AI-DRIVEN INTEGRATIVE ANALYSIS OF CROSS-SECTIONAL BIOLOGICAL DATA TO PREDICT AGING TRAJECTORIES AND AGE-RELATED DISEASES IN THE NOVARA COHORT STUDY

Garro G.;[1,2] Cracas S.;[1,2,3] Antona A.;[1] Venetucci J.; [1,2] Varalda M.; [1,2] Aleni C.; [3] Martorana M.; [3] Pighini I.;[1] Briacca L.; [2] Rolla R.; Faggiano F.;[3] Capello D.;[1,2]

1 Department of Translational Medicine, Centre of Excellence in Aging Sciences, Università del Piemonte Orientale, Novara; 2UPO Biobank, Università del Piemonte Orientale, Novara 3Department for Sustainable Development and Ecological Transition, Novara;

mail: giulia.garro@uniupo.it

BACKGROUND

Aging is a major challenge of the 21st century, marked by rising rates of diseases like cancer, cardiovascular diseases, and neurodegenerative disorders. The aging process is shaped by genetic, environmental, and lifestyle factors, and understanding these could improve health outcomes for the elderly. Although many aging biomarkers have been identified, their practical use is limited due to individual variability, study design differences, and inconsistent results. There is no standardized method for biomarker discovery in aging. Recent multi-omics approaches, combined with clinical and environmental data, provide valuable insights into aging. However, analyzing this data requires advanced techniques like machine learning (ML) and deep learning (DL), which are integral to artificial intelligence (AI).

This study aims to identify predictive biomarkers for aging and age-related diseases by integrating biological and clinical data from the Novara Cohort Study (NCS), using established AI pipelines. First, descriptive statistics of the 66 blood-derived biomarkers and distribution of the principal questionnaire were analyzed, followed by PLS-DA to pinpoint markers linked to conditions like aging and cardiovascular disease.

Although the study's cross-sectional design limits predictive accuracy without longitudinal data, it offers valuable insights into aging profiles and suggests the potential of AI in advancing personalized aging strategies and early disease detection.

RESULTS

DISTRIBUTION OF THE CHARACTERISTICS OF NOVARA COHORT STUDY

A.

CHARACTERISTICS	UNDER 40	40-65	OVER 65
NUMBER OF PARTECIPANTS	54 (14.9)	119 (32.8)	189 (52.2)
AGE (MEAN)	54 (14.9)	119 (32.9)	189 (52.2)
FEMALE (%)	25 (46.3)	78 (65.5)	85 (45)
BMI (MEAN)	24.1	25.3	25.9
EDUCATION LEVEL (GRADE AND POST-GRADE) (%)	37 (10.2)	27 (7.5)	65 (18.0)
SMOKING	16 (29.6)	16 (13.44)	16 (8.5)
FAMILY INCOME (>2500)	30 (12.5)	35 (14.6)	80 (33.3)
MARRIAGE STATUS (%)	15 (31.3)	69 (60.5)	134 (72.4)
RETIRED (%)	0	34 (9.4)	172 (47.5)
EMPLOYED STATUS	44 (83)	71 (59.7)	12 (6.4)
PHYSICAL ACTIVITY, (ACTIVE) (%)	24 (10)	30 (12.5)	72 (30)
ADERENCHE OF MEDITERRANEAN DIET	13 (5.4)	22 (9.2)	74 (31.0)

B.

ICD10 DISEASES CLASSIFICATION	FREQUENCY (%)
CIRCULATORY SYSTEM DISORDERS	184 (50.8)
ENDOCRINE, NUTRITIONAL, AND METABOLIC DISEASES	133 (36.7)
NO PATHOLOGY	72 (19.9)
NEOPLASMS	60 (16.6)
MUSCULOSKELETAL AND CONNECTIVE TISSUE DISEASES	51 (14.1)
GASTROINTESTINAL DISEASES	47 (13.0)
GENITOURINARY DISORDERS	37 (10.2)
RESPIRATORY SYSTEM DISORDERS	30 (8.3)
NEUROPATHIES	25 (6.9)
MENTAL AND BEHAVIORAL DISORDERS	21 (5.8)
OPHTHALMOPATHIES	17 (4.7)
DERMATOPATHIES	17 (4.7)
EAR DISEASES	15 (4.1)
INFECTIOUS AND PARASITIC DISEASES	15 (4.1)
BLOOD AND IMMUNE SYSTEM DISEASES	13 (3.6)
MALFORMATIONS AND CHROMOSOMAL ABNORMALITIES	12 (3.3)
TRAUMA, TOXICOLOGY, AND OTHER EXTERNAL CAUSES OF DISEASE	11 (3.0)
SIGNS, SYMPTOMS, AND ABNORMAL LABORATORY FINDINGS NOT OTHERWISE CLASSIFIED	2 (0.6)
FACTORS THAT MAY INFLUENCE WELL-BEING	2 (0.6)
PREGNANCY, CHILDBIRTH, OR PUERPERIUM	1 (0.3)
CONGENITAL DISORDERS	1 (0.3)

