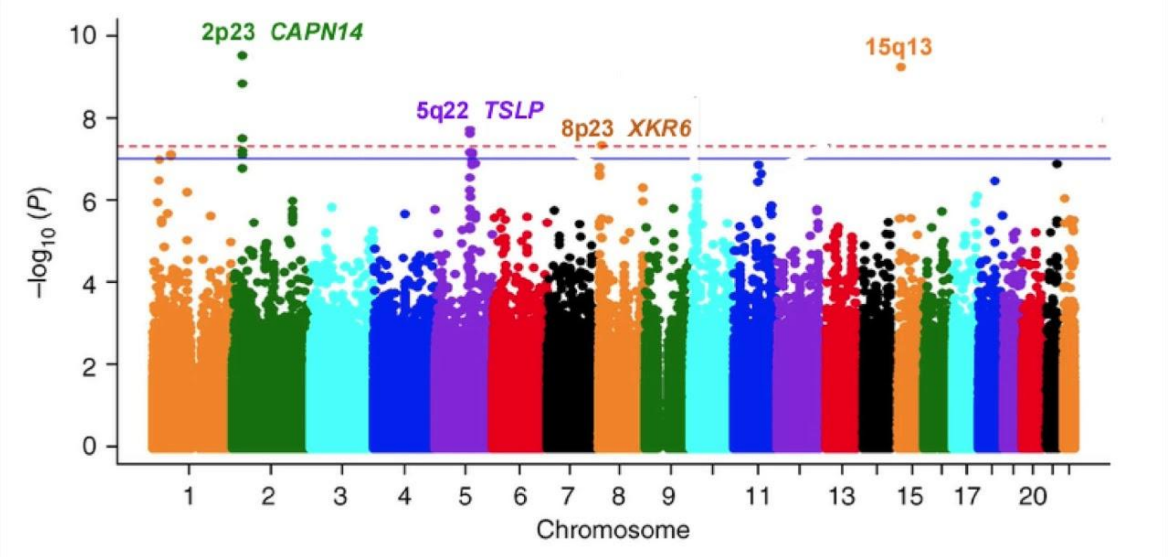
DNA methylation surrogates in epidemiological studies

Giovanni Fiorito
Clinical Bioinformatics Unit
IRCCS ‘Giannina Gaslini’ Institute

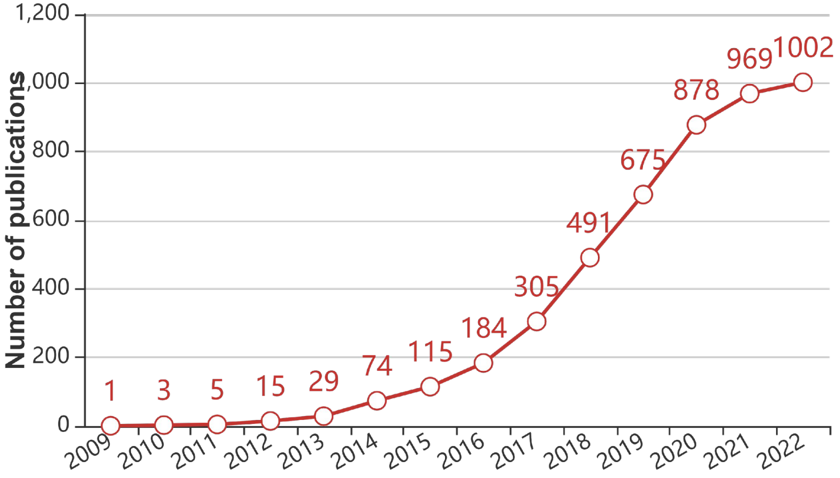
Epigenome wide association studies (EWAS)

- Research approach to identify CpG sites associated with a certain trait/disease.
- Measurement of whole-genome DNAm on individuals discordant for the trait of interest (e.g. healthy vs disease).
- One association test for each CpG site (800 K), correction for multiple testing, and replication in independent studies.

