DNA methylation biomarker selected by an ensemble machine learning approach predicts frailty in an HIV-positive veteran population

Chang Shu[1,2], Amy C. Justice[2,3], Xinyu Zhang[1,2], Vincent C. Marconi[4], Dana B. Hancock[5], Eric O. Johnson[5,6], Ke Xu[1,2]

1. Department of Psychiatry, Yale School of Medicine, New Haven, CT, 06516

2. Connecticut Veteran Healthcare System, West Haven, CT, 06515

3. Department of Internal Medicine, Yale School of Medicine, New Haven, CT, 06516

4. Division of Infectious Disease, Emory University School of Medicine, Atlanta GA

5. Center for Omics Discovery and Epidemiology, Behavioral Health Research Division, RTI International, Research Triangle Park, NC, USA

6. Fellow Program, RTI International, Research Triangle Park, NC, USA.

All corresponding address to

Ke Xu, M.D., Ph.D.

Associate Professor of Psychiatry

Yale School of Medicine

Email: ke.xu@yale.edu

Tel: 203-932-5711x7430

Word count: 2966

Abstract: 332

Figures: 7

Tables: 3

Abstract

Background: With the improved life expectancy in people living with HIV, predicting individual frailty is important for clinical care. DNA methylation (DNAm) has been previously linked to aging and mortality in non-HIV populations. We aim to establish a panel of DNAm biomarkers to predict frailty in HIV-positive population.

Methods: Our samples consist of HIV-positive participants (N_{total}=1,081) from the Veterans Aging Cohort Study (VACS) and were divided into the training set ($N_{training}$=460), the validation set ($N_{validation}$=114), and the testing set ($N_{testing}$=507). We used a well-established score, VACS Index, as a measure of frailty. Model training and fine-tuning were conducted using the ensemble method that aggregated prediction results of four base models. The final number of features were determined in the validation set and the model performance was assessed in the testing set. We conducted a survival analysis to assess whether the predicted frailty was associated with 10-year mortality and gene ontology enrichment analysis of the predictive CpG sites was performed.

Results: We selected a panel of 393 CpGs to build the ensemble machine learning model. The prediction model showed an excellent performance on predicting frailty with Area Under Curve of 0.809 (95%CI: 0.767-0.851) and balanced accuracy of 0.653 in the testing set. The predicted frailty based on our model was significantly associated with 10-year mortality (hazard ratio=1.79, p=4E-05). These 393 CpGs were located within or near 280 genes that were enriched in biological pathways including immune and inflammation response.

Conclusions: We identified a panel of predictive DNAm features associated with frailty and mortality among HIV-positive population. These DNAm features may serve as biomarkers to predict frailty people who live with HIV.

Keywords: DNA methylation, HIV, frailty, mortality, machine learning prediction

Introduction

Combination antiretroviral therapy has significantly improved life expectancy among HIV-positive individuals. The prolonged life expectancy increases frailty risk in the HIV population. The prevalence of frailty in people living with HIV is significantly higher and onset of frailty occurs at an earlier age compared to the general population (1). Frailty is characterized by marked vulnerability and is associated with increased mortality. Thus, prediction of frailty is important to identify vulnerable group and deliver clinical care to highly vulnerable HIV-positive patients. At present, frailty is usually defined by clinical symptoms (e.g., Fried's frailty phenotype, Rockwood and Mitnitski's Frailty Index) or a combination of lab tests that indicate organ damage such as the Veterans Aging Cohort Study index (VACS index) (2). No biomarkers available to capture the early stage of the frailty to predict individual vulnerability in HIV-positive patients.

A large body of evidence has demonstrated that epigenetic modification adapts internal and external environmental changes and that is associated with early stage of pathophysiological processes (3-6). DNAm, one type of the most widely studied epigenetic marks, is strongly related with aging (7-9), substance use (e.g. cigarette smoking and alcohol consumption) (10-16), and a variety of diseases (3-6, 17, 18). Because DNAm is relatively stable and easy to detect from body fluids biospecimen through non-invasive procedure, DNAm marks have emerged as robust biomarkers for cancer diagnosis (19), disease subtype classification (20, 21) and treatment response monitoring (22, 23).

DNAm may play an important role in frailty among HIV-positive individuals. Frailty is associated with elevated inflammation markers such as IL-6 and hsCRP in HIV-positive individuals (24). Genes involved in immune and inflammation processes are also subject to epigenetic modification in immune cells. We previously reported association of two CpG sites in the promoter region of *NLRC5* with HIV infection (25), and *NLRC5* is a major transcriptional activator of MHC class I gene. DNAm was also linked to HIV comorbid medical diseases such as HIV-positive diabetes and kidney function (26, 27). Furthermore, aging, a critical contributor to frailty, is significantly associated with thousands of CpGs in the epigenome, and the epigenetic clock and DNAm age are becoming widely recognized (7-9). DNAm age is significantly correlated with physical frailty in HIV-negative individuals (28) and HIV-positive individuals (29). In HIV-positive individuals, the average DNAm age is accelerated 5 to 10 years faster than HIV-negative individuals (30-33). DNAm accelerated 10 years faster in frailty individuals in comparing to non-frailty individuals in HIV-positive population (29). These age-related DNAm signatures are predictive of mortality (34-39). Therefore, we hypothesize that DNAm is associated with frailty and that DNAm signatures in blood can serve as biomarkers to predict frailty and mortality among HIV-positive individuals.

Machine learning methods have been applied to select informative DNAm features for clinical diagnosis and classification of complex diseases(15, 40, 41). Ensemble machine learning method can aggregate multiple machine learning models (base model) usually provide better prediction outcomes than single base model (42, 43). Ensemble method has been shown to perform well in personalized medicine and disease outcomes such as cancer and diabetes (44, 45).

In this study, by applying an ensemble learning approach, we aimed to identify DNAm features that can serve as biomarkers on frailty and to link frailty-associated DNAm to mortality among HIV-positive individuals. Here, VACS index, a well-established frailty score, was used as a measure of frailty in the HIV-positive population(46-48). Findings from this study provide a set of DNAm biomarkers to predict frailty for potential future clinic use and new insights into the epigenetic mechanism of frailty and mortality in HIV pathology.

Methods

As an overview, our analytical approach is shown in the flowchart in **Figure 1**. Briefly, we first applied an ensemble learning to improve machine learning results on predicting frailty, then examined the association of the selected CpG features with mortality by a survival analysis. Then, we conducted a gene-ontology enrichment analysis of the predictive features. Lastly, we tested a meta-analysis of epigenome-wide association (EWA) on frailty on the entire sample.

Study population

All samples in the sample set 1 and sample set 2 were from the VACS. The VACS is a prospective cohort study of veterans focusing on clinical outcomes in HIV infection (46). DNA samples were extracted from peripheral blood of 1,081 HIV-positive men from the VACS. Samples in sample set 1 was randomly partitioned into training set (80%, N=460) and validation set (20%, N=114), and samples in sample set 2 was used as independent testing set (N=507). Table 1 presents demographic and clinical information on age, race, smoking status, CD4 count, viral load, HIV medication

adherence, VACS index, and mortality in the training set, the validation set, and the testing set. Comparing to the testing set, the training and validation sets included slightly older individuals, more African Americans, and frailty individuals. There were no significant differences in sex, smoking, HIV medication adherence, CD4 count, $\log_{10}$ HIV-1 viral load or 10-year mortality across three sample sets.

Frailty measurement

VACS index is a well-established measure for frailty among HIV population (46-48). The VACS index is a composite score constructed by a sum of pre-assigned points on age, CD4 count, HIV-1 RNA, hemoglobin, platelets, aspartate and alanine transaminase (AST and ALT), creatine, estimated glomerular filtration rate (eGFR), and viral hepatitis C infection (48). Frailty was defined as VACS index > 40 since the predicted 3-year mortality of this group was 10% based on the previous report, (49). Prediction models were developed by machine learning methods to predict frailty (VACS index > 40) and non-frailty group (VACS index ≤ 40).

Genome-wide DNAm profiling and quality control

DNA samples in the sample set 1 were profiled by Infinium Human Methylation 450K BeadChip (HM450K), and DNA samples in sample set 2 were profiled by the Infinium Human Methylation EPIC BeadChip. DNAm for the training and validation sets were evaluated using the same quality control (QC) protocol (50) in the R package minfi (51). In detail, CpG sites on sex chromosomes and within 10 base pairs of a single nucleotide polymorphism were removed. The detection p-value threshold was set at 10^{-12} for both the training and validation sets. After QC, 408,583 common CpG sites

between HM450K and EPIC arrays were used for analysis to ensure the same coverage between two sets. Proportions of 6 cell types (CD4+ T cells, CD8+ T cells, Natural Killer T cells, B cells, monocytes and granulocytes) were estimated using the established method(52) for all samples in the sample set 1 and sample set 2.

Feature selection of CpG sites in the training set

We first pre-selected a panel of CpG sites associated with frailty based on EWA on frailty in the training set. CpG sites with $p<0.001$ were preselected for building predictive models. The marginal significance cutoff of $p=0.001$ was arbitrarily set to ensure sufficient number of CpG sites for building a model. The variable importance of each CpG in the preselected CpG sites was ranked by using Elastic-Net Regularized Generalized Linear Models (GLMNET) variable importance ranking (a score between 0-100) in *caret* based on 100 bootstraps (each bootstrap includes 70% of all samples). In GLMNET, importance of zero means the feature is neglected in model construction. CpG sites with zero variable importance for 80% of the bootstraps were removed from model development. The remaining CpG sites were ranked based on the median importance ranking among 100 bootstraps and were divided as different groups. The CpG sites were grouped based on importance rank of CpG sites. Each CpG group was used to build ensemble machine learning models.

Developing machine learning prediction models for frailty

1) Model development in the training set: We developed an ensemble classification model that aggregated the prediction models from four base machine learning models: random forest (RF), GLMNET, Support Vector Machines (SVM) and k-

Nearest Neighbors (KNN). These four different methods are selected to enlarge the diversity of algorithms (53-56). Ten-fold cross-validation was used in the model training process to minimize data overfitting. The four Models based on the number of CpG groups were separately trained to predict frailty in the training set and then aggregated with greedy ensemble method by using R package *caretEnsemble* (ver. 2.0.1) (57). The prediction performance of each ensemble model was evaluated by using area under receiver operating characteristic curve (auROC) and area under precision-recall curve (auPRC).

2) Select final CpG group in the validation set: The CpG group with both highest auROC and auPRC in the validation set was selected as the final feature group for the ensemble model.

3) Independent evaluation in testing set: using ensemble model and final feature group, we predicted the frailty group in the testing set and evaluated by using auROC and balanced accuracy. Balanced accuracy is defined as the average accuracy obtained on each class as shown in the following formula (58). Balanced accuracy was used in this study to avoid biased accuracy due to imbalanced samples (58).

$$Balanced\ accuracy = \frac{1}{2}\left(\frac{True\ positive}{True\ positive + False\ negative} + \frac{True\ negative}{True\ negative + False\ positive}\right)$$

Survival analysis

By using the final ensemble model we predicted each individual with frailty or non-frailty in the entire samples. Kaplan-Meier survival curves presented 10-year follow-up to visualize the survival differences by participants with predicted frailty or non-frailty.

