Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Senescence and Genotoxicity Signatures Across Clinical and Laboratory Outcomes in Patients at High Cardiovascular Risk: An Automated Machine Learning Study

Vitalii Kyryk

biomedpost@gmail.com

National Scientific Center "M. D. Strazhesko Institute of Cardiology, Clinical and Regenerative Medicine of the National Academy of Medical Sciences of Ukraine"

Dmytro Krasnienkov

D. F. Chebotarev State Institute of Gerontology of the National Academy of Medical Sciences of Ukraine

Alina Ustymenko

National Scientific Center "M. D. Strazhesko Institute of Cardiology, Clinical and Regenerative Medicine of the National Academy of Medical Sciences of Ukraine"

Oleg Tsupykov

O. Bogomoletz Institute of Physiology of the National Academy of Sciences of Ukraine

Kostiantyn Midlovets

D. F. Chebotarev State Institute of Gerontology of the National Academy of Medical Sciences of Ukraine

Oleksandr Skarzhevskyi

National Scientific Center "M. D. Strazhesko Institute of Cardiology, Clinical and Regenerative Medicine of the National Academy of Medical Sciences of Ukraine"

Dmytro Khomiakov

National Scientific Center "M. D. Strazhesko Institute of Cardiology, Clinical and Regenerative Medicine of the National Academy of Medical Sciences of Ukraine"

Oleksandr Parkhomenko

National Scientific Center "M. D. Strazhesko Institute of Cardiology, Clinical and Regenerative Medicine of the National Academy of Medical Sciences of Ukraine"

Research Article

Keywords: cardiovascular risk, COVID-19, biomarkers, cellular senescence, genomic instability, machine learning

Posted Date: November 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11524324/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Senescence and Genotoxicity Signatures Across Clinical and Laboratory Outcomes in Patients at High Cardiovascular Risk: An Automated Machine Learning Study

Vitalii Kyryk[1,2], Dmytro Krasnienkov[3], Alina Ustymenko[1,2], Oleg Tsupykov[1,4], Kostiantyn Midlovets[3], Oleksandr Skarzhevskyi[5], Dmytro Khomiakov[5], Oleksandr Parkhomenko[5]

[1]*National Scientific Center "M. D. Strazhesko Institute of Cardiology, Clinical and Regenerative Medicine of the National Academy of Medical Sciences of Ukraine", Institute of Genetic and Regenerative Medicine, Kyiv, Ukraine*

[2]*Chebotarev State Institute of Gerontology of the National Academy of Medical Sciences of Ukraine, Laboratory of Pathophysiology and Immunology, Kyiv, Ukraine*

[3]*Chebotarev State Institute of Gerontology of the National Academy of Medical Sciences of Ukraine, Laboratory of Epigenetics, Kyiv, Ukraine*

[4]*Bogomoletz Institute of Physiology, National Academy of Sciences of Ukraine, Department of Cytology, Kyiv, Ukraine*

[5]*National Scientific Center "M. D. Strazhesko Institute of Cardiology, Clinical and Regenerative Medicine of the National Academy of Medical Sciences of Ukraine", Department of Intensive Care and Resuscitation, Kyiv, Ukraine*

Corresponding author email (Vitalii Kyryk): biomedpost@gmail.com

Abstract

Background. *Cellular senescence, chromosomal instability, endothelial injury, and altered immune composition may jointly reflect biological aging and cardiometabolic dysfunction. We evaluated whether a fixed 12-biomarker panel provided reproducible feature-level clues for 11 pre-selected clinical and laboratory outcomes in a three-group cohort at high cardiovascular risk.*

Methods. *The dataset included 90 patients suffered from acute coronary syndrome (ACS), divided equally into three groups (n = 30 each): pre-pandemic ACS (Group 1), ACS with*

concomitant laboratory-confirmed acute COVID-19 during the pandemic (Group 2), and ACS without concomitant acute COVID-19 during the COVID-19 pandemic period (Group 3). Four model families were evaluated separately using 5-fold cross-fitting repeated five times. All imputation, scaling, and model fitting were restricted to training folds. Feature evidence integrated held-out permutation importance, cross-fitted SHAP values, linear coefficient stability, Group+SEX-adjusted partial rank associations, within-group sensitivity analyses, and leave-one-group-out transportability.

Results. *The largest median improvements over Group+SEX baselines occurred for history of acute myocardial infarction with coronary revascularization procedures (Δ=0.161), glucose level (Δ=0.118), heart rate (Δ=0.096), and glycated hemoglobin percentage (Δ=0.091). Recurrent feature themes included senescence-associated heterochromatin foci and $CD34^{+}$ progenitor cells for glycemic outcomes, micronuclei for heart rate and prior COVID-19, nucleoplasmic bridges for cardiac-history and inflammatory/immune outcomes, and $CD31^{hi}$ circulating endothelial cells for leukocyte and granulocyte counts. Absolute performance remained weak for several outcomes despite stable feature rankings.*

Conclusions. *Rather than predicting any single outcome with strong absolute accuracy, the same senescence and genotoxicity markers recurred as leading clues across metabolically, cardiac, and immunologically distinct outcomes, and did so consistently across three pathogenetic different cohorts. This cross-outcome stability suggests these markers may index a shared, stressor-independent dimension of cellular burden in ACS patients rather than outcome-specific biology, a hypothesis that warrants mechanistic follow-up.*

Keywords*: cardiovascular risk; COVID-19; biomarkers; cellular senescence; genomic instability; machine learning*

Introduction

Patients at high cardiovascular risk, particularly those presenting with acute coronary syndrome (ACS), are especially vulnerable to systemic stressors capable of altering inflammatory, immune, endothelial, thrombotic, and metabolic processes. The COVID-19 pandemic represented such a complex stressor, combining the direct effects of SARS-CoV-2 infection with substantial changes in healthcare delivery and considerable psychosocial burden. In a multicenter international prospective registry of patients with ACS and COVID-19, substantial delays from symptom onset to hospital admission and higher in-hospital mortality were observed compared with pre-pandemic cohorts, with the increased risk of death persisting after adjustment for confounding factors [1]. Similarly, large observational cohort studies from England and Spain reported less favorable outcomes in patients with COVID-19-associated ACS than in COVID-19-negative patients [2, 3]. Cardiovascular consequences of SARS-CoV-2 infection may also extend beyond the acute phase, with persistent inflammatory, metabolic, endothelial, and thrombotic abnormalities potentially contributing to long-term cardiovascular vulnerability [4]. These findings indicate that SARS-CoV-2 infection may modify the clinical and biological context of ACS beyond conventional cardiovascular risk factors.

These considerations support the search for cellular biomarkers capable of integrating multiple dimensions of systemic stress. Unlike individual biochemical parameters, cytological biomarkers can provide information on cellular senescence, genomic instability, regenerative cell populations, endothelial status, and immune remodeling. Such markers may therefore complement conventional clinical and laboratory assessment and provide additional information about the biological state of patients at high cardiovascular risk.

Cellular senescence is a potential biological link between systemic stress and persistent cellular alterations. Senescent cells exhibit stable cell-cycle arrest, chromatin remodeling, metabolic changes, and a senescence-associated secretory phenotype (SASP), which may contribute to chronic inflammation and impaired tissue and vascular function [5].

Senescence can be induced by diverse stressors, including DNA damage, oxidative stress, inflammation, and viral infection. Experimental studies have demonstrated that SARS-CoV-2 can induce cellular senescence, accompanied by the release of pro-inflammatory and pro-coagulant factors [6]. SARS-CoV-2-induced senescence has also been implicated in endothelial dysfunction and amplification of inflammatory and thrombotic processes [7]. Chronic psychosocial stress may similarly promote a senescence-associated phenotype [8]. Together with evidence linking pandemic-related stress to altered biological aging, these findings provide a rationale for investigating cellular senescence markers in patients exposed to prolonged systemic stress during COVID-19 pandemic.

Because cellular senescence is a heterogeneous phenotype, a combination of complementary markers may provide a more informative assessment than any single marker. Senescence-associated β-galactosidase is one of the most widely used cytological indicators of senescence, although its intensity is graded and influenced by the duration of cell-cycle withdrawal rather than being absolutely specific for senescence [9, 10]. Nuclear chromatin remodeling represents another characteristic of cellular senescence. Senescence-associated heterochromatin foci (SAHF) and broader changes in heterochromatin organization reflect structural and transcriptional alterations accompanying cellular senescence [11, 12]. In parallel, accumulation of DNA damage and chromosomal abnormalities is closely related to the development of senescence. The cytokinesis-block micronuclei (CBMN) cytome assay provides an established cytological approach for assessing chromosomal damage and genomic instability, and increased cytogenetic damage has been reported in patients with COVID-19 [13, 14]. The combination of senescence-associated β-galactosidase (SA-β-gal), chromatin alterations, and micronucleus analysis therefore provides complementary information on cellular senescence and genotoxic stress.

Additional dimensions of cellular homeostasis may be captured by circulating progenitor, endothelial, and immune cell populations. $CD34^+$ cells represent a heterogeneous

population that includes hematopoietic progenitor cells and other cells with regenerative potential. Lower levels of circulating $CD34^+$ progenitor cells have been associated with adverse outcomes in patients with coronary artery disease, supporting their potential relevance as indicators of regenerative disease [15]. Cells expressing vascular endothelial growth factor receptor 2 (VEGFR2/CD309) have traditionally been used to characterize the endothelial progenitor/angiogenic compartment, whereas circulating CD31-positive endothelial cells may reflect endothelial injury, activation, and turnover [16]. Assessment of these populations may therefore provide information complementary to cellular senescence and genotoxicity markers.