EXAMPLE OF MANHATTAN PLOT



EWAS constantly increases

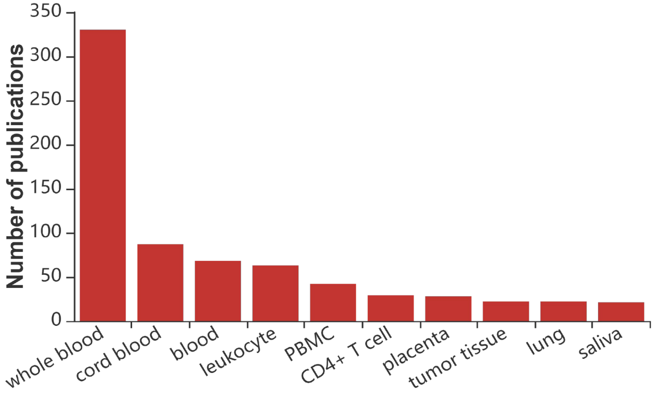


Figures adapted from EWAS ATLAS Open Platform https://ngdc.cncb.ac.cn/ewas/atlas

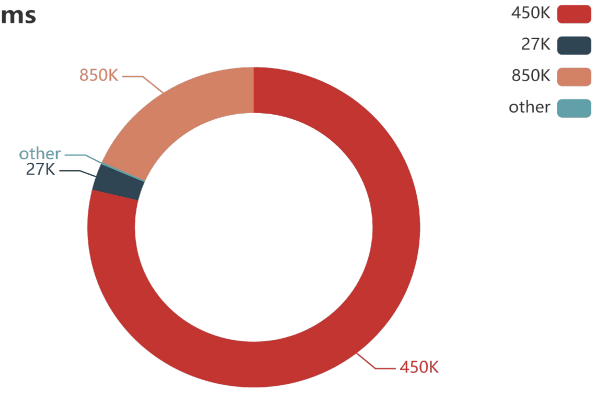
The EWAS ATLAS

- The EWAS ATLAS is database of CpG-trait associations from **‘high-quality’** EWAS:
  - 643,805 associations.
  - 301,524 CpG sites.
  - 36,041 transcripts.
  - 728 traits.
  - 199 tissues/cells.

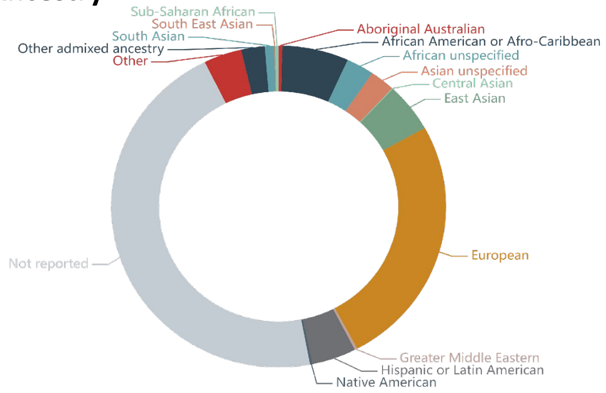
Tissues



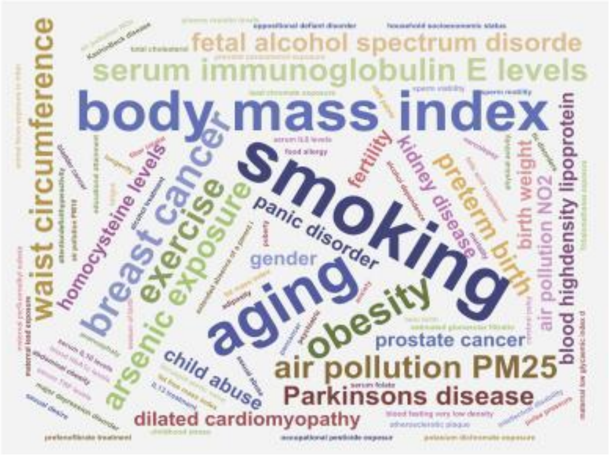
Platforms



Ancestry



Traits

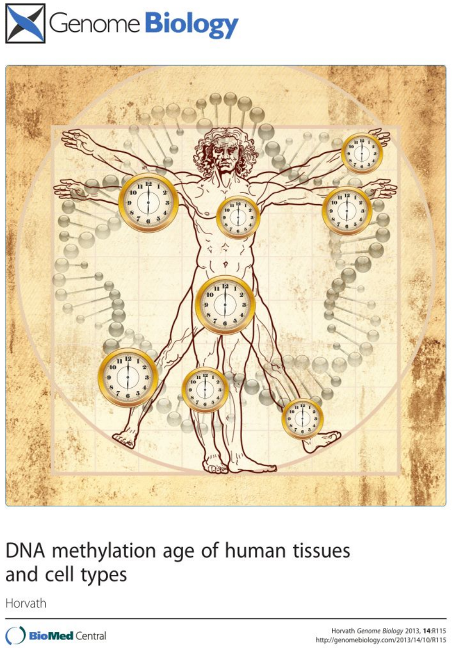


Figures adapted from EWAS ATLAS Open Platform https://ngdc.cncb.ac.cn/ewas/atlas

The concept of DNA methylation (DNAm) surrogate

- DNAm surrogate of a **trait A** (*exposure to risk factor, phenotype, disease-risk*): composite biomarker based on multiple CpG sites correlated with the trait A himself.

<u>Example: Horvath's 'original' (multi-tissue) epigenetic clock is a DNAm surrogate of chronological age</u>



Genome **Biology**

DNA methylation age of human tissues
and cell types

Horvath

BioMed Central

Horvath *Genome Biology* 2013, **14**:R115
http://genomebiology.com/2013/14/10/R115

- **Y** = chronological age; **X** = matrix of DNA methylation data.
- Prediction model (Elastic net penalized) to predict **Y** using **X**.
- Predicted Y ($\hat{\mathbf{Y}}$) is the "epigenetic age".

