RESEARCH ARTICLE

A Deep Survival EWAS approach estimating risk profile based on pre-diagnostic DNA methylation: An application to breast cancer time to diagnosis

Michela Carlotta Massi[1,2]*, **Lorenzo Dominoni**[2], **Francesca Ieva**[1,2], **Giovanni Fiorito**[3]

1 Health Data Science Centre, Human Technopole Foundation, Milan, Italy, **2** MOX Laboratory for Modeling and Scientific Computing, Dept. of Mathematics, Politecnico di Milano, Milan, Italy, **3** Laboratory of Biostatistics, Dept. of Biomedical Sciences, University of Sassari, Sassari, Italy

* michela.massi@fht.org

Abstract

Previous studies for cancer biomarker discovery based on pre-diagnostic blood DNA methylation (DNAm) profiles, either ignore the explicit modeling of the Time To Diagnosis (TTD), or provide inconsistent results. This lack of consistency is likely due to the limitations of standard EWAS approaches, that model the effect of DNAm at CpG sites on TTD independently. In this work, we aim to identify blood DNAm profiles associated with TTD, with the aim to improve the reliability of the results, as well as their biological meaningfulness. We argue that a global approach to estimate CpG sites effect profile should capture the complex (potentially non-linear) relationships interplaying between sites. To prove our concept, we develop a new Deep Learning-based approach assessing the relevance of individual CpG Islands (i.e., assigning a weight to each site) in determining TTD while modeling their combined effect in a survival analysis scenario. The algorithm combines a tailored sampling procedure with DNAm sites agglomeration, deep non-linear survival modeling and SHapley Additive exPlanations (SHAP) values estimation to aid robustness of the derived effects profile. The proposed approach deals with the common complexities arising from epidemiological studies, such as small sample size, noise, and low signal-to-noise ratio of blood-derived DNAm. We apply our approach to a prospective case-control study on breast cancer nested in the EPIC Italy cohort and we perform weighted gene-set enrichment analyses to demonstrate the biological meaningfulness of the obtained results. We compared the results of Deep Survival EWAS with those of a traditional EWAS approach, demonstrating that our method performs better than the standard approach in identifying biologically relevant pathways.

Author summary

Blood-derived DNAm profiles could be exploited as new biomarkers for cancer risk stratification and possibly, early detection. This is of particular interest since blood is a

OPEN ACCESS

Citation: Massi MC, Dominoni L, Ieva F, Fiorito G (2022) A Deep Survival EWAS approach estimating risk profile based on pre-diagnostic DNA methylation: An application to breast cancer time to diagnosis. PLoS Comput Biol 18(9): e1009959. https://doi.org/10.1371/journal.pcbi.1009959

Editor: Eric F. Lock, University of Minnesota Twin Cities, UNITED STATES

Received: February 24, 2022

Accepted: September 10, 2022

Published: September 26, 2022

Data Availability Statement: DNAm data used in this study are available at the GEO repository under the accession number GSE51032. The source code to run the analyses described in the paper is available at github.com/LorenzoDominoni/DeepSurvEWAS.

Funding: GF is funded by the Programma Operativo Nazionale (PON), Ricerca e Innovazione 2014-2020 (http://www.ponricerca.gov.it/), Attrazione e Mobilità Internazionale (AIM), code AIM1874325-2. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

convenient tissue to assay for constitutional methylation and its collection is non-invasive. Exploiting pre-diagnostic blood DNAm data opens the further opportunity to investigate the association of DNAm at baseline on cancer risk, modeling the relationship between sites' methylation and the Time to Diagnosis. Previous studies mostly provide inconsistent results likely due to the limitations of standard EWAS approaches, that model the effect of DNAm at CpG sites on TTD independently. In this work we argue that an approach to estimate single CpG sites' effect while modeling their combined effect on the survival outcome is needed, and we claim that such approach should capture the complex (potentially non-linear) relationships interplaying between sites. We prove this concept by developing a novel approach to analyze a prospective case-control study on breast cancer nested in the EPIC Italy cohort. A weighted gene set enrichment analysis confirms that our approach outperforms standard EWAS in identifying biologically meaningful pathways.

Introduction

DNA methylation (DNAm) is a chemical modification that consists of the addition of a methyl group via a covalent bond to the cytosine ring of DNA, in correspondence of CpG sites (CpGs), and a large body of evidence has demonstrated that CpG islands hypermethylation is implicated in loss of expression of a variety of critical genes in cancer [1].

These alterations can be detected both in the target tissue (e.g., cancer biopsy vs tumor-free adjacent tissue) and in blood-derived DNA. DNAm dysregulation in target tissue are likely the effect of the disease rather than vice versa [2], whereas DNAm alterations in blood are commonly used as biomarkers of long-term exposure and insults to the DNA which includes variability related to genetic predisposition or individual response to risk factors. The above suggest the possibility to use blood-derived DNAm profiles as new biomarkers for cancer risk stratification and possibly, early detection. Investigating DNA methylation data from blood samples is of particular interest since it is a convenient tissue to assay for constitutional methylation and its collection is non-invasive. Moreover, exploiting pre-diagnostic blood DNAm data opens the opportunity to investigate the association of DNAm at baseline on cancer risk, modeling the relationship between sites' methylation and the Time to Diagnosis (TTD). That would indeed be desirable, as to improve the effectiveness of current screening procedures via the definition of novel effective and non-invasive biomarkers (e.g., via DNAm-based scoring methods) is a public health necessity. In this regard, a recent paper by Muse et al. investigating genome-wide DNAm differences of breast tumor tissue with adjacent normal tissue highlights the possibility of identifying early DNAm alteration in breast carcinogenesis and consequently developing epigenetic biomarkers of disease risk [3]. Previous studies investigated the association of breast cancer risk with DNAm biomarkers of age acceleration (epigenetic clocks) and global hypomethylation in prospective studies. In their review paper, Ennour-Idrissi et al. highlight a global trend toward an association of age acceleration and lower global methylation with a higher risk of breast cancer in 17 prospective studies [4]. However, the effect size was modest or relatively weak for most of the studies, with a high risk of bias due to residual confounding for unmeasured (or poorly controlled) risk factors. It is known, in fact, the significant association of both epigenetic clocks and global hypomethylation with breast cancer risk factors like smoking, BMI, parity, age at menarche, breast density, and hormonal risk factors [5, 6]. Further, epigenome wide association studies (EWAS) either ignore the explicit modeling of TTD in a survival analysis setting or mostly provided inconsistent results as there was no overlap among the differentially methylated CpGs identified in different studies [4, 7].

In this work, we focus on the identification of blood DNAm profiles associated with TTD, with the aim to improve the reliability/reproducibility of the results, as well as their biological meaningfulness. The unsatisfactory outcome of previous attempts to find reliable associations may be induced by the limitations of the most traditional approaches for the analysis of whole-genome DNAm data, i.e., Epigenome-Wide Association Studies (EWAS). Indeed, EWAS analyses traditionally comprise multiple independent tests of individual CpG sites or regions, seeking for significant associations by imposing p-value thresholds corrected for multiple comparisons.

The main limitations of this approach are: (i) the extremely high dimensionality of epigenome-wide DNAm data affects the reliability of multiple testing correction, driving p-value thresholds down to extremely low values, (ii) the strong correlation among methylated sites, that is usually not considered in statistical modelling, (iii) the presence of several (likely unmeasured) confounders, since DNAm profiles are influenced by environmental exposures and lifestyle behaviors [8]. Additionally, the context of pre-diagnostic blood DNAm carries further complexities due to the (iv) very low signal-to-noise ratio of differential methylation, as both cases and controls are healthy at the time of DNAm collection. Lastly, (v) these methods based on independent testing do not account for any of the complex and potentially non-linear interactions that might exist between CpG sites or the combined effect of multiple loci together on the phenotype. Indeed, findings from previous studies [9], suggest the need for an epigenome-wide approach, that assigns individual parameters while accounting for sites' combined effect on the phenotype. We refer to this set of parameters as an effects profile.

Nonetheless, exploiting biostatistical approaches s.a. Cox Proportional Hazard (CoxPH) regression, to model survival outcomes and infer this effect profile including all CpG sites as predictors together, would lead to further methodological pitfalls. Firstly, modeling such a large number of covariates leads to effect size overestimation. Moreover, CoxPH models suffer the multi-collinear nature of CpG sites and are based on strong assumptions, such as the additive nature of covariates' effect on the outcome, unless including an even larger number of terms to account for interactions.

These limitations and the complexities of DNAm data can be naturally handled by Machine Learning (ML) approaches, such as Neural Networks (NN) and Deep Learning (DL)-based methods. Indeed, NN are optimized to extract rich latent features from DNAm data, handling multi-collinearity, noise, and considering the complex non-linear interactions between very large amounts of input covariates [10]. Quite a few examples exist of DL-based methods for patients classification [11–13], risk prediction and survival modeling [14–17] from DNAm data. On top of these, some recent works demonstrated the usefulness of AutoEncoders (AE), Variational AE (VAE) and DL models specifically to obtain DL-based EWAS (henceforth, Deep EWAS) [11, 12, 18], i.e. to estimate an effect profile. Oftentimes in this case, DL models are coupled with post-hoc Explanation Methods (EM) [10]. EM like SHapley Additive exPlanations (SHAP) [19, 20] try to overcome the "black box" aspect of these complex models assigning a contribution score (i.e., an importance weight) to each input feature, based on how much it contributed to the model prediction. In Deep EWAS, EMs weights can be considered as an estimation of the aforesaid effects profile on the phenotype. Nonetheless, the highly parametrized deep models are prone to overfitting, unless they are presented with very large training samples, and a suboptimal training may result in unstable and unreliable explanations (i.e., importance weights). Indeed, to obtain effective explanations to derive meaningful conclusions from, both model's accuracy and importance weights stability should be maximized [21].

While Deep EWAS is gaining momentum, there is still a lack of approaches that try to model TTD, inferring effect profiles, or global association parameters, in a survival setting.

Indeed, the search for DNAm associations with a survival outcome (i.e. Survival EWAS) exploiting DL methods has been largely unexplored and there has not been any application yet in the context of blood DNAm. To the best of our knowledge, the existing Deep EWAS literature mainly focuses on classification settings (e.g., cancer type, cancer status, patients' clinical condition, etc.). The most prominent examples of this effort come from the works from Levy et al. [10], that, in its seminal work, proposes an effective Deep EWAS framework based on VAE-based encoding of DNAm data, followed by a prediction model explained through SHAP. Despite the flexibility of the algorithm, no effort has been devoted to tackle the specific facets of a time-to-event setting. Only one related study [11] deals with DNAm data and time of BC recurrence to filter significant latent features. Yousefi et al. [22] propose a general framework for genome-wide data, but their automatic hyperparameter optimization does not account for the stability of their back-propagation based explanations. Despite the lack of efforts towards DL-based EWAS for TTD from blood-derived DNAm, we argue that the search for effective prognostic biomarkers from this type of data would instead benefit from a paradigm shift from standard EWAS approaches. Indeed, we believe that to achieve reliable biomarkers (i) the estimation of CpG sites effects profile requires a global approach rather than independent association testing and that (ii) such approach should exploit the potential of DL-based methods to capture the complex and potentially non-linear relationships interplaying between sites. Finally, (iii) the adopted methodology should be tailored to face the facets of real-life research settings.

In this work, we validate our hypothesis by analyzing blood based DNAm data from a prospective case-control study on Breast Cancer (BC) nested in the EPIC Italy cohort. This cohort, that was previously analyzed in [7, 9], presents all typical complexities of pre-diagnostic DNAm studies, i.e., small sample size, risk of confounders effect due to the long time from recruitment to cancer diagnosis, and the challenge of identifying differentially methylated sites among individuals that are healthy at the time of blood collection.

To tackle this data and prove our concept, we develop a new DL-based approach that assesses the relevance of individual CpG Islands in determining TTD while modeling them all together in a survival analysis setting. The methodology is inspired by previous Deep EWAS approaches [10, 11, 13, 23] and adapted for survival data, therefore we name it Deep Survival EWAS. Our Deep Survival EWAS models the complex relationships among CpG sites and between sites and TTD, and ultimately estimates the desired CpG Islands effects profile through SHAP, in terms of importance weights in determining the hazard rate.