$CD4^+CD8^+$ double-positive (DP) T lymphocytes represent another potentially informative but insufficiently characterized cellular population. T-cell populations undergo substantial quantitative and functional remodeling in response to SARS-CoV-2 infection, systemic inflammation, and senescence. Single-cell transcriptomic analysis has demonstrated substantial heterogeneity among circulating DP cells, including populations with naive, helper, regulatory, cytotoxic, and innate-like characteristics [17]. Their potential relevance to patients at high cardiovascular risk and to individuals exposed to prolonged infectious or psychosocial stress remains insufficiently understood.

Collectively, these observations support the use of a multidimensional blood-cell biomarker panel encompassing cellular senescence, genotoxicity, hematopoietic and endothelial progenitor phenotypes, circulating endothelial cells, and DP T lymphocytes. However, it remains unclear whether such a cellular profile can provide information about other routinely measurable clinical, biochemical, and cytological characteristics of patients at high cardiovascular risk. Importantly, given the large number of potential target variables, only a subset may demonstrate meaningful predictive relationships with the cellular biomarker profile. Automated machine learning (AutoML) provides an opportunity to

systematically examine such relationships and identify the subset of target variables for which the cellular biomarker panel demonstrates meaningful predictive performance [18].

Methods

This study aimed to determine whether a multidimensional panel of blood-cell biomarkers reflecting cellular senescence, genotoxicity, hematopoietic and endothelial progenitor phenotypes, circulating endothelial cells, and $CD4^+CD8^+$ double-positive T lymphocytes could be used to predict selected clinical, biochemical, and cytological parameters in patients at high cardiovascular risk. An automated machine-learning framework incorporating cross-fitted feature discovery was applied to a broad set of candidate target variables to identify those for which the cellular biomarker panel demonstrated meaningful predictive performance.

The study was conducted in accordance with the principles of the Declaration of Helsinki and applicable national regulations governing biomedical research involving human participants. The study protocol was reviewed and approved by the local Ethics Committee of the Strazhesko National Scientific Center (protocol #1-23 dated September 17, 2024). All participants provided written informed consent for participation in the study and for the use of their clinical and laboratory data for research purposes. For participants whose data were obtained retrospectively from medical records, the Ethics Committee approved the use of previously collected data in accordance with applicable regulations. Patient confidentiality was maintained throughout the study. Clinical trial number: not applicable.

Study population. The study cohort comprised 90 patients, who were divided into three groups of 30 patients each according to the timing of the index ACS event in relation to the COVID-19 pandemic and the presence or absence of concomitant COVID-19 at the time of the index ACS event. Potential participants were identified by reviewing the medical records of all patients discharged after treatment for ACS during the predefined study periods (Fig. 1).

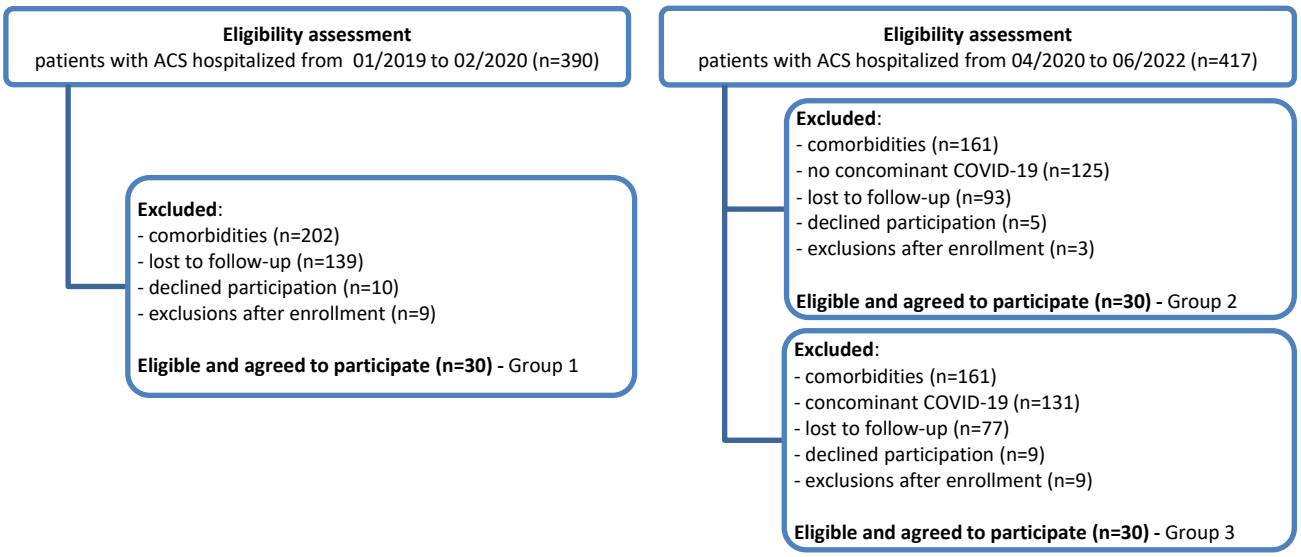


Fig. 1. Flow chart of patient recruitment to the retrospective cohort study.

The index ACS event was defined as the hospitalization for ACS that met the study eligibility criteria. The three groups were comparable in size, while their classification reflected distinct temporal and clinical contexts. The pre-pandemic cohort (Group 1) comprised 30 patients admitted between January 1, 2019, and March 31, 2020, before the onset of the COVID-19 pandemic. The pandemic cohort (Group 2) included 30 patients who developed ACS with concomitant laboratory-confirmed acute COVID-19 and were hospitalized between April 1, 2020, and June 31, 2022. The non-COVID-19 cohort (Group 3) comprised 30 patients admitted between April 1, 2020, and June 31, 2022 who developed ACS in the absence of concomitant acute COVID-19.

The inclusion criteria were: ST-segment elevation acute coronary syndrome (STEMI); hospital admission within 6 hours of symptom onset; age between 35 and 85 years; and laboratory-confirmed acute SARS-CoV-2 infection for patients assigned to Group 2. Patients were excluded if they had diabetes mellitus, active malignancy, severe heart failure, uncontrolled arterial hypertension, chronic inflammatory disease, chronic kidney disease, or anemia.

Clinical follow-up covered the period from hospital discharge to study enrolment. All participants underwent comprehensive clinical and laboratory assessment 61-73 months

after hospitalization. Medical records were reviewed to identify documented COVID-19 episodes and vaccination history (verified through the National Electronic Health System of Ukraine), recurrent acute myocardial infarction (AMI) ischemic events with coronary revascularization procedures included elective percutaneous coronary intervention (PCI) or coronary artery bypass grafting (CABG), and all cardiovascular-related hospital admissions. The demographic and clinical characteristics of both study groups are summarized in supplementary table S1 [see Additional file 1].

Clinical and laboratory parameters assessment. A range of clinical, hematological, cytological, and biochemical parameters was assessed in the patients [see Additional file 2]. All participants underwent a comprehensive clinical evaluation, including standard 12-lead electrocardiography. Transthoracic echocardiography was performed using an SSH-880CV Aplio Artida ultrasound system (Toshiba Medical Systems Corporation, Japan). Endothelial function was assessed by measuring brachial artery flow-mediated dilation (FMD) using the conventional ultrasound-based protocol, as previously described [19].

Peripheral venous blood samples were collected at the follow-up visit for the assessment of routine hematological, biochemical, cellular senescence, and genotoxicity biomarkers. A complete blood count was performed using an Elite 3 hematology analyzer (ERBA, Czech Republic). Serum biochemical parameters were measured on an A-25 automated chemistry analyzer (BioSystems, Spain) using commercially available reagent kits. The analyzed parameters included glycated hemoglobin (HbA1c), glucose, creatinine, N-terminal prohormone of brain natriuretic peptide (NT-proBNP) and lipid profile, comprising total cholesterol, triglycerides, high-density lipoprotein cholesterol, and non-high-density lipoprotein cholesterol. Plasma C-reactive protein (CRP) concentrations were determined using a BTS-350 semi-automated biochemical analyzer (BioSystems, Spain).

The relative proportions of leukocyte subpopulations expressing CD45, CD34, CD31, CD38, CD309, CD3, CD4, and CD8 markers were determined by multiparameter flow

cytometry using fluorochrome-conjugated monoclonal antibodies (all Becton Dickinson, USA) as previously described [19, 20]. SA-β-gal activity was evaluated using the Senescence Assay Kit (cat. no. ab228562, Abcam, USA) according to the manufacturer's instructions. Flow cytometric analyses were performed on a BD FACSAria cell sorter (Becton Dickinson, USA).

The proportion of leukocytes containing senescence-associated heterochromatin foci was determined by fluorescence microscopy following DAPI staining of peripheral blood smears, as previously described [21]. Genotoxicity was assessed on 72h peripheral blood mononuclear cells cultures using the CBMN cytome assay according to the previously published protocol [22]. For each sample, 1,000 binucleated cells were scored to determine the frequencies of micronuclei, nucleoplasmic bridges, and nuclear buds, together with apoptotic and necrotic cells. In addition, mono-, bi-, tri-, and tetranucleated cells were enumerated to calculate the cytokinesis-block proliferation index (CBPI). Microscopic evaluation was performed using an IX71 fluorescence microscope (Olympus, Japan), and image analysis was carried out with QuickPHOTO MICRO version 2.3 software (Olympus, Japan).