The concept of DNA methylation (DNAm) surrogate

DNAm surrogate of TRAIT A

- **Y** = TRAIT A; **X** = matrix of DNA methylation data.

- Prediction model (Elastic net penalized or others) to predict **Y** using **X**.

- Predicted Y ($\hat{\mathbf{Y}}$) is the DNAm surrogate for TRAIT A.

Why DNAm surrogates are useful?

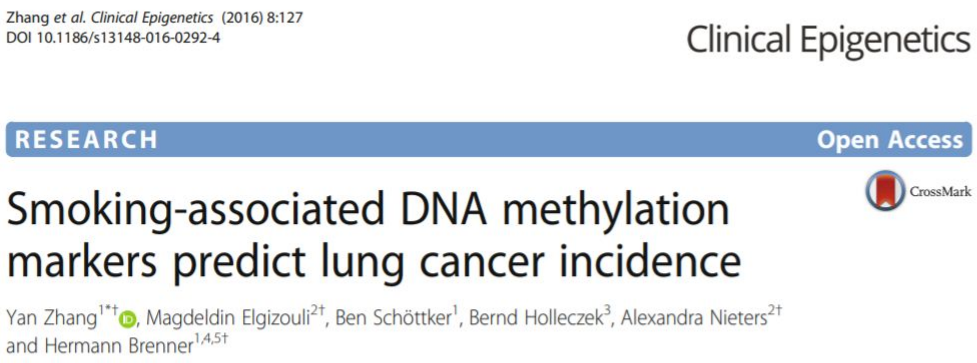
- Epigenetic clocks demonstrate that DNAm surrogates of chronological age predict aging-related diseases and longevity better than chronological age.

- The same concept can be applied to DNAm surrogates of exposure to risk factors and disease-related phenotypes.

- Useful for imputation of missing data and/or for investigating the association of an exposure with a disease, even if the exposure is not directly measured in the population study.

A couple of examples from the literature

DNAm surrogates predict diseases better than their “original” measure



Zhang *et al. Clinical Epigenetics* (2016) 8:127
DOI 10.1186/s13148-016-0292-4

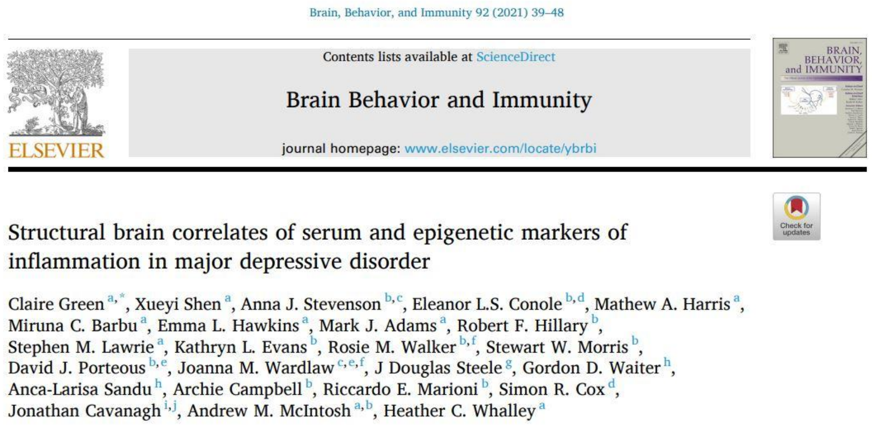
Clinical Epigenetics

RESEARCH Open Access

CrossMark

Smoking-associated DNA methylation markers predict lung cancer incidence

Yan Zhang[1*†], Magdeldin Elgizouli[2†], Ben Schöttker[1], Bernd Holleczek[3], Alexandra Nieters[2†] and Hermann Brenner[1,4,5†]



Brain, Behavior, and Immunity 92 (2021) 39–48

Contents lists available at ScienceDirect

Brain Behavior and Immunity

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

ELSEVIER

Check for updates

Structural brain correlates of serum and epigenetic markers of inflammation in major depressive disorder

Claire Green[a,*], Xueyi Shen[a], Anna J. Stevenson[b,c], Eleanor L.S. Conole[b,d], Mathew A. Harris[a], Miruna C. Barbu[a], Emma L. Hawkins[a], Mark J. Adams[a], Robert F. Hillary[b], Stephen M. Lawrie[a], Kathryn L. Evans[b], Rosie M. Walker[b,f], Stewart W. Morris[b], David J. Porteous[b,e], Joanna M. Wardlaw[c,e,f], J Douglas Steele[g], Gordon D. Waiter[h], Anca-Larisa Sandu[h], Archie Campbell[b], Riccardo E. Marioni[b], Simon R. Cox[d], Jonathan Cavanagh[i,j], Andrew M. McIntosh[a,b], Heather C. Whalley[a]

- DNAm surrogate for smoking predicts lung cancer better than self-reported smoking

- DNAm surrogate for C-reactive protein predicts brain injuries better than blood-measured CRP.

Results interpretation

- **Self-reported exposure** is often **inaccurate** (e.g. smoking, quality of diet, physical exercise), and the DNAm surrogate may be a more reliable indicator.

