MOX-Report No. 24/2022

A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events

Cappozzo, A.; McCrory, C.; Robinson, O.; Freni Sterrantino, A.; Sacerdote, C.; Krogh, V.; Panico, S.; Tumino, R.; Iacoviello, L.; Ricceri, F.; Sieri, S.; Chiodini, P.; Kenny, R.A.; O'Halloran, A.; Polidoro, S.; Solinas, G.; Vineis, P.; Ieva, F.; Fiorito, G.;

MOX, Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

mox-dmat@polimi.it http://mox.polimi.it

A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events

Andrea Cappozzo[1], Cathal McCrory[2], Oliver Robinson[3], Anna Freni Sterrantino[3,4], Carlotta Sacerdote[5], Vittorio Krogh[6], Salvatore Panico[7], Rosario Tumino[8], Licia Iacoviello[9,10], Fulvio Ricceri[11,12], Sabina Sieri[6], Paolo Chiodini[13], Rose Anne Kenny[2], Aisling O'Halloran[2], Silvia Polidoro[14], Giuliana Solinas[15], Paolo Vineis[3], Francesca Ieva[1,16], Giovanni Fiorito[2,3,15, *]

1. Department of Mathematics, Politecnico di Milano, Milan, Italy.
2. Department of Medical Gerontology, Trinity College Dublin, Ireland.
3. MRC-PHE Centre for Environment and Health, Imperial College London, UK.
4. The Alan Turing Institute, London, UK.
5. Unit of Cancer Epidemiology, Città Della Salute e della Scienza University-Hospital, Turin, Italy.
6. Fondazione IRCCS - Istituto Nazionale dei Tumori, Milan, Italy.
7. Department of Clinical Medicine and Surgery, University of Naples Federico II, Naples, Italy.
8. Hyblean Association for Epidemiology Research, AIRE ONLYS, Ragusa, Italy.
9. Department of Epidemiology and Prevention, IRCCS NEUROMED, Pozzilli, Italy.
10. Research Center in Epidemiology and Preventive Medicine (EPIMED), Department of Medicine and Surgery, Torino, Italy.
11. Epidemiology Unit, Regional Health Service TO3, Grugliasco, Italy. Department of Clinical and Biological Sciences, University of Turin, Italy.
12. Centre for Biostatistics, Epidemiology, and Public Health (C-BEPH), Department of Clinical and Biological Sciences, University of Turin, Italy.
13. Department of Mental, Physical Health and Preventive Medicine University of Campania 'Luigi Vanvitelli', Caserta, Italy.
14. Italian Institute for Genomic Medicine (IIGM), Turin, Italy.
15. Laboratory Biostatistics, Department of Biomedical Sciences, University of Sassari, Italy.
16. CHDS - Center for Health Data Science, Human Technopole, Milan, Italy

(*) Corresponding author: Giovanni Fiorito, Laboratory of Biostatistics, Department of Biomedical Sciences, University of Sassari, Via Padre Manzella 4, Sassari, Italy. e-mail: gfiorito@uniss.it and giovanni.fiorito85@gmail.com

Word count: 2,999

Abstract

Background. Evidence highlights the epidemiological value of DNA methylation (DNAm) for predicting cardiovascular diseases (CVDs). DNAm surrogates of exposures and risk factors predict diseases and longevity better than self-reported or measured exposures in many cases. Composite biomarkers based on DNAm surrogates, 'next-generation' epigenetic clocks trained on time-to-death, constitute non-specific biomarkers representing the general health status rather than disease-specific signatures. Training a model on cardiovascular-specific risk factors may improve the identification of high-risk populations for CVD.

Methods. We developed a DNAm-based biomarker predictive of short-term risk for CVD using a two-step approach: 1) development and validation of novel DNAm surrogates for cardiovascular risk biomarkers; 2) development and validation of a *DNAmCVDscore* as a combination of DNAm surrogates. In an independent testing set, we compared the prediction performance of *DNAmCVDscore* with (a) the 'next-generation' epigenetic clock DNAmGrimAge, (b) a DNAm score for CVD derived through a single-step approach, MRS, and (c) the current state-of-the-art prediction model based on traditional CVD risk factors, SCORE2.

Results. We presented novel DNAm surrogates for BMI, blood pressure, fasting glucose and insulin, cholesterol, triglycerides, and coagulation biomarkers, validated in independent datasets. Further, we derived a *DNAmCVDscore* outperforming the model based on traditional CVD risk factors and other epigenetic biomarkers for predicting short-term cardiovascular events.

Conclusions. We provided novel DNAm surrogates useful for future epidemiological research, and we described a DNAm based composite biomarker, *DNAmCVDscore*, predictive of short-term CVD. Our results highlight the usefulness of DNAm surrogate biomarkers of risk factors and exposures to identify high-risk populations.

Keywords: Epigenetics, DNA methylation, surrogate biomarkers, cardiovascular risk.

Key messages:

- We provided novel blood DNA methylation (DNAm) surrogates for cardiovascular risk factors (BMI, cholesterol, triglycerides, fasting glucose and insulin, inflammation and coagulation biomarkers) useful for future epidemiological research.
- We developed a new blood composite biomarker, *DNAmCVDscore*, outperforming models based on traditional cardiovascular risk factors for predicting short-term cardiovascular events (within seven years after blood-collection or less).
- Predictive models based on a two-step training strategy led to more reliable and robust biomarkers for identifying high-risk populations for non-communicable and age-related diseases.
- We encourage testing this two-step approach for predicting other non-communicable and age-related diseases (cancer, mental diseases, neurodegenerative diseases, respiratory problems, hearing and taste loss, etc.) by training and developing DNAm surrogates for disease-specific risk factors and exposures.

Introduction

Emerging epidemiological evidence indicates that composite scores based on blood DNA methylation (DNAm) at different CpG sites are valuable biomarkers to predict complex traits and identify high-risk populations.[1–4] DNAm scores are usually built modelling the association of CpG sites with the trait or disease of interest via epigenome-wide association studies (EWAS). However, EWAS suffer from a lack of replication in independent datasets,[5] with few exceptions like the well-known DNAm CpGs associated with smoking.[6,7] Further, it is unclear how the disease risk tracked by DNAm is complementary or redundant with other risk factors for non-communicable diseases (NCDs). In fact, the inclusion of DNAm scores in prediction models often leads to null or marginal prediction improvement compared with traditional models based on classical risk factors like the Framingham Risk Score and SCORE2 for cardiovascular diseases (CVD).[1,4,8–10]

In contrast, it has been consistently shown that DNAm scores for estimating individual biological age, named epigenetic clocks,[11–15] are associated with several risk factors for NCDs (smoking, alcohol intake, low physical activity, obesity, socio-economic position, and job characteristics),[16,17] and perform very well for predicting ageing-related diseases and all-cause mortality.[18,19] These results may be explained by how 'next-generation' epigenetic clocks like DNAmPhenoAge and DNAmGrimAge have been built.[11,12] Contrary to classical DNAm scores for NCDs, 'next-generation' epigenetic clocks used a two-step approach: 1) development of DNAm surrogates for NCDs risk factors and biomarkers associated with all-cause mortality; 2) development of DNAm epigenetic clocks as a weighted combination of DNAm surrogates. Such a procedure leads DNAm composite scores to be more reliable and reproducible across different cohorts. The best performing epigenetic clock, called DNAGrimAge, incorporates DNAm scores for seven circulating proteins and smoking pack-years, and has been consistently associated with longevity and numerous age-related diseases, and functional and cognitive outcomes.[11,18] Other examples of DNAm surrogate of exposures and risk factors include the DNAm biomarkers by Colicino and colleagues for cumulative lead exposure,[20] the one by Marioni and colleagues for several longevity-related and inflammatory proteins,[21–23] and the classification by Guida and colleagues of current, former (including time since smoking cessation) and never smokers based on blood DNAm biomarkers.[7]

DNAm surrogates can outperform original exposure measurements in predicting diseases in association studies. For example, Zhang and colleagues show that a combination of smoking-associated DNAm markers predicts lung cancer incidence better than self-reported smoking.[24] In addition, Green and colleagues suggest that a DNAm proxy for C-reactive protein (CRP) predicts structural neuroimaging brain measures better than blood measured CRP.[25] DNAm characteristics can explain these counter-intuitive results: 1) DNAm is a more reliable biomarker than self-reported exposure (i.e., in the case of smoking or other exposures measured through self-reported questionnaires); 2) DNAm variability includes individual genetic and metabolic profiles that can influence individual response to exposure and stressors (i.e., the same amount of exposure can be more or less dangerous based on genetic profile and general state of health); 3) DNAm variations reflect long-term exposures and, in some cases, are more stable in time (i.e. in the case of inflammatory status, one of the best blood biomarkers, CRP, has several fluctuations within a single day).

Because of the way 'next-generation' epigenetic clocks have been built (e.g., trained on a set of biomarkers associated with longevity), they are non-specific biomarkers that mirror an individual general state of health rather than risk for specific diseases. This study aims to evaluate the possibility of developing disease-specific blood DNAm biomarkers, training a DNAm score on disease-specific exposure and risk factors (rather than on all-cause mortality, as it has been done for 'next-generation' epigenetic clocks). Specifically, we aim to: 1) develop a DNAm composite biomarker for predicting cardiovascular events trained on CVD-specific risk factors, and 2) to compare its predictive performance for incident CVD events with (a) the 'next-generation' epigenetic clock DNAmGrimAge; (b) a DNAm score for CVD based on a single-step approach developed by Fernández-Sanlés et al., named methylation risk score (MRS);[1] and (c) a prediction model based on traditional CVD risk factors (chronological age, sex, diabetes status, smoking, systolic blood pressure, total and HDL cholesterol levels), named SCORE2.[10]

We selected the most relevant DNAm surrogates for our purpose among those already available from the literature[11,20,21] or newly developed within this study. The initial set of candidate DNAm surrogates includes 60 biomarkers for smoking pack-years, alcohol consumption, obesity indexes, blood pressure, insulin, glucose, blood coagulation biomarkers, cholesterol levels, and several blood-measured (mainly inflammatory) proteins. Our procedure derived a blood DNAm biomarker named '*DNAmCVDscore*' predictive of short-term cardiovascular events as a combination of ten DNAm surrogates, outperforming current state-of-the-art prediction models based on traditional CVD risk factors and DNAm scores based on single-step approaches. Finally, since COVID-19 has several risk factors in common with CVD risk, we investigated the association of *DNAmCVDscore* with COVID-19 susceptibility and severity in an independent case-control study.

