctly predicted chronological age of individuals from three disease groups, Int. J. Legal Med. (2017) https://doi.org/10.1007/s00414-017-1636-0.