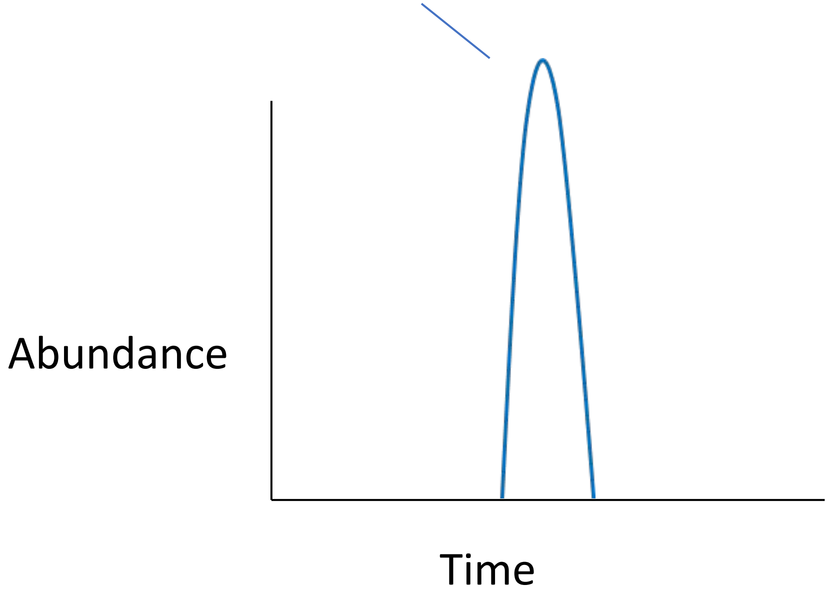
- DNAm surrogates incorporate **variability** due to individual **differential responses to exposures** and/or genetic susceptibility (same exposure - different risk profile).

- DNAm surrogates refer to **long-term and cumulative events** that have affected DNA methylation (as opposed to cross-sectional, volatile measurements of proteins).

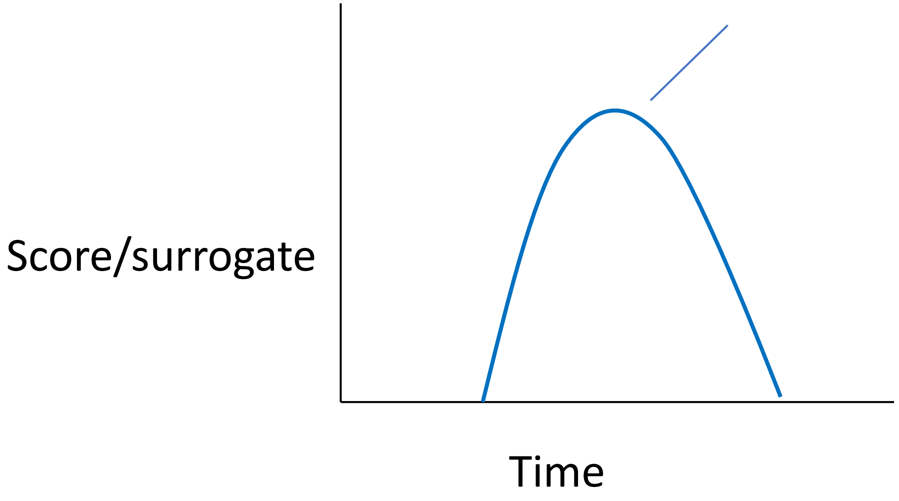
Results interpretation

DNAm surrogate of proteins have more stable longitudinal trajectory

Protein level e.g. CRP changes rapidly after injury



DNAm surrogate of CRP -
- more stable in time



Figures adapted from Gadd et al. *Epigenetic scores for the circulating proteome as tools for disease prediction*, eLife 2022

DNAm surrogates available in the literature

DNAm surrogates for lead exposures in bones

- Colicino, E. *et al.* Blood DNA methylation biomarkers of cumulative lead exposure in adults. *J. Expo. Sci. Environ. Epidemiol.* (2021) doi:10.1038/s41370-019-0183-9.

DNAm surrogates for WBC proportions in blood

- Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* (2012) doi:10.1186/1471-2105-13-86.

DNAm surrogates for ~100 blood-measured proteins

- Gadd, D. A. *et al.* Epigenetic scores for the circulating proteome as tools for disease prediction. Elife 11, (2022). doi:10.7554/eLife.71802

DNAm surrogates for ~600 EHR-derived phenotypes (medications, lab tests, diagnoses)