Survival analysis was conducted by cox proportional hazards model on 10-year mortality comparing frailty and non-frailty groups. We used age as time scale *t*, and our model was adjusted for race, smoking, self-reported HIV medication adherence, log 10 of HIV viral load and CD4 count.

$$h(t) = h_0(t)\exp\,(\beta_1\, methylation\ score\ +\ \beta_2\, race +\ \beta_3\, smoking + \beta_4\, HIV\ medication\ adherence\ +\ \beta_5 \log_{10}(viral\ load) +\ \beta_6\, CD4\ count)$$

Biological interpretation of the predictive panel of CpG sites on frailty

We performed gene ontology (GO) enrichment analysis using *missMethyl* to adjust a bias from differing numbers of CpG site per gene(59, 60). Genes contained at least one predictive CpG site were used for GO analysis.

Epigenome-wide association analysis on frailty in all samples

Since DNA methylation of two sample sets was measured by two different platforms, we performed EWAs on frailty in sample set 1 and sample set 2 separately and then conducted a meta-analysis to detect epigenome-wide signals in the entire samples. In each EWA, we used a two-step linear model approach as previously described (50). The EWA model adjusted for confounding factors including age, race, smoking, cell type proportions and control principle components. Meta-analysis of sample set 1 and sample set 2 EWA was conducted using METAL(61). The weights of effect size estimates were the inverse of the corresponding standard errors for the meta-analysis(61). CpG sites with Bonferroni corrected $p<0.05$ were considered statistically significant.

Results

Feature selection and ensemble model training for frailty

A panel of 856 CpGs with $p<0.001$ was pre-selected based on EWA in the training set. We ranked these candidate predictors by median GLMNET importance ranking among 100 bootstraps using 70% of the training sample. We excluded 178 CpG sites that had variable importance of zero among 80% of the bootstraps. A final panel of 678 CpG sites were selected and were formed into 20 groups based on importance ranking to determine the best performing CpG group for the ensemble model (**Figure 2**).

In the training set, we developed our base models independently for the 20 groups of CpGs using four commonly used machine learning classification models (53-56): RF, GLMNET, SVM, and KNN. The performances of GLMNET, RF and SVM were mostly comparable in terms of auROC and auPRC plateaued to 1 with increasing number of CpGs in the training set (**Figure 3**, **Figure 4**). The performance of KNN was not as good as the other three methods, but its auROC and auPRC remained above 0.9 in the training set. Thus, an ensemble model combined the prediction results of all four base models was used to address imbalance performances from 4 models (**Figure 3**, **Figure 4**).

In the validation set, the performances of four base models varied. Three models of GLMNET, RF, and SVM showed good performance with both auROC and auPRC greater than 0.8, but the performance of KNN was poor with auROC or auPRC less than 0.8 (**Figure 3**, **Figure 4**). The ensemble model showed the best performance at 393 CpG sites with excellent performance of auROC (0.829) and auPRC (0.830). The

accuracy of this prediction model was 0.807, and the balanced accuracy that took into account of class imbalance between frailty and non-frailty participants was 0.782. Thus, the ensemble model with 393 features was determined as the final prediction model (**Table S1**).

In the testing set, the ensemble model with 393 features showed excellent performance with auROC of 0.809 (95%CI: 0.767-0.851), prediction accuracy of 0.761 and balanced accuracy of 0.653 (**Figure 5**), suggesting that an ensemble model of a panel of 393 features enable to predict frailty.

Survival analysis

By using the ensemble model to predict participants with frailty and non-frailty groups, we found that 10-year Kaplan-Meier curves differed between frailty and non-frailty participants (**Figure 6**). The Kaplan-Meier curves showed that participants with frailty showed significantly lower 10-year survival probability compared with non-frailty participants. Using Cox proportional-hazards model to control for sex, baseline age, race, viral load, CD4 count and antiviral medication adherence, this difference remained significant with a hazard ratio (HR) of 1.79 (95%CI: 1.35-2.37, p=4E-05).

Biological interpretations of predictive CpGs by gene ontology enrichment analysis

The 393 predictive CpGs were located in or near 280 genes. The top 8 enriched pathways based these 280 genes included response to virus (p= 4.26E-05), defense response (p=1.29E-04), cytokine receptor binding (p=1.48E-04) and regulation of response to interferon-gamma (p= 4.15E-04) (**Figure 7, Table 2**). Our findings

suggested that the selected 393 CpG sites are biologically relevant to HIV pathogenesis and progression.

Epigenome-wide association (EWA) on frailty

Meta-analysis on EWA of sample set 1 and sample set 2 identified 208 epigenome-wide significant CpGs after Bonferroni correction. These significant CpG sites were located in the 112 genes and included gene previously reported association with HIV pathogenesis. For example, cg07839457 in *NLRC5* had been previously reported to be associated with HIV infection(25). Interestingly, 30 of 208 CpG sites were also in the panel of machine-learning predictive CpGs (**Table 3**). Some of the overlapped CpGs were located in viral response genes such as *IFITM1* and *PARP9*.

Discussion

In this study, we present evidences that DNAm marks in blood are predictive of frailty and are associated with mortality in an HIV-positive population. We identified a panel of 393 CpG sites that were highly predictive for participant with frailty. We also found that ensemble predicted frailty was strongly associated with mortality in the HIV-positive population in our cohort. In addition, these machine learning-selected 393 DNAm features lied in genes involved in HIV pathogenesis and progression. Thus, we identified a panel of 393 DNAm biomarkers that add knowledge to the epigenetic mechanisms underlying frailty and mortality among people living with HIV.

We demonstrated that machine learning approach is robust to identify biologically relevant DNAm marks to predict frailty in HIV-positive individuals. One of the challenges for machine learning method is overfitting model that leads to a bias of prediction. We

took the several steps to address potential overfitting in developing the ensemble prediction models: 1) model development and final model selection were conducted separately in training set and validation set; 2) 10-fold cross-validation was performed during the training; 3) the use of ensemble model that aggregate prediction results from multiple base models instead of arbitrary choice of a specific machine learning prediction model; 4) the final model was evaluated in an independent test set. We further observed that performance of different model generated by individual machine learning method varied. Ensemble-based modeling method appears outperform than individual modeling method and can adjust the imbalanced performance among individual models. Our results suggested that ensemble-selected panel of CpG sites was relatively stable and robust.

Our results show that the frailty predicted by the selected CpG biomarker is strongly associated with mortality in a HIV-positive population. The finding is consistent with previous literature showing that DNAm marks in blood predict mortality in a HIV-negative population (39). Machine learning enables to select informative CpGs under epigenome-wide significant threshold that have predictive values to a phenotype. This panel of 393 CpG sites are biologically meaningful and may shed light on our understanding of the biological mechanisms of frailty for HIV-positive individuals. We found that 30 out of 393 CpGs reached epigenome-wide significance. The majority of the 393 CpG sites are located within or near genes that are involved in known HIV pathology and progression. For example, cg22930808 on *PARP9 and* cg07107453 on *IFI44L* were selected by machine learning and also reached epigenome-wide significance, which both genes are involved in *HIV pathogenesis.* cg12359279 is

located in the *MX1* gene (Interferon-Induced GTP-Binding Protein Mx1). *MX1* encodes a GTP-metabolizing protein induced by interferon I and II and is involved in interferon gamma signaling and Toll-like signaling pathway. The biological relevance of these 393 CpG sites was further supported by the gene ontology enrichment analysis. The top enriched pathways such as response to virus and cytokine receptor binding may point out important biological pathways that leads to frailty and increased risk of mortality among people living with HIV.

We acknowledge several limitations in this study. The generalizability of the selected CpG sites may be limited because our samples were predominantly middle-aged men, which may not generalize to frailty in a different age group. All samples in our study are HIV-positive and the identified CpGs have a limited application to HIV-negative population as we discussed above. Future studies in diverse populations is warranted to validate the selected methylation features.

Conclusions

We identified a panel of 393 predictive DNAm features in blood that was predictive of frailty and mortality among people living with HIV. These DNAm features may serve as biomarkers to identify frailty individuals and may help to prioritize genes to better understand the mechanisms of frailty in HIV-positive population. These DNAm features have potential for monitor HIV progression in future clinical care.

Figures

Figure 1: Flowchart of analytical procedures on selection of CpG sites in peripheral blood methylome, machine learning prediction models to predict frailty, survival analysis, gene ontology enrichment analysis, and epigenome-wide association analysis.

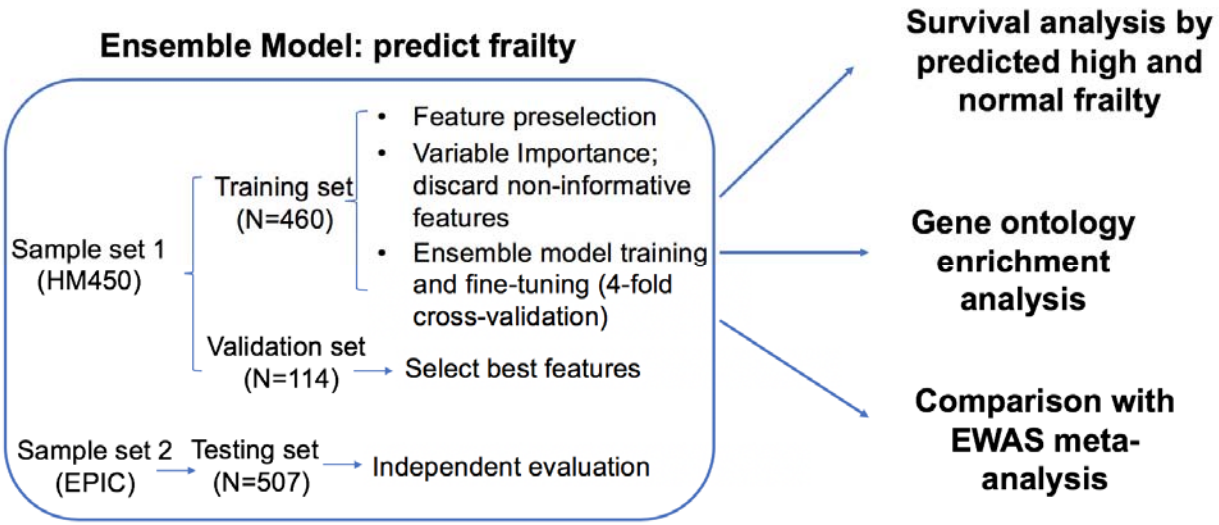


Figure 2: Variable Importance ranking of predictive machine learning CpG sites. Top 20 ranked CpG sites and 20 bootstraps are shown.