Machine learning pipeline. To identify associations between cellular senescence/genotoxicity biomarkers and clinical or laboratory parameters, all patient groups were merged into a single analytical cohort. The original group assignment was intentionally excluded from model development to evaluate biomarker-outcome relationships independently of COVID-19 status.

Twelve biomarkers were selected as fixed predictor variables and represented four biological domains: cellular senescence (SA-β-gal-positive and SAHF-positive leukocytes), endothelial and hematopoietic cell populations ($CD34^+$ hematopoietic progenitor cells, $CD309^+$ endothelial progenitor cells, and $CD31^+$ circulating endothelial cells), immature immune cell phenotype ($CD4^+CD8^+$ double-positive T cells), and genotoxicity indicators

(micronuclei, nuclear buds, nucleoplasmic bridges, necrotic and apoptotic cells, and Cytokinesis-Block Proliferation Index).

Eleven outcomes were pre-selected because they represented overlap or informative disagreement between the original AutoML and strict optimized analyses. Continuous outcomes were analyzed by regression. AMI+PCI+CABG and COVID-19 were analyzed as binary outcomes. Targets were never imputed. Median imputation and scaling of biomarkers, together with encoding of Group and SEX, were performed inside each training fold only. Regularized linear/logistic regression, Extra Trees, Random Forest, and conservative histogram gradient boosting were evaluated as fixed candidates. Five-fold cross-fitting was repeated five times for pooled biomarker-only and Group+SEX-adjusted models. Models were reported separately, avoiding same-fold winner selection. Regression used R^2, RMSE, MAE, and Spearman correlation; binary classification used ROC-AUC, Brier score, and balanced accuracy. Permutation importance was calculated only in held-out observations. SHAP (Shapley Additive Explanations) values were calculated for held-out observations using the best fixed tree model for each target in both adjusted and biomarker-only analyses. SHAP contributions were then averaged across the five cross-fitting repeats for each participant, preventing repeated observations from visually dominating the plots. The direction of a SHAP effect was summarized by the rank correlation between raw feature values and their mean SHAP contributions. Standardized linear coefficients and adjusted partial rank correlations served as independent direction checks. SHAP describes model behavior and does not establish causality.

The AutoML screen was applied to the full panel of candidate outcome variables (n = 50, including the eleven outcomes above), each modeled individually as a regression target. These included demographic characteristics, cardiovascular functional parameters, complete blood count indices, biochemical measurements (glucose, glycated hemoglobin, creatinine,

lipid profile, C-reactive protein, and NT-proBNP), and additional immunophenotyping variables.

Automated model development was performed using H2O AutoML (version 3.46.0.9) [23]. For each target variable, the framework generated and optimized up to 20 candidate models, including gradient boosting machines (GBM) with grid search and learning-rate annealing, distributed random forests (DRF), generalized linear models (GLM), and stacked ensemble models, whereas deep learning models were excluded. Model training and internal validation were performed using two-fold cross-validation, and candidate models were ranked according to the cross-validated root mean square error (RMSE). The maximum training time was limited to 180 seconds for each target variable. Cross-validation models, out-of-fold (OOF) predictions, and fold assignments were retained, and the ten highest-ranked models for each target were archived to ensure reproducibility [24].

Model interpretability was assessed using a strictly out-of-fold SHAP workflow to eliminate information leakage [25]. Because TreeSHAP is not implemented for stacked ensemble and generalized linear models, StackedEnsemble and GLM algorithms were excluded from SHAP analysis. For each remaining tree-based model, SHAP values were calculated separately for each cross-validation fold using only the corresponding holdout samples that had not been used during model training. The SHAP values obtained from both folds were subsequently combined to generate a complete OOF SHAP matrix for the entire study cohort. Biomarker importance was quantified as the mean absolute SHAP value and summarized across the top-performing models using the median feature rank. To verify the robustness of feature attribution, permutation feature importance was additionally calculated using the same out-of-fold evaluation strategy.

Statistical analysis. Predictive model performance was evaluated using the cross-validated coefficient of determination (R^2), calculated as 1 – MSE/Var(y), where MSE represents the mean squared error and Var(y) the variance of the target variable. Models with

a cross-validated R^2 value greater than 0.10 were considered to demonstrate meaningful predictive performance and were retained for further interpretation.

Pairwise associations between variables were assessed using Pearson's correlation analysis. Correlation matrices display only statistically significant correlations ($p < 0.05$), whereas non-significant associations are shown in grey.

Model interpretability was evaluated using OOF SHAP values generated with the SHAP Python package (version 0.51). Global feature effects were visualized using SHAP beeswarm plots. To assess the stability of biomarker importance across models, feature ranking distributions were summarized as boxplots based on the ten highest-performing models and ordered according to their median importance rank.

Results

Predictor and outcome distributions.

After descriptive FDR adjustment, the biomarkers showing the clearest between-group heterogeneity included the percentage of white blood cells positive for SA-β-gal and SAHF, the percentage of binucleated cells with micronuclei, nuclear buds and nucleoplasmic bridges, apoptotic and necrotic cells, the cytokinesis-block proliferation index, and the percentages of $CD34^+$ hematopoietic progenitors, $CD309^+$ endothelial progenitor cells, $CD31^{hi}$ circulating endothelial cells, and $CD4^+CD8^+$ double-positive T lymphocytes. This heterogeneity provides a rationale for evaluating model performance under Group- and sex-adjusted and leave-one-group-out settings rather than relying exclusively on pooled models [see Additional file 3].

Across the full set of clinical, biochemical, and cytological variables considered as candidate outcomes, 11 outcomes demonstrated sufficient predictive signal to warrant further analysis: absolute white blood cell count, absolute neutrophil count, lymphocyte percentage, neutrophil-to-lymphocyte ratio (NLR), percentage of $CD31^-CD38^+$ cells, blood glucose level, C-reactive protein (CRP) level, HbA1c percentage, heart rate, and indicators

reflecting a history of AMI+PCI+CABG or COVID-19 [see Additional file 4]. These tables are descriptive and should not be interpreted as evidence of predictive validity. The between-group comparison q-values quantify the degree of group heterogeneity and do not test whether a biomarker panel can predict an outcome.

Cross-fitted predictive performance was subsequently evaluated for the identified outcomes (Table 1). Improvement over the corresponding nuisance baseline should be distinguished from absolute predictive performance. NLR, absolute neutrophil count, and absolute white blood cell count showed relatively consistent improvement over their respective nuisance baselines; however, their median cross-validated R^2 values remained close to zero or negative, indicating limited absolute predictive performance.

Table 1. Cross-fitted predictive performance of the model.

Outcome	Type	Best model	Metric	P_{10}-P_{90}	Δ baseline	Positive repeats
History of AMI+PCI+CABG	B	random_forest	0.628	0.625 to 0.636	0.161	100%
Glucose	R	random_forest	0.158	0.098 to 0.195	0.118	100%
Heart rate	R	random_forest	0.078	0.013 to 0.127	0.096	100%
HbA1C	R	extra_trees	0.111	0.097 to 0.164	0.091	100%
NLR	R	random_forest	-0.022	-0.055 to 0.043	0.058	100%
Neutrophils	R	random_forest	-0.038	-0.077 to 0.000	0.052	100%
Lymphocytes	R	random_forest	0.003	-0.060 to 0.033	0.048	100%
History of COVID-19 events	B	extra_trees	0.676	0.663 to 0.686	0.047	100%
WBCs	R	random_forest	-0.012	-0.024 to 0.024	0.029	100%

$CD31^{-}CD38^{+}$	R	random_forest	0.223	0.193 to 0.259	0.007	60%
CRP	R	extra_trees	-0.070	-0.103 to -0.024	0.006	60%

Notes: B – binary, CABG – Coronary Artery Bypass Grafting, CRP – C-reactive protein, HbA1c – glycated hemoglobin, AMI – Acute Myocardial Infarction, NLR – Neutrophil-to-Lymphocyte Ratio, PCI – Percutaneous Coronary Intervention, R – regression, WBCs – white blood cells, P_{10}–P_{90} is the 10th-to-90th percentile range of the metric across the repeated cross-validation runs – it's a spread/uncertainty band.

In contrast, blood glucose, HbA1c percentage, heart rate, and the indicators reflecting a history of AMI+PCI+CABG or COVID-19 demonstrated the most interpretable combination of improvement over baseline and non-trivial absolute predictive performance (Fig. 2). These outcomes therefore represented the most robust candidates for subsequent interpretation of predictor contributions and feature stability.