Methods

Study sample

This study sample includes DNAm data from five studies described previously,[17,18,26–29] and summarised in **Table 1**.

EPIC Italy, the training set, contains 1,803 individuals (62% women), age range from 35 to 75 years, including 295 (16.4%) incident CVD cases. The average (standard deviation) time from recruitment to CVD events was 7.6 (3.8) years. The average (standard deviation) follow-up time was 11.3 (5.6) years.

EXPOsOMICS CVD is a case-control study nested in the EPIC Italy cohort, including 160 incident CVD cases and age- and sex-matched controls (not overlapping with EPIC Italy sample), age range from 35 to 70 years (53% women). The average (standard deviation) time from recruitment to CVD events was 9.6 (3.9) years. The average (standard deviation) follow-up time was 12 (4) years.

TILDA includes data for 490 individuals, originally selected to investigate the association of epigenetic biomarkers of biological ageing with intergenerational socio-economic trajectories, with individuals equally distributed among four socio-economic categories, age range from 50 to 80 years (50% women).

The United Kingdom Household Panel Study (UKHLS), also known as Understanding Society, is an ongoing longitudinal, nationally representative study of the UK, designed as a two-stage stratified random sample of the general population. The data used here consist of two pooled cross-sectional waves (waves 2 and 3), age range from 28 to 98 years (59% women).

GSE174818 contains data for 101 COVID-19 cases and 27 age- and sex-matched controls hospitalised with respiratory symptoms, ranging from 21 to 90 years (40% women).

Details of participant recruitment, relevant covariate acquisition, and laboratory methods for DNAm measuring, pre-processing and normalisation procedures are described in **Supplementary Material**.

Statistical analyses

In **Figure 1**, we present the analytical flowchart summarising the main steps for developing the *DNAmCVDscore*:

1) Develop and validate novel DNAm surrogate biomarkers (training set: EPIC Italy study; testing sets: EXPOsOMICS CVD, Understanding Society, TILDA, and GSE174818 studies) through LASSO regularisation for linear regression model.

2) Develop the *DNAmCVDscore* (training set: EPIC Italy study; 60 candidate DNAm surrogate biomarkers) through elastic net for Cox proportional hazards model.

3) Validation of the *DNAmCVDscore* (testing set: EXPOsOMICS CVD study) investigating its prediction performance through ROC curve analysis, right censoring follow-up data at different time points.

4) Comparison of *DNAmCVDscore*, MRS, SCORE2, and DNAmGrimAge predictive value.

The analytical details for each step are presented in **Supplementary Material**.

Results

Estimation and validation of DNAm surrogates. We developed DNAm surrogates for body mass index (BMI), systolic and diastolic blood pressure, and ten blood measured biomarkers. In **Table 2**, we report the number of CpGs whose linear combination best predicted the corresponding marker and the Pearson correlation coefficients of observed (measured) *vs* predicted (DNAm surrogate) in the EPIC Italy testing set (25% of the total sample). The correlation of DNAm surrogates with the corresponding measured marker was always higher than 0.4 (all *P-values* lower than 0.0001), ranging from 0.43 (DNAmPAI-1 *vs* PAI-1) to 0.73 (DNAmTriglycerides *vs* triglycerides). Further, in **Table 2**, we report the Pearson correlation coefficients of observed *vs* predicted values computed in the four validation datasets. The correlation of DNAm surrogates with the corresponding measured marker was always positive, ranging from 0.08 (DNAmHDL *vs* HDL cholesterol) to 0.44 (DNAmInsulin *vs* insulin). The *P-value* was lower than 0.05 for all but D-dimer, diastolic blood pressure, LDL cholesterol, and total cholesterol. Based on the above, we validated nine (out of 13) DNAm surrogates for BMI, CRP, fasting glucose and insulin, HDL cholesterol, triglycerides, PAI-1, Platelet tissue factor (CD142), and systolic blood pressure.

Comparison with previously developed DNAm surrogates. We compared our newly developed DNAm surrogates with previously developed DNAm surrogates for HDL cholesterol, BMI,[21] and PAI-1.[11] The Pearson correlation coefficients of our DNAm surrogates with those previously developed were 0.31 ($P < 0.0001$), 0.45 ($P < 0.0001$), and 0.36 ($P < 0.0001$) for HDL cholesterol, BMI, and PAI-1, respectively.

Development and validation of the DNAmCVDscore. We developed a combined score, *DNAmCVDscore*, predictive of future CVD events by regressing the time to CVD event on 60 DNAm surrogates previously described. The elastic net Cox regression model selected chronological age, sex, and DNAm surrogates for blood measured glucose, HDL cholesterol, systolic blood pressure, PAI-1, CRP (developed within this study), Serine/threonine-protein kinase receptor 3 (SKR3) and hepatocyte growth factor (HGF) (developed in Hillary et al.[21]) growth differentiation factor 15 (GDF15) protein, smoking pack-years (developed in Lu et al.[11]), and lead level measured in patella bone (developed in Colicino et al.[20]). **Table 3** shows the coefficients extracted from the elastic net model, representing weights for computing the *DNAmCVDscore*. Since our validation dataset is a case-control study matched for chronological age and sex, we deliberately chose not to include chronological age and sex in the *DNAmCVDscore* and test its prediction performance in logistic regression models adjusted for chronological age and sex. All the biomarkers but DNAmHDL have a positive regression coefficient (higher risk associated with higher values). The linear combination of standardised values for the ten DNAm surrogates listed in **Table 3** can be interpreted as a standardised (within the population in which it is computed) CVD risk score (named *DNAmCVDscore*).

In the independent test set (EXPOsOMICS CVD dataset), we computed the *DNAmCVDscore* based on the coefficients derived in the training set, and we compared its predictive performance with those of MRS, SCORE2 and DNAmGrimAge through ROC curve analysis of logistic regression models adjusted for age, sex, and centre of recruitment (matching parameters for the EXPOsOMICS CVD case-control study). In **Table 4** and **Figure 2**, we present the area under the ROC curve (AUC), sensitivity, and specificity (best threshold selected according to the minimum distance from the top left corner of the ROC curve) of the four composite biomarkers at different time points. For all the epigenetic biomarkers *DNAmCVDscore*, MRS, and DNAmGrimAge, the AUC increases as the follow-up time decreases, suggesting that epigenetic biomarkers predict short-term events rather than long-term CVD risk (**Table 4** and **Figure 2**). Contrarily, the AUC for SCORE2 was not time-dependent, ranging from 0.678 (seven years follow-up) to 0.785 (four years follow-up). The MRS had the worst performance independently of the follow-up length (**Table 4** and **Figure 2**). SCORE2 outperforms epigenetic biomarkers in predicting CVD events considering follow-up time from 18 to eight years. However, right censoring the follow-up time at seven years or less, *DNAmCVDscore* and

DNAmGrimAge perform better than SCORE2, with *DNAmCVDscore* having a slightly higher AUC than DNAmGrimAge (**Table 4** and **Figure 2**).

Additional results about sensitivity analyses, the correlation of *DNAmCVDscore* with epigenetic clocks, and the association of *DNAmCVDscore* with COVID-19 case-control status and severity are reported in **Supplementary Material**.

Discussion

Emerging evidence highlights the epidemiological value of composite scores based on blood DNAm surrogates of exposures and risk factors, e.g., epigenetic clocks, associated with non-communicable diseases (NCDs) and predictive of mortality.[19] However, since 'next-generation' epigenetic clocks have been trained on time to death, they constitute non-specific biomarkers, representative of the general individual state of health, rather than disease-specific biomarkers. In this work, we present a combined blood DNAm based biomarker for predicting future cardiovascular events, named *DNAmCVDscore*. To the best of our knowledge, this is the first example of a disease-specific biomarker using molecular data only, without the need for additional information (other than age and sex) about the personal history of exposure, general state of health, lifestyle habits, and other commonly used biomarkers. This may be important for future risk prediction avoiding invasive and expensive procedures. Also, a predictive score based on a single experiment reduces the possibility of measurement errors and bias due to self-reported exposure to risk factors. For this aim, DNAm based biomarkers are optimal candidates because DNAm is strongly influenced by long-term exposures, genetic susceptibility, and lifestyle habits.[30] In other words, it is possible to extract information about the history of exposures and susceptibility to complex diseases from whole-genome DNAm data with high accuracy.