In summary, the objectives and contributions of this study are multiple:

- We highlight the need of novel approaches to cancer TTD modeling from blood-derived DNAm. To derive meaningful and generalizable biological insights the proposed method overcomes the limitations of standard EWAS approaches and takes into account the combined effect of DNAm sites on cancer onset.
- We present our original approach to the problem, i.e., Deep Survival EWAS. Besides its natural capability to model complex and potentially non-linear interactions among CpG Islands, the overall algorithm has several valuable methodological details meant to deal with the complexities arising from this crucial real-life biological research context, s.a. small samples, collinearity, noise, and low signal-to-noise ratio of blood-derived DNAm.
- We validate the aforementioned hypothesis presenting the results of our Deep Survival EWAS in an in-depth analysis of a prospective case-control BC study nested in the EPIC Italy cohort. To demonstrate the biological meaningfulness of the obtained effect profile, we perform weighted gene-set enrichment analyses (GSEA), comparing the results of Deep

Survival EWAS with those of a traditional EWAS approach. The GSEA results, indicate that our method performs better than the standard approaches both looking at the biological reliability and the statistical stability of genes and pathways identified.

- We test our proposal by replacing the NN survival model with an alternative state-of-the-art interaction-aware algorithm, i.e. XGBoost, showcasing the superior biological relevance of our results. Finally, we confirm the value of a DL-based model accounting for predictors interactions comparing our method with a simpler Cox model with additive effects only.

Results

DNAm data preprocessing

In this study we exploit blood based DNAm data collected for a BC study nested in the EPIC Italy cohort. After data preprocessing, sample and probe filtering our primary dataset includes DNAm values for 13,499 CpG sites in 248 incident BC cases and one-to-one matched controls. DNAm values are expressed as the ratio of methylated cytosines over total cytosines (named from here on β values). DNAm β values for CpGs pertaining to the same CpG island (according to the UCSC annotation [24]) were grouped as described in Methods section, where we refer the reader to for an in-depth description of the data preparation pipeline (cf. EPIC Breast Cancer Data). The final dataset is composed of DNAm data for 3,807 CpG islands.

Overview of the Deep Survival EWAS approach

To enhance the extraction of relevant information from blood DNAm data, we defined our methodological approach, i.e. the Deep Survival EWAS algorithm, as a set of steps tailored to face the aforementioned facets of these complex dataset and research settings. In Fig 1, we represent the visual schema of the overall procedure.

In brief, as a first step (Panel A, Fig 1), we subdivide the study population between cases (i.e., individuals diagnosed with cancer within the study follow-up) and controls (i.e., individual remained healthy until the end of follow-up). We begin with exploiting the population of cases, and we agglomerate CpG Islands through Hierarchical Feature Clustering based on Euclidean Distance of β-values (Panel B). Once the clusters are defined, we aggregate cluster-specific β-values computing their mean. Through this step, we obtain a feature matrix $F_{cases} \in \mathbb{R}^{N \times J}$, where N is the number of cases and J is the number of feature clusters. Henceforth, these aggregated CpG Islands will be referred to as *features* F_j. We apply the same clustering structure to the CpG Islands of the controls' population to obtain $F_{controls} \in \mathbb{R}^{M \times J}$, where M is the number of controls, in order to maximize inter-group differences in distribution of the aggregated features in case a true difference exists between cases and controls' beta values.

Then, to make a robust and reliable estimation of the EWAS weights associated with the input F_{cases}, we generate K random splits (Panel C) of the dataset into training and test set, with 80/20 ratio and for each training-test split (Panel D): (i) we model the complex non-linear relationship between the input and TTD, by exploiting a deep survival feed-forward neural network predictive of the log-risk function; (ii) we exploit Kernel Shap from SHAP framework [19] to estimate weights ($w_j^{(k)}$) associated to each feature $f \in F$. Notably, SHAP relies on the use of a "background dataset" to estimate the expected model output and estimates individual feature's impact on prediction as their contribution to the difference between the observed and the expected prediction (more details in the following section). Therefore, choosing the right background data is crucial to obtain contextually meaningful explanations. Indeed, to

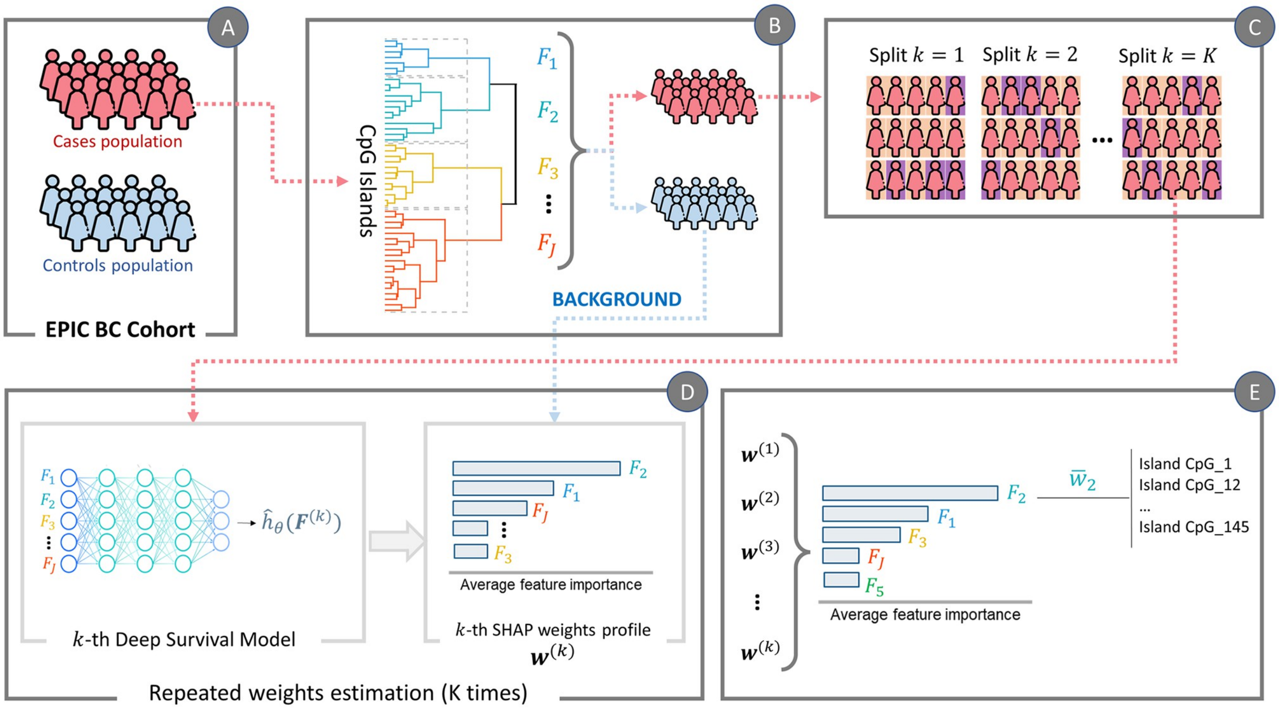


Fig 1. Algorithm pipeline of the methodology applied in this study. **(A)** We started from the EPIC BC cohort, equally split between cases (i.e., patients enrolled healthy but diagnosed with BC within the study time frame) and matched controls (i.e. patients matching cases at baseline, that were not diagnosed with BC within the study time frame). **(B)** The first step is feature aggregation via hierarchical clustering, exploiting CpG Islands continuous β values of the population of cases to infer the J clusters of features. The same clustering structure is then applied to both cases and controls, grouping their CpG Islands accordingly. **(C)** The cases population is exploited to generate K independent and randomly split training and test sets, each of them with 80% patients in training set and 20% patients in the test set. **(D)** Each of the K splits goes through step D independently. In particular, the k-th training set is used to train a Deep Survival Model, that then is used to estimate SHAP weights profiles (w^k) on the k-th test set using the controls' population as background data. **(E)** After generating independently K sets of weights profiles, they are aggregated to obtain the final estimation of the effects profile for BC TTD.

https://doi.org/10.1371/journal.pcbi.1009959.g001

resemble the differential estimation of DNAm effects in EWAS, we compute the importance of the features in F_{cases} using $F_{controls}$ as reference for SHAP. Finally, after generating independently K sets of weights' profiles, they are aggregated averaging them to obtain the final estimation of the effects profile $\bar{\mathrm{w}}$ for BC TTD.

Deep Survival EWAS on the case-control study of breast cancer nested in the EPIC dataset

We applied the described Deep Survival EWAS to the BC sample from EPIC Italy, to infer an effect profile associated to the blood-derived DNAm CpG Islands in the dataset. As many ML or DL based algorithms, our approach comprises several building blocks that require specific choices in terms of hyperparameters and/or implementation details, that need to be optimized to provide a robust estimation of the desired effect profiles. Of note, the term robust is used throughout the paper to describe more generalizable and transferrable parameters, estimated in such a way to avoid overfitting on the small sample they originate from.

First of all, even though the final objective of our algorithm is not per se the prediction of TTD, this performance has to be maximized. Indeed the better the underlying K models will perform in predicting the survival outcome, the more meaningful the feature importance weights derived by SHAP will be. Concurrently, a stable feature importance ranking suggests that the K models trained on different data subsets are consistently capturing and exploiting the information from a specific set of features to obtain such prediction. A satisfactory performance on both aspects together translates in a trustworthy and effective effects profile estimation [25]. For this reason, we seek to identify the best combination of number of feature clusters (J), i.e. the input dimension to the Survival NN, and the architecture and hyperparameters of the best Deep Survival model. These two aspects were jointly optimized to maximize the time to event prediction on the population of cases, averaged across $K = 10$ random splits, and the robustness of the K derived effects profile. The overall best combination is the one we use to compute the final effects profile averaging Shapley values across the 10 splits. The performance for the time to event prediction was measured with the Harrell's Concordance Index (CI), while the robustness of profiles with the Kendall Tau Ranking Stability (KT-stability). To aid both prediction performance and robustness, each of the 10 NN models underwent unsupervised pretraining. This passage was meant to bias the models' parameters towards an optimal configuration, thus potentially aiding the stability and reliability of the derived Shapley values, that are indeed directly impacted by the models' parameters. Indeed, KT-stability was improved significantly for almost all tested models' architectures. Detailed results can be found in Supporting Information S1 Table. Further details on pretraining, best model selection and performance metrics' definitions, including rationale for our choices, can be found in the Methods Section, whereas complete results are available in Supporting Information, S2 Table. The final results presented here, corresponding to the chosen best Deep Survival EWAS configuration, were obtained by grouping CpG Islands into J = 128 clusters. The latter were used as the input for the Deep Survival NN model with a J-dimensional input layer followed by three fully connected layers of 64, 32, and 16 nodes respectively. This model's configuration resulted in an average CI of 0.702 ± 0.019 and an average KT-Stability of 0.669 ± 0.036. In Supplementary Info S1 File we report the PCA plots of the latent components after unsupervised pre-training and after fine-tuning of the survival model. Note that in our algorithm latent components are generated in the process of training the K (with K = 10 in this case) DSNN. Therefore, we have one set of embeddings for each split. As patients with similar TTD after fine-tuning are better clustered together in the embedding space, these plots further validate how the different DSNNs are learning meaningful representations for time to event prediction.

Once selected the best configuration of the algorithm, we exploited the 10 sets of Shapley values weights to derive the final weights profile. Indeed, as mentioned in the previous section, these weights ($\bar{w}_j$) were computed as the average impact on log risk prediction of each feature across the K = 10 random resamples. In Fig 2, left panel, we show the features that correspond to the top 10 highest values in the obtained effects profile on log risk prediction. As for the background dataset, in this case impact was measured w.r.t. the control group of all BC controls. We performed some post-hoc analyses to showcase the tight relationship between the most relevant features according to the estimated effects profile and TTD. In Fig 3, panel A, we split the population of cases into time-to-event classes (i.e., early event $[0 - 3.5y]$, mid-early event $[3.5y - 7y]$, mid-late event $[7y - 10.5y]$ and late event $[10.5 - 16y]$) and we plot the distribution of the aggregated β values of the most important feature (F_{120}). The boxplot shows a decreasing trend of the feature value with increasing TTD. Note that F_{120} groups 20 CpG Islands (cf. Table 1), meaning that higher methylation values on those 20 sites are associated with earlier diagnosis (i.e., higher log risk).

