Supplementary Data

Explainable Machine Learning Framework to Predict Personalized Physiological Aging

David Bernard, Emmanuel Doumard, Isabelle Ader, Philippe Kemoun, Jean-Christophe Pages, Anne Galinier, Sylvain Cussat-Blanc, Felix Furger, Luigi Ferrucci, Julien Aligon, Cyrille Delpierre, Luc Pénicaud, Paul Monsarrat, Louis Casteilla

List of supplementary data

Supplementary Text

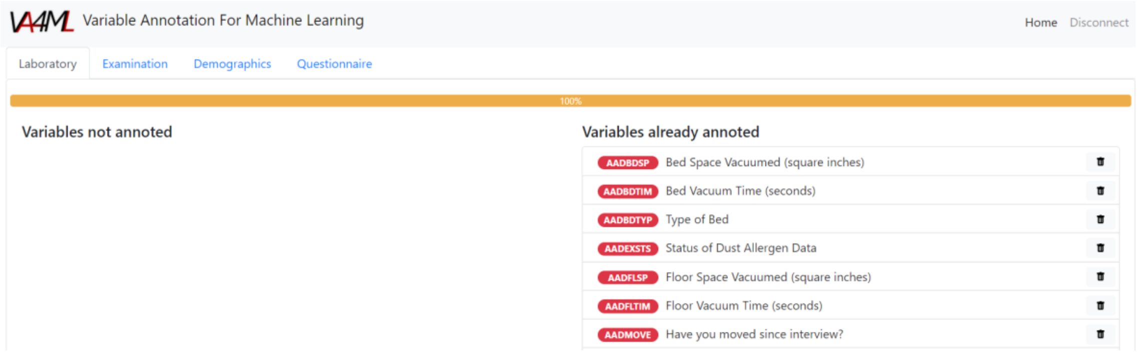
Variable selection and merging

To generate a consistent and large database, with a maximal number of common biological variables for subjects, we performed a manual data cleaning to eliminate redundant outcomes, both within the same year and in different years. This step requires human intervention. For instance, some variables hold different labels and/or codes across years (*e.g.*, LBDHDLSI, "HDL-cholesterol (mmol/L)" from 1999 to 2002 and LBDHDDSI, "Direct HDL-Cholesterol (mmol/L)" from 2003 to 2018), or have different units (*e.g.*, serum glucose in mmol/L LBDSGLSI and in mg/dL LBXSGL). To optimize cleaning by investigators (IA, LC, LP, PK and PM) and ensure reproducibility, a web interface was developed (Fig. S1).

For each laboratory variable available, investigators were independently asked to select the variables based on the inclusion criteria described above. A similarity algorithm (using cosine similarity and Levenshtein distance) based on the "SAS" and "text" labels, proposed a list of potentially synonymous terms to investigators. A manual search tool with autocompletion was also available.

When identical variables were measured several times, the mean value was considered (for example*:* both variables LB2NEPCT and LBXNEPCT corresponded to the same biological variable, *i.e.* segmented neutrophils percent). Biological variables expressed in international units (SI) were privileged over their non-SI counterparts. In case of disagreement by an investigator, a collegial decision was made at the consensus phase. After this step, and considering the distribution of the number of available variables for a given number of subjects, the largest dataset with the minimum amount of missing data was defined. The cut-off for this distribution selected variables with at least 50,000 individuals. Individuals with more than 10% missing values were also dropped from database. After processing, the selected dataset contained 60,322 individuals with 48 laboratory variables (Table S1) and limited missing data (0.6% of data, Fig. S2).