We applied a two-step approach, following the successful example of the epigenetic clocks. First, we developed and validated nine novel DNAm surrogates for CVD risk factors: systolic blood pressure, BMI, CRP, fasting glucose and insulin, HDL cholesterol, triglycerides, PAI-1, and platelet tissue factor (a.k.a. CD142 protein). We provided an *R script* for generating the new DNAm surrogates in independent datasets for future epidemiological research in the **Supplementary material**. Then, we developed a *DNAmCVDscore* starting from 60 candidate DNAm surrogates (nine newly developed within this study plus 51 from previous literature), including surrogate measures for the main risk factors for CVDs (obesity, smoking habits, alcohol consumption, inflammatory proteins, lipid levels, blood pressure, coagulation biomarkers). Our elastic net model extracted ten DNAm surrogate biomarkers whose linear combination constitutes the so-called *DNAmCVDscore*: fasting glucose, HDL cholesterol, systolic blood pressure, smoking pack-years, lead exposure and blood levels of PAI-1, CRP, SKR3, HGF, and GDF15 proteins. We validated the ability of the *DNAmCVDscore* to predict future cardiovascular events in an independent prospective case-control study (EXPOsOMICS CVD). This dataset matches incident CVD cases with healthy controls by age, sex, recruitment centre, and length of follow-up using the incident density sampling method. We showed that existing prediction models based on traditional CVD risk factors (SCORE2,[10] based on chronological age, sex, diabetes, smoking, systolic blood pressure, total and HDL cholesterol) outperform epigenetic biomarkers for predicting long-term CVD risk according to the AUC measure. However, *DNAmCVDscore* predicts short-term (seven years follow-up or less) CVD risk better than SCORE2. According to the AUC measure, the prediction performance of *DNAmCVDscore* and DNAmGrimAge was comparable (slightly higher for *DNAmCVDscore* for short-term events). This is not unexpected considering that four out of ten components of the *DNAmCVDscore* are in common with DNAmGrimAge (DNAmCRP, DNAmPAI1, DNAmPackYears, and DNAmGDF15), and the Pearson correlation coefficient for the two epigenetic biomarkers is R = 0.56 ($P < 0.0001$). It is interesting to observe such commonalities among DNAmGrimAge and *DNAmCVDscore* even if the first was trained on all-cause mortality, whereas the second was trained on the time to cardiovascular events. These results confirm previous research indicating that heightened inflammation (associated with all the four components common in the two scores) plays a major role in biological ageing and the risk of age-related diseases, including CVDs.[31] Finally, we showed that the MRS, built directly modelling the association of CpGs on CVD risk using a single-step approach, had the worst prediction performance independently of the length of follow-up.

The results described above suggest that blood DNAm predictor of diseases may be improved. For example, the *DNAmCVDscore* can be ameliorated in different ways: i) more DNAm surrogates, such as surrogate measures for air pollution exposure, physical activity, dietary quality (e.g., adherence to the Mediterranean diet or consumption of ultra-processed food)[32–35] should be developed and included among the list of candidates in the training model; ii) refined statistical methods can be used to improve DNAm biomarkers reproducibility and reducing noise due to unmeasured batch effect;[36,37] iii) increasing the sample size of the training set by combining data from multiple cohorts and different countries, possibly modelling country-specific risk factors to improve results generalizability.

Also, we showed that, although *DNAmCVDscore* is not directly trained on age, it is correlated with chronological age (R = 0.41, $P < 0.0001$) and epigenetic clocks. These results further support the idea that susceptibility due to increasing ageing is included in the *DNAmCVDscore*, even if chronological age (or epigenetic clocks) does not directly contribute to it. Further, we demonstrated the usefulness of DNAm surrogate biomarkers in investigating COVID-19 susceptibility and severity, showing that *DNAmBMI* was associated with case-control status, while measured BMI was not, and that *DNAmCRP* outperformed blood measured CRP in predicting disease severity. Finally, we showed that *DNAmCVDscore* is higher in COVID-19 patients than in controls (hospitalised with respiratory problems) and that a higher *DNAmCVDscore* is associated with a worse prognosis (according to the

GRAM score) after COVID-19 infection. These results support recent literature suggesting COVID-19 shares direct and indirect determinants (i.e., ethnicity, socio-economic status) with other NCDs, supporting the concept of COVID-19 as a syndemic[38,39] with implications for restrictions and prevention strategies.

Study limitations and conclusions

We developed a combined biomarker as a linear combination of DNAm surrogates, named *DNAmCVDscore*, with high performance in predicting short-term cardiovascular events outperforming current state-of-art CVD prediction models based on traditional risk factors, and DNAm scores based built using a single-step approach. This work has limitations. Both the training and testing sets for the time from recruitment to the cardiovascular events come from Italian population studies, and the predictive performance for long-term CVD was poor. Further, we deliberately chose not to include the effect of chronological age and sex in the *DNAmCVDscore* because of the different study designs of the training and validation datasets. We discussed previously how *DNAmCVDscore* could be refined by re-training the model after increasing the sample size and using updated analytical methods. However, this work provides a proof of concept about the effectiveness of the described methodology based on DNAm surrogate biomarkers. Developing biomarkers for screening, entirely based on blood data, without the need for additional information or invasive measurements, would be significant for disease burden from a public health perspective. Also, our results encourage testing this approach for other NCD diseases (cancer, mental diseases, neurodegenerative diseases, respiratory problems, hearing and taste loss, etc.) by training and developing DNAm surrogates for disease-specific risk factors and exposures.

Fundings

This work was supported by the Programma Operativo Nazionale (PON), Ricerca e Innovazione 2014-2020, Attrazione e Mobilità Internazionale (AIM) code AIM1874325-2 to G.F., the AIRC grant (Progetto IG2013 N.14410, to C.S.) for DNA methylation experiments in EPIC Italy, the H2020 'EXPOsOMICS' and 'Lifepath' grants to P.V. for DNA methylation experiments in EXPOsOMICS CVD and TILDA. O.R. was supported by a UKRI Future Leaders Fellowship (MR/S03532X/1). G.S. is supported by 'Fondo di Ateneo per la ricerca 2019', University of Sassari, Italy. Understanding Society is an initiative funded by the Economic and Social Research Council and various Government Departments, with scientific leadership by the Institute for Social and Economic Research, University of Essex, and survey delivery by NatCen Social Research and Kantar Public. The research data are distributed by the UK Data Service. These data are from Understanding Society: The UK Household Longitudinal Study, which is led by the Institute for Social and Economic Research at the University of Essex and funded by the Economic and Social Research Council (Grant Number: ES/M008592/1). Data governance was provided by the METADAC data access committee, funded by ESRC, Wellcome, and MRC. (2015-2018: Grant Number MR/N01104X/1 2018-2020: Grant Number ES/S008349/1).

Ethics

This study complies with the principles of the Declaration of Helsinki; all participants signed informed consent; national ethics committees approved EPIC Italy and TILDA studies as reported elsewhere.[40] The Understanding Society study has been approved by the University of Essex ethics committee. Approval from the National Research Ethics Service was obtained for the collection of biosocial data by trained nurses in wave 2 and of the main Understanding Society survey. This study follows the STrengthening the Reporting of OBservational studies in Epidemiology (STROBE) guidelines.

Author contribution

A.C. and G.F. conceived the study and performed statistical analyses. F.I. supervised and verified analytical methods. C.M.C., O.R., P.V., G.S., and R.A.K. supervised the findings of this work. S.Po. L.I., and A.H.O. performed and supervised laboratory experiments. All authors discussed the results and contributed writing the final manuscript.

Data availability

DNAm data used in this study are available at the GEO repository under accession numbers GSE51032 (EPIC Italy) and GSE174818 (COVID-19 case-control). DNAm data from TILDA and EXPOsOMICS are available upon request at the studies PIs. Understanding Society data were collected by NatCen, and the genome-wide scan data were analysed by the Wellcome Trust Sanger Institute. Information on how to access the data can be found on the Understanding Society website https://www.understandingsociety.ac.uk/.

An R script for computing the DNAm surrogates for BMI, CRP, fasting glucose and insulin, HDL cholesterol, triglycerides, PAI-1, Platelet tissue factor (CD142), systolic blood pressure, and the composite *DNAmCVDscore* is available in the **Supplementary material**.

Conflict of interest statement

The authors declare no conflict of interest.

References

1. Fernández-Sanlés A, Sayols-Baixeras S, Subirana I, et al. DNA methylation biomarkers of myocardial infarction and cardiovascular disease. *Clin Epigenetics*. 2021;

2. Westerman K, Fernández-Sanlés A, Patil P, et al. Epigenomic assessment of cardiovascular disease risk and interactions with traditional risk metrics. *J Am Heart Assoc*. 2020;

3. Hidalgo BA, Minniefield B, Patki A, et al. A 6-CpG validated methylation risk score model for metabolic syndrome: The HyperGEN and GOLDN studies. *PLoS One*. 2021;

4. Odintsova V V., Rebattu V, Hagenbeek FA, et al. Predicting Complex Traits and Exposures From

Polygenic Scores and Blood and Buccal DNA Methylation Profiles. *Front Psychiatry*. 2021;

5. Zheng Y, Chen Z, Pearson T, Zhao J, Hu H, Prosperi M. Design and methodology challenges of environment-wide association studies: A systematic review. *Environ. Res.* 2020.
6. Allione A, Marcon F, Fiorito G, et al. Novel epigenetic changes unveiled by monozygotic twins discordant for smoking habits. *PLoS One*. 2015;**10**(6).
7. Guida F, Sandanger TM, Castagné R, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum Mol Genet*. 2015;
8. Guarrera S, Fiorito G, Onland-Moret NC, et al. Gene-specific DNA methylation profiles and LINE-1 hypomethylation are associated with myocardial infarction risk. *Clin Epigenetics*. 2015;**7**(1).
9. D'Agostino RB, Vasan RS, Pencina MJ, et al. General cardiovascular risk profile for use in primary care: The Framingham heart study. *Circulation*. 2008;
10. Cardiovascular S working group and E. SCORE2 risk prediction algorithms: New models to estimate 10-year risk of cardiovascular disease in Europe. *Eur Heart J*. 2021;
11. Lu AT, Quach A, Wilson JG, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)*. 2019;
12. Levine ME, Lu AT, Quach A, et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)*. 2018;
13. Belsky DW, Huffman KM, Pieper CF, Shalev I, Kraus WE, Anderson R. Change in the rate of biological aging in response to caloric restriction: Calerie Biobank analysis. *Journals Gerontol - Ser A Biol Sci Med Sci*. 2018;
14. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* 2018.
15. Freni-Sterrantino A, Salerno V. A Plea for the Need to Investigate the Health Effects of Gig-Economy. *Front Public Heal*. 2021;
16. Fiorito G, Polidoro S, Dugué P-A, et al. Social adversity and epigenetic aging: A multi-cohort study on socioeconomic differences in peripheral blood DNA methylation. *Sci Rep*. 2017;**7**(1).