As a further validation, for each CpG Island pertaining to F_{120}, we extracted two sub-population of cases: the hypermethylated population, composed of patients with β value above the 75^{th} percentile for that CpG island, and the hypomethylated population, composed of patients

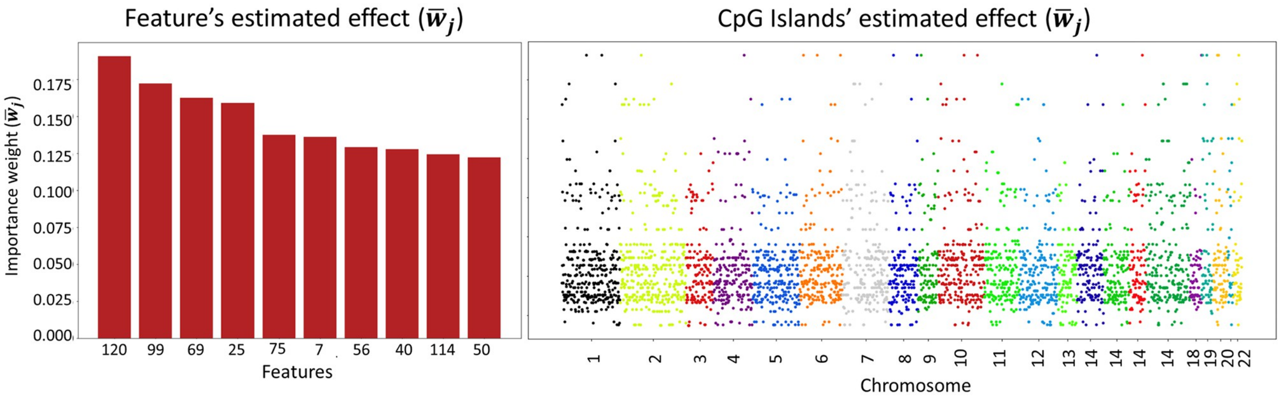


Fig 2. Deep Survival EWAS estimated effects profile. On the left panel, the top 10 aggregated features with the highest associated $\bar{w}_j$ value in the effects profile at Feature's level. On the right, sharing the same y-axis, the effects profile at CpG Island, where each methylated islands in the dataset is associated with the $\bar{w}_j$ value of the feature they are clustered in.

https://doi.org/10.1371/journal.pcbi.1009959.g002

with values below the 25th percentile. Then, we compared the risk profiles of the two groups via Kaplan-Meyer (KM) test. In Fig 3, panel B, we plot the KM curve obtained and the log-rank test p-value for CpG Island 9:34518796–34619343, as representative of the 20 islands in F_{120}. Similar behavior and log-rank test results was obtained for the other 19 islands in F_{120} (all KM curve plots for the CpG Islands in F_{120} are reported in S2 File). As expected, modeling TTD of BC cases only, both KM tend to 0 as t increases. However, it is clear from KM and test results how methylation on those sites is associated with TTD for those patients. It is crucial to note that, despite modeling cases TTD only, the inference of features' relevance exploits data form the controls group used as the reference (i.e. background). In other words, CpG Islands associated with higher weights by our Deep Survival EWAS approach are those with higher

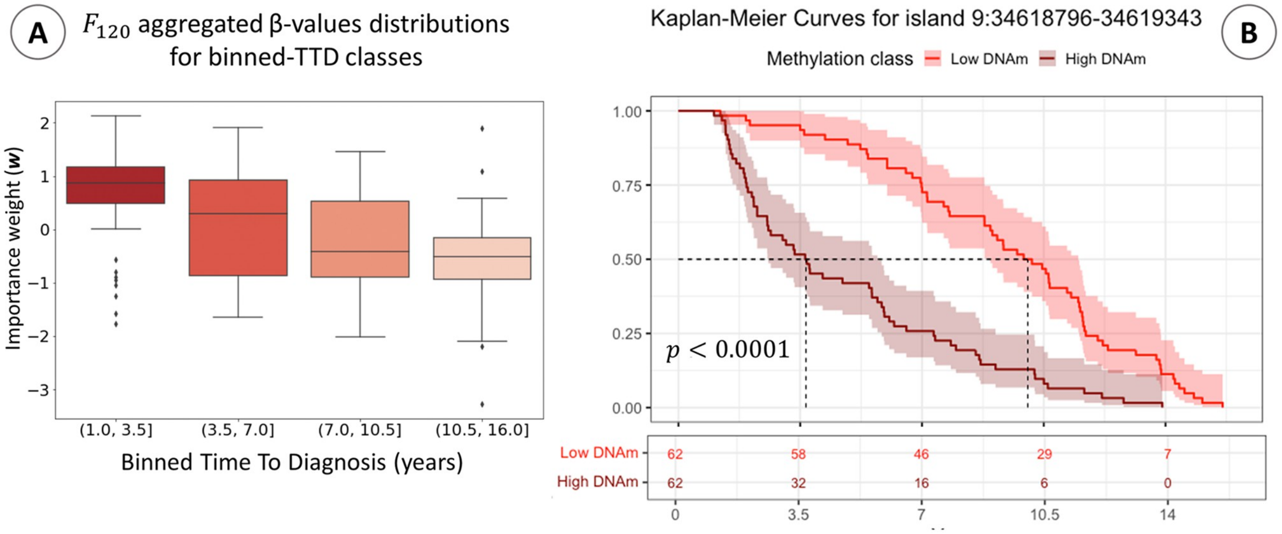


Fig 3. Post-hoc analysis results. (A) Distribution of aggregated DNAm beta-values in the feature (F_{120}) associated with the highest impact in the estimated effect profile. Subjects are binned according to their TTD into four classes. **(B)** Kaplan-Meier curves of low DNAm (below 25th percentile of β-values distribution) and High DNAm (above 75th percentile) populations' in CpG Island 9:34518796–34619343. This island belongs to the cluster that is aggregated into feature F_{120}. The plot reports the Log-Rank test p-value for the difference between the two groups; the lower part of the plot reports the count of subjects in High DNAm and Low DNAm populations for CpG Island 9:34518796–34619343, according to TTD.

https://doi.org/10.1371/journal.pcbi.1009959.g003

Table 1. CpG Islands in feature 120. List of CpG Islands agglomerated in F_{120}, therefore assigned to the highest effect weight.

CpG Island
1:90945518–90945656
1:158090642–158091676
10:102493904–102494072
10:119294070–119294143
14:87862626–87863008
16:85096322–85097146
18:75811758–75814395
19:1704275–1706659
19:13070446–13070515
2:100086548–100088317
20:21438169–21438255
20:21449303–21449404
22:37180713–37182260
4:149584089–149584799
6:1570179–1570756
6:43530362–43531683
6:166137998–166138866
8:21701267–21701566
8:145119282–145120028
9:34618796–34619343

https://doi.org/10.1371/journal.pcbi.1009959.t001

influence on the difference of risk prediction (hazard rate) compared to an expected model built with methylation profiles from the control group. Finally, the right panel of Fig 2 represents the resulting effects profile in terms of CpG Islands, that is the final goal of our Deep Survival EWAS. Specifically, to this end we associated each site to the weight of the feature it was clustered in. The whole list of the ranked features *F*, with the respective CpG Islands they cluster and the resulting effects profile can be found in S3 Table (first sheet for BC controls' reference group).

Validation through gene set enrichment analysis

To validate the biological relevance of the identified effects profile, we performed a Gene Set Enrichment Analyses (GSEA) using a Weighted Kolmogorov-Smirnov (WKS) test [26]. In this work, we have assigned weights to all the features in the dataset rather than performing feature selection. Therefore, we applied a pathway enrichment analysis method which compare the distribution of observed weights with those expected by chance, instead of an enrichment analysis based on Fisher or hypergeometric test (more suitable when feature selection is performed).

For each CpG site, the weight (input for the enrichment analysis) is that of the feature F_j the CpG site belongs to. In Fig 4 we present the KEGG pathways enriched according to the WKS procedure. We found 12 pathways with empirical p-value (1,000 permutations) lower than 0.05, being 'Pathways in Cancer' the most significant (empirical $p < 0.0001$). Interestingly, all the identified pathways were previously described as dysregulated in breast cancer, including 'Human Papilloma virus infection' [27] and 'Epstein-bar virus infection' [28] in addition to well-known BC related pathway like 'Breast Cancer', 'PI3K/Akt/mTOR signalling pathway', 'Calcium Signaling pathway' and 'Mineral absorption' [29, 30], and some cancer generic

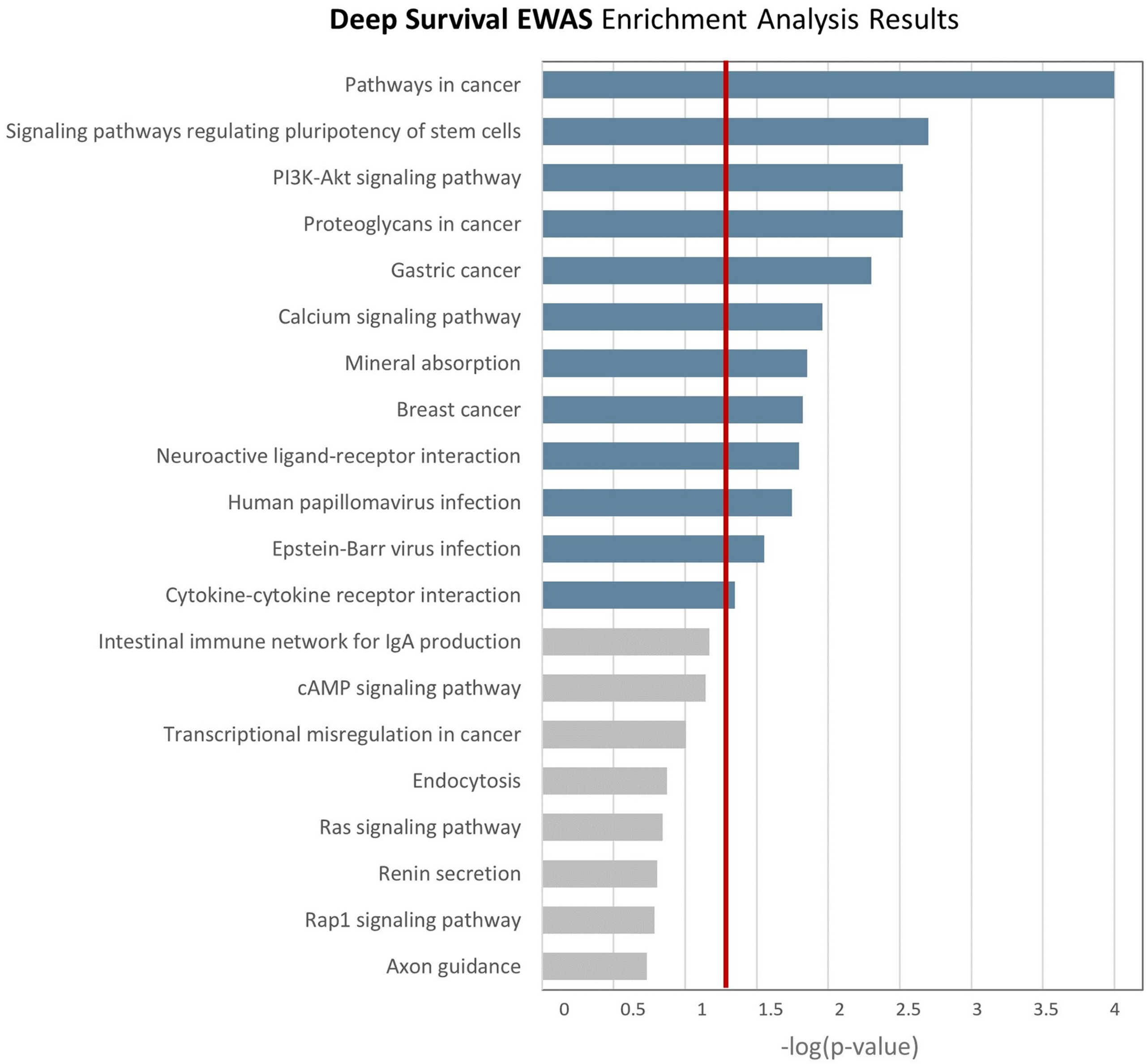


Fig 4. Enrichment analysis results of Deep Survival EWAS on breast cancer controls reference group. In blue are highlighted the significantly associated pathways, i.e. those with empirical p-value above 0.05 (red vertical line), estimated through 10,000 permutations.

https://doi.org/10.1371/journal.pcbi.1009959.g004

pathways like 'Signaling pathways regulating pluripotency of stem cells', 'Proteoglycans in cancer', 'Neuroactive ligand-receptor interaction', and 'Cytokine-cytokine receptor interaction'. As mentioned, the inference of feature importance, and the derived effects profile, can be influenced by the chosen *background*. Thus, different biological insights can be gathered by tailoring the reference group to answer different research question. For this study we wished to investigate both the robustness of the enrichment results changing the background sample, and whether some additional biological associations could be collected on EPIC BC case-control study estimating weights $\bar{\mathbf{w}}$ according to different *background* control groups. Therefore, we created 3 additional reference groups to use as background:

- **’BC Matching Controls’** was defined specifically for each of the *K* splits. In other words, for the k-th test set including 20% of BC cases, we defined the k-th BC Matching Controls group including only the matched controls of those cases. Therefore, we had *K* BC Matching Controls datasets (with potential subjects’ overlap) to use as background data in estimating effect profiles.
- **’All Controls’** sample included all female control subjects collected for breast, lung and colon cancer. It contained 556 healthy individuals supplied together as Background data.
- **’All Controls with Cases’** included all subjects of the previous sample, with the addition of female cases diagnosed with lung or colon cancer during the EPIC follow up period.