A



VAML Variable Annotation For Machine Learning

Home Disconnect

Laboratory | Examination | Demographics | Questionnaire

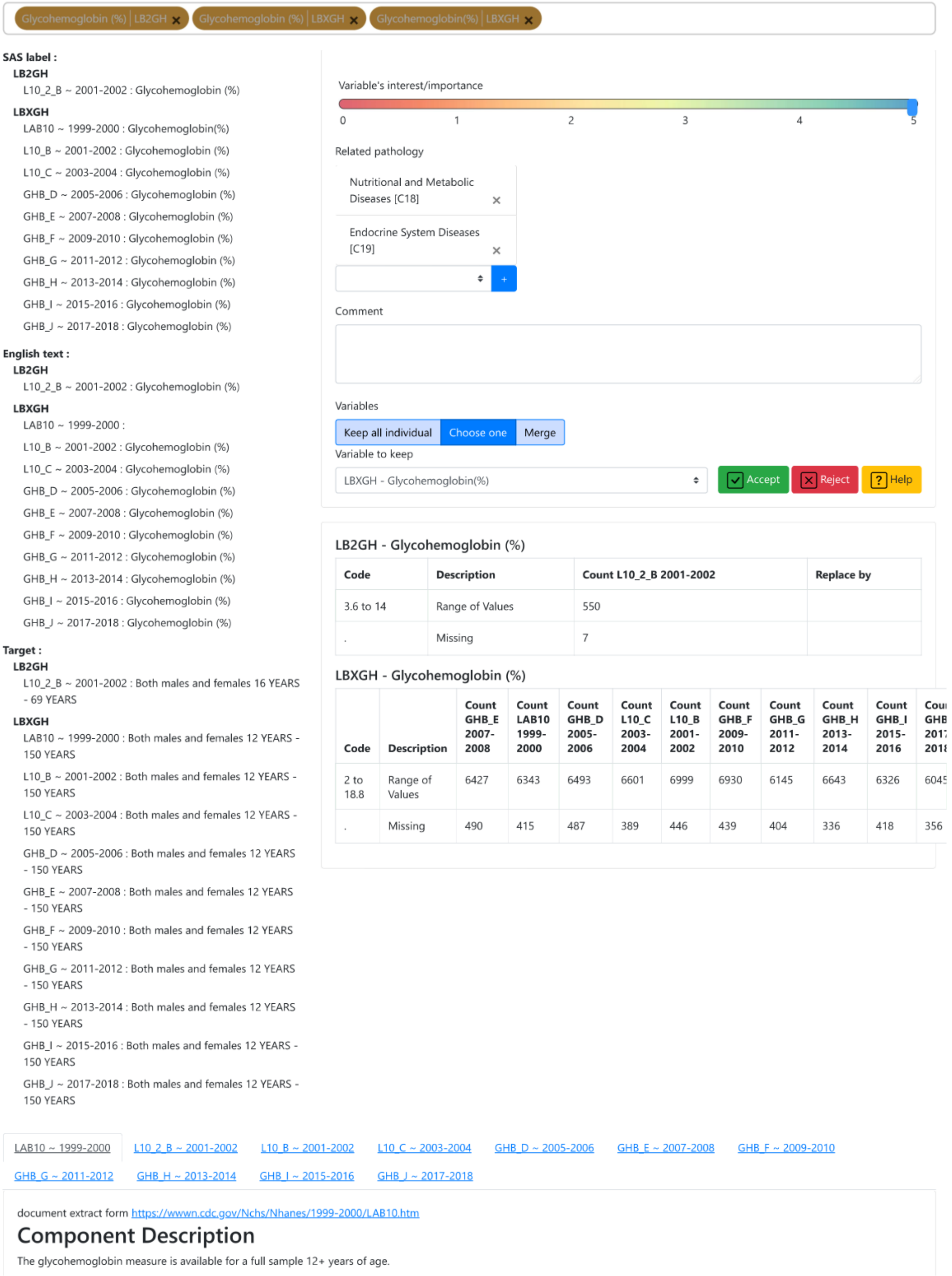
100%

Variables not annoted

Variables already annoted

AADBDSP	Bed Space Vacuumed (square inches)	🗑
AADBDTIM	Bed Vacuum Time (seconds)	🗑
AADBDTYP	Type of Bed	🗑
AADEXSTS	Status of Dust Allergen Data	🗑
AADFLSP	Floor Space Vacuumed (square inches)	🗑
AADFLTIM	Floor Vacuum Time (seconds)	🗑
AADMOVE	Have you moved since interview?	🗑

B



Glycohemoglobin (%) | LB2GH × Glycohemoglobin (%) | LBXGH × Glycohemoglobin(%) | LBXGH ×

SAS label :

LB2GH

L10_2_B ~ 2001-2002 : Glycohemoglobin (%)

LBXGH

LAB10 ~ 1999-2000 : Glycohemoglobin(%)
L10_B ~ 2001-2002 : Glycohemoglobin (%)
L10_C ~ 2003-2004 : Glycohemoglobin (%)
GHB_D ~ 2005-2006 : Glycohemoglobin (%)
GHB_E ~ 2007-2008 : Glycohemoglobin (%)
GHB_F ~ 2009-2010 : Glycohemoglobin (%)
GHB_G ~ 2011-2012 : Glycohemoglobin (%)
GHB_H ~ 2013-2014 : Glycohemoglobin (%)
GHB_I ~ 2015-2016 : Glycohemoglobin (%)
GHB_J ~ 2017-2018 : Glycohemoglobin (%)

English text :

LB2GH

L10_2_B ~ 2001-2002 : Glycohemoglobin (%)

LBXGH

LAB10 ~ 1999-2000 :
L10_B ~ 2001-2002 : Glycohemoglobin (%)
L10_C ~ 2003-2004 : Glycohemoglobin (%)
GHB_D ~ 2005-2006 : Glycohemoglobin (%)
GHB_E ~ 2007-2008 : Glycohemoglobin (%)
GHB_F ~ 2009-2010 : Glycohemoglobin (%)
GHB_G ~ 2011-2012 : Glycohemoglobin (%)
GHB_H ~ 2013-2014 : Glycohemoglobin (%)
GHB_I ~ 2015-2016 : Glycohemoglobin (%)
GHB_J ~ 2017-2018 : Glycohemoglobin (%)

Target :

LB2GH

L10_2_B ~ 2001-2002 : Both males and females 16 YEARS - 69 YEARS

LBXGH

LAB10 ~ 1999-2000 : Both males and females 12 YEARS - 150 YEARS

L10_B ~ 2001-2002 : Both males and females 12 YEARS - 150 YEARS

L10_C ~ 2003-2004 : Both males and females 12 YEARS - 150 YEARS

GHB_D ~ 2005-2006 : Both males and females 12 YEARS - 150 YEARS

GHB_E ~ 2007-2008 : Both males and females 12 YEARS - 150 YEARS

GHB_F ~ 2009-2010 : Both males and females 12 YEARS - 150 YEARS

GHB_G ~ 2011-2012 : Both males and females 12 YEARS - 150 YEARS

GHB_H ~ 2013-2014 : Both males and females 12 YEARS - 150 YEARS

GHB_I ~ 2015-2016 : Both males and females 12 YEARS - 150 YEARS

GHB_J ~ 2017-2018 : Both males and females 12 YEARS - 150 YEARS

Variable's interest/importance

Related pathology

Nutritional and Metabolic Diseases [C18] ×

Endocrine System Diseases [C19] ×

+

Comment

Variables

Keep all individual | Choose one | Merge

Variable to keep

LBXGH - Glycohemoglobin(%)

☑ Accept ☒ Reject ? Help

LB2GH - Glycohemoglobin (%)

Code	Description	Count L10_2_B 2001-2002	Replace by
3.6 to 14	Range of Values	550	
.	Missing	7	

LBXGH - Glycohemoglobin (%)

Code	Description	Count GHB_E 2007-2008	Count LAB10 1999-2000	Count GHB_D 2005-2006	Count L10_C 2003-2004	Count L10_B 2001-2002	Count GHB_F 2009-2010	Count GHB_G 2011-2012	Count GHB_H 2013-2014	Count GHB_I 2015-2016	Cou GHB 201 201
2 to 18.8	Range of Values	6427	6343	6493	6601	6999	6930	6145	6643	6326	6045
.	Missing	490	415	487	389	446	439	404	336	418	356

LAB10 ~ 1999-2000 | L10_2_B ~ 2001-2002 | L10_B ~ 2001-2002 | L10_C ~ 2003-2004 | GHB_D ~ 2005-2006 | GHB_E ~ 2007-2008 | GHB_F ~ 2009-2010 | GHB_G ~ 2011-2012 | GHB_H ~ 2013-2014 | GHB_I ~ 2015-2016 | GHB_J ~ 2017-2018

document extract form https://wwwn.cdc.gov/Nchs/Nhanes/1999-2000/LAB10.htm

Component Description

The glycohemoglobin measure is available for a full sample 12+ years of age.

Fig. S1 : Screenshots of the dedicated interface that has been used to annotate the database.

Each user had his own account and the laboratory variables **(A)** were selected, grouped and/or annotated **(B)**.