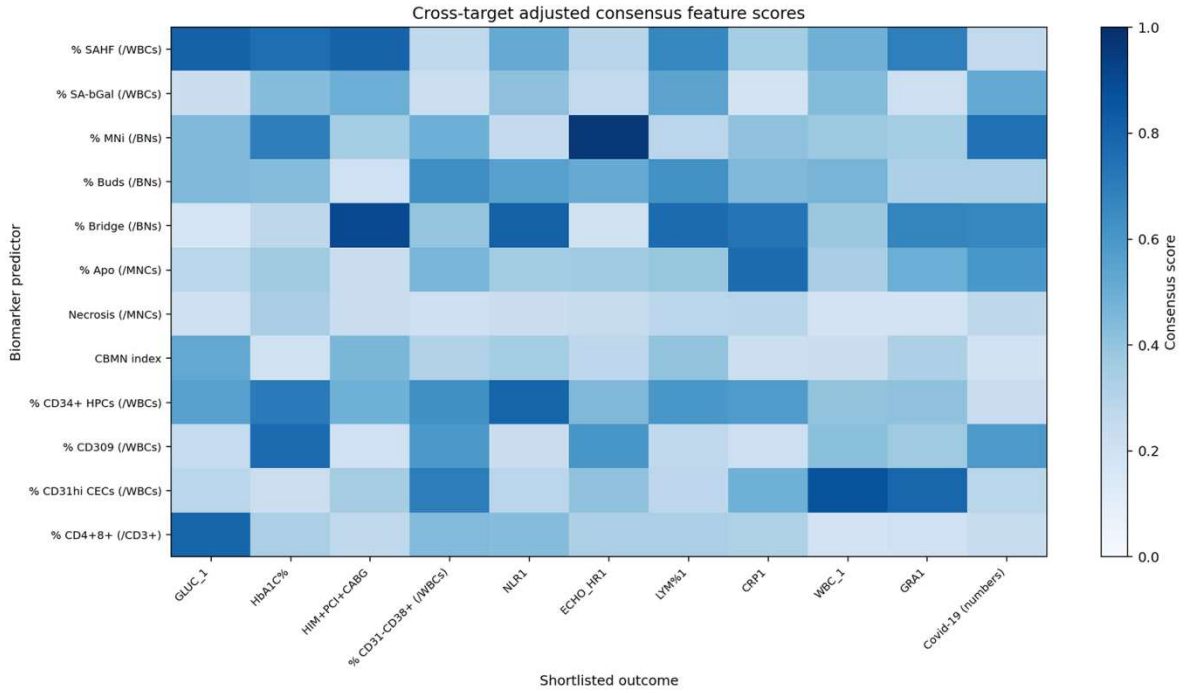


Figure 2. Group+SEX-adjusted consensus feature scores across the 11 shortlisted outcomes.

Notes: Apo – apoptotic cells, BNs – binuclear cells, CBMN – cytokinesis-block micronuclei, CRP – C-reactive protein level, ECHO_HR – heart rate, HbA1C – glycated hemoglobin percentage, HIM+PCI+CABG – history of percutaneous coronary intervention or coronary artery bypass

grafting, GLUC – glucose blood level, GRA – granulocytes absolute count, LYM% – lymphocyte percentage, MNi – micronuclei, NLR – neutrophil-to-lymphocyte ratio, SA-β-gal – senescence-associated β-galactosidase, SAHF – senescence-associated heterochromatin foci, WBC – white blood cells.

Outcome-specific model and SHAP interpretation.

Outcome-specific model performance and SHAP-based feature attribution were evaluated for the 11 outcomes retained after the screening procedure. The five leading cellular biomarkers for each outcome, ranked according to mean absolute SHAP values, together with the direction of the SHAP association, Group+SEX-adjusted partial correlations, and the corresponding clue classification, are summarized in supplementary table [see Additional file 5]. The following sections summarize the principal findings for each outcome, focusing on the leading SHAP features and the concordance between model-derived and adjusted association patterns.

White blood cells count. The best-performing fixed model for absolute WBCs count was a random forest, with a cross-validated R^2 of -0.012 and a median improvement over the Group+SEX-adjusted baseline of 0.029. Although positive improvement was observed in 100% of repeats, the negative R^2 indicates that the model did not provide useful absolute out-of-sample prediction of WBC count. $CD31^{hi}$ circulating endothelial cells had the highest mean absolute SHAP value and ranked first among the model features. Their SHAP association was relatively weak or non-monotonic, whereas the Group+SEX-adjusted partial association was negative (ρ = -0.380, 95% CI -0.531 to -0.202), resulting in a consistent clue classification. $CD34^{+}$ hematopoietic progenitor cells ranked second and showed a negative SHAP direction (ρ = -0.818), with a weaker adjusted partial association (ρ = -0.149, 95% CI -0.381 to 0.100), and were therefore classified as a model-dependent clue. SAHF ranked third, with a positive SHAP direction (ρ = 0.938) and a positive but relatively weak adjusted partial association

(ρ = 0.174, 95% CI -0.033 to 0.355). $CD309^{+}$ endothelial progenitor cells and SA-β-gal also ranked among the five leading features but showed model-dependent patterns.

Neutrophil count. The best-performing fixed model for absolute neutrophil count was a random forest, with a cross-validated R^2 of -0.038 and a median improvement over the Group+SEX-adjusted baseline of 0.052. Although positive improvement was observed in 100% of repeats, the negative R^2 indicates that the model did not provide useful absolute out-of-sample prediction of neutrophil count. The leading held-out SHAP feature was $CD31^{hi}$ CECs, which showed a weak or non-monotonic SHAP relationship and a negative Group+SEX-adjusted partial association (ρ = -0.33). SAHF ranked second and showed a positive SHAP direction, with a positive adjusted partial association (ρ = 0.208). $CD34^{+}$ HPCs ranked third, with higher values associated with lower model predictions and a negative adjusted partial association (ρ = -0.155). Thus, the direction of the model-derived relationships was concordant with the corresponding adjusted associations for the three leading features.

Lymphocyte percentage. The best-performing fixed model for lymphocyte percentage was a random forest, with a cross-validated R^2 of 0.003 and a median improvement over the Group+SEX-adjusted baseline of 0.048. Although positive improvement was observed in 100% of repeats, the near-zero R^2 indicates very limited absolute out-of-sample predictive performance. The leading held-out SHAP feature was SA-β-gal, with higher values associated with lower model predictions and a negative Group+SEX-adjusted partial association (ρ = -0.181). Neoplasmic bridges ranked second, with higher values associated with higher model predictions and a positive adjusted partial association (ρ = 0.191). SAHF ranked third, showing a negative SHAP direction and a negative adjusted partial association (ρ = -0.171). Thus, the direction of the model-derived relationships was concordant with the corresponding Group+SEX-adjusted associations for all three leading features.

Neutrophil-to-Lymphocyte Ratio. The best-performing fixed model for the neutrophil-to-lymphocyte ratio (NLR) was a random forest, with a cross-validated R^2 of -0.022 and a

median improvement over the Group+SEX-adjusted baseline of 0.058. Although positive improvement was observed in 100% of repeats, the negative R^2 indicates that the model did not provide useful absolute out-of-sample prediction of NLR. The leading held-out SHAP feature was $CD34^+$ HPCs, with higher values associated with lower model predictions and a negative Group+SEX-adjusted partial association (ρ = -0.259). SA-β-gal ranked second, with higher values associated with higher model predictions and a positive adjusted partial association (ρ = 0.16). Nucleoplasmic bridges ranked third, with higher values associated with lower model predictions and a negative adjusted partial association (ρ = -0.184). The direction of the model-derived relationships was concordant with the corresponding Group+SEX-adjusted associations for all three leading features.

$CD31^-CD38^+$ cells. The best-performing fixed model for $CD31^-CD38^+$ cells was a random forest, with a cross-validated R^2 of 0.223 and a median improvement over the Group+SEX-adjusted baseline of 0.007. Positive improvement was observed in 60% of repeats. Thus, the model demonstrated modest absolute out-of-sample predictive performance, although its incremental contribution beyond the Group+SEX baseline was limited. The leading held-out SHAP feature was $CD31^{hi}$ CECs, with higher values associated with higher model predictions and a positive Group+SEX-adjusted partial association (ρ = 0.132). $CD309^+$ EPCs ranked second: their SHAP relationship was weak or non-monotonic, with a small positive adjusted partial association (ρ = 0.075). $CD34^+$ HPCs ranked third, with higher values associated with higher model predictions and a positive adjusted partial association (ρ = 0.28). Thus, the direction of the model-derived relationships was concordant with the corresponding adjusted associations for the leading features, although the strength of these associations varied.

Glucose level. The best-performing fixed model for blood glucose level was a random forest, with a cross-validated R^2 of 0.158 and a median improvement over the Group+SEX-adjusted baseline of 0.118. Positive improvement was observed in 100% of repeats, indicating

modest but consistent out-of-sample predictive performance. The leading held-out SHAP feature was SAHF-positive cells, with higher values associated with lower model predictions and a negative Group+SEX-adjusted partial association (ρ = -0.337). $CD34^+$ HPCs ranked second, with higher values associated with higher model predictions and a positive adjusted partial association (ρ = 0.171). $CD4^+CD8^+$ double-positive T lymphocytes ranked third, with higher values associated with lower model predictions and a negative adjusted partial association (ρ = -0.404). The direction of the model-derived relationships was concordant with the corresponding Group+SEX-adjusted associations for all three leading features (Fig.3).

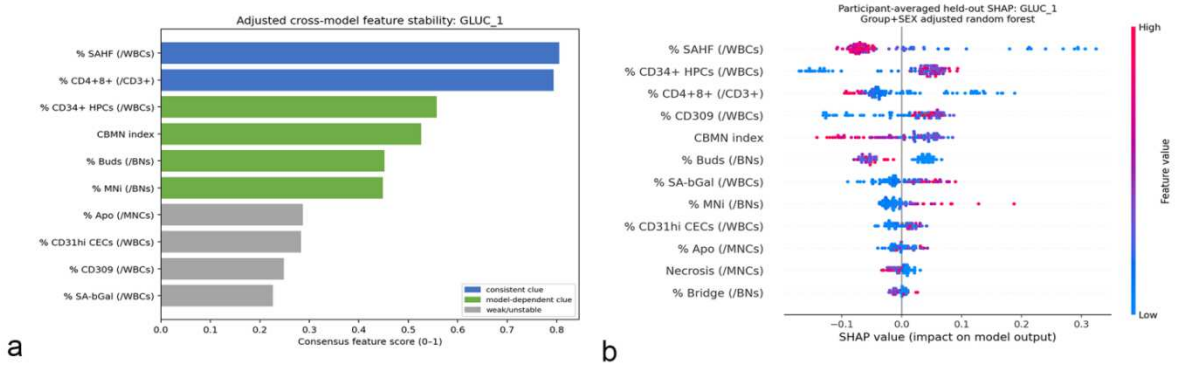


Figure 3. Cross-model feature stability (a) and participant-averaged adjusted held-out SHAP distribution (b) for glucose level.