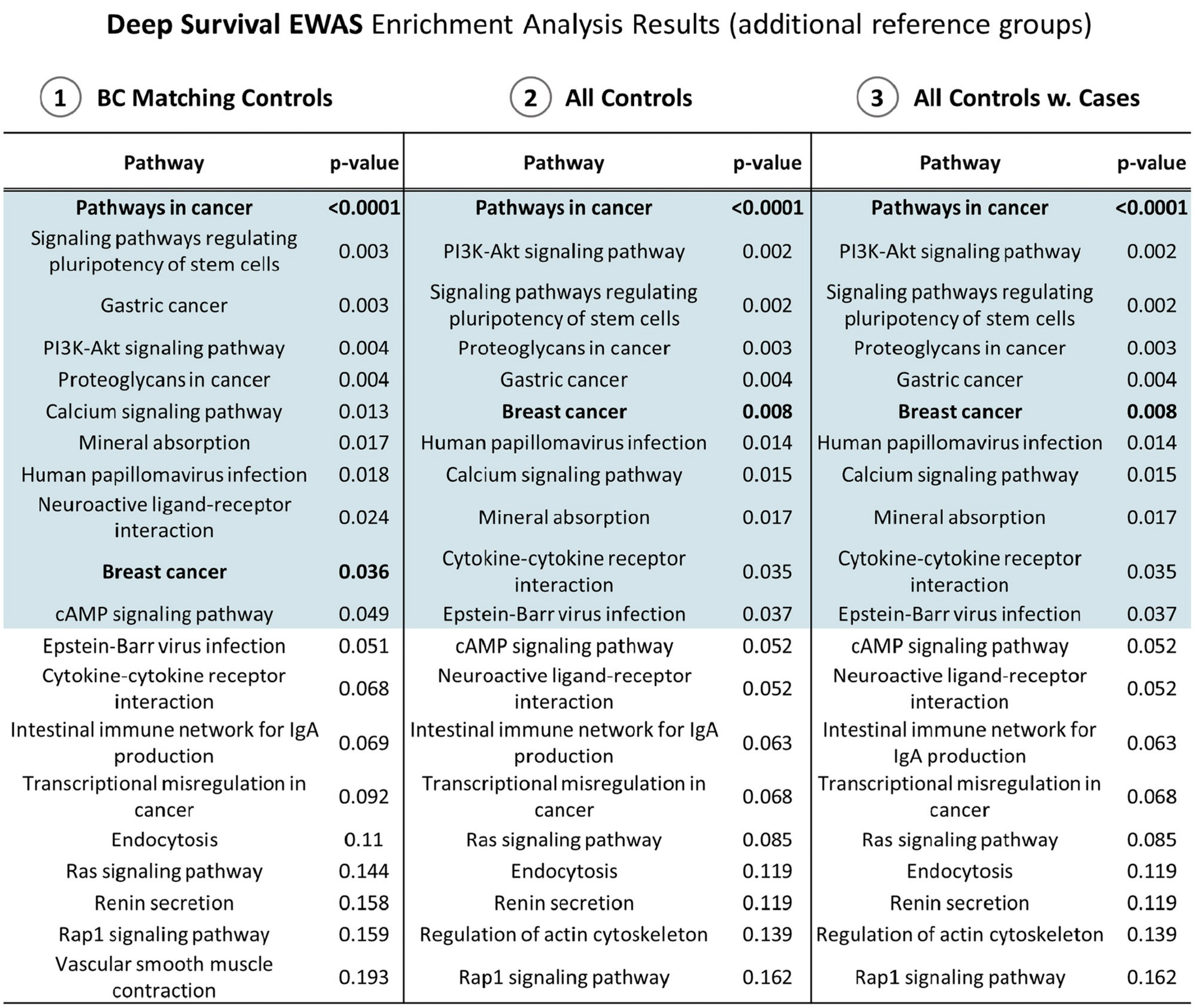
To perform these additional analyses, we kept the same optimal configuration of Deep Survival EWAS, supplying to SHAP applied to the survival NN different background samples to estimate w_j for the 128 features. In Fig 5, we report the results of the three additional enrichment analyses. The full tables of enrichment analyses results for all four reference groups can be found in Supporting Information S4 Table.

Deep Survival outperforms standard EWAS in identifying biologically meaningful effects profiles

In this work we argue that a complex DL-based approach to estimate a global effects profile on TTD from blood-based DNAm can provide better biological insights compared to traditional approaches. Therefore, to prove our concept and investigate whether taking into account the complex interrelationships among features and outcome modeled with Deep Survival leads to more relevant discoveries than Standard EWAS, we compared the enrichment analyses of the two approaches. In Fig 6, we report the results of the weighted GSEA for Standard EWAS, where weights were estimated independently for each CpG Island as the test statistic of a univariate Cox Survival Model. For consistency with our approach, we modeled each univariate CoxPH for the cases population only without adjustment for covariates Detailed tabular results (i.e. p-values and test statistics) are reported in Supporting Information S5 Table). The deriving estimated weights profile (i.e. all CpG Islands p-values, with or without Bonferroni adjustment) are represented in S1 Fig. Finally, Standard EWAS GSEA tabular results are reported in S6 Table. We found eight pathways with empirical p-value lower than 0.05, being ‘Non-small cell lung cancer’ the most significant. Among the identified pathways, two of them are related with the immune system regulation ‘Intestinal immune network for IgA production’ and ‘Chemokine signalling pathway’; two of them with inflammatory processes ‘Leukocyte transendothelial migration’ and ‘C-type lectin receptor signalling pathway’, whereas none of them have been previously described as BC specific. For completeness, we performed the same analysis’ pipeline on CpG sites without aggregating them into islands. Results from site-by-site EWAS and Enrichment Analysis are reported in Supplementary Information S7 and S8 Tables.

The value of complex non-linear modeling of CpG sites interactions

So far, we validated our claim on the need for a global approach to blood-based DNAm. In our Deep Survival EWAS, this global view is obtained by modeling complex and potentially non-linear interactions between DNAm sites via a deep non-linear survival NN. Nevertheless, a global approach can, by definition, be estimated with a simpler model accounting for the effects of all CpG Islands together with additive contribution on the phenotype only. An



Deep Survival EWAS Enrichment Analysis Results (additional reference groups)

① BC Matching Controls		② All Controls		③ All Controls w. Cases	
Pathway	p-value	Pathway	p-value	Pathway	p-value
Pathways in cancer	**<0.0001**	**Pathways in cancer**	**<0.0001**	**Pathways in cancer**	**<0.0001**
Signaling pathways regulating pluripotency of stem cells	0.003	PI3K-Akt signaling pathway	0.002	PI3K-Akt signaling pathway	0.002
Gastric cancer	0.003	Signaling pathways regulating pluripotency of stem cells	0.002	Signaling pathways regulating pluripotency of stem cells	0.002
PI3K-Akt signaling pathway	0.004	Proteoglycans in cancer	0.003	Proteoglycans in cancer	0.003
Proteoglycans in cancer	0.004	Gastric cancer	0.004	Gastric cancer	0.004
Calcium signaling pathway	0.013	**Breast cancer**	**0.008**	**Breast cancer**	**0.008**
Mineral absorption	0.017	Human papillomavirus infection	0.014	Human papillomavirus infection	0.014
Human papillomavirus infection	0.018	Calcium signaling pathway	0.015	Calcium signaling pathway	0.015
Neuroactive ligand-receptor interaction	0.024	Mineral absorption	0.017	Mineral absorption	0.017
Breast cancer	**0.036**	Cytokine-cytokine receptor interaction	0.035	Cytokine-cytokine receptor interaction	0.035
cAMP signaling pathway	0.049	Epstein-Barr virus infection	0.037	Epstein-Barr virus infection	0.037
Epstein-Barr virus infection	0.051	cAMP signaling pathway	0.052	cAMP signaling pathway	0.052
Cytokine-cytokine receptor interaction	0.068	Neuroactive ligand-receptor interaction	0.052	Neuroactive ligand-receptor interaction	0.052
Intestinal immune network for IgA production	0.069	Intestinal immune network for IgA production	0.063	Intestinal immune network for IgA production	0.063
Transcriptional misregulation in cancer	0.092	Transcriptional misregulation in cancer	0.068	Transcriptional misregulation in cancer	0.068
Endocytosis	0.11	Ras signaling pathway	0.085	Ras signaling pathway	0.085
Ras signaling pathway	0.144	Endocytosis	0.119	Endocytosis	0.119
Renin secretion	0.158	Renin secretion	0.119	Renin secretion	0.119
Rap1 signaling pathway	0.159	Regulation of actin cytoskeleton	0.139	Regulation of actin cytoskeleton	0.139
Vascular smooth muscle contraction	0.193	Rap1 signaling pathway	0.162	Rap1 signaling pathway	0.162

Fig 5. Weighted enrichment analysis results from Deep Survival EWAS (additional reference groups). In blue are highlighted the pathways with significant association, i.e. empirical p-value based on 10,000 permutations.

https://doi.org/10.1371/journal.pcbi.1009959.g005

example of such a simpler model is a CoxPH model for TTD with a multivariate input and no interaction terms. Therefore, to justify our choice of including a more complex Survival NN model in our procedure, we tested a multivariate epigenome-wide CoxPH model against the two metrics we deem relevant to evaluate the reliability of the derived effects [21]: model prediction performance and robustness of the estimated effects across the *K* subsamples, using the C-index and KT-stability metrics respectively. For consistency, we supplied as input to the CoxPH model the CpG Islands aggregated in the same features exploited by our Deep Survival EWAS approach. As detailed in the Methods Section, we tested the performance of the CoxPH model for all feature aggregation (*J*) tested when optimizing our approach. To compute CoxPH KT-stability, the weights rank was derived from the regression parameters (i.e., the effect size) of the CoxPH model. Complete results are reported in S9 Table in Supporting Information. For all values of *J*, our approach outperforms CoxPH on both metrics, especially for the CI (cf. S2 and S3 Figs).