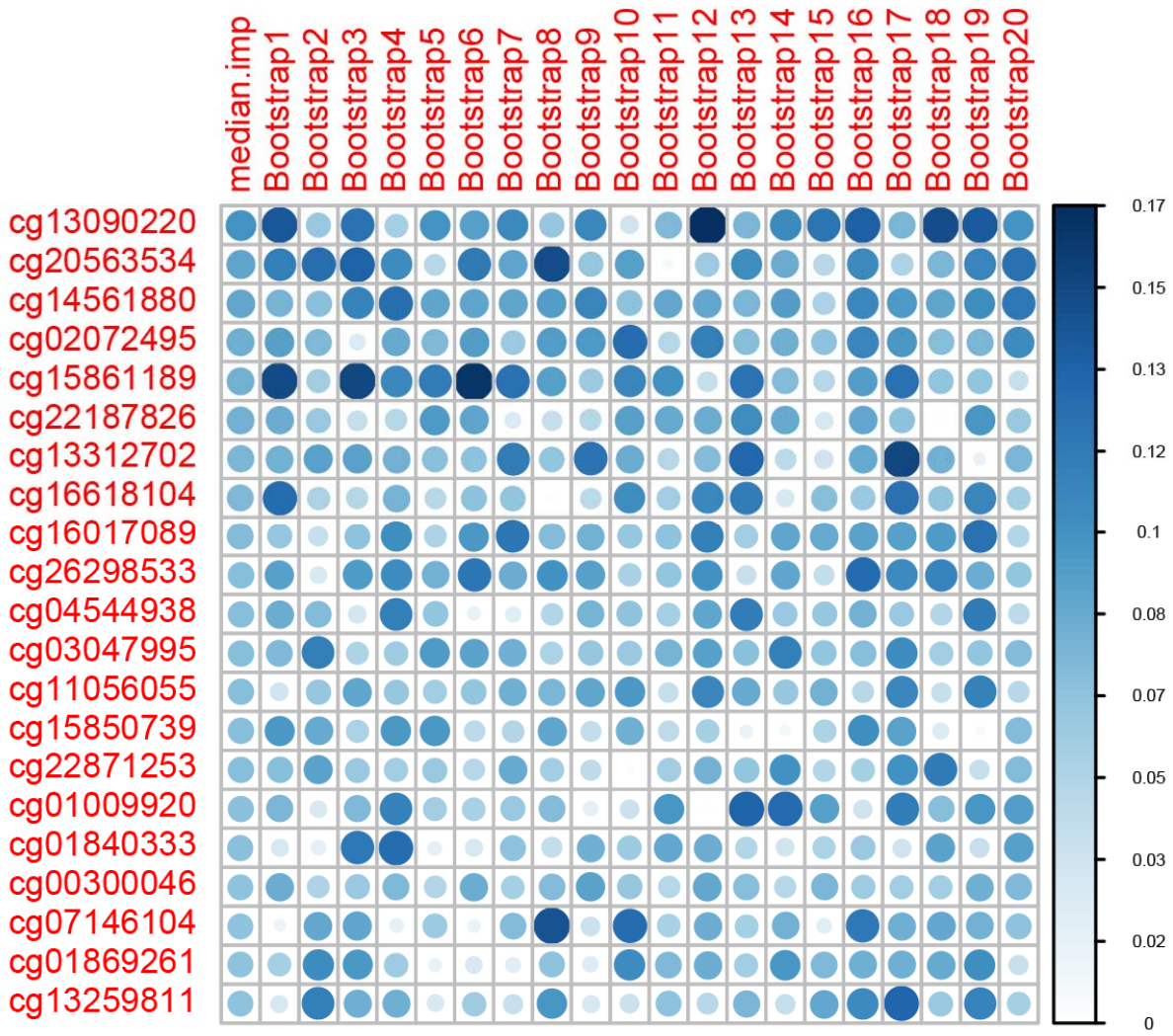


Figure 3: Area under the receiver operating characteristic(auROC) curve in the training set and validation set

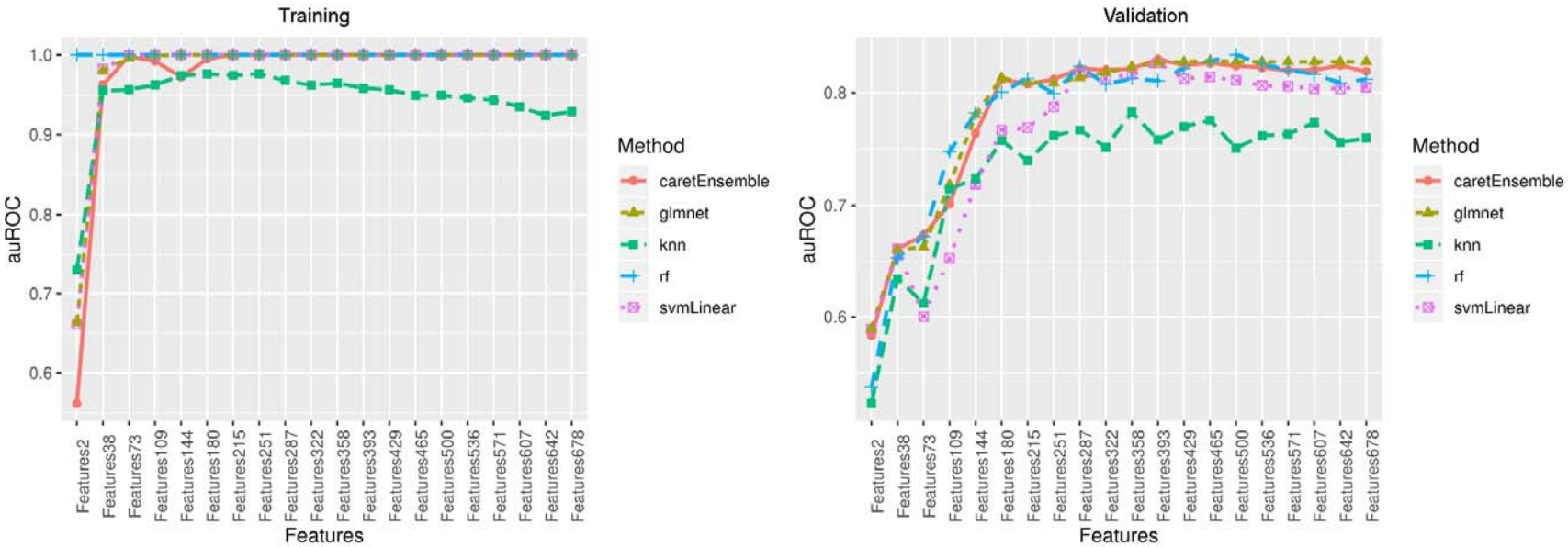


Figure 4: Area under the precision-recall curve(auPRC) in the training set and validation set