17. Fiorito G, McCrory C, Robinson O, et al. Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: A multi-cohort analysis. *Aging (Albany NY).* 2019;**11**(7):2045–2070.

18. McCrory C, Fiorito G, Hernandez B, et al. GrimAge outperforms other epigenetic clocks in the prediction of age-related clinical phenotypes and all-cause mortality. *Journals Gerontol Ser A*. 2020;

19. Oblak L, Zaag J van der, Higgins-Chen AT, Levine ME, Boks MP. A systematic review of biological, social and environmental factors associated with epigenetic clock acceleration. *Ageing Res. Rev.* 2021.

20. Colicino E, Just A, Kioumourtzoglou MA, et al. Blood DNA methylation biomarkers of cumulative lead exposure in adults. *J Expo Sci Environ Epidemiol.* 2021;

21. Hillary RF, Marioni RE. MethylDetectR: a software for methylation-based health profiling. *Wellcome Open Res*. 2020;

22. Stevenson AJ, McCartney DL, Hillary RF, et al. Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clin Epigenetics*. 2020;

23. Stevenson AJ, Gadd DA, Hillary RF, et al. Creating and validating a DNA methylation-based proxy for interleukin-6. *Journals Gerontol Ser A*. 2021;

24. Zhang Y, Elgizouli M, Schöttker B, Holleczek B, Nieters A, Brenner H. Smoking-associated DNA methylation markers predict lung cancer incidence. *Clin Epigenetics*. 2016;

25. Green C, Shen X, Stevenson AJ, et al. Structural brain correlates of serum and epigenetic markers of inflammation in major depressive disorder. *Brain Behav Immun*. 2021;

26. Fiorito G, Vlaanderen J, Polidoro S, et al. Oxidative stress and inflammation mediate the effect of air pollution on cardio- and cerebrovascular disease: A prospective study in nonsmokers. *Environ Mol Mutagen*. 2018;**59**(3):234–246.

27. Benzeval M, Davillas A, Kumari M, Lynn P. Understanding Society: The UK household longitudinal study: Biomarker user guide and glossary. *Inst Soc Econ Res Univ Essex*. 2014;

28. Balnis J, Madrid A, Hogan KJ, et al. Blood DNA methylation and COVID-19 outcomes. *Clin Epigenetics*. 2021;

29. Donoghue OA, McGarrigle CA, Foley M, Fagan A, Meaney J, Kenny RA. Cohort profile update: The

Irish Longitudinal study on ageing (TILDA). *Int J Epidemiol.* 2018;

30. Battram T, Yousefi P, Crawford G, et al. The EWAS Catalog: a database of epigenome-wide association studies. *OSF Prepr.* 2021;

31. Franceschi C, Garagnani P, Parini P, Giuliani C, Santoro A. Inflammaging: a new immune–metabolic viewpoint for age-related diseases. *Nat. Rev. Endocrinol.* 2018.

32. Fiorito G, Vlaanderen J, Polidoro S, et al. Oxidative stress and inflammation mediate the effect of air pollution on cardio- and cerebrovascular disease: A prospective study in nonsmokers. *Environ Mol Mutagen*. 2018;**59**(3):234–246.

33. Fiorito G, Caini S, Palli D, et al. DNA methylation-based biomarkers of aging were slowed down in a two-year diet and physical activity intervention trial: the DAMA study. *Aging Cell.* 2021;

34. Rider CF, Carlsten C. Air pollution and DNA methylation: Effects of exposure in humans. *Clin. Epigenetics* 2019.

35. Maugeri A, Barchitta M. How dietary factors affect dna methylation: Lesson from epidemiological studies. *Med.* 2020;

36. Pain O, Glanville KP, Hagenaars SP, et al. Evaluation of polygenic prediction methodology within a reference-standardized framework. *PLoS Genet.* 2021;

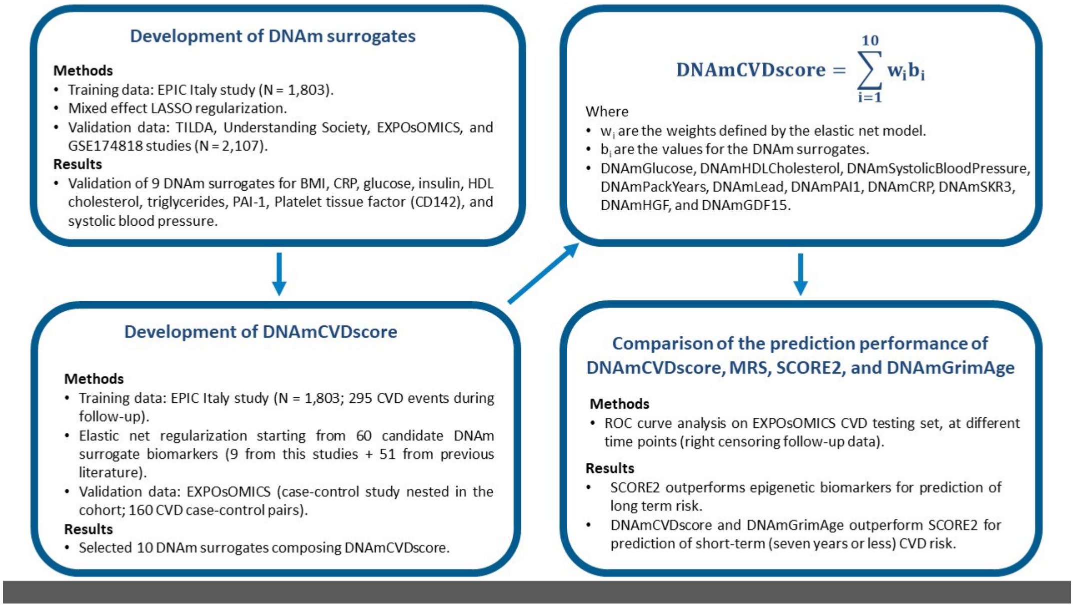
37. Higgins-Chen AT, Thrush KL, Wang Y, et al. A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking. *bioRxiv*. 2021;

38. Courtin E, Vineis P. COVID-19 as a Syndemic. *Front Public Heal.* 2021;

39. Vineis P. COVID-19 as a syndemic: from inequalities to biological embodiment. *Eur J Public Health*. 2021;

40. Fiorito G, McCrory C, Robinson O, et al. Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: A multi-cohort analysis. *Aging (Albany NY)*. 2019;**11**(7):2045–2070.

Figure legends



Development of DNAm surrogates

Methods

- Training data: EPIC Italy study (N = 1,803).
- Mixed effect LASSO regularization.
- Validation data: TILDA, Understanding Society, EXPOsOMICS, and GSE174818 studies (N = 2,107).

Results

- Validation of 9 DNAm surrogates for BMI, CRP, glucose, insulin, HDL cholesterol, triglycerides, PAI-1, Platelet tissue factor (CD142), and systolic blood pressure.

$$\text{DNAmCVDscore} = \sum_{i=1}^{10} w_i b_i$$

Where

- w_i are the weights defined by the elastic net model.
- b_i are the values for the DNAm surrogates.
- DNAmGlucose, DNAmHDLCholesterol, DNAmSystolicBloodPressure, DNAmPackYears, DNAmLead, DNAmPAI1, DNAmCRP, DNAmSKR3, DNAmHGF, and DNAmGDF15.

Development of DNAmCVDscore

Methods

- Training data: EPIC Italy study (N = 1,803; 295 CVD events during follow-up).
- Elastic net regularization starting from 60 candidate DNAm surrogate biomarkers (9 from this studies + 51 from previous literature).
- Validation data: EXPOsOMICS (case-control study nested in the cohort; 160 CVD case-control pairs).

Results

- Selected 10 DNAm surrogates composing DNAmCVDscore.

Comparison of the prediction performance of DNAmCVDscore, MRS, SCORE2, and DNAmGrimAge

Methods

- ROC curve analysis on EXPOsOMICS CVD testing set, at different time points (right censoring follow-up data).

Results

- SCORE2 outperforms epigenetic biomarkers for prediction of long term risk.
- DNAmCVDscore and DNAmGrimAge outperform SCORE2 for prediction of short-term (seven years or less) CVD risk.

Figure 1. Flowchart for development and validation of *DNAmCVDscore*

Step 1: We train prediction models for developing DNAm surrogates for 13 CVD risk factors/biomarkers using data from the EPIC Italy study (*n*=1,803). We tested the validity of DNAm surrogates in four independent studies (*n*=2,107). Nine out of 13 DNAm biomarkers were validated in the testing set.

Step 2: 60 candidate DNAm surrogates (nine newly developed + 51 from the literature) were regressed against the time from recruitment in the study to cardiovascular event in EPIC Italy (*n*=1,803). The elastic net regression model selected ten DNAm surrogates as components of the *DNAmCVDscore*.

Step 3: In EXPOsOMICS CVD dataset (*n*=315), we evaluated the prediction performance of *DNAmCVDscore* at different time points (right censoring follow-up time) using logistic regression models adjusted for chronological age, gender, and centre of recruitment (matching variables in EXPOsOMICS CVD). *DNAmCVDscore* has a higher AUC for short-term cardiovascular events than for long-term CVD.

Step 4: We compared the prediction performance of *DNAmCVDscore* with previously developed composite biomarkers: MRS, DNAmGrimAge, and SCORE2. SCORE2 outperforms epigenetic predictors for long-term CVD risk (occurred more than eight years after recruitment), whereas *DNAmCVDscore* predicts short-term events (occurred within seven years after recruitment) better than other biomarkers.