Notes: for abbreviations, see notes to Fig. 2

Glycated hemoglobin (HbA1C). The best-performing fixed model for glycated hemoglobin (HbA1c) was an extra-trees model, with a cross-validated R^2 of 0.111 and a median improvement over the Group+SEX-adjusted baseline of 0.091. Positive improvement was observed in 100% of repeats, indicating modest but consistent out-of-sample predictive performance. The leading held-out SHAP feature was SAHF, with higher values associated with lower model predictions and a negative Group+SEX-adjusted partial association (ρ = −0.297). Micronuclei ranked second, with higher values associated with higher model predictions and a positive adjusted partial association (ρ = 0.174). $CD34^+$ HPCs ranked third, also showing a positive SHAP direction and a positive adjusted partial association (ρ = 0.171).

The direction of the model-derived relationships was concordant with the corresponding Group+SEX-adjusted associations for all three leading features (Fig. 4).

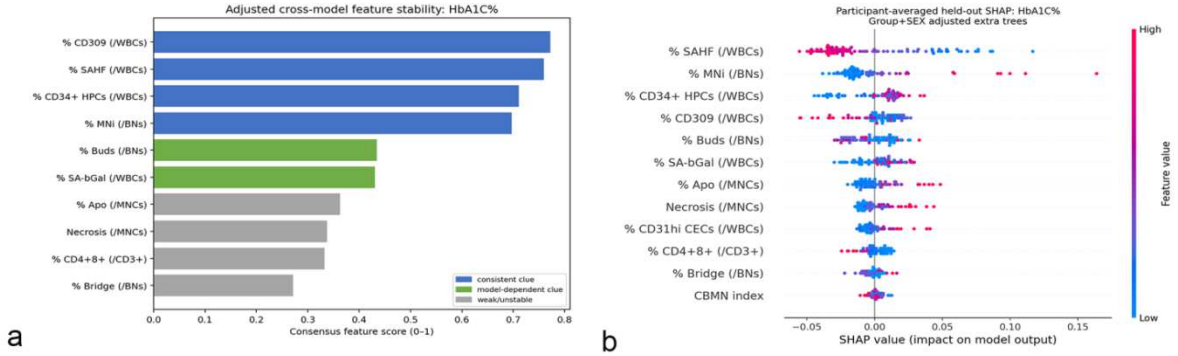


Figure 4. Cross-model feature stability (a) and participant-averaged adjusted held-out SHAP distribution (b) for HbA1C.

Notes: for abbreviations, see notes to Fig. 2

C-reactive protein. The best-performing fixed model for CRP level was an extra-trees model, with a cross-validated R^2 of -0.070 and a median improvement over the Group+SEX-adjusted baseline of 0.006. Positive improvement was observed in only 60% of repeats. The negative R^2 and limited improvement over baseline indicate that the model did not provide useful absolute out-of-sample prediction of CRP levels. The leading held-out SHAP feature was nucleoplasmic bridges, with higher values associated with lower model predictions and a negative Group+SEX-adjusted partial association (ρ = -0.086). Apoptotic cells ranked second, also showing a negative SHAP direction and a negative adjusted partial association (ρ = -0.149). Nuclear buds ranked third, with higher values associated with lower model predictions and a negative adjusted partial association (ρ = -0.070). Thus, the direction of the model-derived relationships was concordant with the corresponding Group+SEX-adjusted associations for all three leading features (Fig. 5).

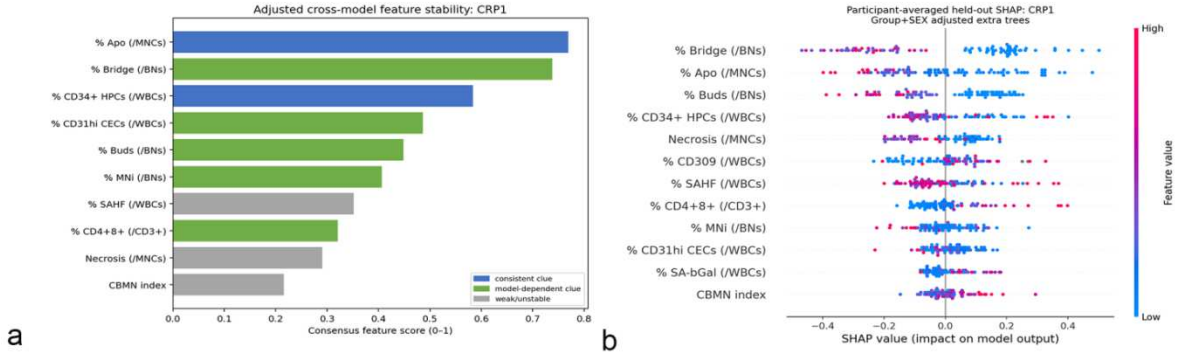


Figure 5. Cross-model feature stability (a) and participant-averaged adjusted held-out SHAP distribution (b) for C-reactive protein.

Notes: for abbreviations, see notes to Fig. 2

Heart rate. The best-performing fixed model for heart rate was a random forest, with a cross-validated R^2 of 0.078 and a median improvement over the Group+SEX-adjusted baseline of 0.096. Positive improvement was observed in 100% of repeats. Although the absolute R^2 remained below the prespecified threshold for non-trivial predictive performance, the consistent improvement over baseline suggests a reproducible but limited predictive signal. The leading held-out SHAP feature was micronuclei, with higher values associated with higher model predictions and a positive Group+SEX-adjusted partial association (ρ = 0.243). $CD34^+$ HPCs ranked second, with higher values associated with lower model predictions and a negative adjusted partial association (ρ = -0.175). $CD309^+$ EPCs ranked third; their SHAP relationship was weak or non-monotonic, with a positive adjusted partial association (ρ = 0.118). The direction of the model-derived relationships was concordant with the corresponding adjusted associations for the two leading features (Fig. 6).

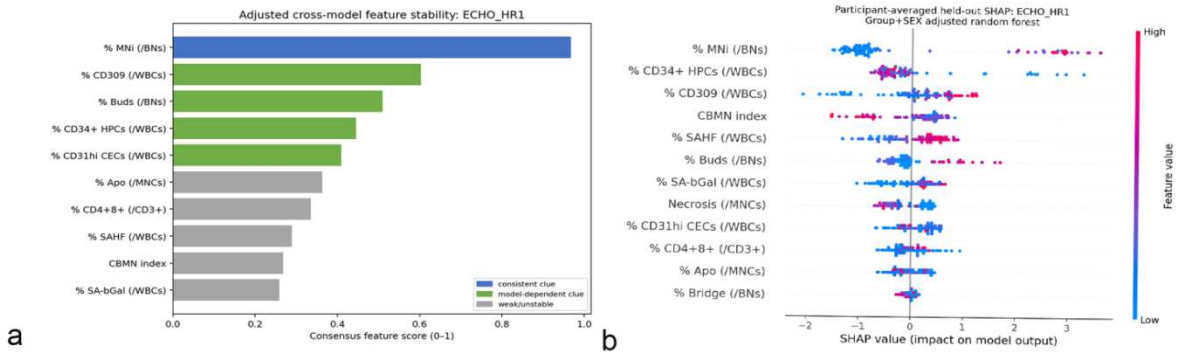


Figure 6. Cross-model feature stability (a) and participant-averaged adjusted held-out SHAP distribution (b) for heart rate.

Notes: for abbreviations, see notes to Fig. 2

History of AMI+PCI+CABG. The best-performing fixed model for the history of AMI+PCI+CABG was a random forest, with a ROC-AUC of 0.628 and a median improvement over the Group+SEX-adjusted baseline of 0.161. Positive improvement was observed in 100% of repeats, indicating weak-to-modest discrimination. The leading held-out SHAP feature was nucleoplasmic bridges, with higher values associated with higher model predictions and a positive Group+SEX-adjusted partial association (ρ = 0.266). SAHF ranked second, also showing a positive SHAP direction and a positive adjusted partial association (ρ = 0.112). The cytokinesis-block proliferation index ranked third, with higher values associated with higher model predictions and a positive adjusted partial association (ρ = 0.118). The direction of the model-derived relationships was concordant with the corresponding adjusted associations for all three leading features (Fig. 7).

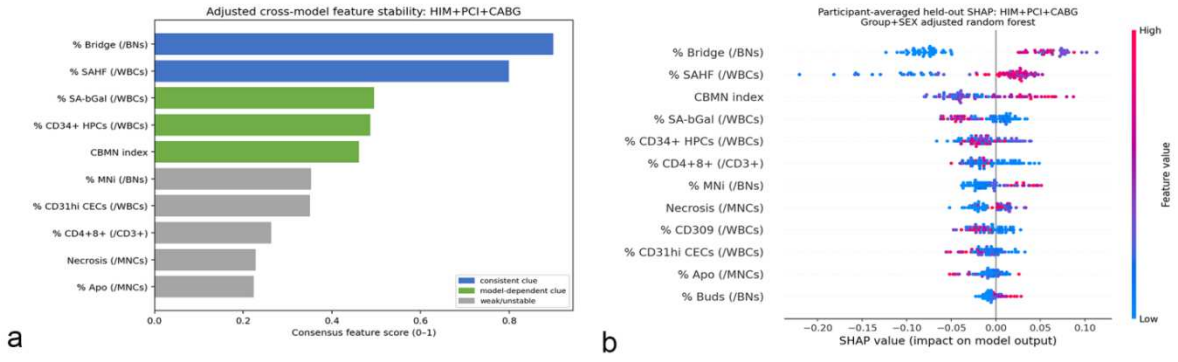


Figure 7. Cross-model feature stability (a) and participant-averaged adjusted held-out SHAP distribution (b) for AMI+PCI+CABG history.