A. Description of the principal questionnaire that investigate the lifestyle of our participants. Categorical variables were expressed as frequencies (percentages). BMI, body mass index. **B.** The disease are expressed following the ICD10 classification.

EXPLORING BLOOD-DERIVED BIOMARKERS IN AGE AND AGE ASSOCIATED DISEASE

A.

VARIABLES	Mean (SD)/Median (IQR) <65	Mean (SD)/Median (IQR) >65	p_value
eGFR (IQR)	93 (20.5)	73 (26)	0.00
CysC (IQR)	0.86 (0.25)	1.03 (0.27)	0.00
DD (IQR)	326 (230.75)	463 (411.5)	0.00
RDW-SD (IQR)	43 (3.15)	44.3 (0.9)	0.00
ALB (IQR)	4.6 (0.35)	4.4 (0.3)	0.00
BUN (IQR)	14.6 (5.75)	16.6 (5.1)	0.00
HbA1c% (IQR)	5.4 (0.5)	5.5 (0.5)	0.00
HbA1c (IQR)	36 (6)	37 (5)	0.00
K (IQR)	3.9 (0.3)	4 (0.49	0.00
LYMPH% (SD)	33.56 (6.90)	30.18 (8.76)	0.00
RDW-CV (IQR)	13.2 (0.9)	13.5 (0.9)	0.00
MCV (IQR)	89.1 (5.45)	91.1 (4.7)	0.00
FER (IQR)	91.8 (105.65)	124(139.4)	0.00
NEU% (SD)	56.33 (7.12)	59.29 (9.16)	0.00
CR (IQR)	0.81 (0.24)	0.89 (0.28)	0.00
UA (IQR)	4.70 (1.25)	5.07 (1.14)	0.00

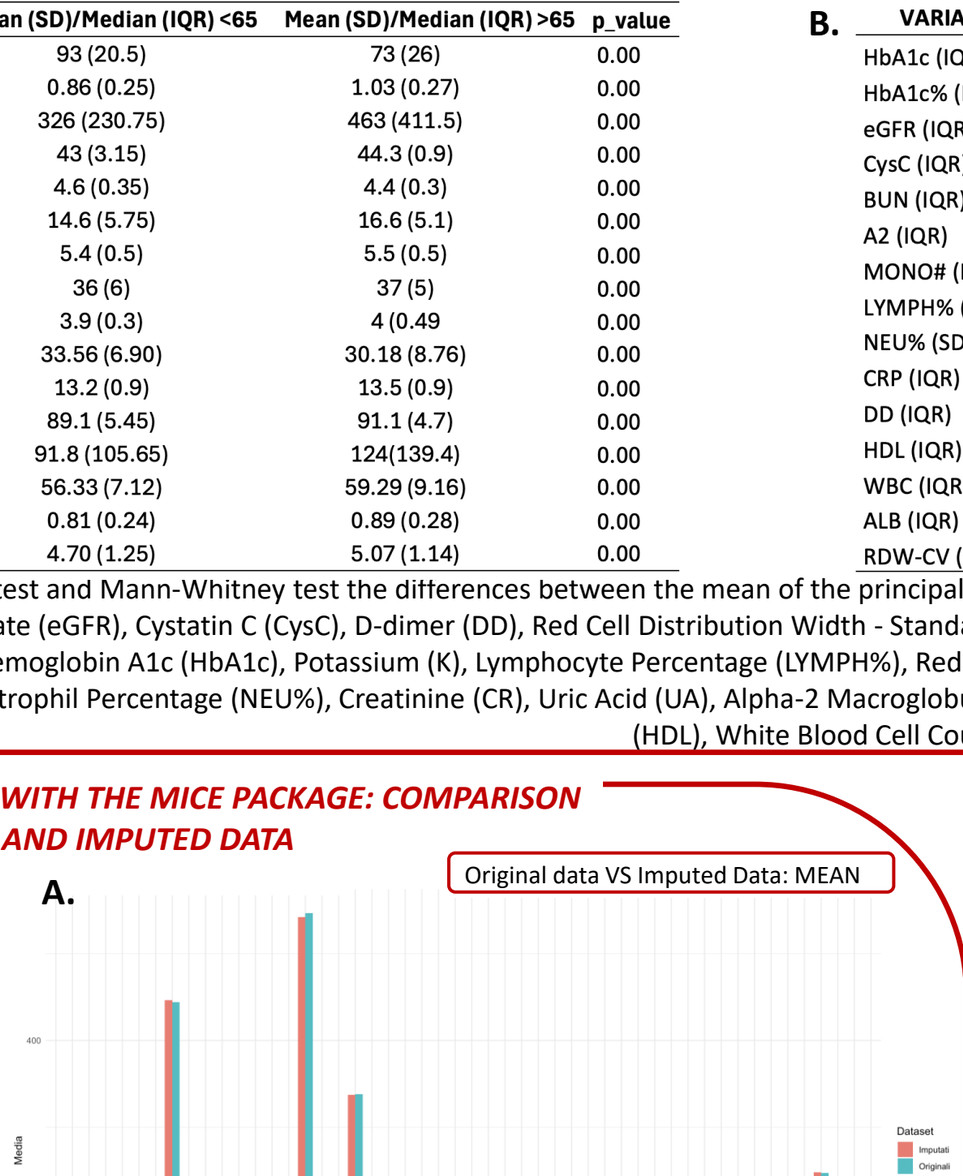
B.