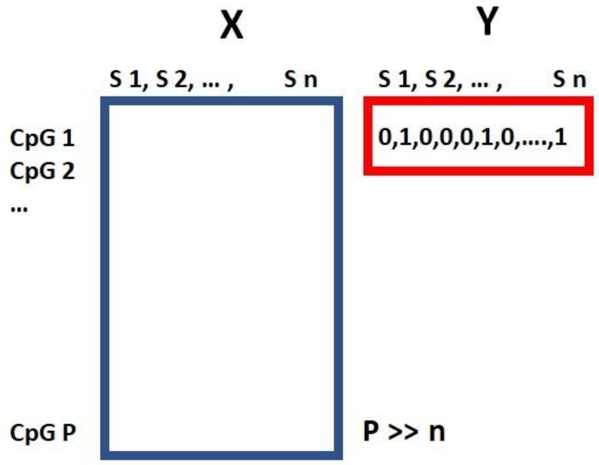
- Thompson, M. *et al.* Methylation risk scores are associated with a collection of phenotypes within electronic health record systems. *Genomic Medicine* (2022). doi:10.1038/s41525-022-00320-1

DNAm surrogates for cholesterol, insulin, glucose, blood pressure, BMI, CRP, and coagulation biomarkers.

- Cappozzo, A. *et al.* A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events. *Clinical Epigenetics* (2022). doi:10.1186/s13148-022-01341-4

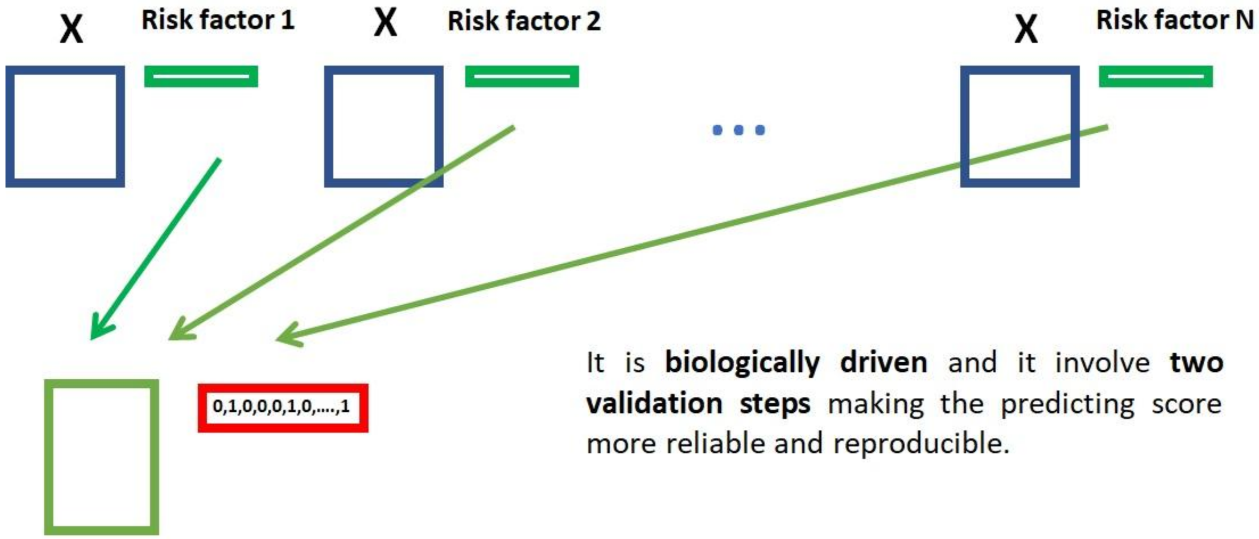
DNAm surrogates to develop disease-specific risk scores: One-step vs two-step approach