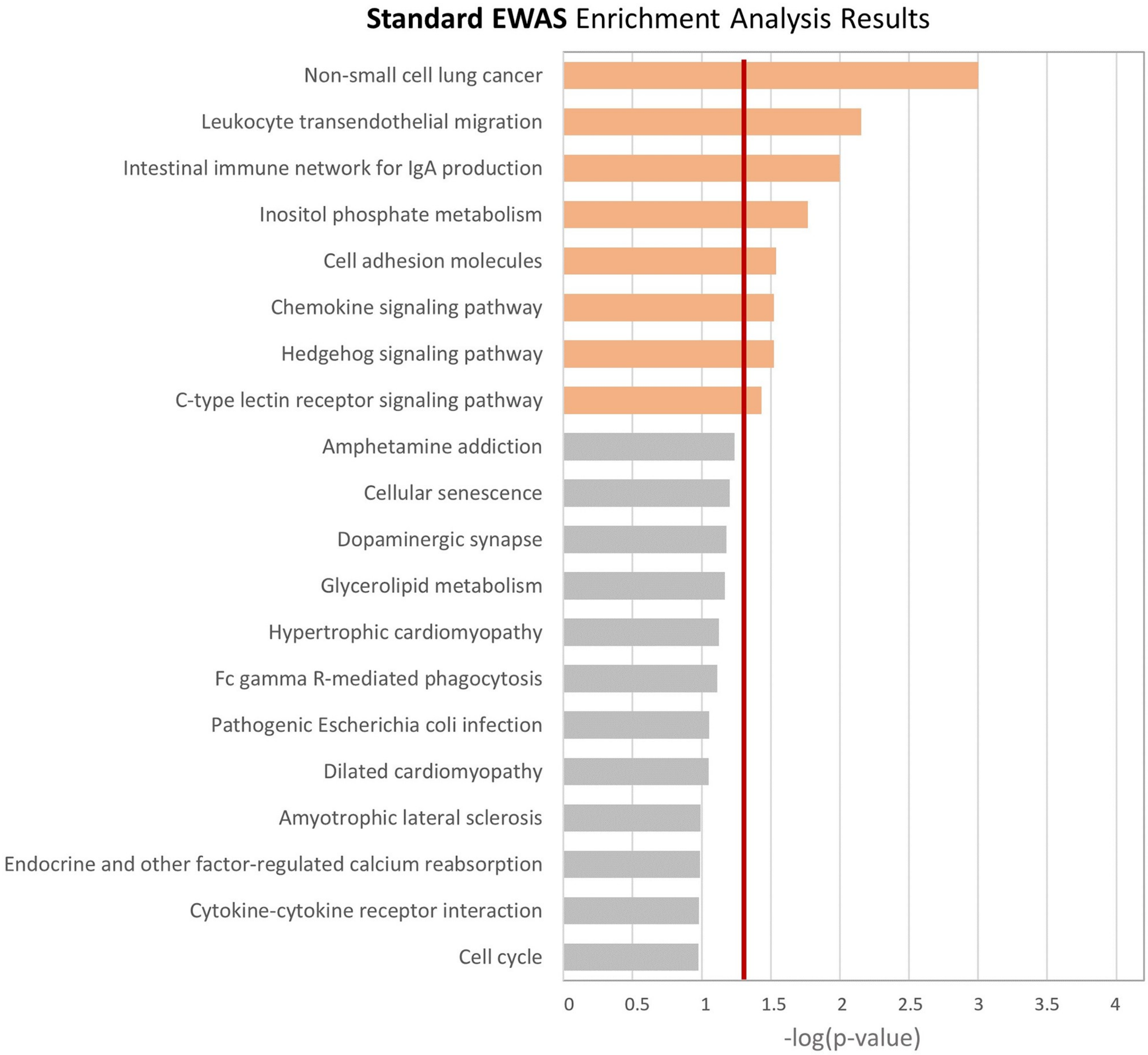


Fig 6. Weighted enrichment analysis results from standard EWAS approach.

https://doi.org/10.1371/journal.pcbi.1009959.g006

On the choice of the interaction-aware survival prediction model

In the previous section we validated the relevance of modeling complex non-linear interactions to predict TTD, rather than exploiting simpler additive models like CoxPH. In this study we chose a DSNN for this task, but other methods exist in the Machine Learning literature that can capture the effect of interactions while modeling a time to event outcome. One quite relevant state-of-the-art example is represented by XGBoost for survival prediction. This tree-based algorithm could in principle replace the DSNN in the framework proposed by our algorithm, as it similarly allows for the estimation of Shapley values through SHAP, both via KernelSHAP and the efficient model-specific SHAP TreeExplainer. To verify whether our preference for the DSNN was supported by evidence, we ran an experiment comparing the two predictors both in terms of survival modeling and biological relevance of the identified effects profiles, which is, notably, the final objective of our proposed approach. We trained both alternatives on the same 10 splits, for all the usual four input dimensions *J*. XGBoost was trained with standard hyper-parameters, kept consistent across input dimensions for a fair comparison. In Supplementary Information S10 Table we report the results in terms of CI and KT-stability. The two non-linear predictors have a comparable performance in terms of CI, with the DSNN slightly outperforming XGBoost on larger *J* values, suggesting a better scalability of our approach. Of note, XGBoost significantly outperforms the DSNN in terms of KT-stability on the same dimensions where our proposed approach shows better prediction performance, which might imply that XGBoost is consistently picking suboptimal interaction paths in its tree construction. However, none of the alternative state-of-the-art methods demonstrate an evident superiority, while both are clearly better then the simpler CoxPH model, highlighting once again the need for interactions to aid TTD prediction. As a final comparison among the two, in Supporting Information S4 Fig we report the results of the enrichment analysis run on the effects profile estimated by SHAP from XGBoost. Comparing these results with Deep Survival EWAS, we observed that the most relevant pathways like 'Pathways in cancer', 'Proteoglycan in cancer', 'Breast cancer' and 'Gastric cancer' were correctly identified by the XGBoost as well. However, some of the pathways previously described as dysregulated in breast cancer that our algorithm found significant were not identified by the tree based method, such as 'Human Papilloma virus infection', 'Epstein-bar virus infection', 'PI3K/Akt/mTOR signalling pathway', 'Calcium Signaling pathway' and 'Mineral absorption'. In addition, XGBoost identified a set of immune response-related pathways like 'Complement and coagulation cascade', 'Glycosphingolipids biosynthesis', 'NK mediated cytotoxicity, and 'Viral carcinogenesis', whose role in breast cancer development is not yet well characterized, suggesting that some degree of noise might have been introduced in weight assignment. Therefore, from a biological meaningfulness standpoint, the Deep Survival EWAS exploiting the DSNN model definitely outperforms the benchmark classifier.

Discussion

In this work, we presented our approach to solve the complex task of determining the effect of pre-diagnostic blood DNAm in prospective epidemiological studies. Specifically, we focused on the time from recruitment to cancer diagnosis outcome, with the aim of modeling the effect of DNAm CpG sites on the phenotype as in a survival (time to event) analysis setting. To demonstrate the need of a paradigm shift from Standard EWAS approaches (i.e. multiple independent tests) for the analysis of blood based DNAm, we developed and applied a novel approach, Deep Survival EWAS, that robustly estimates weights associated to each CpG Island by considering the combined (global) effect of CpG islands and their complex interactions on the phenotype. To this aim, we modeled the non-linear relationships and interplay between CpG sites and TTD by first grouping them into aggregated features and then feeding them as input to a non-linear deep survival NN, deriving the effects profile through SHAP. We validated our claims analyzing pre-diagnostic blood-derived DNAm from a BC case-control study nested in the EPIC Italy cohort. This is the first attempt to analyze a dataset from an epidemiological prospective study trying to model explicitly the TTD using a DL approach with the aim of obtaining EWAS weights. Notably, BC research is one of the areas that could mostly benefit from a proper modeling of TTD from blood based DNAm, as well as one that suffered the most from inconclusive outcomes (e.g. [7] and references therein). This makes the case study presented in this work an interesting yet challenging testing ground for our proof of concept.

In the present study, we decided to filter our DNAm predictors on a restricted number of CpG Islands, focusing on Polycomb Repressive Complex 2 (PCR2) regulated sites based on previous evidence of the importance of these genomic regions in cancer. In addition to the biological rationale, this choice was motivated by methodological/analytical considerations. Indeed, with the objective of obtaining reliable feature weights, rather than prediction performance, a much larger input space may harm the reliability of the parametrized model, incur in overfitting on such small sample, and in turn affect the estimation of the importance weights. Similarly, keeping the same number of clustered features *J*, with a much larger islands base, may have a negative impact on the aggregation of beta values per cluster (reducing the effect of the signal carried by each site when computing the average) and harm the reliability of assigning the same weight to each site in the same cluster. Given the above, a user defined feature selection appear to be the most reasonable way to proceed, and instead of an outcome-driven selection (that once again might be biased by the small sample size and reduce generalizability of the obtained biological insights), we went for a selection driven by external knowledge (i.e. specific transcription factors).

By estimating the desired global effects profile for BC TTD on EPIC DNAm data, we note that the CpG islands grouped in the most relevant features show a clear association of blood DNA methylation with the TTD, with higher methylation values associated with a higher risk of BC in the short term, as shown in the KM curves. Moreover, the overall biological meaningfulness of our procedure was confirmed by the results of a GSEA that identified pathways previously described as associated with cancer onset (with some BC-specific pathway). Instead, a GSEA analysis based on the results of a Standard EWAS (one association test for each CpG island, or one test for each CpG site in the site-specific version) identified molecular pathways indirectly associated with cancer onset (via immune system dysregulation or inflammatory processes), whereas none of them was BC specific. This results suggest how a global model of methylation profile captures the relationship with the phenotype better than considering DNAm variables one-by-one.

Then, we compared the survival modeling and weights' stability performances (i.e., CI and KT-stability) of our approach against a multidimensional CoxPH model, demonstrating that even after CpG islands aggregation into features clusters, a global yet simpler additive effects-based model could not obtain comparable results, further supporting the need to account for non-linear interactions among DNAm predictors as well. Indeed, XGBoost algorithm that naturally accounts for interactions while modeling the outcome performed comparably to the DSNN in terms of CI. However, the resulting weights profile identified just a subset of the relevant pathways' associations found through our NN-based algorithm, accompanied by some biologically less meaningful associations.

Besides Deep Survival EWAS is computationally more expensive than Standard EWAS, or more parametrized and complex than a more traditional survival model, the above results provide strong evidence about the advantages of using such an approach in Epigenome-Wide prospective studies using blood DNAm data. Additionally, we compared the results from GSEA by varying the reference group. This is the first time this peculiarity of SHAP is exploited to gain more insights from a biological perspective, comparing the biological function of the different weights associated to CpG Islands. Specifically, we did not observe significant differences when including women who developed other type of cancers (lung and colon) within the control group. These results suggest our findings provide a specific DNAm signature of BC cancer risk rather than a DNAm signature for all cancer risk.

Whilst we believe the most relevant highlight of this study lies in Deep Survival EWAS achievement of an improved biological interpretability of blood-derived DNAm data on TTD, the tailored choices we made in algorithm design deserve some attention as well. First of all, we decided to focus on modeling TTD for BC cases only. This approach resembles the case-only analysis performed through a standard EWAS approach in [31]. From a methodological standpoint, it allowed to improve survival modeling accuracy, increasing the reliability of the derived explanations [25]. Other methodological details were included to account for all the complexities and peculiarities of both DNAm data facets, and real-life research settings. In particular, clustering CpG Islands based on similarity had the objective of reducing noise and dimensionality simultaneously. The latter aspect reduces the effort in parametrizing the downstream survival model based on NN, alleviating the risk of overfitting on very small sample size, and obtaining suboptimal training results. Besides, this step had a biological justification in that previous studies identified potentially non-contiguous genetically controlled methylation clusters significantly associated to several diseases [32]. Furthermore, the rationale for the metrics exploited in CpG Islands clustering (i.e. Euclidean Distance and mean) was inspired by the results presented in Gagliardi et al. [9], where the association between blood-based DNAm and the phenotype was modeled under the hypothesis that extreme methylation values (i.e., epimutations) are significantly associated with the outcome. The Euclidean Distance would first identify Islands with similar beta values distribution across patients, while computing their mean (a metric that is sensitive to outlier values) we wish to preserve the effect of extreme values. Finally, the choice of clustering on the basis of cases group's distribution of beta values and then transfer the clustering structure to the controls' population is meant to maximize inter-class separability, and aid SHAP algorithm to identify truly relevant features when using controls' group as background. Likewise islands agglomeration, the *K* training-test split and subsequent aggregation of results, aims at alleviating the risk of overfitting when the sample size is small, providing a final weight profile that potentially generalizes better on unseen patients. Finally, NN optimal parametrization is aided by pretraining and network regularization (see Methods). This care for attaining a model with carefully estimated parameters is indeed extremely relevant for the estimation of reliable weight profiles through SHAP. Notably, most of the aforementioned methodological cautions meant to tackle the real-life complexities of epidemiological studies and blood-based DNAm data, can be considered per se valuable algorithmic suggestions whose rationale apply to a broader systems biology research context. For instance, the risk of overfitting resonates with any analysis trying to model complex high-dimensional omic data (s.a., genomic, transcriptomic, metabolomic, gut microbiome, etc.) with ML or DL-based methods when sample size is small. To the best of our knowledge, closely related methods for DNAm data, irrespectively of the specific task, mostly test on large datasets from public biobanks. Similarly, the attentions devoted to the reliability of SHAP-derived explanations, or the peculiar use of the background data to answer precise research question, are relevant aspects that have been largely ignored.

Conclusion

In conclusion, the contributions of this work are twofold: on the one side, by presenting the results of a case study in the context of BC research, we defend our claims and highlight the need for a paradigm shift from Standard EWAS approach when modeling blood-derived DNAm data in the search for effective TTD biomarkers; on the other, we describe the methodology we applied to effectively achieve the desired effect profile, that has per se several notable and transferrable points of attention. This study presents some limitations that should be accounted for. One is related to the lack of adjustments for clinical covariates in our analyses. However, this choice was supported by recent literature, suggesting that DNAm signature of risk factors may predict diseases better than measured risk factor itself [33–35]. Therefore, adjustment for clinical covariates and lifestyle related risk factors that is usually implemented in EWAS studies may remove significant variability due to genetic and metabolic profiles that can influence the response to exposures and stressors. Similarly, we did not adjust for estimated cell type proportions, which are usually estimated directly from DNAm data [36], although it is a common practice in EWAS studies. Current literature provides discordant opinions about the reliability and biological interpretability of the results after such a procedure [37]. Further, previous works using this study datasets report no significant differences in estimated cell type proportion between BC cases and healthy controls [38], supporting our analytical choice.