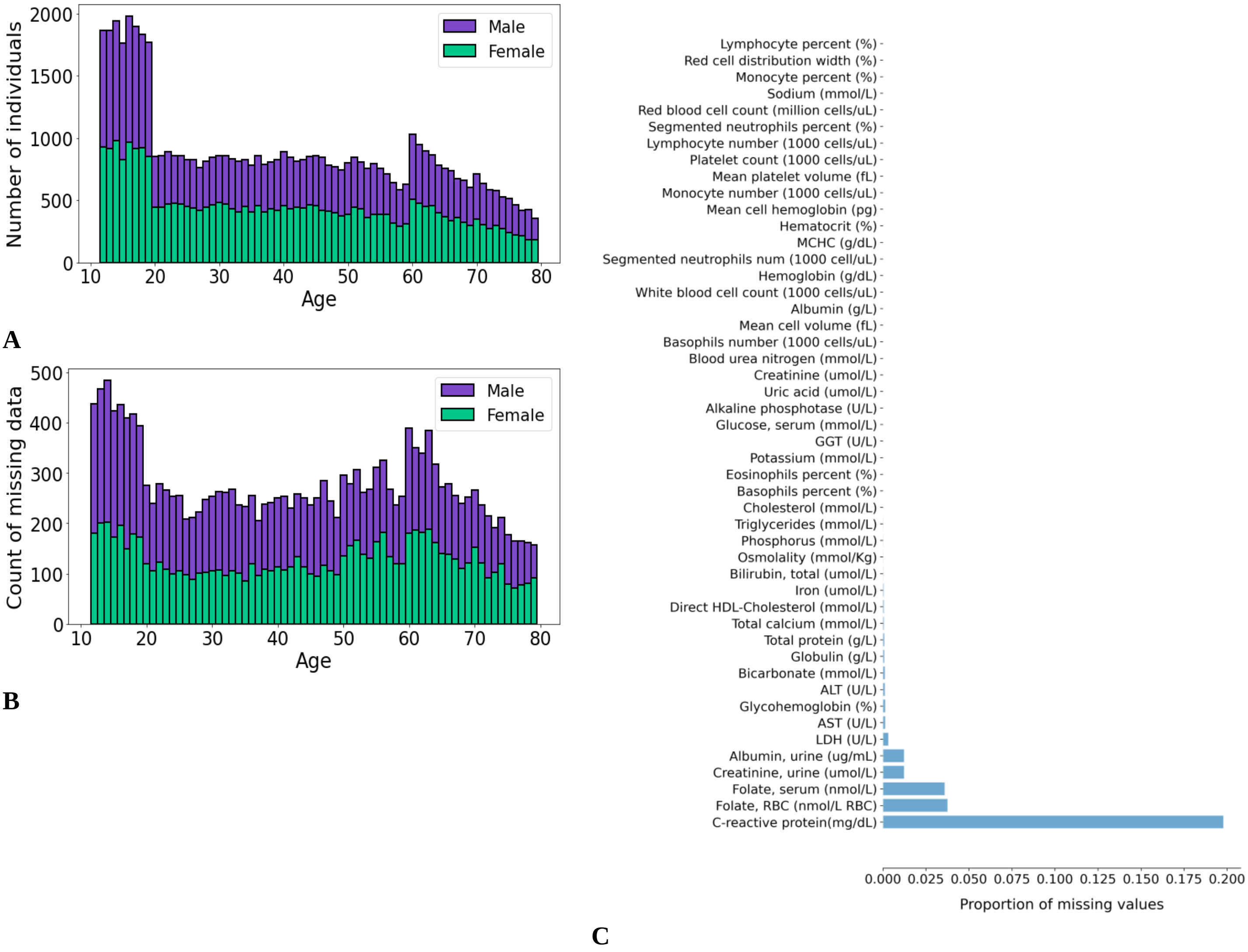
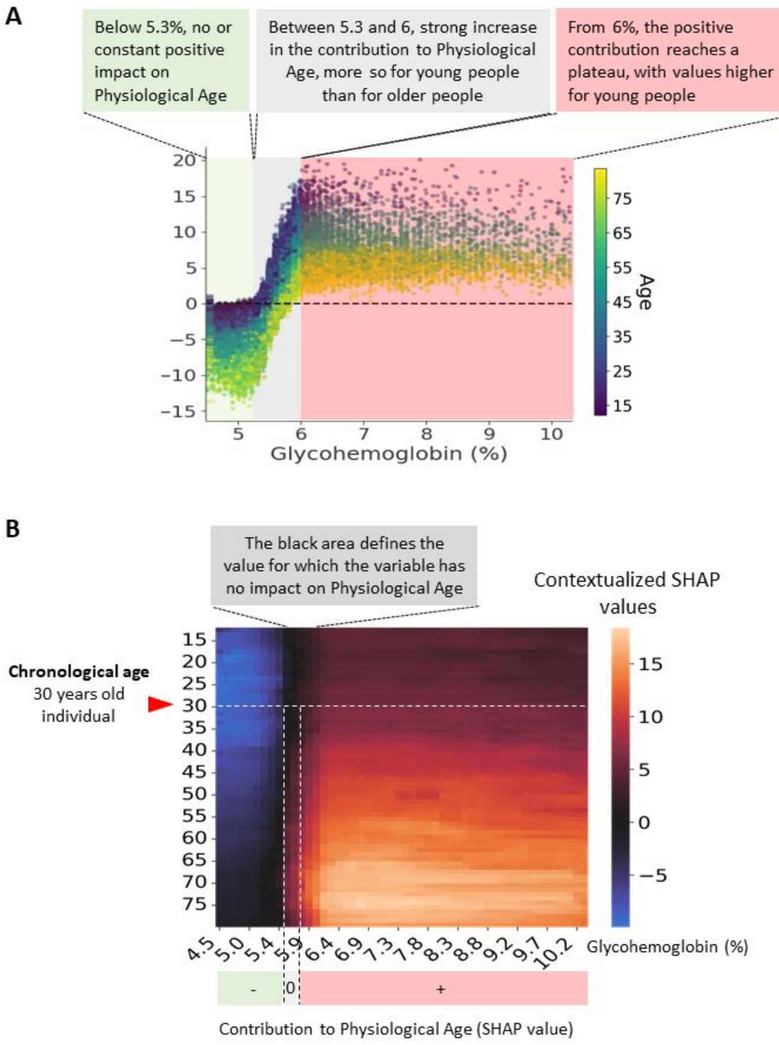


Fig. S2: Distribution of data within the final dataset.

(A) Distribution of the number of individuals by chronological age and gender. The amount of data from 12 to 20 years was twice those of other age and a 25% decrease of available subject number from 70 to 79 years old. No major gender imbalance was pointed out across age groups. **(B)** Uniform distribution of missing data among chronological age and gender. **(C)** Proportion of missing data by variable. The amount of missing data was low (25% of individuals with one missing value representing 0.6% of the total values). They were mainly related to the lack of C-reactive protein, folate, albumin, and creatinine data.