Notes: for abbreviations, see notes to Fig. 2

History of COVID-19. The best-performing fixed model for history of COVID-19 was an extra-trees model, with a ROC-AUC of 0.676 and a median improvement over the Group+SEX-adjusted baseline of 0.047. Positive improvement was observed in 100% of repeats, indicating

weak-to-modest discrimination. The leading held-out SHAP feature was micronuclei, with higher values associated with higher model predictions and a positive Group+SEX-adjusted partial association (ρ = 0.262). Nucleoplasmic bridges ranked second, also showing a positive SHAP direction and a positive adjusted partial association (ρ = 0.202). SA-β-gal-positive cells ranked third, with higher values associated with higher model predictions and a positive adjusted partial association (ρ = 0.212). The direction of the model-derived relationships was concordant with the corresponding adjusted associations for all three leading features (Fig.8).

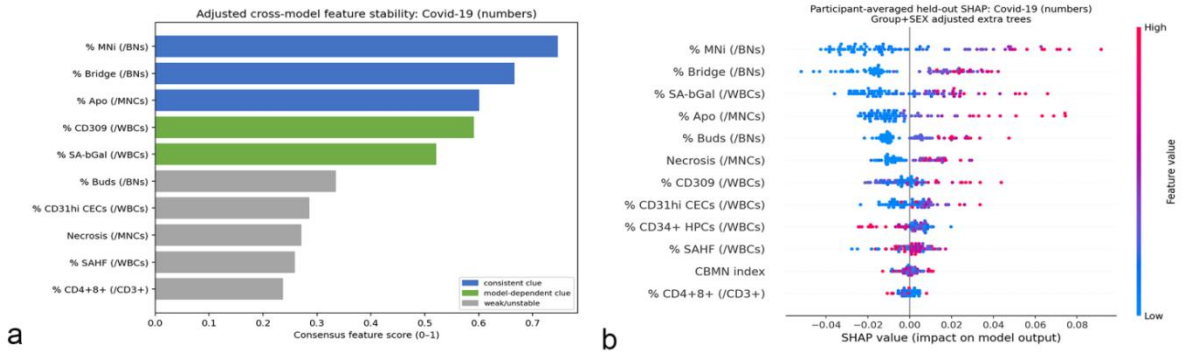


Figure 8. Cross-model feature stability (a) and participant-averaged adjusted held-out SHAP distribution (b) for COVID-19 history.

Notes: for abbreviations, see notes to Fig. 2

Agreement with prior analyses. The intermediate results generally yielded lower performance estimates than the original AutoML analysis and were consistent with the stricter nested-CV analysis, in which no pooled outcome met the prespecified confirmatory FDR criterion (Table 2). Nevertheless, the repeated held-out feature rankings identified a limited set of biomarkers that showed relatively stable contributions across resampling and adjustment settings. These recurring patterns provide a basis for prioritizing mechanistic hypotheses for subsequent validation in larger and independent cohorts.

Table 2. Predictive performance of the AutoML models across study outcomes.

Outcome	Original AutoML R^2	Strict metric	Intermediate metric	Intermediate Δ
WBCs	0.149	-0.041	-0.012	0.029
Neutrophils	0.118	-0.039	-0.038	0.052
Lymphocytes	0.623	0.062	0.003	0.048
NLR	0.116	0.051	-0.022	0.058
$CD31^{-}CD38^{+}$	0.257	0.175	0.223	0.007
Glucose	0.147	0.178	0.158	0.118
HbA1C	0.090	0.074	0.111	0.091
CRP	0.621	-0.078	-0.070	0.006
Heart rate	0.199	0.049	0.078	0.096
AMI+PCI+CABG	0.089	0.649	0.628	0.161
COVID-19	0.041	0.648	0.676	0.047

Notes: CABG – Coronary Artery Bypass Grafting, CRP – C-reactive protein, HbA1c – glycated hemoglobin, AMI – Acute Myocardial Infarction, NLR – Neutrophil-to-Lymphocyte Ratio, PCI – Percutaneous Coronary Intervention, WBCs – white blood cells.

Discussion

The present study addressed a methodological and biological gap related to the systemic effects of SARS-CoV-2 infection in high-risk cardiovascular patients. Rather than considering COVID-19 as a direct determinant of clinical outcomes, we examined whether a compact panel of circulating cellular biomarkers could identify reproducible relationships with clinical, biochemical, and cytological characteristics. This approach is relevant because long-COVID may affect multiple biological systems simultaneously, whereas conventional

clinical measures capture only selected components of this response [26, 27, 28]. These findings support the investigation of cellular biomarkers that may integrate vascular, immune, regenerative, and senescence-related responses in high-risk cardiovascular populations.

The senescence-related predictors, SA-β-gal and SAHF, showed outcome-specific rather than generalized predictive value. SA-β-gal was among the leading features for lymphocyte percentage, NLR, and COVID-19 history, whereas SAHF contributed prominently to WBC count, neutrophils, lymphocytes, glucose, HbA1c, and history of AMI+PCI+CABG. The repeated appearance of these markers across different outcomes suggests that the senescence-related component of the panel contained information that was reused by the models, but the corresponding absolute model performance differed markedly between outcomes.

The strongest evidence for SAHF-related predictive information was observed for glucose and HbA1c. SAHF ranked first for glucose and HbA1c and showed concordant negative adjusted partial correlations. Importantly, these associations were accompanied by non-trivial cross-validated performance: $R^2 = 0.158$ for glucose and $R^2 = 0.111$ for HbA1c. The fact that the same senescence-associated feature contributed to two related measures of glucose metabolism strengthens the reproducibility of this signal, although it does not establish a mechanistic relationship between SAHF and glycemic control.

The biological plausibility of this observation is supported by evidence that cellular senescence is associated with metabolic remodeling and that systemic stress can influence biological-aging pathways. Chronic catecholaminergic stimulation of human peripheral blood mononuclear cells has been shown experimentally to induce DNA damage, telomere shortening, increased SA-β-gal activity, and a senescence-like phenotype [29]. However, these experimental findings do not demonstrate that SAHF or SA-β-gal directly determines glucose or HbA1c levels. The present results should therefore be interpreted as evidence that

senescence-associated cellular information may contribute to the prediction of selected metabolic phenotypes.

SA-β-gal was also the leading feature for lymphocyte percentage and the second-ranked feature for NLR, but these models had R^2 values of 0.003 and -0.022, respectively. Thus, although the marker was repeatedly selected by the models, its contribution did not translate into useful absolute prediction. This distinction illustrates why feature rank alone should not be interpreted as evidence of clinical utility. Similarly, SA-β-gal ranked third in the COVID-19-history classifier, where the model achieved a ROC-AUC of 0.676. The association is compatible with the established experimental relationship between SARS-CoV-2 and cellular senescence, but the modest discrimination prevents interpretation as a clinically useful marker of previous COVID-19.[6]

SAHF also contributed to the recurrent ischemic events (AMI+PCI+CABG-history), with a low-to-moderate ROC-AUC of 0.628. Because this outcome was derived from clinical history, the observed contribution of SAHF should not be interpreted as evidence that chromatin remodeling identifies previous coronary events independently of conventional clinical characteristics. Instead, it represents an exploratory cellular correlate of a pre-existing clinical phenotype.

The genotoxicity-related component showed a similarly selective pattern. Micronuclei were among the leading features for HbA1c, heart rate, and COVID-19 history, whereas nucleoplasmic bridges contributed to several inflammatory and clinical-history outcomes. These findings are biologically plausible because micronuclei, nuclear buds, and nucleoplasmic bridges represent different manifestations of chromosome instability and abnormal nuclear material segregation. Experimental SARS-CoV-2-induced syncytia have been shown to generate micronuclei accompanied by DNA-damage responses and cGAS–STING activation, while preliminary cytogenetic studies in patients with COVID-19 have reported increased micronuclei and other nuclear abnormalities [14, 30].

In the present dataset, however, the strongest interpretation of these markers depends on the outcome. Micronuclei ranked first for heart rate, but the overall model explained only 7.8% of out-of-sample variance. The same marker ranked second for HbA1c, and first for COVID-19 history. Thus, the recurrent contribution of micronuclei across metabolically, physiologically, and infection-related outcomes suggests that genomic-instability information may have broader systemic correlates, but the magnitude of prediction remained modest.

Nucleoplasmic bridges were the leading feature for AMI+PCI+CABG history and CRP, whereas ranked second for COVID-19 history the third-ranked feature for NLR. The CRP model, however, had an R^2 of -0.070, indicating that the apparent feature contribution did not provide useful absolute prediction. The COVID-19 classifier performed somewhat better, but its ROC-AUC of 0.676 remained insufficient for clinical discrimination. These results illustrate an important distinction between biological plausibility and predictive usefulness: cytogenetic abnormalities may be associated with systemic inflammatory or infectious processes without providing sufficient information to predict an individual outcome.