Another limit of the present study is the lack of external validation. Unfortunately, no comparable dataset with pre-diagnostic blood-based DNAm data and TTD outcome was available at this time. Despite being aware that the biological insights we obtained in this case study should not be considered general findings before a proper validation has been performed, note that the aim of this work is not that of making general discoveries from a clinical/biological point of view. Rather, we aim at demonstrating that a multivariate global approach can describe the role of CpG Islands methylation on BC cancer onset better than univariate approaches, in particular when a time-to-event outcome is considered. We demonstrate this by comparing our method with the standard procedures, and constructing through our algorithm a CpG Islands weights ranking that significantly associates to BC pathways, where traditional methods fail. As mentioned in the Introduction to this paper, previous studies oftentimes presented inconsistent results. While we cannot currently verify whether better generalizability on new samples can be achieved through our approach, we proposed a solution that overcomes most of the theoretical limitation of standard EWAS methods and the specificity of the findings with regards to BC is promising in that direction. Nevertheless, the real impact of this algorithm on generalizability has still to be verified and will be the next research effort as soon as proper external data becomes available. Moreover, the complex multi-step algorithm based on highly parametrized models we applied here opens the way for future works, that will investigate on what of these algorithms should be transferred as-is on new data (e.g., feature clusters, model parameters, etc.) or retrained from scratch. Nevertheless, the strength of the biological results obtained on the highly challenging case study presented here, and the generalizable methodological cautions, make the present work a relevant milestone in the advancement of blood-based DNAm studies and epidemiological studies in general, whenever reliable inference of global effect profiles on a time to event outcome is needed.

An interesting future perspective for this study might be the development of an easily interpretable risk prediction model extracting the most important features and their interactions from our algorithm, obtaining a kind of methylation-based Polygenic Risk model. One similar example can be found in [39], but further research would be needed to avoid one such model to incur in the curse of dimensionality and include complex interactions of potentially very high order. Moreover, while we discussed the theoretical and technical rationale behind a knowledge-driven selection of input CpG islands, one future evolution of our algorithm might entail the application of the same pipeline in parallel on different CpG Islands subsets (different binding sites, for example) to then aggregate results downstream. This might be a promising way to scale-up our approach without the risks we mentioned in the Discussion.

Materials and methods

Ethics statement

This study was performed according to the principles of the Declaration of Helsinki; all EPIC Italy participants provided written informed consent; the Human Genetics Foundation (HuGeF) Ethics Committee approved the study as reported elsewhere [40].

Study sample description

The European Prospective Investigation into Cancer and Nutrition (EPIC) is a large European study on diet and cancer and has been previously described elsewhere [41]. The Italian component of EPIC (EPIC-IT) [42] recruited 47,749 adult volunteers (men and women) at five centres. It is a prospective cohort study with blood samples collected from healthy participants at recruitment. After recruitment, participants were then observed for over 15 years for the insurgence of cancer, cardiovascular diseases and all-cause mortality. At the end of the follow up, all breast, colon and lung cancer cases with available blood sample suitable for epigenetic analyses were paired with an equal number of controls, individually matched on age (±5 years), sex, season of blood collection, center, and length of follow-up (incident density sampling method) to perform case-control studies nested in the cohort. In this study we focus on the breast cancer case-control dataset, which includes 248 incident cases and matched controls. The average time from recruitment to diagnosis was 7.22 years (min = 1.02 years; max = 15.64 years) for breast cancer cases, whereas the average time from recruitment to the last follow up was 14.88 years (min = 14.11 years; max = 17.68 years) for matched controls.

A detailed description of data collection, DNA methylation measurements and pre-processing and sample filtering can be found in S1 Appendix.

In this work, we focused on the BC case-control study, composed of 248 BC cases and an equal number of matching controls. Additionally, we exploited samples collected for lung cancer (168 control subjects) and colon cancer (140 control subjects) to enlarge our controls group exploited for effects profile weight estimation as described in section Validation through Gene Set Enrichment Analysis.

Each subject was described in terms of CpG sites methylation scores, reported as β-values, i.e., the proportion of methylated cells per site over the total. Therefore, each CpG site is a continuous value bounded between 0 (no methylation) and 1 (complete methylation). After DNAm data preprocessing, quality control and filtering, we had data for 313,324 CpG sites per subject. In this work we focus on CpG sites corresponding to transcription factor binding sites of EZH2 and SUZ12: two proteins pertaining to the Polycomb Repressive Complex 2 (PRC2). This choice was motivated by previous evidence of the accumulation of DNAm outliers values in these genomic regions, considering also previously described association of the total number of DNAm outliers with BC risk [9]. After filtering on EZH2 and SUZ12 CpG sites, we decided to perform an additional pre-processing by grouping CpG sites into CpG Islands. CpG islands are regions of the genome with a high proportion of CpG dinucleotide repeats in which DNAm is generally conserved. To derive CpG islands β-values we aggregated all single sites falling in a specific island by computing their mean. By doing that, we eventually obtained a DNAm representation of 3,807 methylated islands for each subject, that was the input dataset for the algorithm described below.

As shown in Figure S5 Fig, 38% of CpG islands fall in intergenic regions (intergenic CpG islands), 48% in intragenic regions without overlap of different gene elements, and only 14% of CpG Islands are in intragenic regions with both CpG sites in the promoter and the body of the same gene. Besides, for the latter class of CpG islands, we have investigated the within-island correlation structure. Only a negligible proportion of the pairwise comparison have a negative Pearson correlation coefficient, consistently with previous observation about the tendency of DNAm of CpG sites on the same island to be positively correlated [43].

Details of the proposed approach

The algorithm we crafted for Deep Survival EWAS on blood DNAm comprises several steps: (i) the Feature Agglomeration of CpG Island methylation profiles, (ii) the non-linear survival modeling of the aggregated features to model the complex relationship between the CpG Islands-derived features and BC TTD and (iii) the estimation of the relevance of each of those features in determining TTD (i.e. their importance in predicting survival risk). A schematic graphical representation of the overall process flow is reported in Fig 1. In this section we will provide a more detailed description of the applied algorithm with theoretical underpinnings when needed.

Feature agglomeration. The first step of our approach comprises a hierarchical feature clustering that is spatially independent. This technique is similar to a hierarchical agglomerative clustering procedure, but recursively merges features instead of samples.

Specifically, given the initial input data $\mathbf{X}^{(cases)} \in \mathbb{R}^{N\times Q}$, where N is the total number of cases and Q the total number of CpG Islands, to identify J clusters of CpG Islands we exploited the Euclidean Distance with Ward linkage. Then, we computed the representative value of the j-th agglomerated feature as

$$F_j^{(cases)} = \frac{1}{|C_j|}\sum_{q\in C_j} x_q$$

Where C_j is the j-th cluster and $x_q^{(cases)}$ is the q-th CpG Island, expressed as β-value bounded in [0,1], in the sample of cases. We exploit the same clustering structure (i.e., the same groups C_j of indexes q) to compute $F_j^{(controls)}$ from $\mathbf{X}^{(cases)} \in \mathbb{R}^{N\times Q}$. As mentioned, throughout this paper we refer to the aggregated CpG Islands as *features*.

Deep non-linear survival modeling. To model the relationship between the just defined input features and BC TTD we exploit a Deep Survival NN, closely related to DeepSurv [44], to account for the complex non-linear interactions determining the phenotype. In particular, our Deep Survival model is a multi-layer feed-forward NN which predicts the effects of a patient's covariates on their hazard rate parameterized by the weights of the network θ. The input to the network for patient i is its baseline data in terms of DNAm features ($F^{(i)}$), while the output $\widehat{h_\theta}(F^{(i)})$ is a single node with a linear activation which estimates its log-risk function. Let T be the times to disease and E the event indicator, the objective function to optimize becomes:

$$\mathcal{L}(\theta) = -\sum_{i:E_i=1}\left(\widehat{h_\theta}(F^{(i)}) - log\sum_{j\in\mathcal{R}(T_i)} e^{\widehat{h_\theta}(F^{(j)})}\right)$$

where the risk set $\mathcal{R}(t) = \{i : T_i \geq t\}$ is the set of patients still at risk of disease at time t.

Deep Survival model architecture and training. The architecture of the Survival NN was inspired by DeepSurv containing a set of consecutive fully connected layers of decreasing dimensionality (i.e., number of nodes), each followed by a batch normalization layer. The output layer is composed of one node only, that makes a linear combination of the nodes in the second-last layer to predict the log-risk function $\widehat{h_\theta}(F)$. The number of layers and the number of nodes in each layer were optimized as described in Section Algorithm Design and Optimization, below.

Among other choices to improve Deep Survival Model training and mitigate the risk of sub-optimal parametrization, we included an *unsupervised pre-training* step before training our model to minimize the survival loss function. In particular, given a NN model architecture from input to second-last layer (embedding layer) before the single output node, hereby defined encoder, we initialized its parameters through AutoEncoder (AE) Layer-wise pretraining. In this work, we exploit the concept of Stacked AEs to identify an initial set of parameters for the Survival NN. To do that, we train progressively deeper AEs: starting from the input dimensionality of the model (input layer), we build an AE with the next layer as bottleneck and we train it to reconstruct the input; then, we retain the parameters from input layer to first layer to build and train the next AE from input to second layer of our Survival NN, and we progress until we have included all available layers in the encoder as bottlenecks.

At the end of this process, the parameters of the encoder are retained as initialization for the Survival NN.

While pretraining is known to aid network regularization, the set of parameters from second-last to output node are initialized at random, therefore we foster their regularization adding a Drop-out layer in between.

This pretraining procedure is meant to aid the extraction of stable effects profile, by biasing the initial parameters towards similar and beneficial for learning values. This would in turn improve both models' performance and stability of the derived Shapley values, computed by predicting with the parametrized models, as described in the following section. To validate this hypothesis, as mentioned in the Results section, we run an experiment comparing the performance of the model with and without pretraining (cf. Supporting Information S1 Table).

CpG Islands effects profile estimation through SHAP. To estimate the effect of the CpG Island-derived features on the prediction of the log-risk function we exploit a post-hoc explanation method applied to the deep survival NN. Among the existing methods, SHapley Additive exPlanation (SHAP), by [19], an algorithm that aims at explaining the prediction of an instance by computing the contribution of each feature to the prediction. This contribution is estimated in terms of Shapley regression values, an additive feature attribution method inspired by the coalitional game theory [45]. In the original Shapley formulation, feature impact is defined as the change in the expected value of the model's output when a feature is observed versus unknown. Given a specific prediction $f(x)$, we can compute the Shapley values $\varphi_i(f, x)$ using a weighted sum that represents the impact of each feature being added to the model averaged over all possible orders of features being introduced:

$$\varphi_i(f,x) = \sum_{S \subseteq S_{all} \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)]$$

where S is a subset of all the features (S_{all}) used in the model, M is the number of features and $f_x(S)$ is the prediction for feature values in set S that are marginalized over features that are not included in that set. This computation is prohibitive, therefore SHAP's authors proposed several sampling-based alternatives to estimate these values. Among them, we chose KernelSHAP [19], that is model agnostic. Of note, model-specific gradient-based implementations exist for Deep Neural Networks, such as DeepSHAP [19]. However, we decided to exploit KernelSHAP to allow for future evolutions of our algorithm, that might include a different survival prediction model within the proposed framework. However, for the sake of completeness, in Supporting Information S6 Fig we include the comparison of feature importance ranking extracted with the two algorithms, to demonstrate their interchangeability. Nevertheless, to improve computational time for NN models, we would suggest to exploit DeepSHAP's efficient implementation.

In KernelSHAP, "Missing features" in a sampled feature coalition are simulated by averaging the model's prediction over bootstrapped samples with values for these features sampled from the so called ***background*** **dataset**. This makes the estimation dependent on the choice of such reference, an aspect that we exploited to gather different biological insights by changing the BC cases composition of the background dataset supplied to the method (cf. Results Section).

In general, in our algorithm, to derive the importance weights that constitute the global effects profile, for each of the K splits, we trained the model on the training data and we computed SHAP values of test data w.r.t. the background control group. As SHAP values are computed locally for each observation, we estimate the features' impact ($w^{(k)}$) as the mean of the absolute value of local estimates.