VARIABLES	Mean(SD)/Median (IQR) NoCVD	Mean(SD)/Median (IQR) CVD	p_value
HbA1c (IQR)	35 (4)	38 (5)	0.00
HbA1c% (IQR)	5.4 (0.4)	5.6 (0.4)	0.00
eGFR (IQR)	87 (21)	78 (29.5)	0.00
CysC (IQR)	0.89 (0.23)	1.01 (0.3)	0.00
BUN (IQR)	15 (5.9)	16.1 (5.3)	0.00
A2 (IQR)	9.1 (1.8)	9.8 (1.9)	0.00
MONO# (IQR)	0.33 (0.15)	0.37 (0.13)	0.00
LYMPH% (SD)	33.56 (6.89)	30.18 (8.76)	0.00
NEU% (SD)	56.33 (7.11)	59.29 (9.16)	0.00
CRP (IQR)	0.09 (0.13)	0.13 (0.19)	0.01
DD (IQR)	372 (245.75)	415 (321)	0.01
HDL (IQR)	63 (22.5)	57 (22)	0.01
WBC (IQR)	5.09 (1.65)	5.34 (1.85)	0.02
ALB (IQR)	4.5 (0.4)	4.4 (0.3)	0.02
RDW-CV (IQR)	13.3 (0.85)	13.5 (1)	0.03