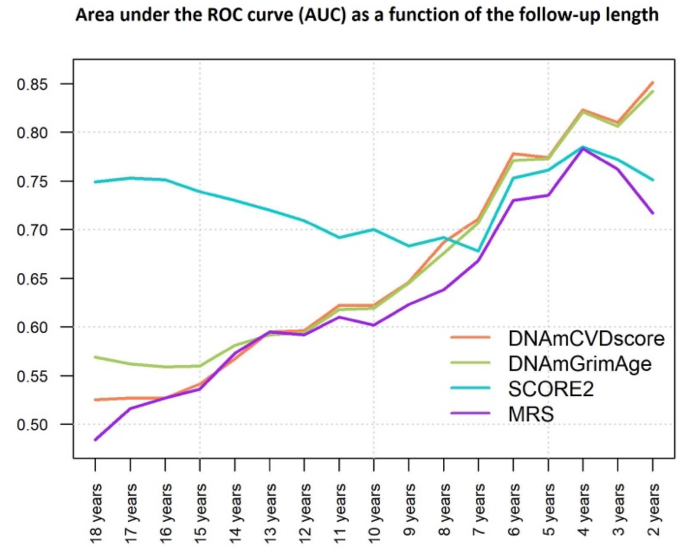


Figure 2. Comparison of the prediction performance of *DNAmCVDscore*, MRS, DNAmGrimAge, and SCORE2

Area under the ROC curve (AUC), on the y-axis, as a function of the follow-up length (x-axis) for the four composite biomarkers investigated in this study. MRS has the worst prediction performance at each time points. SCORE2 outperforms epigenetic predictors for long-term CVD risk (occurred more than eight years after recruitment), whereas *DNAmCVDscore* and DNAmGrimAge predict short-term risk (CVD events within seven years after recruitment or less) better than the other biomarkers.

Tables

Table 1: Study sample descriptive table.

Study name	Description	Country	N	Age means (min; max)	Female %	Training/Testing set
EPIC Italy	Italian sub-sample of the European Investigation into Cancer and Nutrition study	Italy	1,803	53.3 (34.7; 74.9)	62%	Training set for DNAm surrogates and *DNAmCVDscore*
EXPOsOMICS CVD	Case-control study on CVD nested in the EPIC Italy cohort	Italy	315	54.9 (35.2; 69.3)	53%	Validation set for DNAm surrogates and *DNAmCVDscore*
Understanding Society	The United Kingdom Household Panel Study (UKHLS)	UK	1,174	58.0 (28.0; 98.0)	59%	Validation set for DNAm surrogates
TILDA	The Irish Longitudinal Study on Ageing	Ireland	490	62.1 (50.0; 80.0)	50%	Validation set for DNAm surrogates
GSE174818	Case-control study on COVID-19 susceptibility and progression	USA	127	61.8 (21.0; 90.0)	40%	Validation set for DNAm surrogates

Table 2: List of newly developed DNAm surrogate biomarkers. For each candidate marker, we reported: the model used to extract significant CpGs (LASSO or mixed effect LASSO depending on the association with the centre of recruitment), the number of CpGs whose linear combination constitute the best marker prediction, the Pearson correlation coefficient and p-value in the primary test set (random 25% of EPIC Italy samples), the Pearson correlation coefficient and p-value in independent test sets (random effect meta-analysis across studies). Nine out of 13 DNAm surrogates for CVD risk factors/markers were validated in independent testing set (p-value for the Pearson correlation test lower than 0.05). R script to compute DNAm surrogates in independent datasets (together with the list of CpGs) is provided in the **Supplementary material**.

Model training, EPIC Italy training set, *n*=1,352			**Results on EPIC ITALY test set *n*=451**		**Results on the validation set**			
Risk factor/biomarker	**Model type**	**Number of CpGs**	**Pearson R**	**P**	**Validation datasets (N)**	**Pearson R**	**P**	**Validated DNAm surrogate**
BMI	Mixed-effect LASSO	405	0.59	<0.0001	US, TILDA, EXPOsOMICS, GSE174848 (2,045)	0.27	<0.0001	Yes
CRP	LASSO	265	0.57	<0.0001	US, TILDA, EXPOsOMICS, GSE174849 (1,893)	0.23	<0.0001	Yes
D-dimer	LASSO	483	0.72	<0.0001	EXPOsOMICS, GSE174848 (248)	0.17	0.56	No
Diastolic blood pressure	Mixed-effect LASSO	401	0.57	<0.0001	EXPOsOMICS, TILDA (772)	0.10	0.36	No
Glucose	Mixed-effect LASSO	354	0.67	<0.0001	EXPOsOMICS, TILDA, US (1,810)	0.28	0.007	Yes
HDL cholesterol	Mixed-effect LASSO	151	0.58	<0.0001	EXPOsOMICS, TILDA, US (1,829)	0.08	0.001	Yes
Insulin	Mixed-effect LASSO	574	0.66	<0.0001	EXPOsOMICS (170)	0.44	<0.0001	Yes
LDL cholesterol	Mixed-effect LASSO	368	0.62	<0.0001	EXPOsOMICS, TILDA (661)	0.15	0.36	No
PAI-1	LASSO	90	0.43	<0.0001	EXPOsOMICS (171)	0.28	0.0001	Yes
Systolic blood pressure	Mixed-effect LASSO	275	0.64	<0.0001	EXPOsOMICS, TILDA (772)	0.28	0.001	Yes
Tissue Factor (CD142)	Mixed-effect LASSO	197	0.62	<0.0001	EXPOsOMICS (171)	0.16	0.03	Yes
Total Cholesterol	Mixed-effect LASSO	257	0.53	<0.0001	EXPOsOMICS, TILDA, US (1,830)	0.13	0.14	No
Triglycerides	LASSO	471	0.73	<0.0001	EXPOsOMICS, TILDA (661)	0.22	0.0003	Yes

Table 3: DNAm surrogates composing the DNAmCVD score. *DNAmCVDscore* is computed as a linear combination of standardised (mean=0, variance=1) DNAm surrogates with weights listed in the coefficient column. All biomarkers but DNAmHDL have positive coefficients (higher CVD risk associated with a higher value for the biomarker).

Study	DNAm surrogate biomarker	*DNAmCVDscore* coefficient	Original biomarker/risk factor
This study	DNAmGlucose	0.0329	Blood glucose
This study	DNAmHDL	-0.4473	Blood HDL cholesterol
This study	DNAmSBP	0.1420	Systolic blood pressure
This study	DNAmCRP	0.0276	Blood C-reactive protein
This study	DNAmPAI1	0.1679	Blood Plasminogen activator inhibitor 1
Hillary et al. 2020	DNAmSKR3	0.0362	Blood Serine/threonine-protein kinase receptor R3
Hillary et al. 2020	DNAmHGF	0.0371	Blood Hepatocyte growth factor
Colicino et al. 2021	DNAmLeadPatella	0.0402	Lead levels in Patella's bone
Lu et al. 2019	DNAmGDF15	0.0947	Blood Growth Differentiation Factor 15
Lu et al. 2019	DNAmPACKYRS	0.1192	Smoking pack years

Table 4: Results from the ROC curve analyses. For each composite biomarker, we report the AUC (95% CI), sensitivity and specificity according to the best threshold (minimising the distance from the top left corner of the ROC curve) derived from logistic regression model adjusted for matching parameters (age, sex, and centre of recruitment). Predictive performance was evaluated at different time points, right-censoring the follow-up time in the range of 18 to two years, with one year interval. The highest AUC for each time point is highlighted in red.

Follow-up time	# events	*DNAmCVDscore*		DNAmGrimAge		SCORE2		MRS	
		AUC (95% CI)	sensitivity; specificity	AUC (95% CI)	sensitivity; specificity	AUC (95% CI)	sensitivity; specificity	AUC (95% CI)	sensitivity; specificity
18 years	160	0.525 (0.461; 0.589)	0.477; 0.580	0.569 (0.505; 0.632)	0.647; 0.482	0.749 (0.696; 0.803)	0.719; 0.698	0.484 (0.420; 0.548)	0.523; 0.500
17 years	159	0.527 (0.463; 0.591)	0.520; 0.553	0.562 (0.498; 0.626)	0.669; 0.472	0.753 (0.700; 0.806)	0.753; 0.671	0.516 (0.452; 0.580)	0.545; 0.484
16 years	158	0.527 (0.463; 0.591)	0.513; 0.554	0.559 (0.496; 0.623)	0.583; 0.572	0.751 (0.698; 0.805)	0.718; 0.692	0.527 (0.463; 0.591)	0.449; 0.610
15 years	149	0.541 (0.477; 0.604)	0.518; 0.577	0.560 (0.496; 0.623)	0.681; 0.450	0.739 (0.684; 0.794)	0.741; 0.671	0.536 (0.472; 0.600)	0.509; 0.477
14 years	139	0.567 (0.504; 0.630)	0.483; 0.662	0.581 (0.518; 0.644)	0.494; 0.662	0.730 (0.674; 0.786)	0.705; 0.691	0.573 (0.509; 0.636)	0.682; 0.475
13 years	123	0.595 (0.531; 0.659)	0.625; 0.537	0.592 (0.528; 0.656)	0.526; 0.667	0.720 (0.663; 0.777)	0.662; 0.699	0.595 (0.531; 0.659)	0.531; 0.650
12 years	111	0.596 (0.530; 0.662)	0.529; 0.649	0.594 (0.529; 0.659)	0.510; 0.694	0.709 (0.650; 0.768)	0.588; 0.775	0.592 (0.525; 0.658)	0.588; 0.613
11 years	95	0.622 (0.554; 0.690)	0.536; 0.705	0.618 (0.549; 0.686)	0.568; 0.663	0.693 (0.630; 0.755)	0.582; 0.758	0.610 (0.540; 0.680)	0.536; 0.695
10 years	84	0.622 (0.551; 0.693)	0.580; 0.607	0.619 (0.548; 0.691)	0.576; 0.631	0.700 (0.636; 0.765)	0.550; 0.845	0.602 (0.530; 0.673)	0.541; 0.667
9 years	67	0.647 (0.571; 0.722)	0.601; 0.612	0.645 (0.569; 0.721)	0.649; 0.582	0.683 (0.611; 0.754)	0.609; 0.746	0.623 (0.546; 0.700)	0.581; 0.627
8 years	55	0.687 (0.611; 0.762)	0.569; 0.709	0.676 (0.601; 0.752)	0.627; 0.673	0.692 (0.618; 0.766)	0.623; 0.746	0.638 (0.554; 0.722)	0.635; 0.618
7 years	37	0.711 (0.626; 0.796)	0.615; 0.703	0.707 (0.625; 0.789)	0.554; 0.811	0.678 (0.587; 0.770)	0.622; 0.703	0.668 (0.566; 0.770)	0.669; 0.622
6 years	28	0.778 (0.702; 0.854)	0.673; 0.786	0.771 (0.688; 0.853)	0.634; 0.821	0.753 (0.668; 0.839)	0.686; 0.786	0.730 (0.616; 0.844)	0.669; 0.750
5 years	23	0.774 (0.695; 0.853)	0.623; 0.826	0.773 (0.694; 0.851)	0.634; 0.826	0.761 (0.665; 0.856)	0.723; 0.783	0.735 (0.618; 0.852)	0.688; 0.783
4 years	16	0.823 (0.746; 0.899)	0.799; 0.688	0.821 (0.736; 0.906)	0.632; 0.938	0.785 (0.653; 0.917)	0.722; 0.875	0.783 (0.649; 0.917)	0.689; 0.875
3 years	13	0.811 (0.720; 0.901)	0.642; 0.846	0.806 (0.724; 0.888)	0.652; 0.846	0.772 (0.620; 0.923)	0.772; 0.846	0.762 (0.628; 0.896)	0.768; 0.769
2 years	7	0.851 (0.767; 0.935)	0.799; 0.857	0.842 (0.745; 0.939)	0.737; 0.857	0.751 (0.575; 0.926)	0.656; 0.857	0.717 (0.562; 0.872)	0.779; 0.571