Moreover, to reduce computational time of the nested sampling to estimate SHAP values, as suggested by the authors [20], we supplied to the algorithm the background data grouped into 20 centroids, i.e. a sample of 20 representative observations derived from the application of k-means algorithm (with $k = 20$) to the reference sample.

Performance measures

Time-to-event prediction performance. To evaluate the modeling performance of both our Deep Survival NN model and CoxPH we exploited the traditional Harrel's Concordance Index (CI) [46]. The Harrel CI is a measure of rank correlation between the models' predicted risk scores and the observed time points. It quantifies how well a model predicts the ordering of patients' diagnosis times. It can be computed by the following formula:

$$CI = \frac{\sum_{i,j} 1_{T_j < T_i} \cdot 1_{\eta_j > \eta_i} \cdot E_j}{\sum_{i,j} 1_{T_j < T_i} \cdot E_j}$$

where η_i is the risk score of unit i.

Weights profiles robustness. To estimate the robustness of the effects profile estimated across K iterations in our Deep Survival EWAS, we exploited the metric developed in [25]. Specifically, for each split k, with $k = 1, \ldots, K$, we get a different $w_j^{(k)}$ and we rank order features based on these importance scores. To measure weights robustness we measure the dispersion of these K SHAP-derived importance rankings $\mathbf{r}_1, \ldots, \mathbf{r}_K$. To do that, for each pair of ranked vectors $\mathbf{r}_i$ and $\mathbf{r}_j$ we compute their dissimilarity as the **Kendall Tau Distance between two rank vectors** [47]. This distance can be expressed as the minimum number of bubble swaps needed to convert one rank vector to the other, divided by the total number of pairs in the vector. Additionally, we impose a $1/j$ penalty on any comparison involving the j-th rank as it is more relevant to identify the top predictors of a model, rather than accurately rank all features. Finally, we truncate the rank vectors after the top 10 ranks to reduce computation time. The mean of these $\binom{K}{2}$ pairwise distances represent the stability of our measurements, that we define **Kendall Tau Stability** (**KT-stability**).

Algorithm design and optimization

To optimize the design of the overall pipeline of the algorithm we needed to make several architectural and hyperparameter choices. In particular, we had to optimize jointly the dimensionality of the input to the Survival NN model (i.e. the number of features clusters) and the architecture of the NN itself. To do that, we tested a set of combinations and compared their performances on CI and KT-stability, with the ultimate goal of achieving a reliable set of weights for the effects profile. In particular, we defined a set of input dimensions $J \in [128, 256, 512, 1024]$, and for each of those dimensions we tested two alternative NN architecture, reducing the number of nodes in half at each subsequent layer down to a final embedding layer of either 16 or 32 nodes. Therefore, in total, we tested 8 NN configurations. The dimensions of NN layers were defined as mentioned in order to aid memory efficiency in model training, and to make these modular architectures more comparable. In total, we tested 8 NN configurations, as reported in S2 Table in Supporting Information. For pre-training, model parameters were initialized randomly following the approach in Glorot et al. The MSE loss function was optimized by mini-batch stochastic gradient descent with momentum with a learning rate of 0.1 and a momentum parameter of 0.9, in accordance with the most common hyperparameters definition for this optimizer. The mini-batch size was set to 32, counting for around 15% of the training set. We trained the model for 150 epochs. After pre-training, the resulting parameters were exploited as initialization parameters for the overall survival NN, fine-tuned through gradient descent of the previously defined survival loss function, while the last layer was initialized randomly. The optimization was performed with adaptive moment estimation, inspired by DeepSurv, with a learning rate of 0.0001, a mini-batch size of 32 samples and a drop-out probability of the last fully connected layer of 0.1. We fine-tuned the whole network for 150 epochs leaving all other hyperparameters to default. To identify the best combination of input dimensionality and survival model, we first defined the 4 grouped input datasets by varying J in feature agglomeration, then we run the entire pipeline from NN pretraining to SHAP weights estimation (using all BC controls as background data) repeated on 10 random data splits, comparing the average results (cf. S2 Table). As an additional decision support resource, we compared the distributions of CpG Islands grouped in each aggregated feature for the different K values (S7 Fig). As none of the alternative showed significantly superior performances on the two metrics together, we opted for the smaller model (i.e. 128 nodes in the input and 16 nodes in the last fully connected layer before output) that granted comparable results to the others with significantly less parameters to train. Moreover, the size distribution with input granularity of 128 features (cf. S7 Fig) shows an average of features' size between 19 and 35 CpG Islands per feature, which seemed a balanced and reasonable clustering considering the input of almost 4,000 Islands.

Cox Proportional Hazards model implementation

As mentioned, to validate the need for a complex non-linear model such as a DSNN as predictor in the framework of our algorithm, we compared its performance with that of a simpler CoxPH model trained on the aggregated feature sets.

We trained one CoxPH model for all feature aggregation values (*J*) tested when optimizing our approach. Models were implemented in Python 3 using scikit-learn library. In order to make a fair comparison with the DSNN, we wanted to find a common set of hyperparameters for all input dimensions, as we did for the nonlinear model. The fitting of CoxPH models and the hyperparameters we defined, as detailed in the following, were determined mostly by convergence issues and extremely long running times (over 4 hours), especially at higher dimensionalities. In particular, we set the number of iterations to 100, and a learning rate of 0.1. Additionally, we set the initialization to all zeros, since that was the only initialization avoiding gradient explosion after a few iterations. For what concerns penalties, we kept the default value of the python CoxPH implementation, which is a light L2 penalty of 1e-4, to avoid excessively strong shrinking of models' parameters. Indeed, as we had to evaluate the models' applicability as a building block for EWAS, we needed an effect magnitude associated to each input dimension. Nevertheless, we still allowed some regularization to aid the model convergence and reduce the variability of its estimates.

Supporting information

S1 Appendix. Detailed data description. Supporting information on cohort description and DNA methylation data extraction and preprocessing.
(PDF)

S1 File. Latent space PCA plots. PCA plots in 2D of the embedded points (i.e. each data point is a patient) in the 16-dimensional latent spaces defined by the best model exploited for Deep Survival EWAS (i.e. input 128 nodes, bottleneck 16 nodes). Each row of plots represents one of the K = 10 splits, i.e. one of the trained Deep Survival models. The plots on the left represent the latent space after pre-training, the plots on the right represent the latent space after supervised fine-tuning of the model to predict the survival outcome. Patients are grouped and colored according to 4 time-to-event classes.
(PDF)

S2 File. Kaplan-Meyer plots of F_{120} CpG Islands. Entire set of Kaplan-Meyer curves for all CpG Islands grouped in feature F_{120}.
(PDF)

S1 Table. Performance comparison with and without pretraining. Performance comparison of Deep Survival EWAS algorithm with and without pretraining each of the K survival models. Each row reports results (in terms of KT stability and Concordance Index) for one of the tested model architectures.
(PDF)

S2 Table. Optimization results. Performance in terms of Kendall-Tau Stability (robustness) and Harrell C-Index (survival prediction performance) for all the Deep Survival Network architecture trained in optimization. Performance is averaged across K splits; 95% confidence intervals are reported in parentheses. In the first column (Input) the input shape (J), referring to the granularity of the CpG Island agglomeration. In the second column (Latent) the dimensionality of the Survival NN layer before output. In the second column (Architecture), the list of layers with respective number of nodes, up until before output layer (one single output node for all architectures).
(PDF)

S3 Table. Deep Survival EWAS effects profile estimation w.r.t. reference groups. Estimated CpG islands effects profiles. Each sheet in the file contains the results for one reference group (respectively: (1) BC controls, (2) Matching Controls, (3) All controls, (4) All controls with cases). The first column reports the CpG Island name, the second reports the Feature they are agglomerated into. Third column reports the effect weight (w), while last column reports the ranking of the feature in terms of importance (i.e. ordered by descending magnitude of effect weight associated to the feature).
(XLSX)

S4 Table. Deep Survival EWAS Gene Set enrichment analyses results. Deep Survival EWAS GSEA results for all four reference groups. Each sheet in the file contains the results for one reference group (respectively: (a) BC controls, (b) Matching Controls, (c) All controls, (d) All controls with cases). The first column reports the KEGG pathway, followed by pathway code and empirical p-value (last column).
(XLSX)

S5 Table. Standard EWAS association results. Results of independent modeling of CpG Islands w.r.t. TTD. The table reports p-values and test statistics obtained by modeling one univariate CoxPH for each island.
(XLSX)

S6 Table. Standard EWAS Gene Set enrichment analysis results. Results of wighted GSEA for Standard EWAS approach, where weights were determined by the test statistics in the univariate CoxPH performed independently for each CpG Island. The first column reports the KEGG pathway, followed by pathway code and empirical p-value (last column).
(XLSX)

S7 Table. Standard EWAS CpG site-specific association results. Results of independent modeling of single CpG sites w.r.t. TTD. The table reports p-values and test statistics obtained by modeling one univariate CoxPH for each CpG site.
(CSV)

S8 Table. Standard EWAS CpG site-specific Gene Set enrichment analysis results. Results of wighted GSEA for Standard EWAS approach, where weights were determined by the test statistics in the univariate CoxPH performed independently for each CpG site. The first column reports the KEGG pathway, followed by pathway code and empirical p-value (last column).
(CSV)

S9 Table. Multivariate Cox Proportional Hazards results. Performance in terms of Kendall-Tau Stability (robustness) and Harrell C-Index (survival prediction performance) for all Multivariate CoxPH models with different input dimensions. Performance is averaged across K splits; 95% confidence intervals are reported in parentheses. In the first column (Input) the input shape (*J*), referring to the granularity of the CpG Island agglomeration.
(PDF)

S10 Table. XGBoost algorithm results. Performance in terms of Kendall-Tau Stability (robustness) and Harrell C-Index (survival prediction performance) for all XGBoost algorithms with different input dimensions. Performance is averaged across K splits; 95% confidence intervals are reported in parentheses. In the first column (Input) the input shape (*J*), referring to the granularity of the CpG Island agglomeration.
(PDF)

S1 Fig. Standard EWAS weights profile. Weights profile for Standard EWAS approach, where each CpG Island is associated with the p-value of the test statistic in an independent CoxPH model. Panel A reports the p-values without Bonferroni adjustment, panel B reports the same p-values after Bonferroni adjustment. The red line denotes the p-value threshold of 0.05.
(PNG)

S2 Fig. Predictive Performance comparison. Predictive Performance (Harrel CI) of all the tested Deep Survival NN architectures (blue) and all multivariate CoxPH models fitted. The values in parenthesis for Deep Survival NNs represent the number of input nodes (i.e. the granularity of features' clusters) and the number of nodes in the last layer before the output. Whereas the parenthesis for CoxPH models report the granularity of the input features' clusters. Dots represent the average performance value, while bands report the confidence intervals around the mean computed on the K = 10 splits.
(PNG)

S3 Fig. Weights stability comparison. Importance weights stability performance (KT-stability) of all the tested Deep Survival NN architectures (blue) and all multivariate CoxPH models fitted. The values in parenthesis for Deep Survival NNs represent the number of input nodes (i.e. the granularity of features' clusters) and the number of nodes in the last layer before the

output. Whereas the parenthesis for CoxPH models report the granularity of the input features' clusters. Dots represent the average performance value, while bands report the confidence intervals around the mean computed on the K = 10 splits.
(PNG)

S4 Fig. Enrichment analysis on XGBoost weights profile. Results for the enrichment analysis performed on weights' profile estimated via TreeSHAP from XGBoost classifier.
(JPG)

S5 Fig. Analysis of CpG sites and their correlation. (a) Pie chart: frequencies of the Intergenic (red) and Intragenic (green and blue) CpG Islands. Most (48% of the total) intragenic islands intersect one gene element (gene body or gene promoter), whereas 14% of CpG Islands are intragenic and overlap gene body and promoter of the same gene. (b) Histogram: distribution of the pairwise Pearson correlations of CpGs pairs in the same island for the 'blue' category. Only 2% of pairwise comparisons have a negative Pearson correlation coefficient.
(JPG)

S6 Fig. Comparison of DeepSHAP and Kernel SHAP. Top ten features in terms of highest estimated weights from Deep Survival EWAS algorithm: on the left, the weights were estimated exploiting the model-agnostic Kernel SHAP: on the right, the weights were estimated with the model-specific DeepSHAP algorithm.
(PNG)

S7 Fig. Features' dimension distributions for varying *J*. Distribution of dimensions of the features for each input granularity *J* (i.e. 128, 256, 512, 1024), in terms of number of CpG Island they group.
(PNG)

Acknowledgments

We thank the EPIC Italy research group (Carlotta Sacerdote, Vittorio Krogh, Domenico Palli, Salvatore Panico, Rosario Tumino, Paolo Vineis and their collaborators) for giving us access to the data of the BC study nested in the cohort.