One-step approach



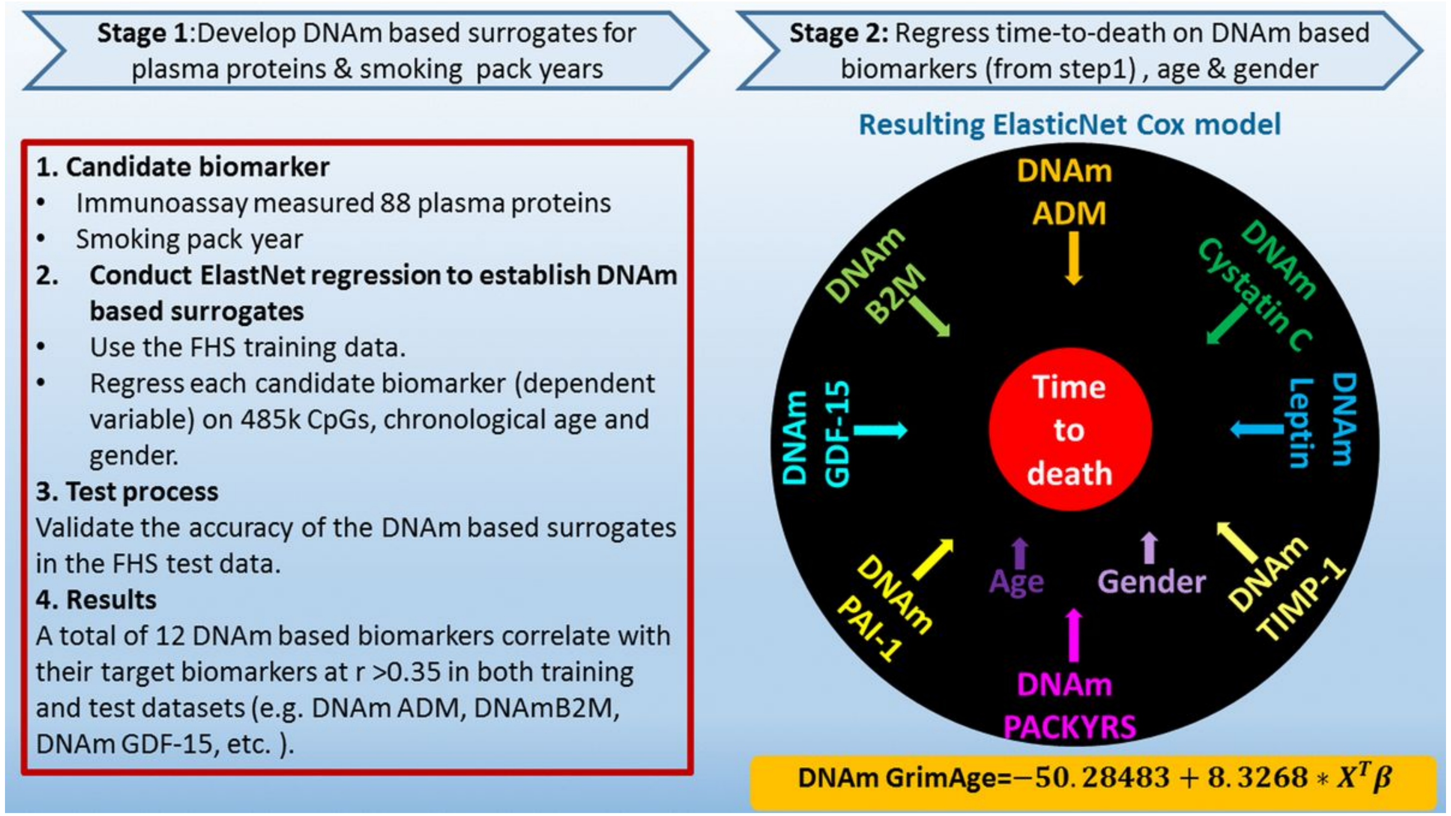
Lack of replication in independent datasets. **Negligible additional values** compared with currently used models based on traditional risk factors.

Two-step approach



It is **biologically driven** and it involve **two validation steps** making the predicting score more reliable and reproducible.

The two-step method outperforms one-step approach: example 1 (DNAmGrimAge)



Stage 1: Develop DNAm based surrogates for plasma proteins & smoking pack years

Stage 2: Regress time-to-death on DNAm based biomarkers (from step1) , age & gender

1. Candidate biomarker

- Immunoassay measured 88 plasma proteins
- Smoking pack year

2. Conduct ElastNet regression to establish DNAm based surrogates

- Use the FHS training data.
- Regress each candidate biomarker (dependent variable) on 485k CpGs, chronological age and gender.

3. Test process

Validate the accuracy of the DNAm based surrogates in the FHS test data.

4. Results

A total of 12 DNAm based biomarkers correlate with their target biomarkers at r >0.35 in both training and test datasets (e.g. DNAm ADM, DNAmB2M, DNAm GDF-15, etc.).

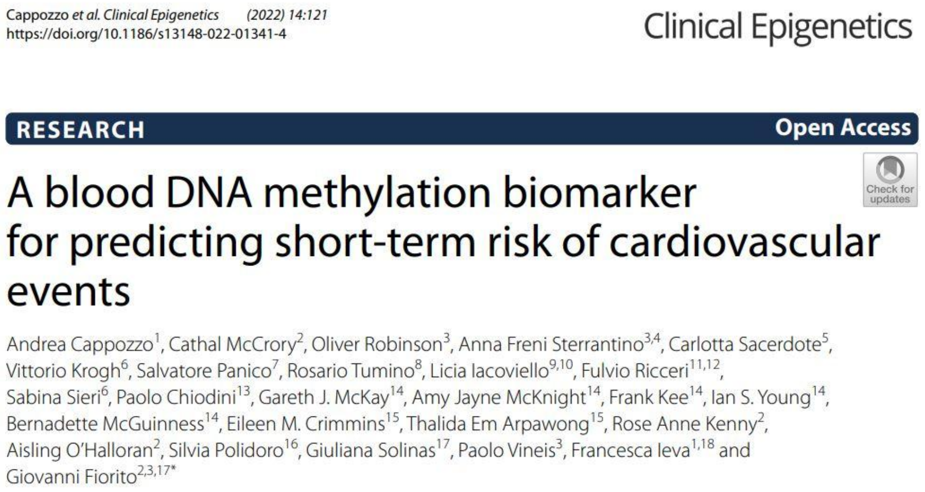
Resulting ElasticNet Cox model

DNAm GrimAge$=-50.28483 + 8.3268 * X^T\beta$

- Linear combination of DNAm surrogates trained on time to death (**Y**).
- It predicts mortality (and age-related clinical phenotypes) better than chronological age and previous epigenetic clocks.