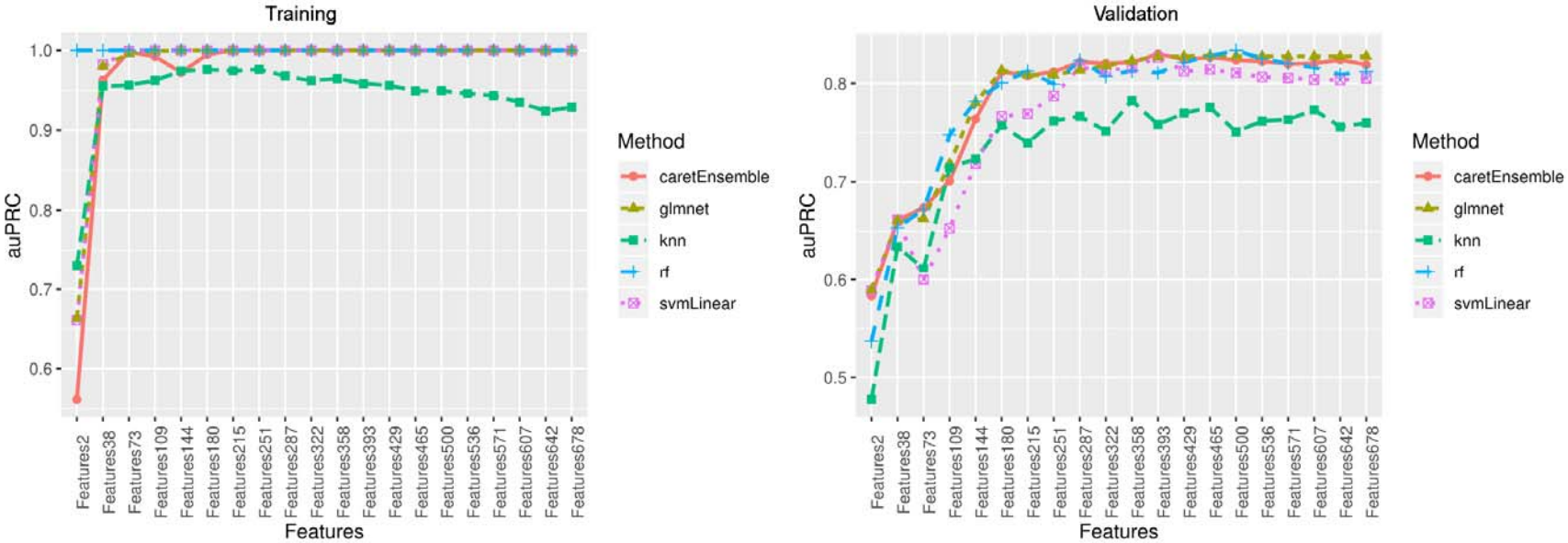


Figure 5: Receiver operating characteristic(auROC) curve in the testing set

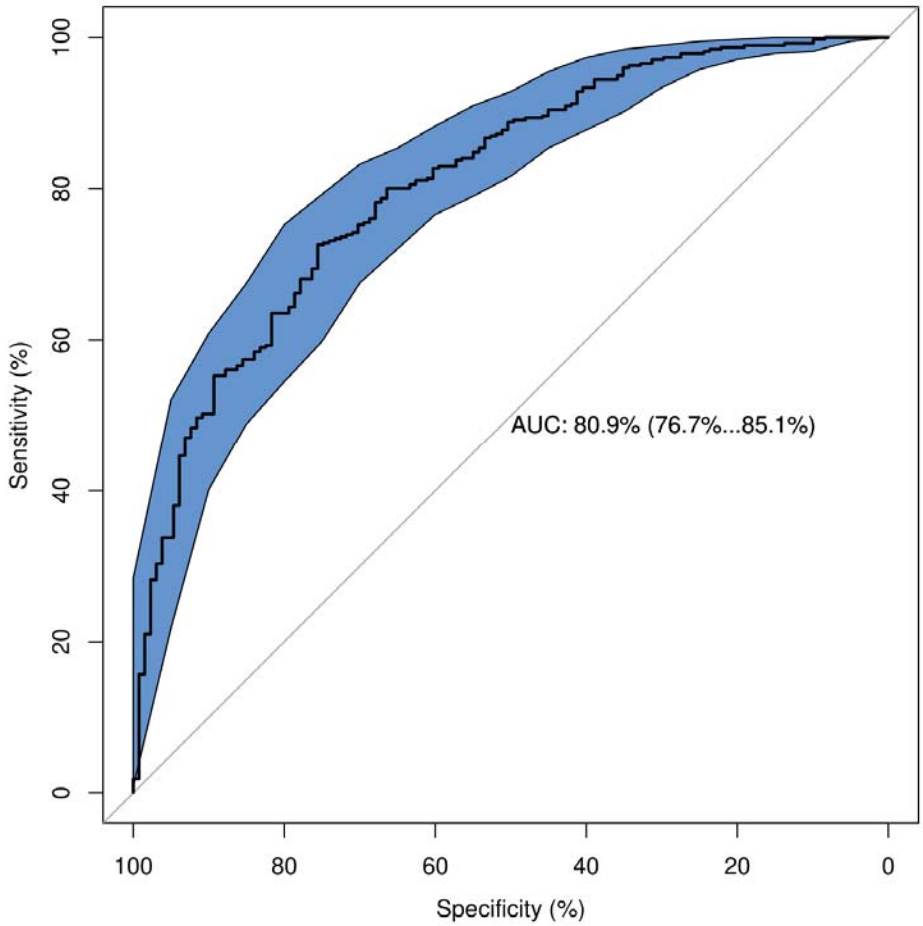


Figure 6: Kaplan-Meier curves of frailty and non-frailty groups

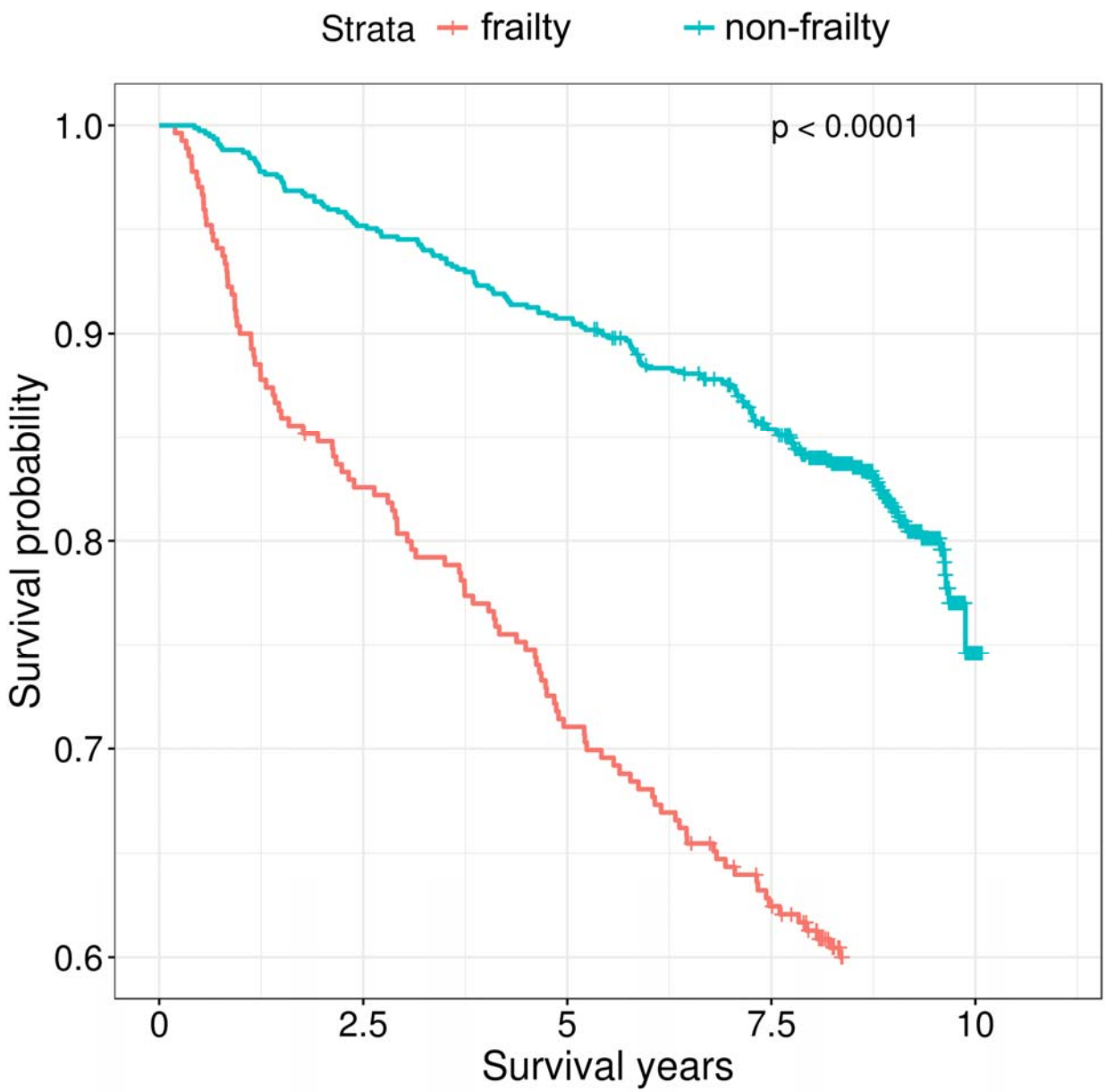
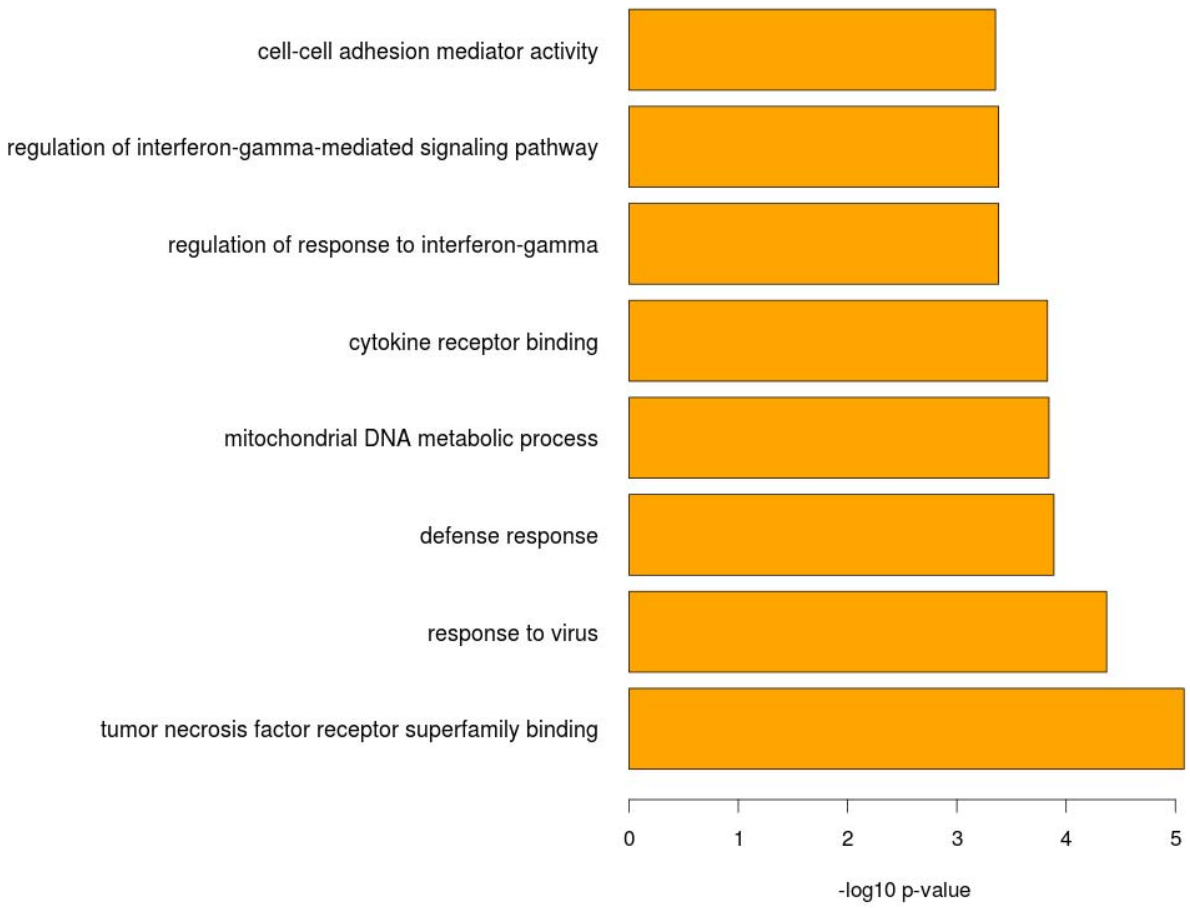


Figure 7: Top 8 enriched pathways by gene ontology term enrichment analysis of the selected 393 CpG sites



Tables

Table 1: Study sample characteristics

	Training set (N=460)	Validation set (N=114)	Testing set (N=507)	p value*
Age (year, mean +/-sd)	52.56 (7.54)	51.21 (8.09)	50.86 (7.67)	0.002
Female (%)	6 (1.3)	1 (0.9)	11 (2.2)	0.452
Race (%)				
Caucasian	45 (9.8)	9 (7.9)	46 (9.1)	
African Americans	392 (85.2)	103 (90.4)	409 (80.7)	0.003
Other	23 (5.0)	2 (1.8)	52 (10.3)	
Smokers (%)	360 (59.4)	309 (58.4)	294 (58.0)	0.719
HIV treatment adherence (%)	362 (78.7)	85 (74.6)	382 (75.3)	0.399
CD4 count	432.68 (291.43)	411.82 (281.55)	450.61 (280.05)	0.287
log 10 HIV-1 viral load	2.76 (1.22)	2.68 (1.17)	2.83 (1.22)	0.399
High frailty (%)	162 (35.2)	40 (35.1)	131 (25.8)	0.004
10-year mortality (%)	123 (26.7)	32 (28.1)	121 (23.9)	0.477

***ANOVA test is used for continous variables, chi-square test is used for categorical variables**

Table 2: Gene ontology term enrichment analysis of the selected 393 CpG sites that predict frailty in a HIV-positive population

Term	Total genes	Predictive Genes	P value
tumor necrosis factor receptor superfamily binding	46	7	8.33E-06
response to virus	303	15	4.26E-05
defense response	1505	39	1.29E-04
mitochondrial DNA metabolic process	18	4	1.43E-04
cytokine receptor binding	267	12	1.48E-04
regulation of response to interferon-gamma	25	4	4.15E-04
regulation of interferon-gamma-mediated signaling pathway	25	4	4.15E-04
cell-cell adhesion mediator activity	50	6	4.43E-04
intrinsic component of the cytoplasmic side of the plasma membrane	7	3	4.76E-04
immune response	1896	45	4.97E-04