Author Contributions

Conceptualization: Michela Carlotta Massi, Francesca Ieva, Giovanni Fiorito.

Formal analysis: Michela Carlotta Massi, Lorenzo Dominoni.

Methodology: Michela Carlotta Massi, Francesca Ieva, Giovanni Fiorito.

Software: Michela Carlotta Massi, Lorenzo Dominoni.

Supervision: Francesca Ieva, Giovanni Fiorito.

Visualization: Lorenzo Dominoni.

Writing – original draft: Michela Carlotta Massi, Giovanni Fiorito.

References

1. Yang X, Yan L, Davidson NE. DNA methylation in breast cancer. Endocrine-related cancer. 2001; 8 (2):115–127. https://doi.org/10.1677/erc.0.0080115 PMID: 11446343
2. Das PM, Singal R. DNA methylation and cancer. Journal of clinical oncology. 2004; 22(22):4632–4642. https://doi.org/10.1200/JCO.2004.07.151 PMID: 15542813

3. Muse ME, Titus AJ, Salas LA, Wilkins OM, Mullen C, Gregory KJ, et al. Enrichment of CpG island shore region hypermethylation in epigenetic breast field cancerization. Epigenetics. 2020; 15(10):1093–1106. https://doi.org/10.1080/15592294.2020.1747748 PMID: 32255732

4. Ennour-Idrissi K, Dragic D, Durocher F, Diorio C. Epigenome-wide DNA methylation and risk of breast cancer: a systematic review. BMC cancer. 2020; 20(1):1–10. https://doi.org/10.1186/s12885-020-07543-4 PMID: 33129307

5. Chen M, Wong EM, Nguyen TL, Dite GS, Stone J, Dugué PA, et al. DNA methylation-based biological age, genome-wide average DNA methylation, and conventional breast cancer risk factors. Scientific Reports. 2019; 9(1):1–10. https://doi.org/10.1038/s41598-019-51475-4 PMID: 31636290

6. Caini S, Fiorito G, Palli D, Bendinelli B, Polidoro S, Silvestri V, et al. Pre-diagnostic DNA methylation patterns differ according to mammographic breast density amongst women who subsequently develop breast cancer: a case-only study in the EPIC-Florence cohort. Breast Cancer Research and Treatment. 2021; p. 1–10. PMID: 34101077

7. Bodelon C, Ambatipudi S, Dugué PA, Johansson A, Sampson JN, Hicks B, et al. Blood DNA methylation and breast cancer risk: a meta-analysis of four prospective cohort studies. Breast Cancer Research. 2019; 21(1):1–9. https://doi.org/10.1186/s13058-019-1145-9 PMID: 31101124

8. Hüls A, Czamara D. Methodological challenges in constructing DNA methylation risk scores. Epigenetics. 2020; 15(1-2):1–11. https://doi.org/10.1080/15592294.2019.1644879 PMID: 31318318

9. Gagliardi A, Dugué PA, Nøst TH, Southey MC, Buchanan DD, Schmidt DF, et al. Stochastic epigenetic mutations are associated with risk of breast cancer, lung cancer, and mature b-cell neoplasms. Cancer Epidemiology and Prevention Biomarkers. 2020; 29(10):2026–2037. https://doi.org/10.1158/1055-9965.EPI-20-0451 PMID: 32788174

10. Levy JJ, Titus AJ, Petersen CL, Chen Y, Salas LA, Christensen BC. MethylNet: an automated and modular deep learning approach for DNA methylation analysis. BMC bioinformatics. 2020; 21(1):1–15. https://doi.org/10.1186/s12859-020-3443-8 PMID: 32183722

11. Macías-García L, Martínez-Ballesteros M, Luna-Romera JM, García-Heredia JM, García-Gutiérrez J, Riquelme-Santos JC. Autoencoded DNA methylation data to predict breast cancer recurrence: Machine learning models and gene-weight significance. Artificial Intelligence in Medicine. 2020; 110:101976. https://doi.org/10.1016/j.artmed.2020.101976 PMID: 33250148

12. Zheng C, Xu R. Predicting cancer origins with a DNA methylation-based deep neural network model. PloS one. 2020; 15(5):e0226461. https://doi.org/10.1371/journal.pone.0226461 PMID: 32384093

13. Mallik S, Seth S, Bhadra T, Zhao Z. A linear regression and deep learning approach for detecting reliable genetic alterations in cancer using dna methylation and gene expression data. Genes. 2020; 11 (8):931. https://doi.org/10.3390/genes11080931 PMID: 32806782

14. Bichindaritz I, Liu G, Bartlett C. Integrative survival analysis of breast cancer with gene expression and DNA methylation data. Bioinformatics. 2021; 37(17):2601–2608. https://doi.org/10.1093/bioinformatics/btab140 PMID: 33681976

15. Azher ZL, Vaickus LJ, Salas LA, Christensen BC, Levy JJ. Development of biologically interpretable multimodal deep learning model for cancer prognosis prediction. In: Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing; 2022. p. 636–644.

16. Poirion OB, Jing Z, Chaudhary K, Huang S, Garmire LX. DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data. Genome medicine. 2021; 13 (1):1–15. https://doi.org/10.1186/s13073-021-00930-x PMID: 34261540

17. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver CancerUsing Deep Learning to Predict Liver Cancer Prognosis. Clinical Cancer Research. 2018; 24(6):1248–1259. https://doi.org/10.1158/1078-0432.CCR-17-0853 PMID: 28982688

18. Liu B, Liu Y, Pan X, Li M, Yang S, Li SC. DNA methylation markers for pan-cancer prediction by deep learning. Genes. 2019; 10(10):778. https://doi.org/10.3390/genes10100778 PMID: 31590287

19. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems; 2017. p. 4768–4777.

20. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature machine intelligence. 2020; 2(1):56–67. https://doi.org/10.1038/s42256-019-0138-9 PMID: 32607472

21. Liu H, Wu X, Zhang S. Feature selection using hierarchical feature clustering. In: Proceedings of the 20th ACM international conference on Information and knowledge management; 2011. p. 979–984.

22. Yousefi S, Amrollahi F, Amgad M, Dong C, Lewis JE, Song C, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Scientific reports. 2017; 7(1):1–11. https://doi.org/10.1038/s41598-017-11817-6 PMID: 28916782

23. Levy JJ, Chen Y, Azizgolshani N, Petersen CL, Titus AJ, Moen EL, et al. MethylSPWNet and MethylCapsNet: Biologically Motivated Organization of DNAm Neural Network, Inspired by Capsule Networks. bioRxiv. 2021; p. 2020–08. https://doi.org/10.1038/s41540-021-00193-7 PMID: 34417465

24. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome research. 2002; 12(6):996–1006. https://doi.org/10.1101/gr.229102 PMID: 12045153

25. Liu B, Udell M. Impact of Accuracy on Model Interpretations. arXiv preprint arXiv:201109903. 2020;.

26. Charmpi K, Ycart B. Weighted Kolmogorov Smirnov testing: an alternative for gene set enrichment analysis. Statistical applications in genetics and molecular biology. 2015; 14(3):279–293. https://doi.org/10.1515/sagmb-2014-0077 PMID: 26030794

27. Khodabandehlou N, Mostafaei S, Etemadi A, Ghasemi A, Payandeh M, Hadifar S, et al. Human papilloma virus and breast cancer: the role of inflammation and viral expressed proteins. BMC cancer. 2019; 19(1):1–11. https://doi.org/10.1186/s12885-019-5286-0

28. Su J, Yan D, Wu S, et al. Epstein-Barr virus infection and increased sporadic breast carcinoma risk: a meta-analysis. Medical Principles and Practice. 2020; 29(2):195–200. https://doi.org/10.1159/000502131 PMID: 31311020

29. Ortega MA, Fraile-Martínez O, Asúnsolo Á, Buján J, García-Honduvilla N, Coca S. Signal transduction pathways in breast cancer: the important role of PI3K/Akt/mTOR. Journal of oncology. 2020; 2020. https://doi.org/10.1155/2020/9258396 PMID: 32211045

30. Azimi I, Roberts-Thomson S, Monteith G. Calcium influx pathways in breast cancer: opportunities for pharmacological intervention. British journal of pharmacology. 2014; 171(4):945–960. https://doi.org/10.1111/bph.12486 PMID: 24460676

31. Xu Z, Sandler DP, Taylor JA. Blood DNA methylation and breast cancer: a prospective case-cohort analysis in the sister study. JNCI: Journal of the National Cancer Institute. 2020; 112(1):87–94. https://doi.org/10.1093/jnci/djz065 PMID: 30989176

32. Liu Y, Li X, Aryee MJ, Ekström TJ, Padyukov L, Klareskog L, et al. GeMes, Clusters of DNA Methylation under Genetic Control, Can Inform Genetic and Epigenetic Analysis of Disease. The American Journal of Human Genetics. 2014; 94(4):485–495. https://doi.org/10.1016/j.ajhg.2014.02.011 PMID: 24656863

33. Green C, Shen X, Stevenson AJ, Conole EL, Harris MA, Barbu MC, et al. DNA methylation signatures of C-reactive protein associations with structural neuroimaging measures and major depressive disorder. medRxiv. 2020;.

34. Zhang Y, Elgizouli M, Schöttker B, Holleczek B, Nieters A, Brenner H. Smoking-associated DNA methylation markers predict lung cancer incidence. Clinical epigenetics. 2016; 8(1):1–12. https://doi.org/10.1186/s13148-016-0292-4 PMID: 27924164

35. Cappozzo A, McCrory C, Robinson O, Sterrantino AF, Sacerdote C, Krogh V, et al. A blood DNA methylation biomarker for predicting short-term risk of cardiovascular events. 2022;.

36. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC bioinformatics. 2012; 13(1):1–16. https://doi.org/10.1186/1471-2105-13-86 PMID: 22568884

37. Barton SJ, Melton PE, Titcombe P, Murray R, Rauschert S, Lillycrop KA, et al. In epigenomic studies, including cell-type adjustments in regression models can introduce multicollinearity, resulting in apparent reversal of direction of association. Frontiers in genetics. 2019; p. 816. https://doi.org/10.3389/fgene.2019.00816 PMID: 31552104

38. van Veldhoven K, Polidoro S, Baglietto L, Severi G, Sacerdote C, Panico S, et al. Epigenome-wide association study reveals decreased average methylation levels years before breast cancer diagnosis. Clinical epigenetics. 2015; 7(1):1–12. https://doi.org/10.1186/s13148-015-0104-2 PMID: 26244061

39. Levy JJ, O'Malley AJ. Don't dismiss logistic regression: the case for sensible extraction of interactions in the era of machine learning. BMC medical research methodology. 2020; 20(1):1–15. https://doi.org/10.1186/s12874-020-01046-3 PMID: 32600277

40. Fiorito G, McCrory C, Robinson O, Carmeli C, Rosales CO, Zhang Y, et al. Socioeconomic position, lifestyle habits and biomarkers of epigenetic aging: a multi-cohort analysis. Aging (Albany NY). 2019; 11 (7):2045. https://doi.org/10.18632/aging.101900 PMID: 31009935

41. Gonzalez CA. The European prospective investigation into cancer and nutrition (EPIC). Public health nutrition. 2006; 9(1a):124–126. https://doi.org/10.1079/PHN2005934 PMID: 16512959

42. Palli D, Berrino F, Vineis P, Tumino R, Panico S, Masala G, et al. A molecular epidemiology project on diet and cancer: the EPIC-Italy Prospective Study. Design and baseline characteristics of participants. Tumori Journal. 2003; 89(6):586–593. https://doi.org/10.1177/030089160308900602 PMID: 14870823

43. Wu H, Caffo B, Jaffee HA, Irizarry RA, Feinberg AP. Redefining CpG islands using hidden Markov models. Biostatistics. 2010; 11(3):499–514. https://doi.org/10.1093/biostatistics/kxq005 PMID: 20212320

44. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC medical research methodology. 2018; 18(1):1–12. https://doi.org/10.1186/s12874-018-0482-1 PMID: 29482517

45. Shapley LS. A value for *n*-person games. In: Contributions to the Theory of Games. vol. 2. Princeton University Press; 1953. p. 307–317.

46. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. Jama. 1982; 247(18):2543–2546. https://doi.org/10.1001/jama.1982.03320430047030 PMID: 7069920

47. Kumar R, Vassilvitskii S. Generalized distances between rankings. In: Proceedings of the 19th international conference on World wide web; 2010. p. 571–580.