Supplementary material for: 'A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events' by Andrea Cappozzo[1], Cathal McCrory[2], Oliver Robinson[3], Anna Freni Sterrantino[3,4], Carlotta Sacerdote[5], Vittorio Krogh[6], Salvatore Panico[7], Rosario Tumino[8], Licia Iacoviello[9,10], Fulvio Ricceri[11,12], Sabina Sieri[6], Paolo Chiodini[13], Rose Anne Kenny[2], Aisling O'Halloran[2], Silvia Polidoro[14], Giuliana Solinas[15], Paolo Vineis[3], Francesca Ieva[1,16], Giovanni Fiorito[2,3,15, *]

Subject recruitment, demographic/lifestyle variables acquisition and DNA methylation measurements

EPIC Italy - Study participants were drawn from the Italian component of the European Prospective Investigation into Cancer and Nutrition (EPIC) cohort,[1] a large general population cohort consisting of ~520,000 individuals, with standardized lifestyle and personal history questionnaires, measured anthropometric data and blood samples collected for DNA extraction. Smoking habits data were collected at study enrolment through the use of a questionnaire, and participants were categorized as 'never', 'former' and 'current' smokers. Height and weight were measured at enrolment with a standardized protocol, and body mass index (BMI) was calculated as the ratio between weight in kg and squared height in meters, treated as a continuous variable. Methods for measurements of blood pressure, cholesterol levels, triglycerides, and PAI1, D-dimer, and CRP are reported elsewhere.[2]

This study sample includes individuals from five nested case-control studies on breast, colon, and lung cancer, lymphomas, and myocardial infarction.[3,4] Participants were sampled from the 47,749 participants of the EPIC Italy cohort and included 354 incident breast cancer cases, 169 incident colon cancer cases, 192 incident lung cancer cases, 72 incident lymphoma cases, 295 incident myocardial infarction cases and their 1,079 matched controls. Controls were individually matched on age (±5 years), sex, the season of blood collection, centre, and length of follow-up. Since the disease diagnoses were made years after the blood draw, all the subjects were treated as healthy at recruitment. In the time to CVD event analyses, the follow-up time was (right) censored at the time of diagnosis for incident cancer cases. Overall, after DNA methylation data quality controls and sample filtering 1,803 EPIC Italian subjects were used in this analysis.

EXPOsOMICS CVD - Study participants pertain to the EPIC Italy cohort. 160 incident CVD cases and one-to-one matched controls (not overlapping with the EPIC Italy dataset described hereafter) were extracted using the incident density sampling method.[5] After DNAm data quality control and sample filtering, 315 individuals were included in this study.

For the microarray, DNA samples were extracted from buffy coats using the QIAsymphony DNA Midi Kit (Qiagen, Crawley, UK). Bisulphite conversion of 500 ng of each sample was performed using the EZ-96 DNA Methylation-Gold™ Kit according to the manufacturer's protocol (Zymo Research, Orange, CA). Then, bisulphite-converted DNA was used for hybridization on the Infinium HumanMethylation 450 BeadChip, following the Illumina Infinium HD Methylation protocol. Briefly, a whole genome amplification step was followed by enzymatic end-point fragmentation and hybridization to HumanMethylation 450 BeadChips at 48 ° C for 17 h, followed by single nucleotide extension. The incorporated nucleotides were labeled with biotin (ddCTP and ddGTP) and 2,4-dinitrophenol (DNP) (ddATP and ddTTP). After the extension step and staining, the BeadChip was washed and scanned using the Illumina HiScan SQ scanner. The intensities of the images were extracted using the GenomeStudio (v.2011.1) Methylation module (1.9.0) software, which normalizes within-sample data using different internal controls that are present on the HumanMethylation 450 BeadChip and internal background probes. The methylation score for each CpG was represented as a β-value according to the fluorescent intensity ratio representing any value between 0 (unmethylated) and 1 (completely methylated).

The Irish Longitudinal Study on Ageing (TILDA) is a large prospective cohort study examining the social, economic and health circumstances of 8,175 community-dwelling older adults aged 50 years and over resident in the Republic of Ireland. The sample was generated using a 3-stage selection process and the Irish Geodirectory as the sampling frame. The Irish Geodirectory is a comprehensive listing of all addresses in the Republic of Ireland, which is compiled by the national post service and ordnance survey Ireland. Subdivisions of district electoral divisions pre-stratified by socio-economic status, age, and geographical location, served as the primary sampling units. The second stage involved the selection of a random sample of 40 addresses from within each PSU resulting in an initial sample of 25,600 addresses. The third stage involved the recruitment of all members of the household aged 50 years and over. Consequently, the response rate was defined as the proportion of households including an eligible participant from whom an interview was successfully obtained. A response rate of 62% was achieved at the household level. There were three components to the survey. Respondents completed a computer-assisted personal interview and a separate self-completion paper and pencil module which collected information that was considered sensitive. All participants were invited to undergo an independent health assessment at one of two national centres using trained nursing staff. Blood samples were taken during the clinical assessment with the consent of participants. A more detailed exposition of study design, sample selection and protocol is available elsewhere. The present study sample included 500 healthy individuals: 125 for each of the four SES classes: stable professional, any downward mobility, any upward mobility, and stable unskilled (see socioeconomic position assessment). Buffy coat or peripheral blood mononuclear cells (PBMC) samples were available for all the individuals. Overall, after DNA methylation data quality controls and sample filtering, 490 subjects were analysed in this study.

For the microarray, DNA samples were extracted from buffy coats using the QIAGEN GENTRA AUTOPURE LS (Qiagen, Crawley, UK). Bisulphite conversion of 500 ng of each sample was performed using the EZ DNA Methylation-Lightning™ Kit according to the manufacturer's protocol (Zymo Research, Orange, CA). Then, bisulphite-converted DNA was used for hybridization on the Infinium HumanMethylation 850k BeadChip, following the Illumina Infinium HD Methylation protocol. Briefly, a whole-genome amplification step was followed by enzymatic end-point fragmentation and hybridization to HumanMethylation EPIC Chip at 48 ° C for 17 h, followed by single nucleotide extension. The incorporated nucleotides were labeled with biotin (ddCTP and ddGTP) and 2,4-dinitrophenol (DNP) (ddATP and ddTTP). After the extension step and staining, the BeadChip was washed and scanned using the Illumina HiScan SQ scanner. The intensities of the images were extracted using the GenomeStudio (v.2011.1) Methylation module (1.9.0) software, which normalizes within-sample data using different internal controls that are present on the HumanMethylation 850k BeadChip and internal background probes. The methylation score for each CpG was represented as a β-value according to the fluorescent intensity ratio representing any value between 0 (unmethylated) and 1 (completely methylated).

Understanding Society: The study sample consisted of participants from the United Kingdom Household Panel Study (UKHLS), also known as Understanding Society,[6] an ongoing longitudinal, nationally representative study of the UK, designed as a two-stage stratified random sample of the general population. While Understanding Society is a panel survey, the data used here consist of two pooled cross-sectional waves where a nurse collected blood samples from the respondents, among other physiological measures. The eligibility criteria for collecting blood samples were: (i) participation in the previous main interviews in England (had participated in all annual interviews between 1999 (BHPS wave 9) and 2011–2013 (Understanding Society wave 2 and 3); (ii) age 16 and over; (iii) living in England, Wales, or Scotland. From the potential pool of 6,337 survey respondents, eligibility requirements for epigenetic analyses meant that the samples for DNA methylation measurement were restricted to participants of white ethnicity, resulting in 1,175 subjects; more details can be found elsewhere.[7] Details about laboratory analyses for DNAm and how to access raw data can be found at the Understanding Society web site (https://www.understandingsociety.ac.uk/documentation/mainstage/dataset-documentation/variable/epigenetics).