Table 3: Overlapping CpG sites between machine learning selected CpG sites for frailty and epigenome-wide significant CpG sites on frailty

probe	Chr	Position	Nearest gene	Variable Importance	Meta Effect (SE)	Meta P	Refgene group	Relation to CpG island
cg01971407	11	313624	*IFITM1*	9.9	-0.0399 (0.0039)	8.05E-25	TSS1500	N_Shelf
cg22930808	3	122281881	*PARP9;DTX3L*	13.7	-0.1057 (0.0105)	7.24E-24	5UTR;TSS1500	N_Shore
cg23570810	11	315102	*IFITM1*	13.0	-0.0638 (0.0065)	1.15E-22	Body	N_Shore
cg14864167	8	66751182	*PDE7A*	12.0	-0.0832 (0.0085)	1.54E-22	Body	N_Shelf
cg01190666	20	62204908	*PRIC285*	19.4	-0.0343 (0.0035)	1.79E-22	5UTR	N_Shore
cg11702942	8	144102584	*LY6E;LY6E*	7.8	-0.0382 (0.004)	7.52E-22	Body;Body	S_Shore
cg03607951	1	79085586	*IFI44L*	23.0	-0.0683 (0.0072)	3.58E-21	TSS1500	
cg03848588	9	32525008	*DDX58*	14.4	-0.0274 (0.0029)	6.02E-21	Body	N_Shore
cg04582010	11	313120	*IFITM1*	22.1	-0.0455 (0.0052)	2.19E-18	TSS1500	S_Shore
cg18394552	5	159428643		24.4	0.0342 (0.0043)	1.57E-15		
cg03753191	13	43566902	*EPSTI1;EPSTI1*	4.1	-0.018 (0.0023)	2.43E-15	TSS1500	S_Shore
cg17267239	1	173640200	*ANKRD45*	21.8	-0.0186 (0.0025)	1.22E-13	TSS1500	S_Shore
cg12461141	11	5710654	*TRIM22*	37.6	-0.0274 (0.0037)	1.37E-13	TSS1500	
cg09251764	17	6659070	*XAF1;XAF1*	13.8	-0.0142 (0.002)	3.93E-13	TSS200	
cg05626226	4	106515450	*FLJ20184*	11.5	0.0279 (0.004)	2.45E-12	Body	
cg22959742	10	13913931	*FRMD4A*	16.4	0.0289 (0.0041)	3.08E-12	Body	
cg16936953	17	57915665	*TMEM49*	11.3	-0.0379 (0.0056)	1.95E-11	Body	
cg18181703	17	76354621	*SOCS3*	32.6	-0.0215 (0.0033)	1.15E-10	Body	N_Shore
cg07107453	1	79114976	*IFI44*	4.8	-0.0292 (0.0046)	2.26E-10	TSS1500	
cg25114611	6	35696870	*FKBP5;LOC285847*	12.8	-0.0125 (0.002)	4.78E-10	TSS1500;Body	S_Shore
cg01059398	3	172235808	*TNFSF10*	26.7	-0.0289 (0.0047)	7.85E-10	Body	
cg19459791	15	65363022		15.1	0.0159 (0.0026)	1.20E-09		S_Shelf
cg26282236	12	1025755	*RAD52*	30.2	0.0243 (0.0041)	3.63E-09	Body	

cg06357748	12	1025529	*RAD52*	10.7	0.0277 (0.0047)	4.12E-09	Body	
cg04442417	20	62191507	*PRIC285;PRIC285*	26.6	0.0201 (0.0035)	1.22E-08	Body	Island
cg14602222	12	1025663	*RAD52*	13.2	0.0241 (0.0042)	1.42E-08	Body	
cg26724018	11	5716255	*TRIM22*	11.9	-0.0171 (0.003)	1.49E-08	5UTR	
cg03084350	3	38065265	*PLCD1;PLCD1;PLCD1*	18.0	0.0114 (0.002)	2.05E-08	Body	N_Shore
cg00569896	4	204382		0.2	0.0254 (0.0047)	5.78E-08		N_Shore
cg12126344	1	12207564		8.4	-0.011 (0.002)	7.36E-08		

Legends

Figure 1: Flowchart of analytical procedures on selection of CpG sites in peripheral blood methylome, machine learning prediction models to predict frailty, survival analysis, gene ontology enrichment analysis, and epigenome-wide association analysis.

Figure 2: Variable Importance ranking of predictive machine learning CpG sites

Figure 3: Area under the receiver operating characteristic(auROC) curve in the training set and validation set

Figure 4: Area under the precision-recall curve(auPRC) in the training set and validation set

Figure 5: Receiver operating characteristic(auROC) curve in the testing set

Figure 6: Kaplan-Meier curves of frailty and non-frailty groups

Figure 7: Top 8 enriched pathways by gene ontology term enrichment analysis of the selected 393 CpG sites

Figure 8: Epigenome-wide association (EWA) on frailty: quantile-quantile plot, Manhattan plot and overlapping CpG sites between EWA significant CpGs and predictive CpGs.

Figure S1: Distribution of Veteran Aging Cohort Study index (VACS index) between training set and testing set

Table 1: Study sample characteristics

Table 2: Gene ontology term enrichment analysis of the selected 393 CpG sites that predict frailty

Table 3: Overlapping CpG sites between machine learning selected CpG sites for frailty and epigenome-wide significant CpG sites on frailty.

Table S1: Predictive CpG sites by variance importance

Table S2: Epigenome-wide significant CpG sites in meta-analysis of EWAS on frailty

List of abbreviations

AUC: Area Under Curve

CI: Confidence interval

DMR: differentially methylated region

DNA: Deoxyribonucleic acid

DNAm: DNA methylation

DAVID: Database for Annotation, Visualization and Integrated Discovery

EWA: epigenome-wide association

FDR: False discovery rate

FWER: Family-wise error rate

GLMNET: Lasso and Elastic-Net Regularized Generalized Linear Models

GO: Gene ontology

HIV: Human immunodeficiency virus

HM450K: Human Methylation 450K BeadChip

NK: Natural killer

PC: Principal component

QC: Quality control

SVM: Support Vector Machines

VACS: Veterans Aging Cohort Study

VACS index: Veterans Aging Cohort Study index

XGBoost: Extreme Gradient Boosting Tree

Declarations

Ethics approval and consent to participate

The study was approved by the committee of the Human Research Subject Protection at Yale University and the Institutional Research Board Committee of the Connecticut Veteran Healthcare System. All subjects provided written consents.

Availability of data and materials

Demographic and clinical variables and DNAm data for the VACS samples were submitted to GEO dataset (GSE117861) and are available to the public. All codes for analysis are also available upon a request to the corresponding author.

Competing interests

The authors declare that they have no competing interests.

Funding

The project was supported by the National Institute on Drug Abuse (R03DA039745, R01DA038632, R01DA047063, R01DA047820).

Authors' contributions

CS was responsible for data analysis and manuscript preparation. ACJ provided DNA samples, clinical data, and contributed to manuscript preparation. XZ was responsible for the bioinformatics data processing. VM involved clinical data collection and manuscript preparation. DH and EJ contributed to analytical approach and the manuscript preparation. KX was responsible for the study design, study protocol, sample preparation, data analysis, interpretation of findings, and manuscript preparation.

Acknowledgements

The authors appreciate the support of the Veteran Aging Study Cohort Biomarker Core and Yale Center of Genomic Analysis.