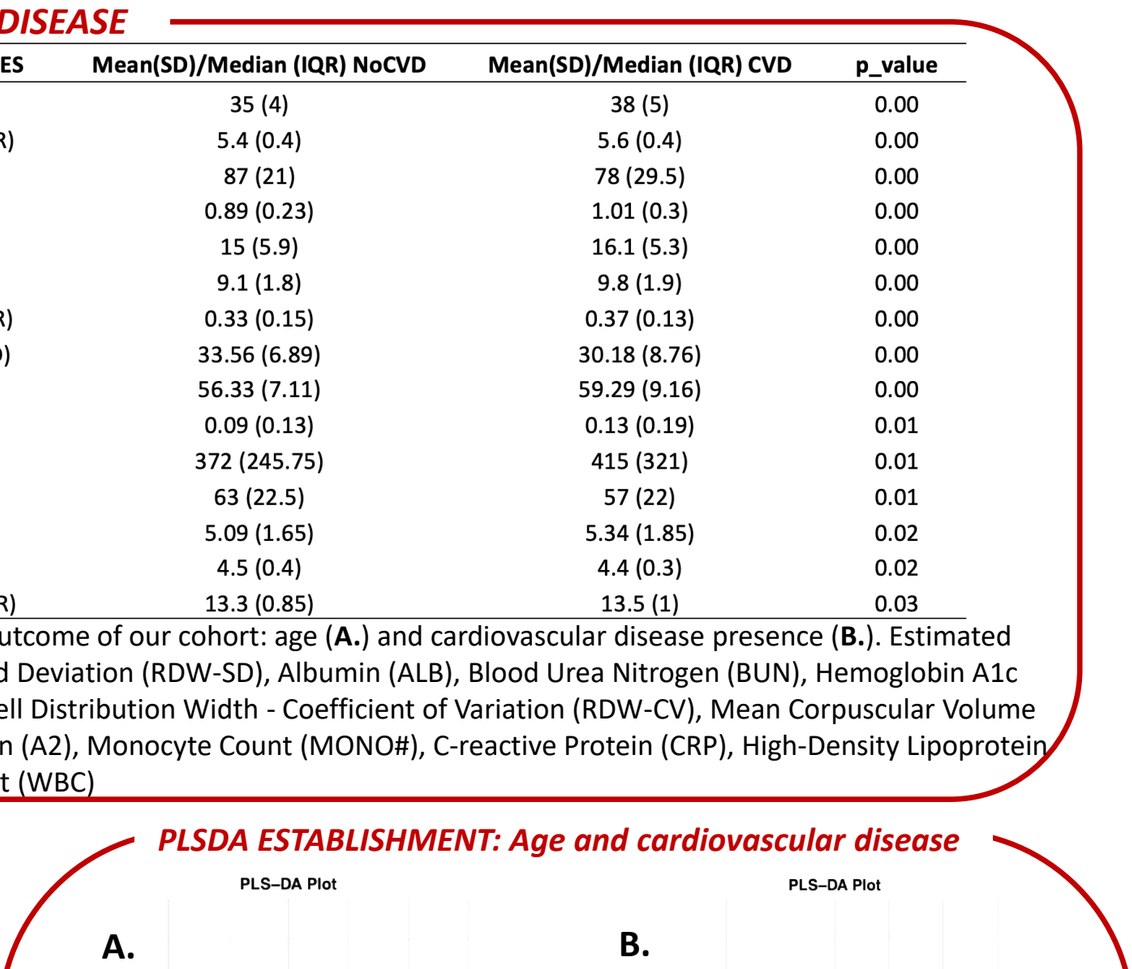
Exploration through T-test and Mann-Whitney test the differences between the mean of the principal outcome of our cohort: age (**A.**) and cardiovascular disease presence (**B.**). Estimated Glomerular Filtration Rate (eGFR), Cystatin C (CysC), D-dimer (DD), Red Cell Distribution Width - Standard Deviation (RDW-SD), Albumin (ALB), Blood Urea Nitrogen (BUN), Hemoglobin A1c Percentage (HbA1c%), Hemoglobin A1c (HbA1c), Potassium (K), Lymphocyte Percentage (LYMPH%), Red Cell Distribution Width - Coefficient of Variation (RDW-CV), Mean Corpuscular Volume (MCV), Ferritin (FER), Neutrophil Percentage (NEU%), Creatinine (CR), Uric Acid (UA), Alpha-2 Macroglobulin (A2), Monocyte Count (MONO#), C-reactive Protein (CRP), High-Density Lipoprotein (HDL), White Blood Cell Count (WBC)

DATA IMPUTATION WITH THE MICE PACKAGE: COMPARISON OF ORIGINAL DATA AND IMPUTED DATA

Missing data were imputed using the MICE package in R. Before conducting the classification analysis, it is crucial comparing the descriptive parameters of the original dataset with those of the imputed one. **A.** Compare the mean. **B.** Compare the standard deviation.



PLSDA ESTABLISHMENT: Age and cardiovascular disease



PLS-DA analysis aim to classify and discriminate the biomarkers most closely associated with the outcomes. **A.** describe the different analytes between over and under 65 years old groups. **B.** describe the different analytes between the participant that has CVDs and not. The VIPscore>1 (variable important in projection) explain the variable that discriminate better the group.

CONCLUSION

The NCS study shows that participants generally have a healthy lifestyle, but the voluntary recruitment process introduces a selection bias, limiting the population's representativeness. Future recruitment should target more diverse groups with varied habits.

Biomarkers related to aging and disease were identified, particularly kidney function markers like eGFR and cystatin C, which are closely linked to aging and cardiovascular disease.

While the PLS-DA model successfully classified aging, it struggled with cardiovascular disease classification, suggesting the need for alternative methods like random forest or logistic regression for better group differentiation.

In the future, integrating additional omics data, such as proteomics and metabolomics, will provide a more comprehensive understanding of biological processes linked to aging. The long-term goal is to create an aging phenotype specific to the population and develop a personalized "biological clock" to predict biological age and age-related diseases.

UPO AGING PROJECT UNIVERSITÀ DEL PIEMONTE ORIENTALE