For the GSE174818 (Covid-19 case-control) study, details of sample characteristics and laboratory methods for DNAm and biomarker analyses are described in the original publication.[8]

DNA methylation data pre-processing and quality controls

For all the studies, raw DNAm data were pre-processed and normalized using in-house software written for the R statistical computing environment, including background and colour bias correction, quantile normalization, and BMIQ procedure to remove type I/type II probes bias, as described elsewhere.[9] DNAm levels were expressed as the ratio of the intensities of methylated cytosines over the total intensities (β values). Samples were excluded if the bisulphite conversion control fluorescence intensity was less than 10,000 for both type I and type II probes. Methylation measures were set to missing if the detection

P-value was greater than 0.01. Additionally, the set of cross-reactive and/or polymorphic (with minor allele frequency greater than 0.01 in Europeans) CpGs (*n*=39,238) described by Chen et al.[10] was excluded due to the low reliability of methylation measure.

The Fernández-Sanlés methylation risk score (MRS) was computed as a standardised weighted sum of 34 CpG sites, with weights defined by the estimates described by the authors in the Supplementary material of their original publication.[11] DNAmGrimAge and other epigenetic clocks were computed using Steve Horvath online DNAmAge calculator (https://horvath.genetics.ucla.edu/html/dnamage/).

Additional statistical analyses

Development and validation of DNAm surrogates. We developed DNAm surrogates for BMI, systolic and diastolic blood pressure, and ten blood measured biomarkers: total cholesterol, HDL cholesterol, LDL cholesterol, triglycerides, Plasminogen activator inhibitor-1 (PAI-1), C-reactive protein (CRP), D-dimer, Platelet tissue factor (a.k.a. CD142 protein), fasting glucose and insulin. We used the EPIC Italy dataset randomly split into training (*n*=1,352; 75% of the sample) and test set (*n*=451; 25% of the sample). For each risk factor/biomarker, we created a DNAm surrogate through a three-step procedure:

i) we identify risk factors/biomarkers showing significant differences across EPIC Italy centres (Turin, Varese, Naples, Ragusa) via ANOVA analyses. We employed a linear model with a random intercept component, accounting for differences across centres for this subset of biomarkers, consisting of all but PAI-1, CRP, D-dimer, and triglycerides. We used a fixed-effect linear model for the other biomarkers.

ii) log-transformed risk factors/biomarkers were regressed on DNAm through a linear model adjusted for age, gender (fixed effect), and centre of recruitment (random effect, where necessary) to identify the top 1% ranked CpGs based on the *P-value*.

iii) DNAm surrogates of risk factors/biomarkers were constructed, regressing the response variables on the top 1% CpG sites, adjusting for sex and age. Finally, we applied L1 penalised estimation for enforcing sparsity in the regression coefficients employing the LASSO procedure[12] or the corresponding penalised mixed model[13] (for the biomarkers showing difference by centre) depending on the biomarker. For the latter, method ad-hoc R routines were devised: the source code is freely available in the form of an R package at

https://github.com/AndreaCappozzo/mixedelnet.

We validated the DNAm surrogates investigating their association (Pearson correlation coefficients) with the corresponding measured risk factor/biomarker in the EPIC Italy testing set (*n*=451, 25% of the sample), and four additional independent studies: Understanding Society (*n*=1,174), TILDA (*n*=490), EXPOsOMICS CVD (*n*=315), and GSE174818 (*n*=128). We used fixed-effect meta-analysis (inverse variance weights) to combine the results across the four validation datasets into a single estimate. As a result, we defined as 'validated' DNAm surrogates with significant associations ($P < 0.05$) in both EPIC Italy and the combined validation sets. As further validation, we investigated the correlation of our newly developed DNAm surrogates with those previously developed for BMI, HDL cholesterol,[14] and PAI-1.[15]

Derivation of DNAmCVDscore. We developed a blood DNAm based biomarker (that integrates several DNAm surrogates) for predicting the risk of future CVD events named *DNAmCVDscore*. We used Cox regression model with elastic net regularisation to regress the time from recruitment to CVD event, and for selecting the most critical features from 60 (standardised: mean=0, standard deviation=1) blood DNAm surrogates: nine newly developed within this study, 32 DNAm surrogates for blood measured (mainly inflammatory) proteins produced by Hillary and colleagues;[14,16,17] three epigenetic clocks (HorvathDNAmAge, HannumDNAmAgem and DNAmPhenoAge);[18] two DNAm surrogates for lead exposure;[19] six 'Houseman' DNAm surrogates for white blood cell (WBC) proportion;[20] and the nine components of the DNAmGrimAge clock (DNAm surrogates for smoking pack-years, telomere length, and seven blood measured proteins).[15] The best λ parameter was derived from ten-fold cross-validation to minimise the Harrel concordance C-index. The overall procedure includes 1,000 permutations using each time 80% of the whole EPIC Italy dataset (*n*=1,443). The DNAm surrogates comprising the *DNAmCVDscore* were selected among those with non-zero coefficients in at least half of the permutations. Finally, *DNAmCVDscore* was computed as a linear combination of the selected DNAm surrogates where weights correspond to the average (non-zero) coefficient among the 1,000 permutations.

Validation of DNAmCVDscore and comparison with MRS, SCORE2 and DNAmGrimAge. We validated the *DNAmCVDscore* in an independent dataset (EXPOsOMICS CVD, *n*=315), including incident CVD cases and matched controls. Since the testing set is designed as a case-control study nested in a cohort, we ran logistic regression analyses, and we evaluated the predictive performance of *DNAmCVDscore* through ROC curve analysis. We compared the performance of four logistic regression models: i) based on *DNAmCVDscore* adjusted for matching parameters (age, sex, and centre of recruitment); ii) based on MRS adjusted for matching parameters; iii) the SCORE2 prediction model based on chronological age, sex, diabetes, smoking, systolic blood pressure, total and HDL cholesterol, adjusting for matching parameters; iv) based on DNAmGrimAge adjusted for matching parameters.

To investigate the predictive performance of the four composite biomarkers at different time points, we computed the area under the ROC curve (AUC), sensitivity, and specificity as a function of the time from recruitment to diagnosis, right-censoring follow-up at constant intervals of one year from 18 to two years. Confidence intervals for AUC were computed according to De Long et al.[21]

DNAm surrogates and DNAmCVDscore vs COVID-19 case-control status and severity. As an additional sensitivity analysis, despite being out of the main scope of this work, we investigated the usefulness of using DNAm surrogate biomarkers in epidemiological studies on COVID-19 using the GSE174818 dataset (101 patients with COVID-19 infection and 26 controls hospitalised with respiratory problems). Specifically, we investigated the association of BMI and blood measured CRP with COVID-19 case-control status and severity (using the GRAM score as a proxy), and we compared the results with those obtained using their DNAm surrogates (DNAmBMI and DNAmCRP). Finally, since CVDs and COVID-19 share several risk factors[22] we investigated the association of the *DNAmCVDscore* with COVID-19 case-control status and severity. We used logistic and linear regression models adjusted for age and gender to investigate the association with case-control status and GRAM score, respectively.

Additional results

Sensitivity leave-one-out analysis. We performed a sensitivity analysis to evaluate whether one of the ten DNAm surrogate biomarkers comprising the *DNAmCVDscore* drives the results described in the previous section. First, the *DNAmCVDscore* was re-computed ten times, excluding one DNAm surrogate each time. Then, AUC and 95% CI were calculated at different time points right censoring follow-up length as described previously. The results presented in **Table S1** show that the AUC obtained using ten DNAm surrogates is generally higher than those obtained excluding one of them. However, none of the biomarkers significantly reduces the AUC when excluded (according to the DeLong test), suggesting that all the ten DNAm surrogates contribute predicting CVD events.

Correlation of DNAmCVDscore with epigenetic clocks. We computed the Pearson correlation coefficients (meta-analysis of the five studies) between *DNAmCVDscore* and previously developed epigenetic clocks. Although *DNAmCVDscore* was not explicitly trained on chronological age, it is highly correlated with age (R = 0.41) and four epigenetic clocks (R range from 0.35 to 0.56), DNAmGrimAge being the one with the highest correlation. In **Supplementary Figure 1,** we present the correlation heatmap.

DNAm surrogates and DNAmCVDscore vs COVID-19 case-control status and severity. In the GSE174818 dataset, BMI was not significantly different when comparing COVID-19 cases with matched controls whereas DNAm surrogate for BMI was significantly associated with COVID-19 case-control status: OR per one standard deviation increase = 2.64 (95% CI 1.56; 4.77, *P* = 0.0006, **Table S2**). Among COVID-19 cases, DNAmCRP outperforms blood measured

CRP in predicting disease severity (GRAM score). The increase in the GRAM score were 17.1 (8.6; 25.6, *P* = 0.0002, **Table S2**) and 9.9 (1.1; 18.7, *P* = 0.03, **Table S2**) for DNAmCRP and blood measured CRP, respectively. Finally, *DNAmCVDscore* was (borderline significantly) associated with COVID-19 case-control status: OR per one standard deviation increase = 1.84 (95% CI 0.96; 3.65, *P* = 0.07, **Table S2** and **Supplementary Figure 2A**), and GRAM score severity index. The estimate from the linear regression (interpretable as an increase in the GRAM score for one standard deviation increase in *DNAmCVDscore*) was 16.35 (95% CI 1.36; 31.04, *P* = 0.03, **Table S2** and **Supplementary Figure 2B**).