Below 5.3%, no or constant positive impact on Physiological Age

Between 5.3 and 6, strong increase in the contribution to Physiological Age, more so for young people than for older people

From 6%, the positive contribution reaches a plateau, with values higher for young people

The black area defines the value for which the variable has no impact on Physiological Age

Fig. S3: Interpretation of partial dependance plot (A) and heatmap of contextualized SHAP values (B).

(A) Each dot represents an individual. The color indicates the corresponding chronological age. X-axis corresponds to the real value of the variable, while the y-axis corresponds to the SHAP value given to this individual for this variable. While the contextualized SHAP values are negative in low values of glycohemoglobin, a sharp increase occurs between 5.3 and 6%. This transition zone, characterized by the passage from zero, is different according to age and clearly visualized in (B) as a dark zone in the heatmap. **(B)** Heatmap of contextualized SHAP values as a function of chronological age. The color of each pixel indicates the average SHAP value of a variable (x-axis) as a function of chronological age (y-axis). For an individual of 30 y.o., the normal range is about 5.5 – 5.8%.

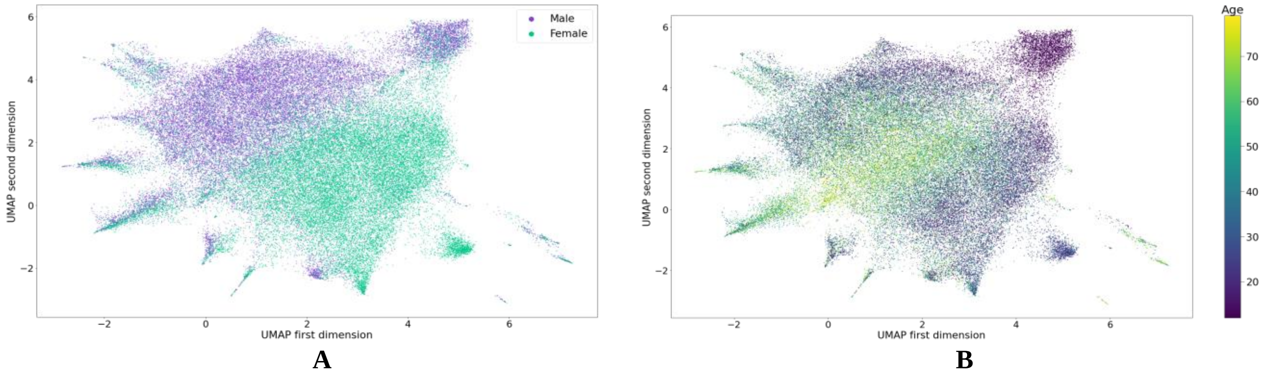


Fig. S4: Gender and chronological age distribution of individuals

UMAP 2-dimensional projection of the 48 variables of the dataset, colored by chronological age **(A)** and gender **(B)**. UMAP revealed some clustering across the second dimension by gender with mostly males in the upper part of the UMAP and females in the lower part. In addition, the first dimension mainly contains chronological age information, with a gradient from youngest to oldest from right to left.