Figure adapted from Lu et al. *DNA methylation GrimAge strongly predicts lifespan and health span*; Aging 2019

The two-step method outperforms one-step approach: example 2 (DNAmCVDscore)



Cappozzo *et al. Clinical Epigenetics* (2022) 14:121
https://doi.org/10.1186/s13148-022-01341-4

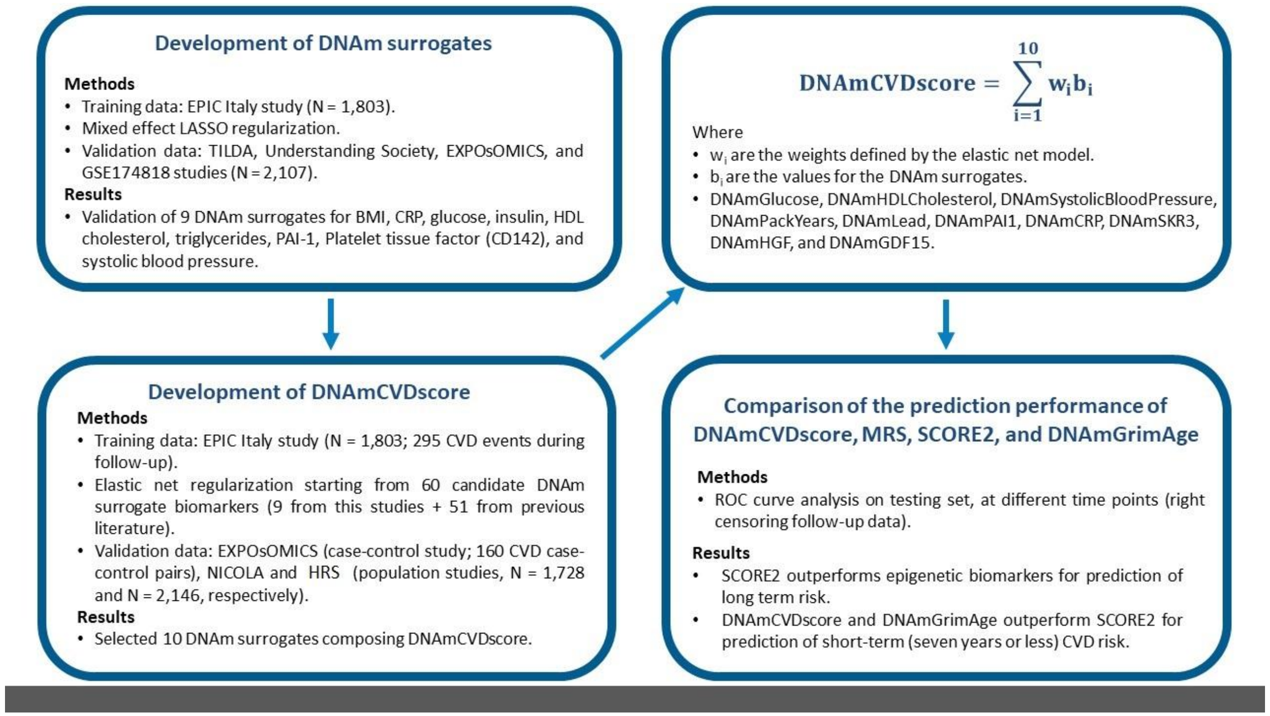
Clinical Epigenetics

RESEARCH **Open Access**

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], Gareth J. McKay[14], Amy Jayne McKnight[14], Frank Kee[14], Ian S. Young[14], Bernadette McGuinness[14], Eileen M. Crimmins[15], Thalida Em Arpawong[15], Rose Anne Kenny[2], Aisling O'Halloran[2], Silvia Polidoro[16], Giuliana Solinas[17], Paolo Vineis[3], Francesca Ieva[1,18] and Giovanni Fiorito[2,3,17*]

Workflow



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.

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

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; 160 CVD case-control pairs), NICOLA and HRS (population studies, N = 1,728 and N = 2,146, respectively).

Results

- Selected 10 DNAm surrogates composing DNAmCVDscore.