Additional references

1. Riboli E, Hunt K, Slimani N, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr*. 2002;
2. Iacoviello L, Agnoli C, Curtis A De, et al. Type 1 plasminogen activator inhibitor as a common risk factor for cancer and ischaemic vascular disease: The EPICOR study. *BMJ Open*. 2013;
3. Gagliardi A, Dugué P-A, Nøst TH, et al. Stochastic Epigenetic Mutations Are Associated with Risk of Breast Cancer, Lung Cancer, and Mature B-cell Neoplasms. *Cancer Epidemiol Biomarkers Prev*. 2020;
4. Fiorito G, Guarrera S, Valle C, et al. B-vitamins intake, DNA-methylation of One Carbon Metabolism and homocysteine pathway genes and myocardial infarction risk: The EPICOR study. *Nutr Metab Cardiovasc Dis*. 2014;**24**(5):483–488.
5. Fiorito G, Vlaanderen J, Polidoro S, et al. Oxidative stress and inflammation mediate the effect of air pollution on cardio- and cerebrovascular disease: A prospective study in nonsmokers. *Environ Mol Mutagen*. 2018;**59**(3):234–246.
6. Understanding Society: design overview. *Longit Life Course Stud*. 2012;
7. Benzeval M, Davillas A, Kumari M, Lynn P. Understanding Society: The UK household longitudinal study: Biomarker user guide and glossary. *Inst Soc Econ Res*

Univ Essex. 2014;

8. Balnis J, Madrid A, Hogan KJ, et al. Blood DNA methylation and COVID-19 outcomes. *Clin Epigenetics*. 2021;
9. Campanella G, Polidoro S, Gaetano C Di, et al. Epigenetic signatures of internal migration in Italy. *Int J Epidemiol*. 2015;44(4):1442–1449.
10. Chen YA, Lemire M, Choufani S, et al. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*. 2013;
11. Fernández-Sanlés A, Sayols-Baixeras S, Subirana I, et al. DNA methylation biomarkers of myocardial infarction and cardiovascular disease. *Clin Epigenetics*. 2021;
12. Tibshirani R. Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Ser B*. 1996;
13. Rohart F, San Cristobal M, Laurent B. Selection of fixed effects in high dimensional linear mixed models using a multicycle ECM algorithm. *Comput Stat Data Anal*. 2014;
14. Hillary RF, Marioni RE. MethylDetectR: a software for methylation-based health profiling. *Wellcome Open Res*. 2020;
15. Lu AT, Quach A, Wilson JG, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY)*. 2019;
16. Stevenson AJ, Gadd DA, Hillary RF, et al. Creating and validating a DNA methylation-based proxy for interleukin-6. *Journals Gerontol Ser A*. 2021;
17. Stevenson AJ, McCartney DL, Hillary RF, et al. Characterisation of an inflammation-related epigenetic score and its association with cognitive ability. *Clin Epigenetics*. 2020;
18. Oblak L, Zaag J van der, Higgins-Chen AT, Levine ME, Boks MP. A systematic review of biological, social and environmental factors associated with epigenetic clock acceleration. *Ageing Res. Rev.* 2021.
19. Colicino E, Just A, Kioumourtzoglou MA, et al. Blood DNA methylation biomarkers of cumulative lead exposure in adults. *J Expo Sci Environ Epidemiol*. 2021;
20. Houseman EA, Accomando WP, Koestler DC, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012;
21. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*. 1988;
22. Courtin E, Vineis P. COVID-19 as a Syndemic. *Front Public Heal*. 2021;

Figure S1: Correlation heatmap with mutual Pearson correlation coefficients among *DNAmCVDscore* and epigenetic clocks

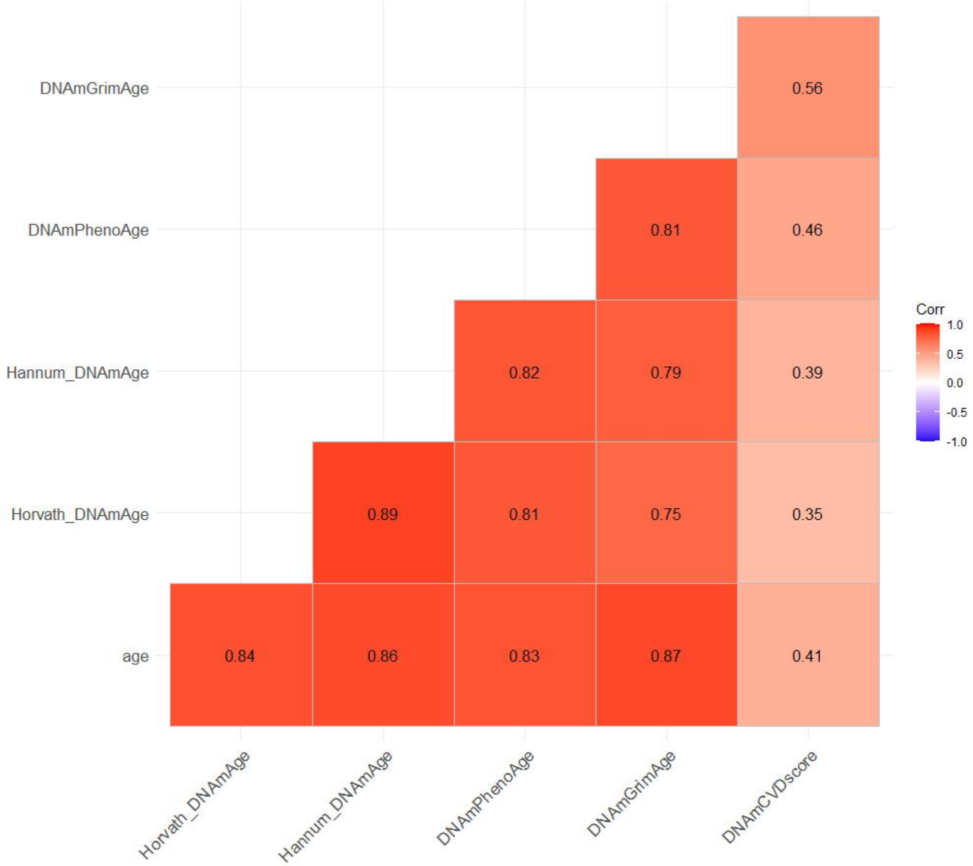
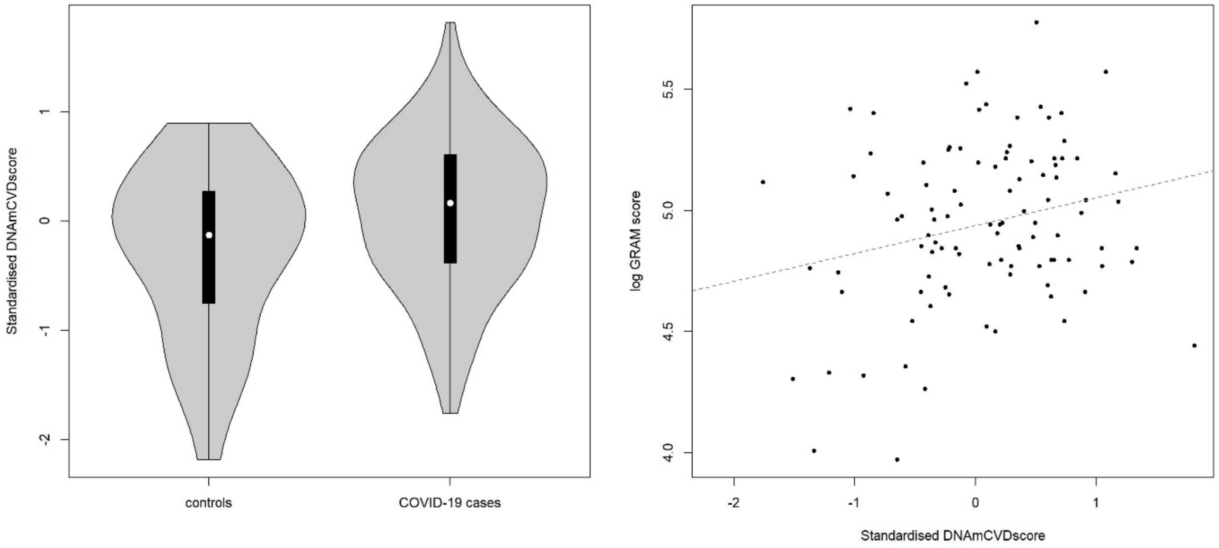


Figure S2: a) Violin plot of standardised *DNAmCVDscore* vs COVID-19 case-control status; b) scatterplot and regression line for *DNAmCVDscore* vs COVID-19 severity (GRAM score).



MOX Technical Reports, last issues

Dipartimento di Matematica
Politecnico di Milano, Via Bonardi 9 - 20133 Milano (Italy)

23/2022 Masci, C.; Ieva, F.; Paganoni, A.M.
A multinomial mixed-effects model with discrete random effects for modelling dependence across response categories

22/2022 Regazzoni, F.; Pagani, S.; Quarteroni, A.
Universal Solution Manifold Networks (USM-Nets): non-intrusive mesh-free surrogate models for problems in variable domains

21/2022 Cappozzo, A.; Ieva, F.; Fiorito, G.
A general framework for penalized mixed-effects multitask learning with applications on DNA methylation surrogate biomarkers creation

20/2022 Clementi, L.; Gregorio, C; Savarè, L.; Ieva, F; Santambrogio, M.D.; Sangalli, L.M.
A Functional Data Analysis Approach to Left Ventricular Remodeling Assessment

19/2022 Lupo Pasini, M.; Perotto, S.
Hierarchical model reduction driven by machine learning for parametric advection-diffusion-reaction problems in the presence of noisy data

18/2022 Bennati, L; Vergara, C; Giambruno, V; Fumagalli, I; Corno, A.F; Quarteroni, A; Puppini, G; L
An image-based computational fluid dynamics study of mitral regurgitation in presence of prolapse

17/2022 Regazzoni, F.
Stabilization of staggered time discretization schemes for 0D-3D fluid-structure interaction problems

14/2022 Zappon, E.; Manzoni, A.; Quarteroni A.
Efficient and certified solution of parametrized one-way coupled problems through DEIM-based data projection across non-conforming interfaces

16/2022 G. Ciaramella, T. Vanzan
Substructured Two-grid and Multi-grid Domain Decomposition Methods

15/2022 G. Ciaramella, T. Vanzan
Spectral coarse spaces for the substructured parallel Schwarz method