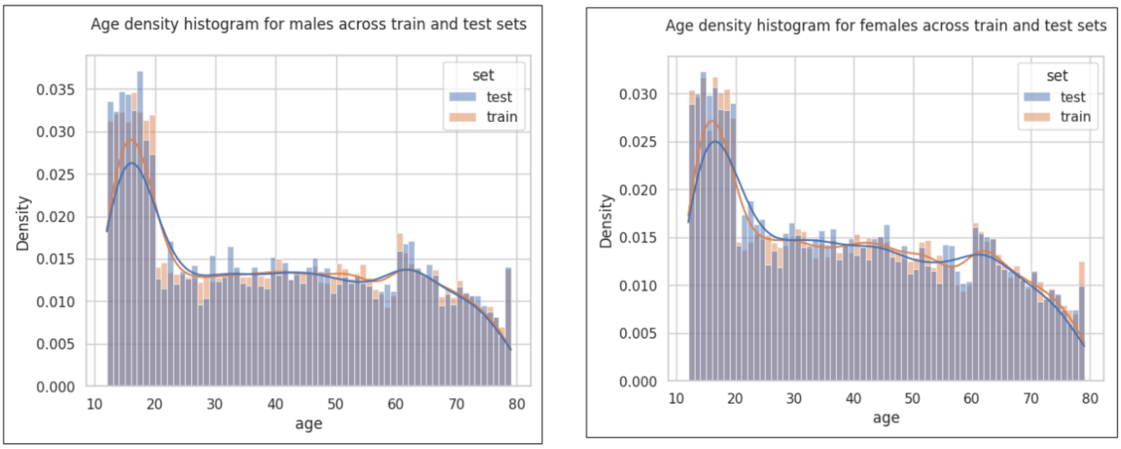


Fig. S5: No gender and age group imbalance between train and test datasets

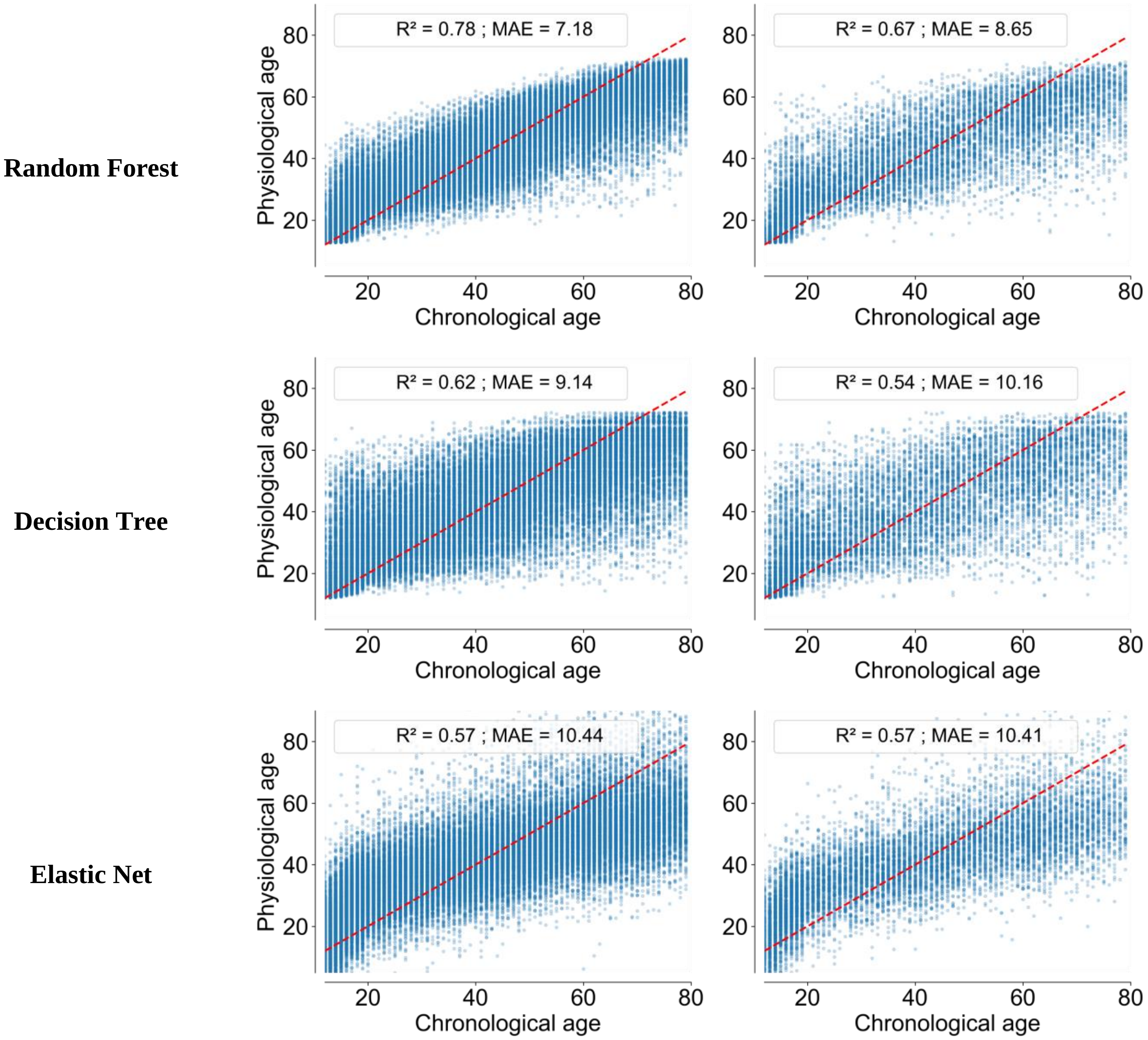


Fig. S6 : Distribution of residuals for the prediction of chronological age by Random Forest, Decision Tree and Elastic Net.

Left scatter plot illustrates the distribution of residuals on the train dataset and right scatter plot on the test dataset. The performances are largely inferior to those obtained with XGBoost or MultiLayer Perceptron with a strong performance discrepancy across the age group (younger people are predicted to be older and conversely).

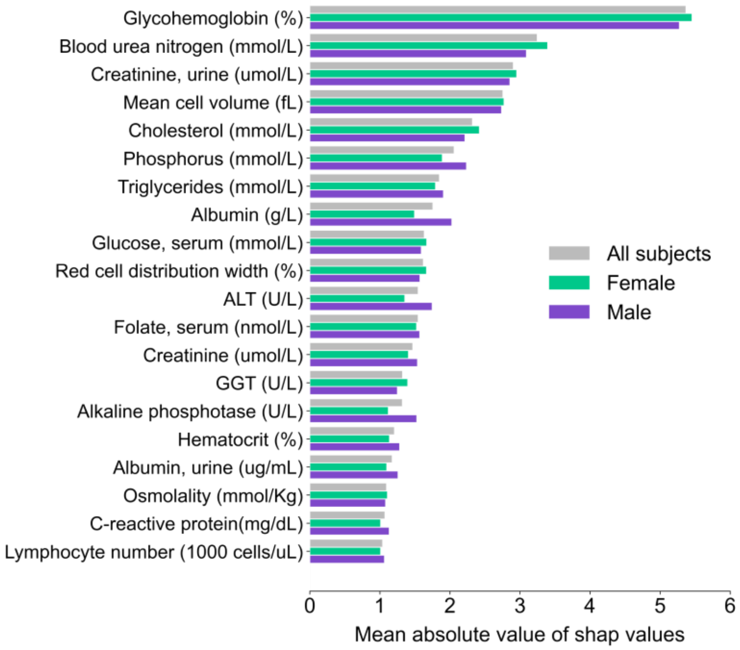


Fig. S7: Relative importance of the most important variables in physiological age prediction.

The mean absolute value of shap values for the 20 most important variables are shown for the whole population (gray), female (green) and male (purple) populations. A similar importance can been shown according gender.

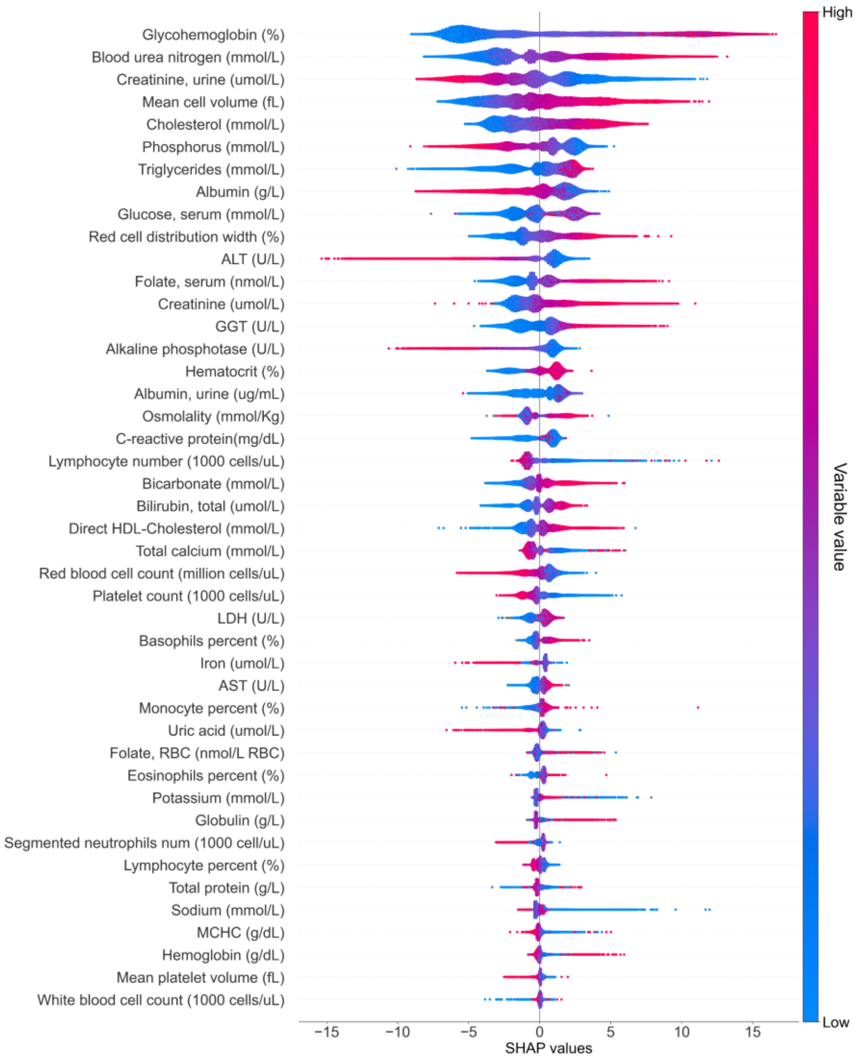


Fig. S8: Global explainability of the PPA model in importance order of mean of absolute SHAP values.