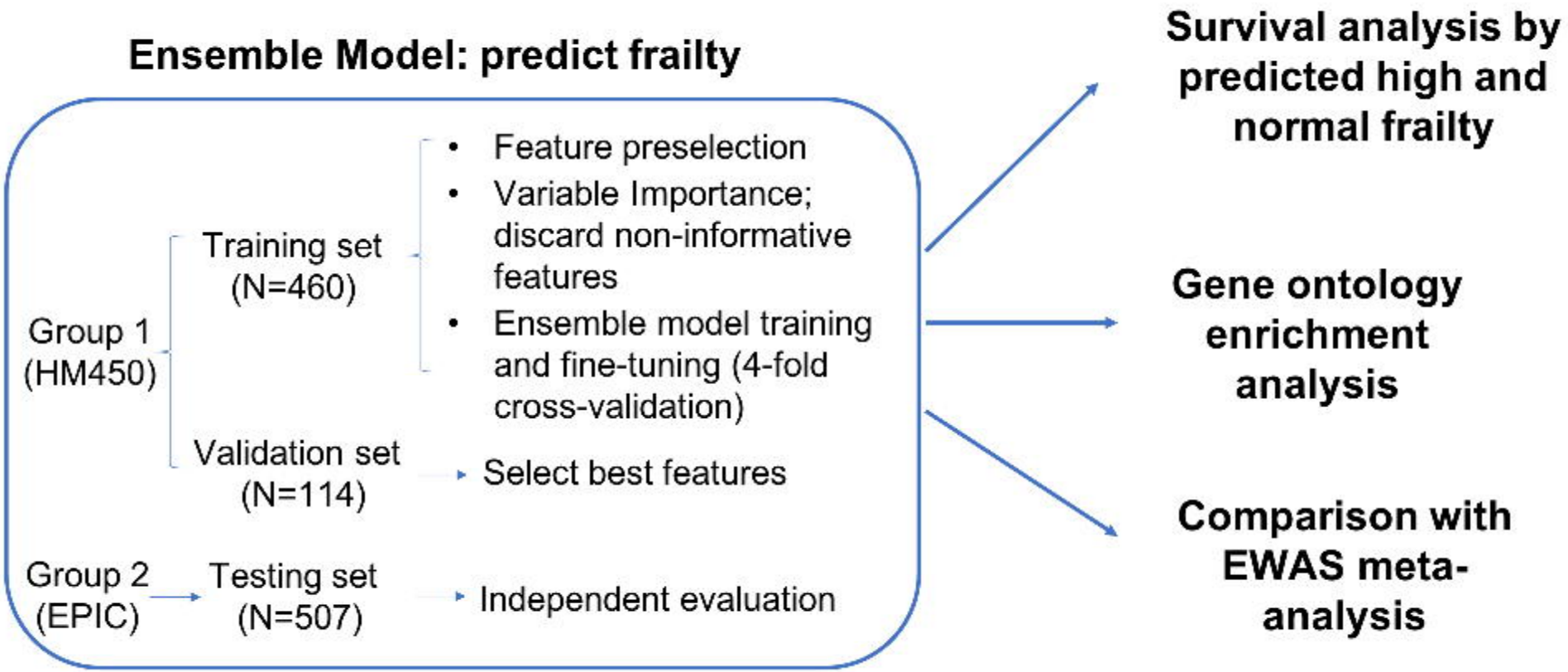
References

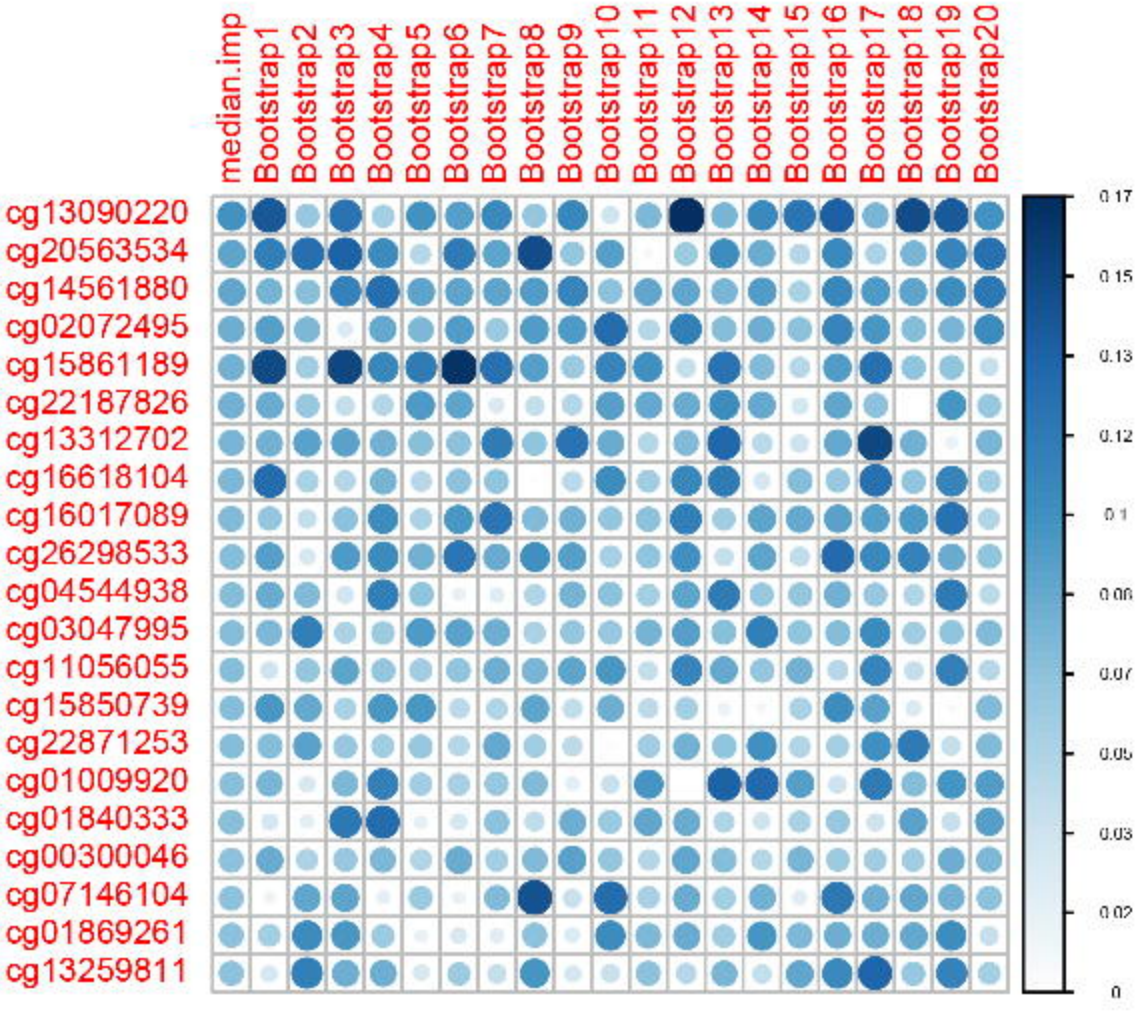
1. Kooij KW, Wit FW, Schouten J, van der Valk M, Godfried MH, Stolte IG, et al. HIV infection is independently associated with frailty in middle-aged HIV type 1-infected individuals compared with similar but uninfected controls. AIDS. 2016;30(2):241-50.
2. Dent E, Kowal P, Hoogendijk EO. Frailty measurement in research and clinical practice: A review. Eur J Intern Med. 2016;31:3-10.
3. Lam K, Pan K, Linnekamp JF, Medema JP, Kandimalla R. DNA methylation based biomarkers in colorectal cancer: a systematic review. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer. 2016;1866(1):106-20.
4. Teroganova N, Girshkin L, Suter CM, Green MJ. DNA methylation in peripheral tissue of schizophrenia and bipolar disorder: a systematic review. BMC genetics. 2016;17(1):27.
5. Bakusic J, Schaufeli W, Claes S, Godderis L. Stress, burnout and depression: A systematic review on DNA methylation mechanisms. Journal of Psychosomatic Research. 2017;92:34-44.
6. Li M, D'Arcy C, Li X, Zhang T, Joober R, Meng X. What do DNA methylation studies tell us about depression? A systematic review. Translational psychiatry. 2019;9(1):68.
7. Horvath S. DNA methylation age of human tissues and cell types. Genome biology. 2013;14(10):3156.
8. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nature Reviews Genetics. 2018:1.
9. Hannum G, Guinney J, Zhao L, Zhang L, Hughes G, Sadda S, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular cell. 2013;49(2):359-67.
10. Liu C, Marioni RE, Hedman ÅK, Pfeiffer L, Tsai P-C, Reynolds LM, et al. A DNA methylation biomarker of alcohol consumption. Molecular psychiatry. 2018;23(2):422.
11. Breitling LP, Yang R, Korn B, Burwinkel B, Brenner H. Tobacco-smoking-related differential DNA methylation: 27K discovery and replication. The American Journal of Human Genetics. 2011;88(4):450-7.
12. Gao X, Zhang Y, Breitling LP, Brenner H. Relationship of tobacco smoking and smoking-related DNA methylation with epigenetic age acceleration. Oncotarget. 2016;7(30):46878-89.
13. Joubert BR, Håberg SE, Nilsen RM, Wang X, Vollset SE, Murphy SK, et al. 450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. Environmental health perspectives. 2012;120(10):1425-31.
14. Lee KW, Pausova Z. Cigarette smoking and DNA methylation. Frontiers in genetics. 2013;4:132.
15. Zhang X, Hu Y, Aouizerat BE, Peng G, Marconi VC, Corley MJ, et al. Machine learning selected smoking-associated DNA methylation signatures that predict HIV prognosis and mortality. 2018;10(1):155.
16. Zhang R, Miao Q, Wang C, Zhao R, Li W, Haile CN, et al. Genome-wide DNA methylation analysis in alcohol dependence. Addiction biology. 2013;18(2):392-403.
17. Kraiczy J, Nayak KM, Howell KJ, Ross A, Forbester J, Salvestrini C, et al. DNA methylation defines regional identity of human intestinal epithelial organoids and undergoes dynamic changes during development. Gut. 2019;68(1):49-61.

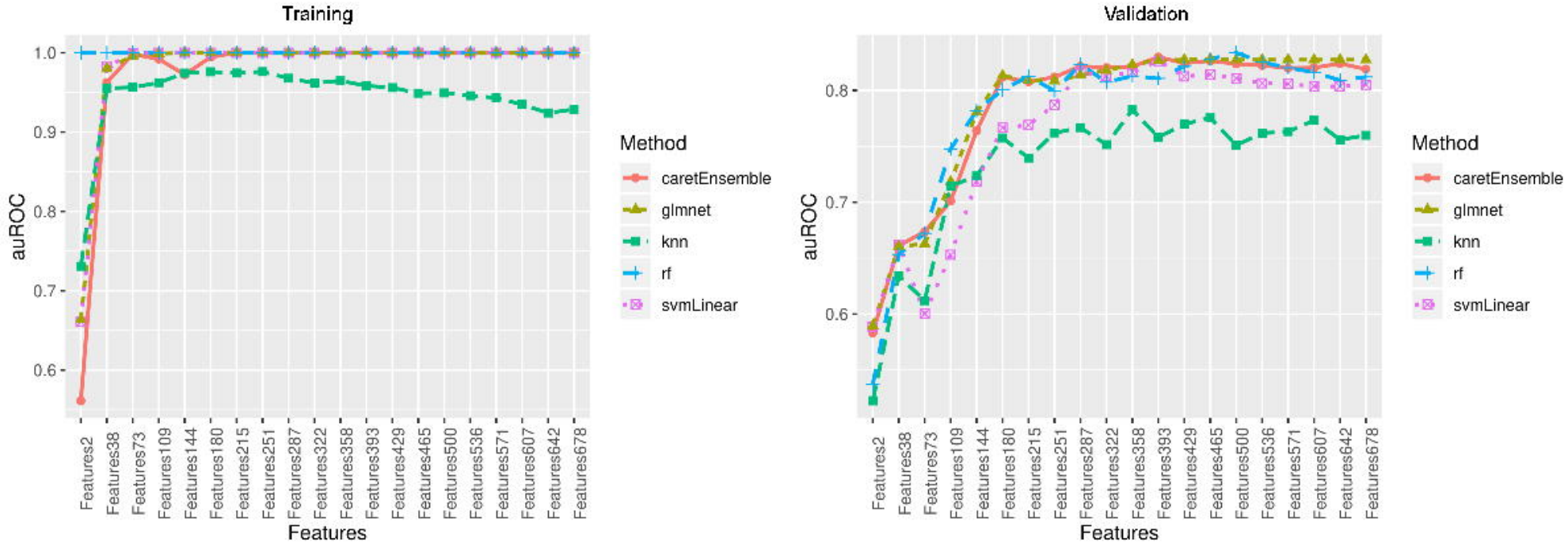
18. Nano J, Ghanbari M, Wang W, de Vries PS, Dhana K, Muka T, et al. Epigenome-Wide Association Study Identifies Methylation Sites Associated With Liver Enzymes and Hepatic Steatosis. Gastroenterology. 2017;153(4):1096-106.e2.
19. Delpu Y, Cordelier P, Cho W, Torrisani J. DNA methylation and cancer diagnosis. International journal of molecular sciences. 2013;14(7):15029-58.
20. Figueroa ME, Lugthart S, Li Y, Erpelinck-Verschueren C, Deng X, Christos PJ, et al. DNA methylation signatures identify biologically distinct subtypes in acute myeloid leukemia. Cancer cell. 2010;17(1):13-27.
21. Holm K, Hegardt C, Staaf J, Vallon-Christersson J, Jönsson G, Olsson H, et al. Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns. Breast cancer research. 2010;12(3):R36.
22. Berdasco M, Esteller M. Clinical epigenetics: seizing opportunities for translation. Nature Reviews Genetics. 2019;20(2):109-27.
23. Mohammad HP, Barbash O, Creasy CL. Targeting epigenetic modifications in cancer therapy: erasing the roadmap to cancer. Nat Med. 2019;25(3):403-18.
24. Erlandson KM, Ng DK, Jacobson LP, Margolick JB, Dobs AS, Palella Jr FJ, et al. Inflammation, immune activation, immunosenescence, and hormonal biomarkers in the frailty-related phenotype of men with or at risk for HIV infection. The Journal of infectious diseases. 2016;215(2):228-37.
25. Zhang X, Justice AC, Hu Y, Wang Z, Zhao H, Wang G, et al. Epigenome-wide differential DNA methylation between HIV-infected and uninfected individuals. Epigenetics. 2016;11(10):750-60.
26. Chen J, Huang Y, Hui Q, Mathur R, Gwinn M, So-Armah K, et al. Epigenetic Associations with Estimated Glomerular Filtration Rate (eGFR) among Men with HIV Infection. Clin Infect Dis. 2019.
27. Mathur R, Hui Q, Huang Y, Gwinn M, So-Armah K, Freiberg MS, et al. DNA Methylation Markers of Type 2 Diabetes Mellitus Among Male Veterans With or Without Human Immunodeficiency Virus Infection. J Infect Dis. 2019;219(12):1959-62.
28. Breitling LP, Saum K-U, Perna L, Schöttker B, Holleczek B, Brenner H. Frailty is associated with the epigenetic clock but not with telomere length in a German cohort. Clinical Epigenetics. 2016;8(1):21.
29. Sanchez-Conde M, Rodriguez-Centeno J, Dronda F, Lopez JC, Jimenez Z, Berenguer J, et al. Frailty phenotype: a clinical marker of age acceleration in the older HIV-infected population. Epigenomics. 2019;11(5):501-9.
30. Horvath S, Levine AJ. HIV-1 infection accelerates age according to the epigenetic clock. The Journal of infectious diseases. 2015;212(10):1563-73.
31. Rickabaugh TM, Baxter RM, Sehl M, Sinsheimer JS, Hultin PM, Hultin LE, et al. Acceleration of age-associated methylation patterns in HIV-1-infected adults. PloS one. 2015;10(3):e0119201.
32. Gross AM, Jaeger PA, Kreisberg JF, Licon K, Jepsen KL, Khosroheidari M, et al. Methylome-wide analysis of chronic HIV infection reveals five-year increase in biological age and epigenetic targeting of HLA. Molecular cell. 2016;62(2):157-68.