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

Methods

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

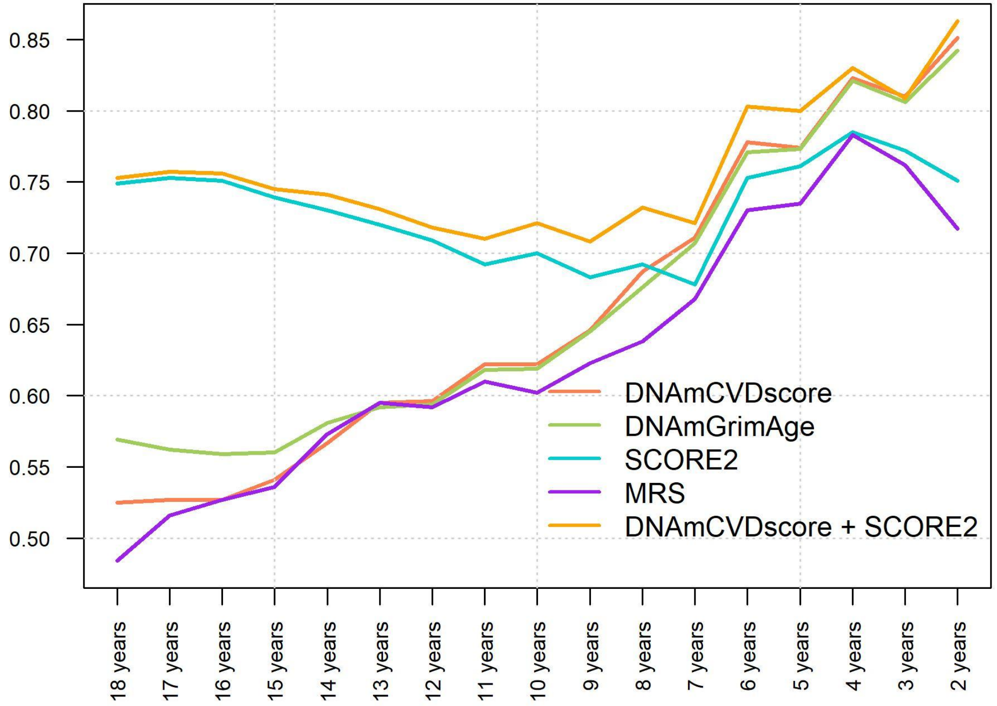
Results

- 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 from Cappozzo et al. *A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events*; CLEP 2019

Results

AUC as a function of the follow-up length

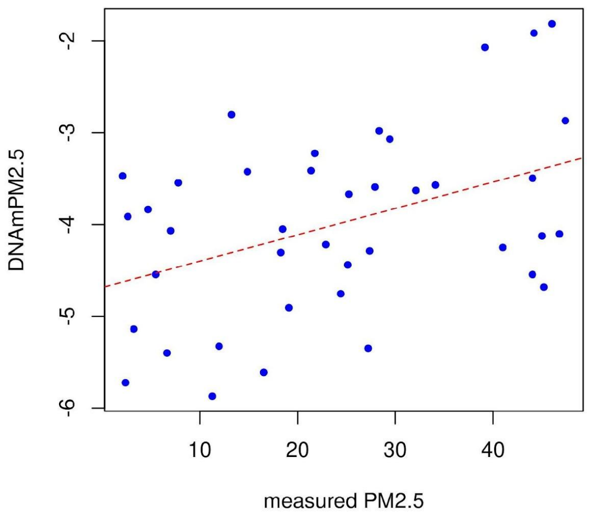


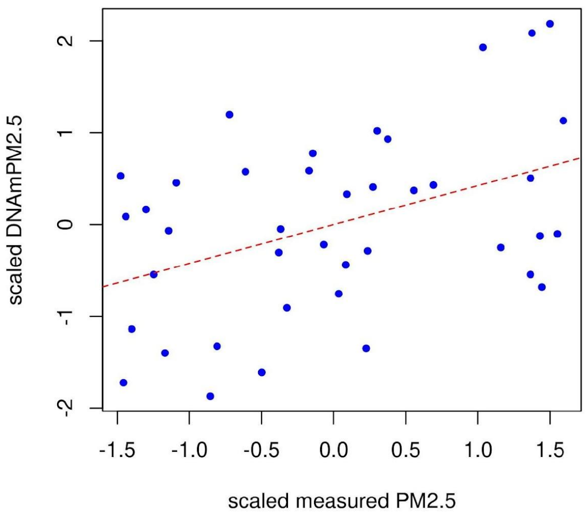
- A risk score derived using **DNAm surrogates** involves a **double validation** and **outperforms** risk scores derived using a **single-step approach**, and those based on **traditional risk factors** (SCORE2).

- **Similar performance** for DNAmCVDscore and **DNAmGrimAge** (4 inflammation-related common components).

Figure from Cappozzo et al. *A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events*; CLEP 2019

Limitations of DNAm surrogates
Example of DNAmPM2.5





- **Negative** (unreliable) **values** for DNAm surrogates of exposure to air pollution.
- But still, **R = 0.42** for measured PM2.5 vs DNAmPM2.5.
- The DNAm surrogate **works on a relative scale**, it is not able to provide an absolute measure of exposure to PM2.5.

Conclusions:
Strengths and limitations of DNAm surrogates

- DNAm surrogates allow investigating the associations with **multiple exposures** **even if** those specific exposures **were not directly measured** in the cohort (but DNA methylation data is available).

- In some cases, they **predict diseases better than the original** (measured) **biomarkers**.

- **Lack of validation** and need for calibration in independent cohorts **for some exposures/proteins** (see Gadd et al. eLife 2022).

- They provide a **'relative' measure**, not an absolute one (example of DNAmPM2.5).

Thank you for your attention !!!