Each point color encodes the SHAP value of each variable for each individual, red and blue colors for high and low values of the variable respectively. On the x-axis, a positive or negative SHAP value means that the variable, for one individual contributes to the estimation of physiological age positively or negatively respectively.

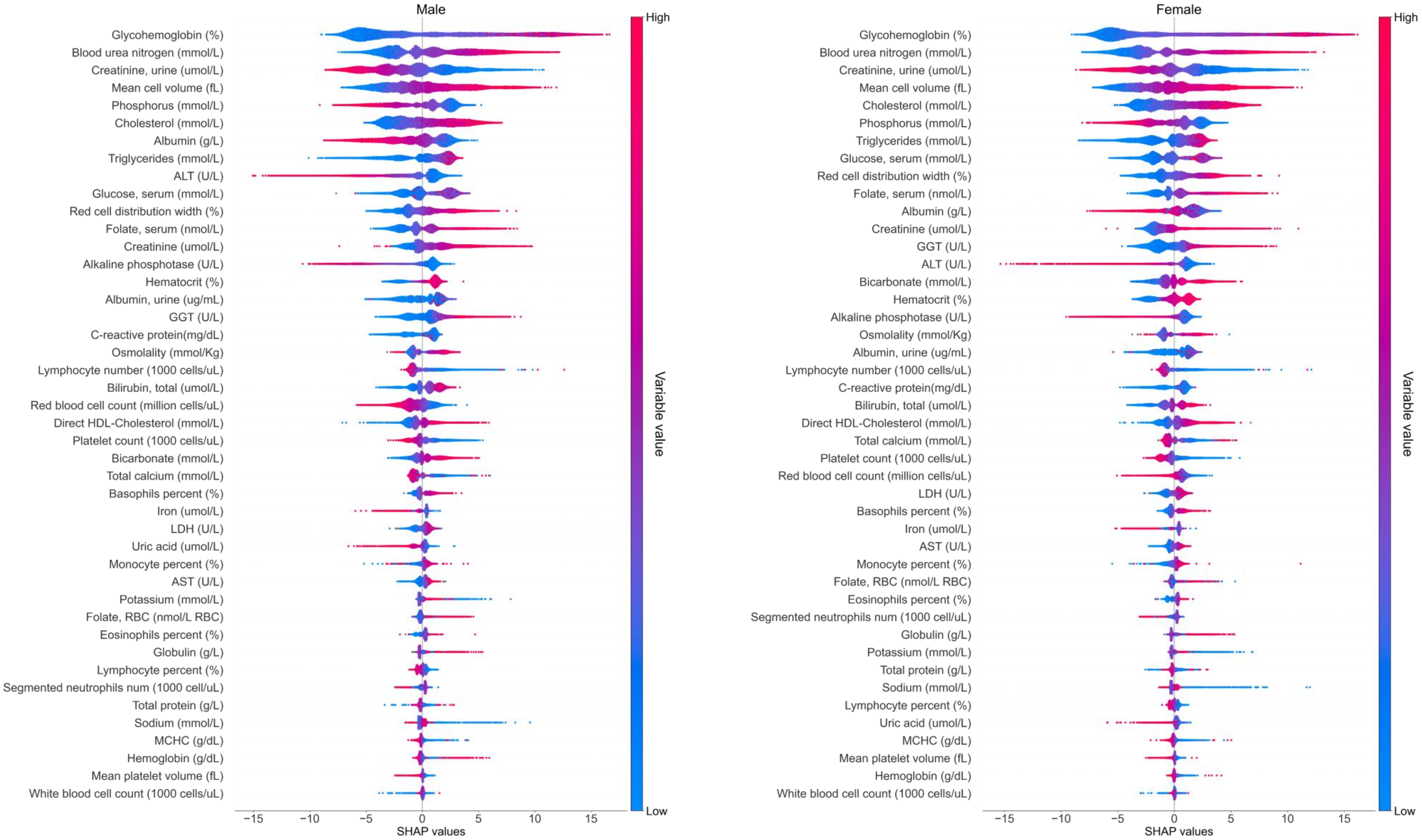


Fig. S9: Global explainability of the PPA model in importance order of mean of absolute SHAP values for male (A) and female individuals (B).

Each point color encodes the SHAP value of each variable for each individual, red and blue colors for high and low values of the variable respectively. On the x-axis, a positive or negative SHAP value means that the variable, for one individual contributes to the estimation of physiological age positively or negatively respectively. Similar explainability profile can be found between male and female.