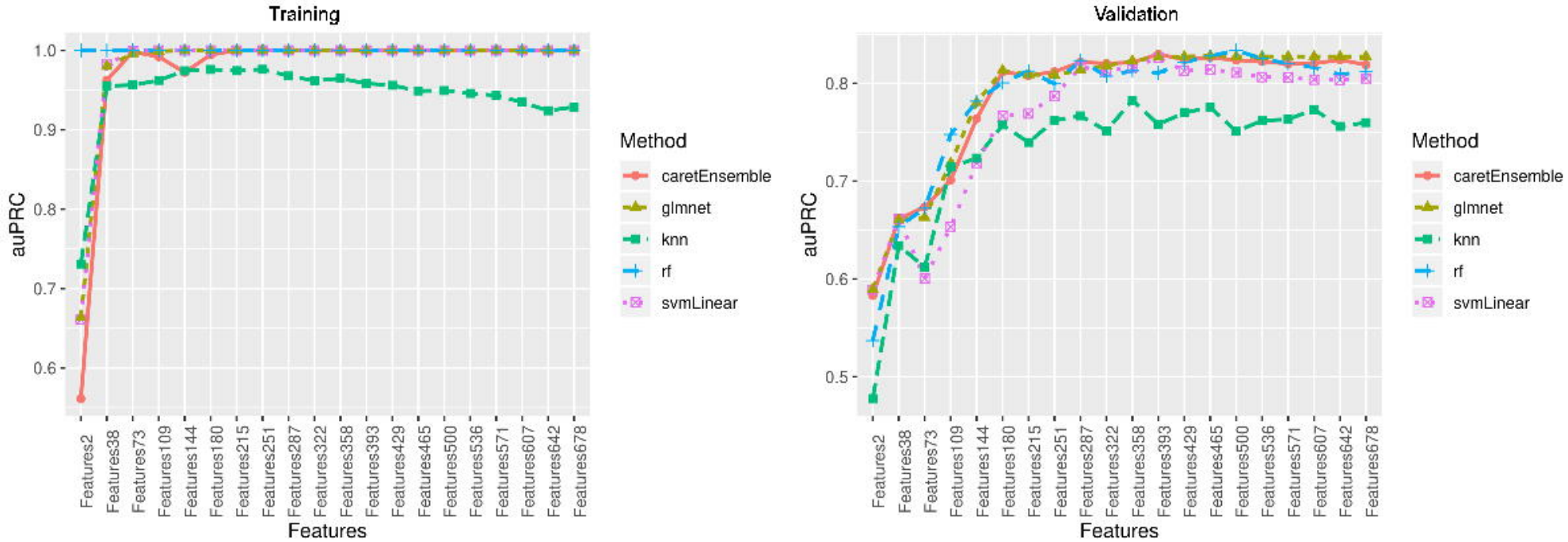
33. Nelson KN, Hui Q, Rimland D, Xu K, Freiberg MS, Justice AC, et al. Identification of HIV infection-related DNA methylation sites and advanced epigenetic aging in HIV-positive, treatment-naive U.S. veterans. AIDS. 2017;31(4):571-5.
34. Marioni RE, Shah S, McRae AF, Chen BH, Colicino E, Harris SE, et al. DNA methylation age of blood predicts all-cause mortality in later life. Genome biology. 2015;16(1):25.
35. Marioni RE, Harris SE, Shah S, McRae AF, von Zglinicki T, Martin-Ruiz C, et al. The epigenetic clock and telomere length are independently associated with chronological age and mortality. International journal of epidemiology. 2016;45(2):424-32.
36. Perna L, Zhang Y, Mons U, Holleczek B, Saum K-U, Brenner H. Epigenetic age acceleration predicts cancer, cardiovascular, and all-cause mortality in a German case cohort. Clinical epigenetics. 2016;8(1):64.
37. Christiansen L, Lenart A, Tan Q, Vaupel JW, Aviv A, McGue M, et al. DNA methylation age is associated with mortality in a longitudinal Danish twin study. Aging Cell. 2016;15(1):149-54.
38. Fransquet PD, Wrigglesworth J, Woods RL, Ernst ME, Ryan J. The epigenetic clock as a predictor of disease and mortality risk: a systematic review and meta-analysis. Clin Epigenetics. 2019;11(1):62.
39. Zhang Y, Wilson R, Heiss J, Breitling LP, Saum K-U, Schöttker B, et al. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. Nature communications. 2017;8:14617.
40. Holder LB, Haque MM, Skinner MK. Machine learning for epigenetics and future medical applications. Epigenetics. 2017;12(7):505-14.
41. Crowgey EL, Marsh AG, Robinson KG, Yeager SK, Akins RE. Epigenetic machine learning: utilizing DNA methylation patterns to predict spastic cerebral palsy. BMC Bioinformatics. 2018;19(1):225.
42. Dietterich TG, editor Ensemble methods in machine learning. International workshop on multiple classifier systems; 2000: Springer.
43. Dietterich TG. Ensemble learning. The handbook of brain theory and neural networks. 2002;2:110-25.
44. Moon H, Ahn H, Kodell RL, Baek S, Lin C-J, Chen JJ. Ensemble methods for classification of patients for personalized medicine with high-dimensional data. Artificial Intelligence in Medicine. 2007;41(3):197-207.
45. Alghamdi M, Al-Mallah M, Keteyian S, Brawner C, Ehrman J, Sakr S. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. PloS one. 2017;12(7).
46. Justice AC, Dombrowski E, Conigliaro J, Fultz SL, Gibson D, Madenwald T, et al. Veterans aging cohort study (VACS): overview and description. 2006;44(8 Suppl 2):S13.
47. Justice AC, Modur S, Tate JP, Althoff KN, Jacobson LP, Gebo K, et al. Predictive accuracy of the Veterans Aging Cohort Study (VACS) index for mortality with HIV infection: a north American cross cohort analysis. Journal of acquired immune deficiency syndromes (1999). 2013;62(2):149.
48. Tate JP, Justice AC, Hughes MD, Bonnet F, Reiss P, Mocroft A, et al. An internationally generalizable risk index for mortality after one year of antiretroviral therapy. AIDS (London, England). 2013;27(4):563-72.

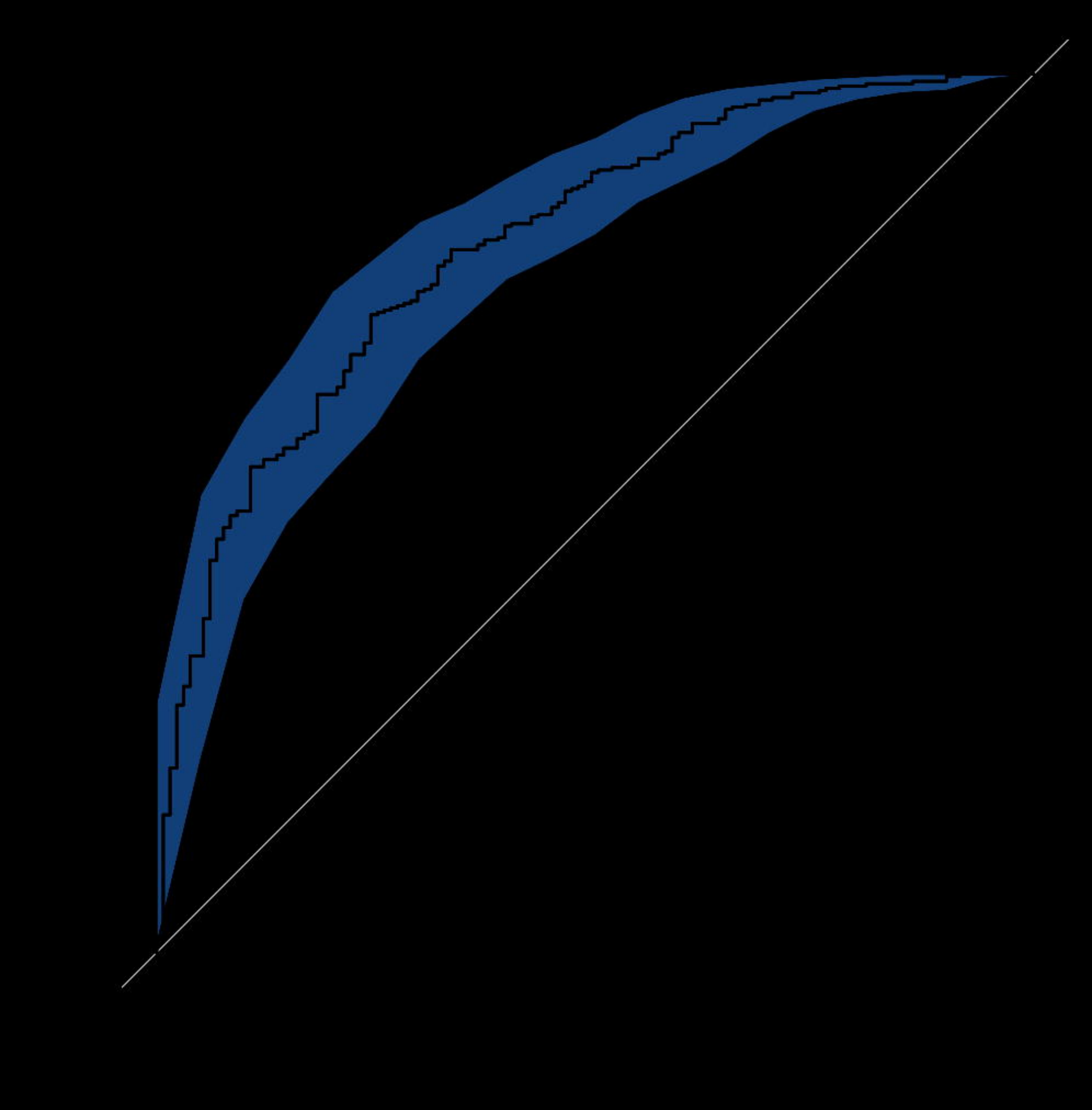
49. Justice AC, Modur SP, Tate JP, Althoff KN, Jacobson LP, Gebo KA, et al. Predictive accuracy of the Veterans Aging Cohort Study index for mortality with HIV infection: a North American cross cohort analysis. Journal of acquired immune deficiency syndromes (1999). 2013;62(2):149-63.
50. Lehne B, Drong AW, Loh M, Zhang W, Scott WR, Tan S-T, et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. 2015;16(1):37.
51. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. 2014;30(10):1363-9.
52. 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):86.
53. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. The Mathematical Intelligencer. 2005;27(2):83-5.
54. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering. 2007;160:3-24.
55. Chen T, Guestrin C, editors. Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016: ACM.
56. Ogutu JO, Piepho H-P, Schulz-Streeck T. A comparison of random forests, boosting and support vector machines for genomic selection. BMC Proc. 2011;5 Suppl 3(Suppl 3):S11-S.
57. Kuhn MJJoss. Building predictive models in R using the caret package. 2008;28(5):1-26.
58. Brodersen KH, Ong CS, Stephan KE, Buhmann JM, editors. The balanced accuracy and its posterior distribution. 2010 20th International Conference on Pattern Recognition; 2010: IEEE.
59. Phipson B, Maksimovic J, Oshlack A. missMethyl: an R package for analyzing data from Illumina's HumanMethylation450 platform. Bioinformatics. 2015;32(2):286-8.
60. Geeleher P, Hartnett L, Egan LJ, Golden A, Raja Ali RA, Seoighe C. Gene-set analysis is severely biased when applied to genome-wide methylation data. Bioinformatics. 2013;29(15):1851-7.
61. Willer CJ, Li Y, Abecasis GRJB. METAL: fast and efficient meta-analysis of genomewide association scans. 2010;26(17):2190-1.