The convergence of micronuclei, nuclear buds, and nucleoplasmic bridges with several outcomes supports their consideration as hypothesis-generating indicators of systemic cellular stress. Nevertheless, because no outcome associated with these markers met the strict confirmatory criteria in the pooled analysis, these findings should not be interpreted as evidence of a validated genotoxicity-based prediction signature.

The hematopoietic progenitor and endothelial markers provided the most consistent evidence that different cellular compartments may contribute complementary information. $CD34^+$ HPCs repeatedly appeared among the leading predictors for WBC count, neutrophils, NLR, $CD31^-CD38^+$ cells, glucose, HbA1c, and heart rate. This recurrent contribution is consistent with previous cardiovascular studies showing that lower circulating $CD34^+$ progenitor-cell levels are associated with adverse cardiovascular outcomes and mortality [15, 31]. Importantly, however, those studies examined prognostic associations with

cardiovascular events, whereas the present analysis addressed prediction of cross-sectional laboratory and clinical characteristics. The biological relevance of $CD34^+$ cells therefore supports the plausibility of our findings but does not constitute independent validation of the specific outcome relationships observed here.

The most convincing $CD34^+$-related signals occurred for glucose and HbA1c, where $CD34^+$ HPCs ranked second and third, respectively. For glucose and HbA1c, the adjusted partial correlations were positive ($\rho = 0.171$). The recurrence of $CD34^+$ HPCs in both glycemic outcomes is therefore more informative than its isolated appearance in the weaker WBC, neutrophil, NLR, or heart-rate models. Nevertheless, the modest effect sizes indicate that $CD34^+$ HPCs represent only one component of the information underlying these metabolic outcomes.

$CD34^+$ HPCs were also the leading feature for NLR and a leading feature for WBC and neutrophil counts, but all three models had negative or near-zero R^2 values. Thus, the repeated selection of $CD34^+$ HPCs does not mean that this marker can be used to predict routine leukocyte parameters. Rather, it indicates that the cellular progenitor compartment was repeatedly informative to the fitted models relative to the available predictors, even when the total information content was insufficient for useful out-of-sample prediction.

$CD31^{hi}$ CECs ranked first for WBC count and neutrophils and first for $CD31^-CD38^+$ cells, while $CD309^+$ EPCs ranked second for $CD31^-CD38^+$ cells and third for heart rate. The strongest absolute performance involving these markers was observed for $CD31^-CD38^+$ cells. The simultaneous contribution of $CD31^{hi}$ CECs, $CD309^+$ EPCs, and $CD34^+$ HPCs to this outcome is biologically coherent because these populations occupy related hematopoietic and vascular cellular domains. Previous cardiovascular studies have demonstrated prognostic associations of circulating $CD34^+/VEGFR2^+$ progenitor populations with cardiovascular events and mortality [32, 33].

$CD4^+CD8^+$ double-positive T cells contributed less frequently than the senescence, genotoxicity, and progenitor markers, but their appearance in the glucose model is noteworthy. They ranked third for glucose, with a negative adjusted partial correlation, while the glucose model showed $R^2 = 0.158$ and positive improvement in all repeats. This makes the DP T cell signal potentially relevant rather than merely a high-ranking feature in an otherwise uninformative model. The biological interpretation should nevertheless remain cautious. Circulating $CD4^+CD8^+$ cells are heterogeneous, encompassing phenotypically and functionally distinct populations, as demonstrated by single-cell transcriptomic analysis [17]. Their contribution to the glucose model may therefore represent an aspect of immune remodeling not captured by total lymphocyte counts or conventional $CD4^+/CD8^+$ measurements. Whether this reflects metabolic-immune coupling, systemic stress, or another unmeasured factor cannot be determined from the present data. The finding is therefore best considered a hypothesis-generating signal for independent investigation.

Several limitations of our study should be considered. First, the sample comprised only 90 participants, which limits statistical power and increases uncertainty in model performance estimates. Cross-fitting reduces but does not eliminate the risk of instability in small datasets. Second, the study was observational and does not permit causal inference. SHAP values describe the contribution of a feature to a fitted prediction model; they do not establish biological causation, mediation, or directionality in the underlying disease process. Concordance between SHAP direction and adjusted rank association strengthens reproducibility of the observed pattern but does not convert an association into a causal relationship.

Consequently, temporal effects, changes in medical care, treatment patterns, and other cohort-specific factors may contribute to between-group differences. Medication use, comorbidities, disease severity, lifestyle factors, time since ACS, and other clinical characteristics may influence both cellular biomarkers and laboratory outcomes. Adjustment

for Group and sex reduces only part of this potential confounding. Although cross-fitting and subsequent stricter analyses were used to reduce optimism, outcome selection remains a potential source of multiplicity and selection bias. Finally, the study is cross-sectional and does not support causal or prognostic claims.

The principal theoretical contribution of this study is the demonstration that cellular biomarkers representing distinct biological dimensions can be evaluated as a coordinated predictor system rather than as isolated markers. The results suggest that cellular senescence, genomic instability, progenitor-cell status, endothelial turnover, and immune remodeling may contain partially overlapping but non-redundant information about selected systemic phenotypes. The most compelling pattern was the convergence of SAHF, $CD34^+$ HPCs, and $CD4^+CD8^+$ DP T cells in the glucose model and of SAHF, micronuclei, and $CD34^+$ HPCs in the HbA1c model. Likewise, the repeated contribution of micronuclei and nuclear abnormalities to COVID-19 history is consistent with experimental evidence linking SARS-CoV-2 to genomic stress, but the modest classifier performance indicates that these features are not sufficient to reconstruct exposure history reliably.

From a practical perspective, the study provides a framework for prioritizing biomarkers for subsequent validation rather than proposing an immediately deployable clinical prediction tool. The combination of a predefined cellular panel, automated model comparison, cross-fitted performance assessment, and held-out SHAP interpretation can be applied to small multidimensional biomedical datasets in which the number of potential outcomes is large relative to sample size.

Conclusions

The present study demonstrates that a multidimensional circulating-cell panel can provide selective rather than universal predictive information across clinical, biochemical, and cytological outcomes in patients at high cardiovascular risk. Among approximately 50

candidate outcomes, only 11 showed sufficient signal for cross-fitted evaluation, and their performance varied substantially.

The most informative cross-domain findings were observed for glucose and HbA1c, for which senescence-associated chromatin features, $CD34^+$ hematopoietic progenitor cells, and, for glucose, $CD4^+CD8^+$ double-positive T cells repeatedly contributed to held-out predictions. Genotoxicity-related markers, particularly micronuclei and nucleoplasmic bridges, showed recurrent contributions to several outcomes, including COVID-19 history, but their predictive performance remained modest. $CD31^{hi}$ CECs, $CD309^+$ EPCs, and $CD34^+$ HPCs provided the strongest cellular-domain signal for $CD31^-CD38^+$ cells, although this finding may partly reflect measurement-domain proximity.

Thus, the principal contribution of this study is not the establishment of a clinical prediction model, but the identification of a restricted set of reproducible biomarker–outcome relationships from a much larger exploratory search space. Cross-fitted AutoML combined with held-out SHAP interpretation provides a useful framework for prioritizing such relationships for independent validation and mechanistic investigation in larger cohorts exposed to infectious and prolonged psychosocial stress.

List of abbreviations

ACS – acute coronary syndrome;

AMI – acute myocardial infarction;

AutoML – automated machine learning;

CABG – coronary artery bypass grafting;

CBMN – cytokinesis-block micronuclei;

CRP – C-reactive protein;

DRF – distributed random forests;

EPCs – endothelial progenitor cells;

GBM – gradient boosting machines;

GLM – generalized linear models;

HbA1C – glycated hemoglobin;

HPCs – hematopoietic progenitor cells;

NLR – neutrophil-to-lymphocyte ratio;

PCI – percutaneous coronary intervention;

SHAP – Shapley Additive Explanations;

SA-β-gal – senescence-associated β-galactosidase;

SAHF – senescence-associated heterochromatin foci;

WBC – white blood cells.

Declarations

Ethics approval and consent to participate. The study protocol was reviewed and approved by the local Ethics Committee of the Strazhesko National Scientific Center (protocol #1-23 dated September 17, 2024). All participants provided written informed consent for participation in the study and for the use of their clinical and laboratory data for research purposes.

Clinical trial number. Not applicable.

Consent for publication. Not applicable.

Availability of data and material. The primary datasets generated and/or analysed during the current study are not publicly available due to the presence of patients' personal data and the need to protect their confidentiality but are available from the corresponding author on reasonable request, subject to applicable data protection requirements.

Competing interests. The authors declare that they have no competing interests.

Funding. This study was conducted as part of project No. 2023.03/0048 "Development of personalized criteria for assessing high cardiovascular risk based on markers of cellular senescence" with grant support from the National Research Foundation of Ukraine.

Authors' contributions. VK – flow cytometry, conceptualization, writing and original draft submission; DKr – machine learning and analysis; AU – CBMN assay; OT – SAHF analysis; KM – machine learning; OS – patient recruitment, clinical observation and blood sampling; DKh – patient clinical and instrumental observation; OP – supervision, review and editing. All authors read and approved the final manuscript.

Acknowledgements. Not applicable.