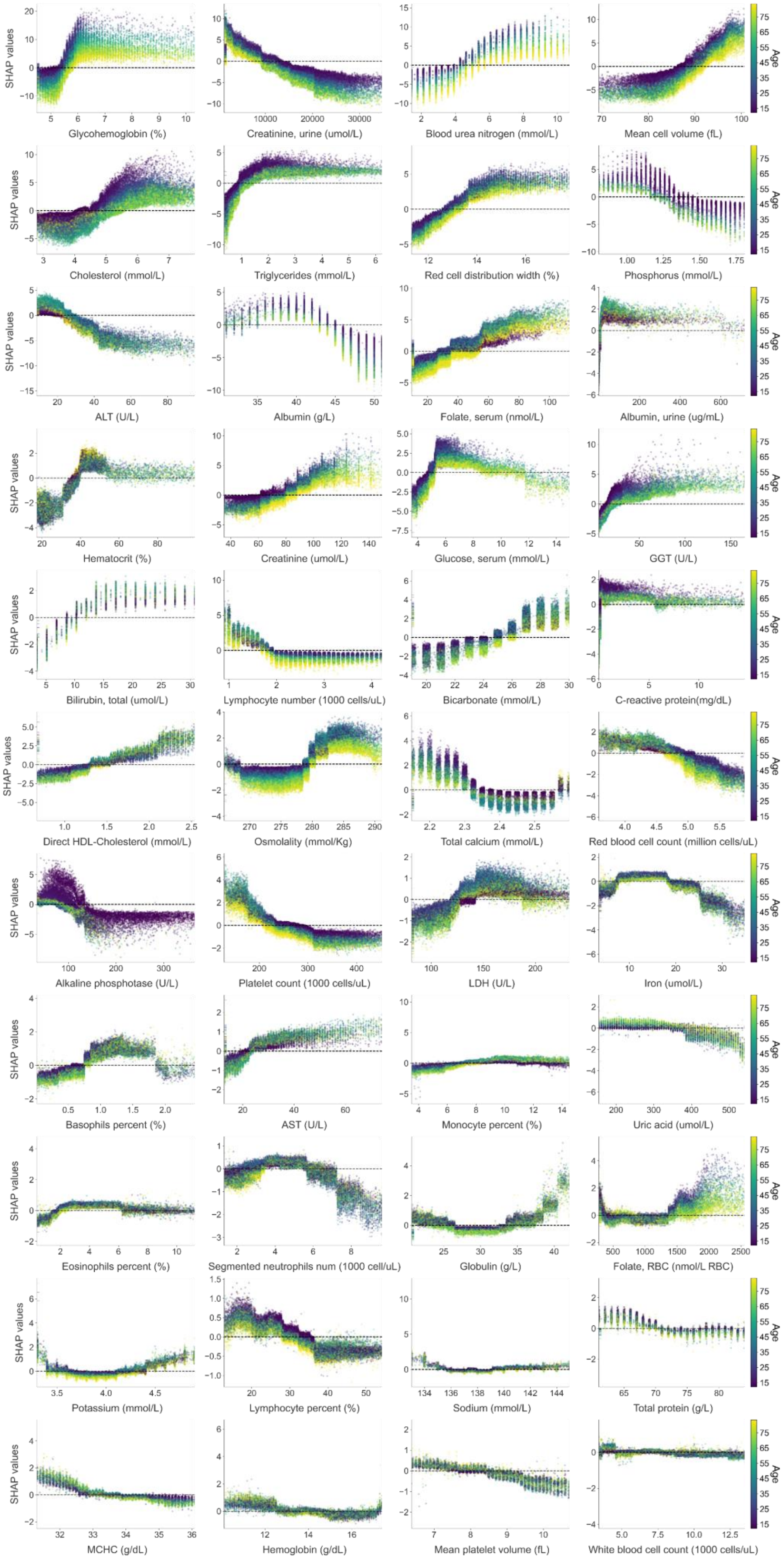
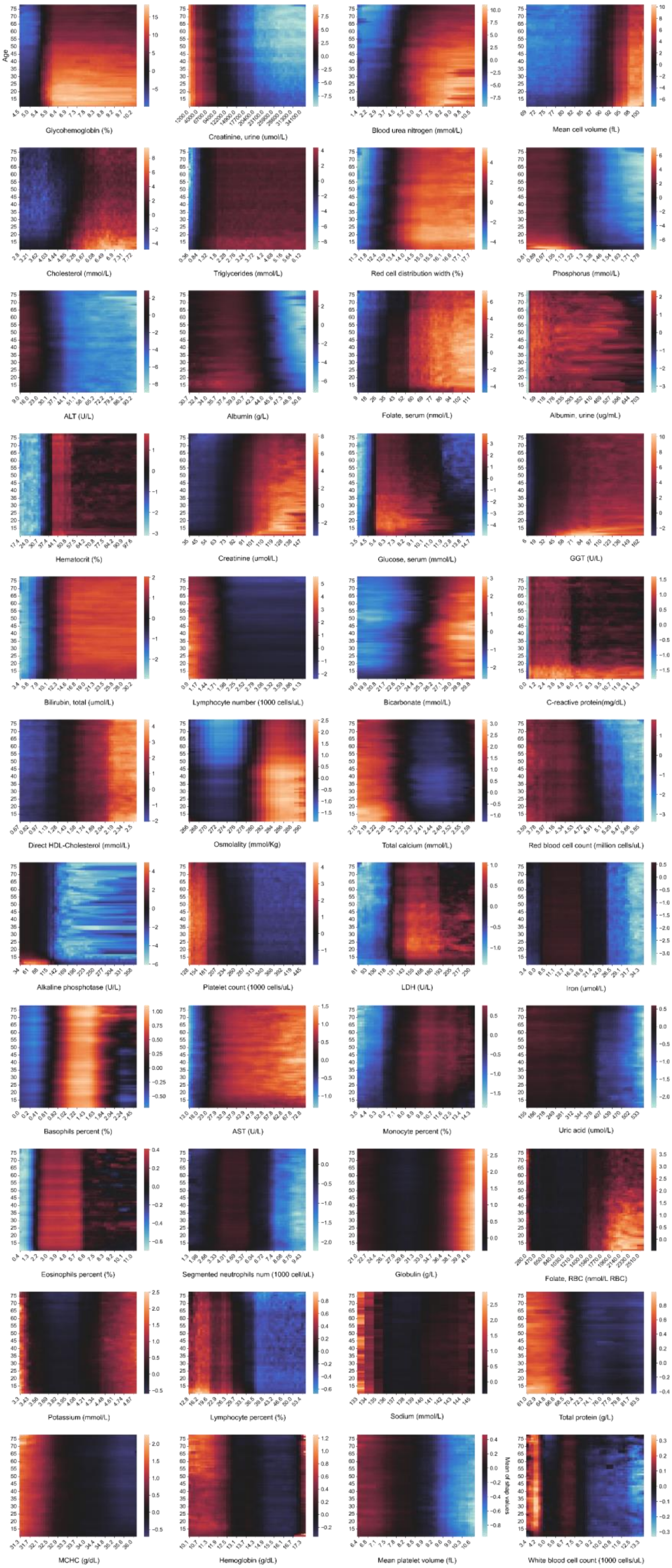


Fig. S10: Partial Dependence Plots of contextualized SHAP values.

(A) Contextualized SHAP values as a function of variable values. Each dot represents an individual. The color indicates the corresponding chronological age (scale on the right). X-axis corresponds to the real value of the variable, while the y-axis corresponds to the SHAP value given to this individual for this variable. The dotted line corresponds to the SHAP value of 0, which means that when the individual displays a variable value for which the SHAP value is 0, the variable has no impact on the physiological age.



(B) Heatmap of contextualized SHAP values as a function of chronological age. The color of each pixel indicates the average SHAP value of a variable (x-axis) as a function of chronological age (y-axis).