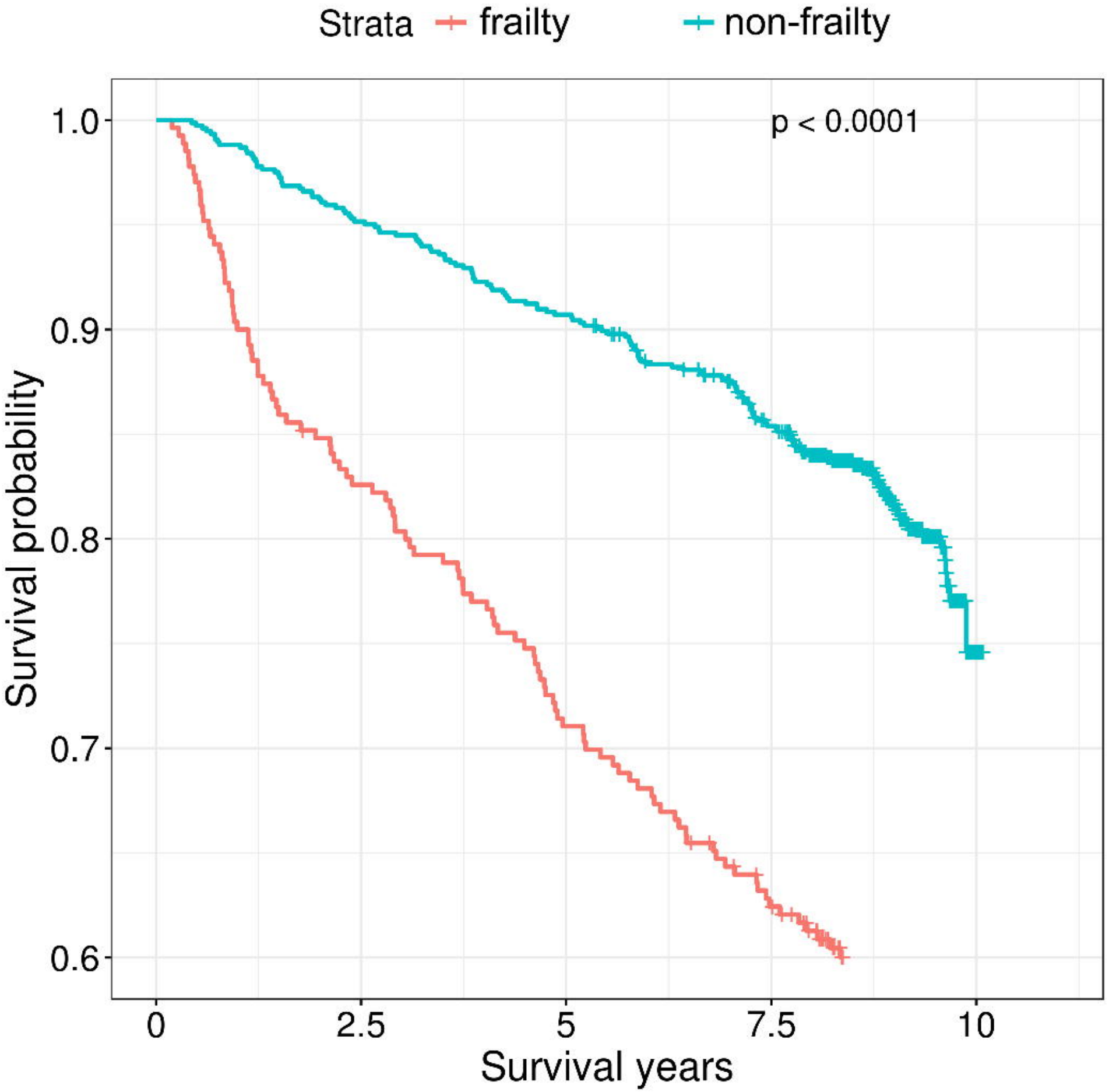

Ensemble Model: predict frailty
Group 1 (HM450)
Training set (N=460)
Feature preselection
Variable Importance; discard non-informative features
Ensemble model training and fine-tuning (4-fold cross-validation)
Validation set (N=114)
Select best features
Group 2 (EPIC)
Testing set (N=507)
Independent evaluation
Survival analysis by predicted high and normal frailty
Gene ontology enrichment analysis
Comparison with EWAS meta-analysis

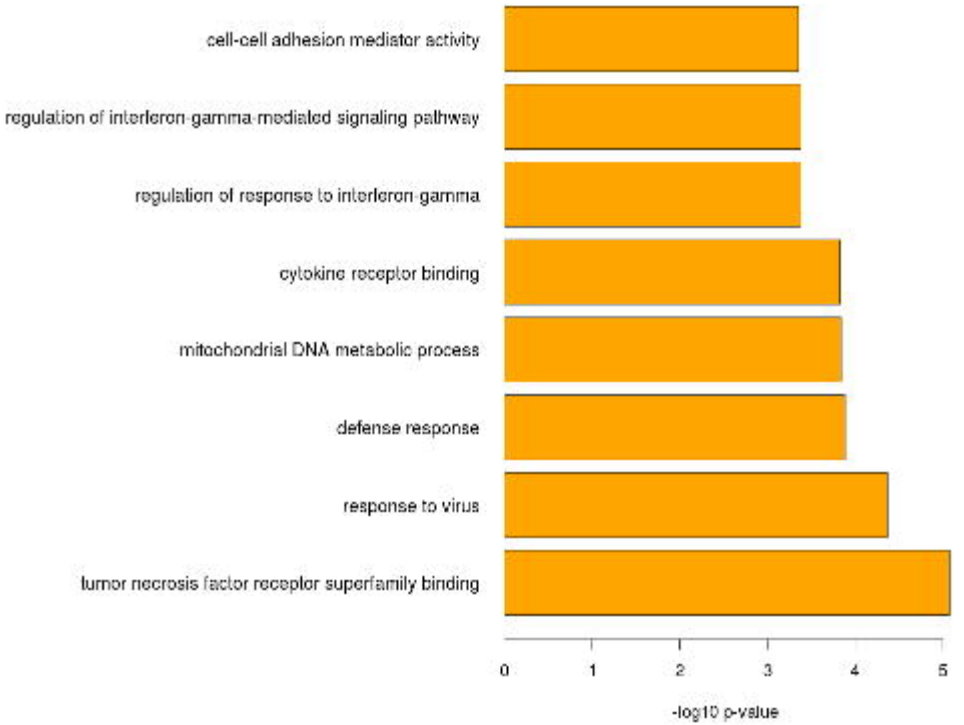

median.imp
Bootstrap1
Bootstrap2
Bootstrap3
Bootstrap4
Bootstrap5
Bootstrap6
Bootstrap7
Bootstrap8
Bootstrap9
Bootstrap10
Bootstrap11
Bootstrap12
Bootstrap13
Bootstrap14
Bootstrap15
Bootstrap16
Bootstrap17
Bootstrap18
Bootstrap19
Bootstrap20
cg13090220
cg20563534
cg14561880
cg02072495
cg15861189
cg22187826
cg13312702
cg16618104
cg16017089
cg26298533
cg04544938
cg03047995
cg11056055
cg15850739
cg22871253
cg01009920
cg01840333
cg00300046
cg07146104
cg01869261
cg13259811
0.17
0.15
0.13
0.12
0.1
0.08
0.07
0.05
0.03
0.02
0


Training
Validation
auROC
1.0
0.9
0.8
0.7
0.6
Features
Features2
Features38
Features73
Features109
Features144
Features180
Features215
Features251
Features287
Features322
Features358
Features393
Features429
Features465
Features500
Features536
Features571
Features607
Features642
Features678
Method
caretEnsemble
glmnet
knn
rf
svmLinear


Training
Validation
auPRC
Features
Method
caretEnsemble
glmnet
knn
rf
svmLinear
1.0
0.9
0.8
0.7
0.6
0.5
Features2
Features38
Features73
Features109
Features144
Features180
Features215
Features251
Features287
Features322
Features358
Features393
Features429
Features465
Features500
Features536
Features571
Features607
Features642
Features678




Strata
frailty
non-frailty
p < 0.0001
1.0
0.9
0.8
0.7
0.6
Survival probability
0
2.5
5
7.5
10
Survival years


cell-cell adhesion mediator activity
regulation of interferon-gamma-mediated signaling pathway
regulation of response to interferon-gamma
cytokine receptor binding
mitochondrial DNA metabolic process
defense response
response to virus
tumor necrosis factor receptor superfamily binding
0
1
2
3
4
5
-log10 p-value

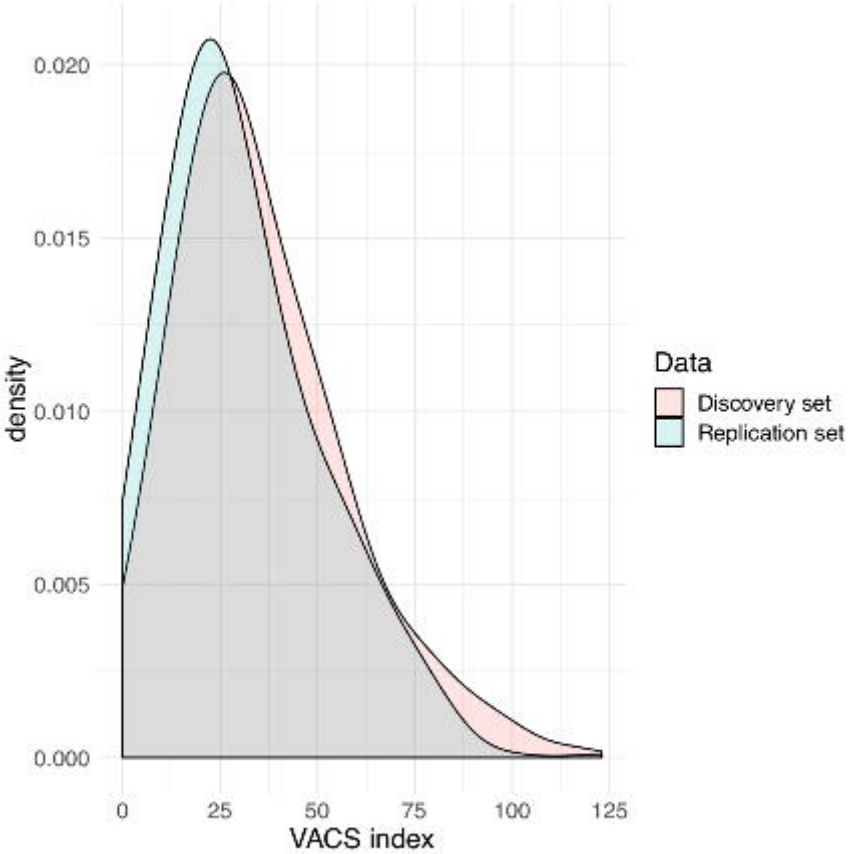

0.020
0.015
0.010
0.005
0.000
density
0
25
50
75
100
125
VACS index
Data
Discovery set
Replication set

Figure S1