References

1. Kite TA, Ludman PF, Gale CP, Wu J, Caixeta A, Mansourati J, et al. International Prospective Registry of Acute Coronary Syndromes in Patients With COVID-19. J Am Coll Cardiol. 2021;77(20):2466-2476. doi:10.1016/j.jacc.2021.03.309
2. Rashid M, Wu J, Timmis A, Curzen N, Clarke S, Zaman A, et al. Outcomes of COVID-19-positive acute coronary syndrome patients: A multisource electronic healthcare records study from England. J Intern Med. 2021;290(1):88-100. doi:10.1111/joim.13246
3. Alquézar-Arbé A, Miró Ò, Castillo JGD, Jiménez S, Llorens P, Martín A, et al. Incidence, Clinical Characteristics, Risk Factors and Outcomes of Acute Coronary Syndrome in Patients With COVID-19: Results of the UMC-19-S1010. J Emerg Med. 2022;62(4):443-454. doi:10.1016/j.jemermed.2021.10.046
4. Lindberg P, Wiqvist S, Juszczyk M, Lee S, Kisiel MA, Wachtler C, et al. Long COVID and risk of incident cardiovascular disease: a prospective cohort study using the Multimorbidity Integrated Registry Across Care Levels in Stockholm (MIRACLE-S) cohort. EClinicalMedicine. 2026;94:103846. doi:10.1016/j.eclinm.2026.103846

5. Ajoolabady A, Pratico D, Bahijri S, Tuomilehto J, Uversky VN, Ren J. Hallmarks of cellular senescence: biology, mechanisms, regulations. Exp Mol Med. 2025;57(7):1482-1491. doi:10.1038/s12276-025-01480-7
6. Schmitt CA, Tchkonia T, Niedernhofer LJ, Robbins PD, Kirkland JL, Lee S. COVID-19 and cellular senescence. Nat Rev Immunol. 2023;23(4):251-263. doi:10.1038/s41577-022-00785-2
7. Lee S, Yu Y, Trimpert J, Benthani F, Mairhofer M, Richter-Pechanska P, et al. Virus-induced senescence is a driver and therapeutic target in COVID-19. Nature. 2021;599(7884):283-289. doi: doi:10.1038/s41586-021-03995-1. Epub 2021 Sep 13. Erratum in: Nature. 2025;638(8052):E46. doi:10.1038/s41586-025-08718-4
8. Polsky LR, Rentscher KE, Carroll JE. Stress-induced biological aging: A review and guide for research priorities. Brain Behav Immun. 2022;104:97-109. doi:10.1016/j.bbi.2022.05.016
9. de Mera-Rodríguez JA, Álvarez-Hernán G, Gañán Y, Martín-Partido G, Rodríguez-León J, Francisco-Morcillo J. Is Senescence-Associated β-Galactosidase a Reliable in vivo Marker of Cellular Senescence During Embryonic Development? Front Cell Dev Biol. 2021;9:623175. doi:10.3389/fcell.2021.623175
10. Ashraf HM, Fernandez B, Spencer SL. The intensities of canonical senescence biomarkers integrate the duration of cell-cycle withdrawal. Nat Commun. 2023;14(1):4527. doi:10.1038/s41467-023-40132-0
11. Sun L, Yu R, Dang W. Chromatin Architectural Changes during Cellular Senescence and Aging. Genes (Basel). 2018;9(4):211. doi: doi:10.3390/genes9040211
12. Zhang X, Liu X, Du Z, Wei L, Fang H, Dong Q, et al. The loss of heterochromatin is associated with multiscale three-dimensional genome reorganization and aberrant transcription during cellular senescence. Genome Res. 2021;31(7):1121-1135. doi:10.1101/gr.275235.121

13. Fenech M. Cytokinesis-block micronucleus cytome assay. Nat Protoc. 2007;2(5):1084-104. doi:10.1038/nprot.2007.77
14. Pinto TG, Alpire MES, Ribeiro DA. Cytogenetic Biomonitoring in Buccal Mucosa Cells of COVID-19 Patients: Preliminary Findings. In Vivo. 2021;35(6):3495-3499. doi:10.21873/invivo.12651
15. Dhindsa DS, Desai SR, Jin Q, Sandesara PB, Mehta A, Liu C, et al. Circulating progenitor cells and outcomes in patients with coronary artery disease. Int J Cardiol. 2023;373:7-16. doi:10.1016/j.ijcard.2022.11.047
16. Edsfeldt A, Osterlund J, Sun J, Pan M, Tengryd C, Nitulescu M, et al. Circulating CD31 reflects endothelial activity and is associated with a lower risk for cardiovascular complications. European Heart Journal. 2022; 43 (Suppl. 2):ehac544.3054. doi:10.1093/eurheartj/ehac544.3054
17. Choi SM, Park HJ, Choi EA, Jung KC, Lee JI. Cellular heterogeneity of circulating $CD4^+CD8^+$ double-positive T cells characterized by single-cell RNA sequencing. Sci Rep. 2021;11(1):23607. doi:10.1038/s41598-021-03013-4
18. Ikemura K, Bellin E, Yagi Y, Billett H, Saada M, Simone K, et al. Using Automated Machine Learning to Predict the Mortality of Patients With COVID-19: Prediction Model Development Study. J Med Internet Res. 2021 Feb 26;23(2):e23458. doi:10.2196/23458
19. Parkhomenko O, Skarzhevskyi O, Kyryk V, Shumakov O, Khomiakov D. The Impact of Concomitant COVID-19 in patients with acute myocardial infarction on endothelial function, markers of systemic inflammation, and blood Cell subpopulations in the long-term follow-up period. Ukrainian Journal of Cardiology. 2025; 32(6):7-18. doi:10.31928/2664-4479-2025.6.718
20. Kyryk V, Ustymenko A, Tsupykov O, Skarzhevskyi O, Parkhomenko O. Analysis of Cytological Markers of Cellular Senescence in High-Risk Cardiovascular Patients After

797 COVID-19. BioMed Research International. 2026; 5553698: 11 pages.
798 doi:10.1155/bmri/5553698
799 21. Tsupykov O, Kyryk V, Ustymenko A, Skarzhevskyi O, Khomyakov D, Parkhomenko O.
800 Optimized protocol for identifying senescence-associated heterochromatin foci (SAHF) as
801 markers of cellular senescence in blood samples. Cell Organ Transpl. 2025;
802 13(2):e2025132183. doi:10.22494/cot.v13-2.183
803 22. Ustymenko A, Kyryk V, Tsupykov O, Parkhomenko O. Prospects for cytokinesis-block
804 micronucleus cytome assay in the comprehensive evaluation of cellular senescence
805 markers. Cell Organ Transpl. 2024; 12(2):126-135 (e2024122172).
806 doi:10.22494/cot.v12i2.172
807 23. H2O AutoML (version 3.46.0.9). https://h2o.ai/platform/h2o-automl. Accessed 10
808 October 2026.
809 24. A Romero RA, Y Deypalan MN, Mehrotra S, Jungao JT, Sheils NE, Manduchi E,et al.
810 Benchmarking AutoML frameworks for disease prediction using medical claims. BioData
811 Min. 2022 Jul 26;15(1):15. doi:10.1186/s13040-022-00300-2
812 25. Chen H, Lundberg SM, Lee SI. Explaining a series of models by propagating Shapley values.
813 Nat Commun. 2022;13(1):4512. doi:10.1038/s41467-022-31384-3
814 26. DePace NL, Colombo J. Long-COVID Syndrome and the Cardiovascular System: A Review
815 of Neurocardiologic Effects on Multiple Systems. Curr Cardiol Rep. 2022;24(11):1711-
816 1726. doi:10.1007/s11886-022-01786-2
817 27. Skevaki C, Moschopoulos CD, Fragkou PC, Grote K, Schieffer E, Schieffer B. Long COVID:
818 Pathophysiology, current concepts, and future directions. J Allergy Clin Immunol.
819 2025;155(4):1059-1070. doi:10.1016/j.jaci.2024.12.1074
820 28. Gomes Z, Tulloch H, Bouchard K, Mulvagh SL. The Bidirectionality of Psychosocial Stress
821 and Cardiovascular Disease Risk in Women; Recognition and Strategies for Management.
822 Curr Atheroscler Rep. 2025;27(1):87. doi:10.1007/s11883-025-01338-3

29. Laumann M, Palombo P, Fieres J, Thomas M, Saretzki G, Bürkle A, et al. Senescence-like phenotype after chronic exposure to isoproterenol in primary quiescent immune cells. Biomolecules. 2024;14(12):1528. doi:10.3390/biom14121528
30. Ren H, Ma C, Peng H, Zhang B, Zhou L, Su Y, et al. Micronucleus production, activation of DNA damage response and cGAS-STING signaling in syncytia induced by SARS-CoV-2 infection. Biol Direct. 2021;16(1):20. doi:10.1186/s13062-021-00305-7
31. Muggeridge D, Dodd J, Ross MD. $CD34^{+}$ progenitors are predictive of mortality and are associated with physical activity in cardiovascular disease patients. Atherosclerosis. 2021;333:108-115. doi:10.1016/j.atherosclerosis.2021.07.004
32. Rigato M, Fadini GP. Circulating Stem/Progenitor Cells as Prognostic Biomarkers in Macro- and Microvascular Disease: A Narrative Review of Prospective Observational Studies. Curr Med Chem. 2018;25(35):4507-4517. doi:10.2174/0929867324666170920154020
33. Tomić I, Zeljko I, Brizić I, Šoljić V, Ivančić I, Tomić M, et al. Decreased Endothelial Progenitor Cells Are Associated with Severe Coronary Artery Disease: Insights from a Clinical Study. J Cardiovasc Dev Dis. 2025;12(4):132. doi:10.3390/jcdd12040132

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile4.docx
- Additionalfile2.docx
- Additionalfile5.docx
- Additionalfile3.docx