Table S1: List of the 48 biological variables, by alphabetical order

SAS label	Excluded during feature selection
Albumin (g/L)	
Albumin, urine (ug/mL)	
Alkaline phosphotase (U/L)	
ALT (U/L)	
AST (U/L)	
Basophils number (1000 cells/uL)	
Basophils percent (%)	
Bicarbonate (mmol/L)	
Bilirubin, total (umol/L)	
Blood urea nitrogen (mmol/L)	
Cholesterol (mmol/L)	
C-reactive protein(mg/dL)	
Creatinine (umol/L)	
Creatinine, urine (umol/L)	
Direct HDL-Cholesterol (mmol/L)	
Eosinophils percent (%)	Yes
Folate, RBC (nmol/L RBC)	
Folate, serum (nmol/L)	
GGT (U/L)	
Globulin (g/L)	
Glucose, serum (mmol/L)	
Glycohemoglobin (%)	
Hematocrit (%)	
Hemoglobin (g/dL)	
Iron (umol/L)	Yes
LDH (U/L)	Yes
Lymphocyte number (1000 cells/uL)	
Lymphocyte percent (%)	
MCHC (g/dL)	
Mean cell hemoglobin (pg)	
Mean cell volume (fL)	
Mean platelet volume (fL)	Yes
Monocyte number (1000 cells/uL)	
Monocyte percent (%)	
Osmolality (mmol/Kg)	
Phosphorus (mmol/L)	
Platelet count (1000 cells/uL)	
Potassium (mmol/L)	
Red blood cell count (million cells/uL)	
Red cell distribution width (%)	
Segmented neutrophils num (1000 cell/uL)	
Segmented neutrophils percent (%)	
Sodium (mmol/L)	
Total calcium (mmol/L)	
Total protein (g/L)	
Triglycerides (mmol/L)	
Uric acid (umol/L)	
White blood cell count (1000 cells/uL)	

Table S2: List of hyperparameters used during model tuning.

The grid search is presented for each model together with the best hyperparameters found for each model.

Model	Grid search parameters	Best hyperparameters found
Elastic Net	l1_ratio: [0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99] alpha: uniform (-4, -2, 0.5)	l1_ratio: 0.99 alpha: 0.0001
Random Forest	n_estimators: loguniform (100, 1000) max_features: [auto, sqrt] max_depth: randint (3,12) min_samples_split: [2,5,10] min_samples_leaf: [1,2,4] bootstrap: [True, False]	n_estimators: 598 max_features: auto max_depth: 11 min_samples_split: 5 min_samples_leaf: 2 bootstrap: True
Decision Tree	max_depth: int(2, 50) min_samples_split: int(2, 12) min_samples_leaf: int(2, 50)	max_depth: 28 min_samples_split: 6 min_samples_leaf: 24
Multilayer Perceptron	n_layers: [2,3,4] with hidden_layer_sizes [16,32,64,128,256] activation: [relu, identity] beta_1: [0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99] beta_2: [0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99] alpha: uniform (-4, -1, 0.5)	n_layers: 2 with hidden_layer_sizes (16,64,32,64) activation: relu beta_1: 0.1 beta_2: 0.4 alpha: 0.003
XGBoost Model	max_depth: [3,4] subsample: uniform(0.2, 0.8, 0.05) colsample_bytree: uniform(0.2, 1.0, 0.05) colsample_bylevel: uniform(0.2, 1.0, 0.05) learning_rate: 10^(uniform(-4.0, -1.0, 0.5))	max_depth: 3 subsample: 0.7 colsample_bytree: 0.85 colsample_bylevel: 0.9 learning_rate: 0.1
XGBoost Model with custom loss		max_depth: 3 subsample: 0.8 colsample_bytree: 1.0 colsample_bylevel: 0.5 learning_rate: 0.01

Table S3: Details of the validation variables

Socio-demographic variables	**Gender** (RIAGNDR, Male - Female) **Ethnicity** (RIDRETH1, Non-Hispanic white - Mexican American - Non-Hispanic black - Other) **Family income** : Annual Family Income (INDFMINC and INDFMIN2) divided into quartile **Poverty Index Ratio** (INDFMPIR) divided into Poverty (<1) and No poverty (≥1)
Medical variables	**Body Mass Index** : BMXMI divided into Obesity (≥ 30), Overweight (≥25 and <30), **Normal weight**: (≥18.5 and <25) and Underweight (<18.5) **Tobacco exposure** : LBXCOT (Cotinine) divided into No exposure (≤13) and Exposure (>13) **Sedentarity**: combination of PAD020, PAQ635 (Walked or bicycled over past 30 days, Yes/No), PAQ180 (Avg level of physical activity each day, from no activity to high activity) and PAD320, PAQ620 (Moderate activity over past 30 days, Yes/No). Active defined as at least one activity among the previous ones **AAC24 score:** DXXAAC24 (AAC Total 24 Score) recoded into Low score (0-1), Med score (2-5), and High score (6+) **Pathologies** Liver diseases: MCQ160L, MCQ500 (Ever told you had any liver condition) Coronary Heart Diseases: presence of a condition among MCQ160C (Ever told you had coronary heart disease), MCQ160F (Ever told you had a stroke), MCQ160B (Ever told had congestive heart failure), MCQ160D (Ever told you had angina/angina pectoris) and MCQ160E (Ever told you had heart attack) Diabetes: presence of a condition among DIQ010 (Doctor told you have diabetes) and DIQ160 (Ever told you have prediabetes) Thyroid diseases: presence of a condition among MCQ160H (Ever told you had a goiter), MCQ160M (Ever told you had a thyroid problem) and MCQ160I (Ever told you had thyroid disease) Arthritis: presence of a condition among MCQ160N (Doctor ever told you that you had gout?), MCQ160A (Doctor ever said you had arthritis) and ARQ125E (Ever told had Ankylosing Spondylitis) Cancer: MCQ220 (Ever told you had cancer or malignancy) Kidney diseases: presence of a condition among KIQ020, KIQ022 (Ever told you had weak/failing kidneys) and OHQ144 (Have kidney disease w/ renal dialysis?) Bronchitis: presence of a condition among MCQ160K (Ever told you had chronic bronchitis), MCQ160o (Ever told you had COPD?) and MCQ010 (Ever been told you have asthma) Auto-immune digestive disease: presence of a condition among ARQ125C (Ever told you had Ulcerative Colitis) and ARQ125D (Ever told you had Crohns Disease) Digestive ulcer: MCQ200 (Ever told had stomach/duod/peptic ulcer) Eye disease: presence of a condition among VIQ090 (Ever told had glaucoma) and VIQ310 (Told had macular degeneration) Dermatologic disease: presence of a condition among DEQ053, MCQ070 (Ever told had Psoriasis?) and AGQ180 (Doctor told have eczema)