Aging and Disease
www.aginganddisease.org
Early access date: February 14, 2026
http://dx.doi.org/10.14336/AD.2025.0828

Original Article

Longevity *vs*. Healthy Longevity: Different Outcomes Underlain by Different Mechanisms

Daria Kashtanova[1], Aleksandra Mamchur[1*], Maria Bruttan[1], Veronika Daniel[1], Mikhail Ivanov[1], Elena Zelenova[1], Irina Dzhumaniiazova[1], Lorena Matkava[1], Mikhail Terekhov[1], Antonina Rumyantzeva[1], Konstantin Grammatikati[1], Sergey Mitrofanov[1], Vladimir Yudin[1], Valentina Maksyutina[1], Ekaterina Maralova[1], Aleksey Ivashechkin[1], Aleksandra Nekrasova[1], Irina Strazhesko[2], Valentin Makarov[1], Anton Keskinov[1], Olga Tkacheva[2], Sergey Yudin[1], Veronika Skvortsova[3]

[1]Federal State Budgetary Institution «Centre for Strategic Planning and Management of Biomedical Health Risks» of the Federal medical and biological agency, Moscow, Russia. [2]Russian Clinical Research Center for Gerontology of Pirogov Russian National Research Medical University of the Ministry of Healthcare of the Russian Federation, Moscow, Russia. [3]The Federal Medical Biological Agency, Moscow, Russia.

[Received July 2, 2025; Revised January 12, 2026; Accepted January 13, 2026]

ABSTRACT: Most genomic studies compare the genomes of long-living adults to those of the general population to identify potential genetic markers of longevity. We propose a refined approach: focusing on the genetic makeup of healthy, long-living adults to detect mechanisms promoting both longer lifespan and improved quality of life. To this end, we analyzed medical and genomic data from 3,703 long-living adults aged ≥90 years and 22,354 individuals aged 18–75 years (total N = 26,057). Using whole-genome sequencing (WGS) and a genome-wide association study (GWAS), we found that variants with significant and negative associations with longevity in the GWAS were located in genes such as *APOE*, *APOC1*, and *CFAP46*, which are implicated in an increased risk of age-related diseases. However, the presence or absence of these variants should not be considered a definitive determinant of longevity or sustained health after the age of 90. We found that healthy longevity was positively associated with variants within the *MYO18B, TBC1D28,* and *LOC105376454* genes. To demonstrate the multifactorial nature of the examined phenotypes, we constructed polygenic score models that accounted for nonlinear interactions among the predictors. Trial registration: Clinical Trials NCT06268132 (for long-living adults). Registered 22 February 2024 (retrospectively registered).

Keywords: Longevity, Healthy longevity, Long-living adults, GWAS, Polygenic scores

INTRODUCTION

As the global population ages, countries are facing increasing healthcare and economic strains. Therefore, it is imperative to identify the factors that promote longevity and successful aging, defined as a longer life span (over 90 years) coupled with healthy physical and cognitive functioning [1]. Longitudinal studies have established that lifestyle choices, such as a nutrient-rich diet, physical activity, and the avoidance of smoking and excessive drinking, significantly increase the likelihood of longevity [2]. However, longevity is a complex phenotype determined not only by these environmental and lifestyle factors but also by genetics [3], with the heritability of age at death estimated at 25–40% [4,5].

More than 100 candidate genes have been associated with longevity [6], including *APOE*, *FOXO1A*, *GF1*, *IGF1*, *INCR*, and *TP53* [7]. Although high-throughput

***Correspondence should be addressed to:** Dr. Aleksandra Mamchur, Federal State Budgetary Institution «Centre for Strategic Planning and Management of Biomedical Health Risks» of the Federal Medical and Biological Agency, Moscow, Russia. Email: amamchur@cspfmba.ru.

sequencing has identified numerous variants linked to this trait, replication efforts have often produced inconsistent results. For instance, a 65-study meta-analysis by Revelas et al. showed that specific polymorphisms, including rs4340 (*ACE*), rs7412 (*APOE*), rs429358 (*APOE*), rs2802292 (*FOXO3A*), and rs1800795 (*IL6*), as well as the Klotho KL-VS haplotype, significantly contributed to longevity [8]. Importantly, variants that fall below the genome-wide significance threshold may also play a role. These complex interactions have led to the suggestion that assessing the non-linear combined genetic effect of multiple genetic variants on the phenotype could be a valuable approach, as demonstrated by Yashin et al. [9].

Despite extensive research, the molecular genetic mechanisms of longevity are not fully understood. Many genes associated with longevity are, in fact, linked to aging-related diseases, suggesting that a low risk for these diseases may increase the likelihood of longevity. In this study, we investigated whether longevity is driven solely by this reduced disease risk or by additional genetic factors. To address the heterogeneity of the longevity phenotype, we applied a previously proposed stratification approach for older adults [10]. This approach enabled us to identify a cohort of most successfully aging individuals and to examine the genetic basis of healthy longevity.

METHODS

The sample of long-living adults

This study included 3,703 long-living adults aged 90 years and older. All participants provided informed consent and underwent a comprehensive geriatric assessment. A detailed description of the participants and study design is provided in our previous publication [11]. The study protocol was approved by the Russian Clinical Research Center for Gerontology of the Pirogov Russian National Research Medical University of the Ministry of Healthcare of the Russian Federation (excerpt from Protocol No. 30 dated December 24, 2019). The trial including long-living adults is registered on Clinical Trials (Identifier: NCT06268132). Even though the study was registered retrospectively, the study protocol was finalized before participant enrollment to ensure transparency and prevent reporting bias.

The representative population sample

A comparison group of 22,354 individuals aged 18–75 was derived from a database curated by the Federal Medical Biological Agency, containing a randomized, de-identified representative sample of the Russian population (n = 75,144) collected as part of an epidemiological study on the prevalence of hereditary genetic variants associated with chronic diseases across 79 regions. Where applicable, diagnoses are indicated. The epidemiological study was approved by the ethics committee of the Centre for Strategic Planning of the FMBA of Russia (excerpt from Protocol No. 5 dated December 28, 2020).

Stratification of the sample of long-living adults

Aging phenotypes were determined for 3,525 long-living adults. To identify the healthiest individuals (n = 378), we applied a machine learning-based aging phenotype calculator developed in our previous study [10]. The calculator—a multiclass support vector machine (SVM) classifier—was trained by first performing a cluster analysis using 15 parameters. These included cardiovascular diseases (CVDs), diabetes mellitus (DM), chronic obstructive pulmonary disease (COPD), cancer, polypharmacy, body mass index (BMI), right-hand grip strength, and scores from the Mini-Mental State Examination (MMSE), Frontal Assessment Battery (FAB), Short Physical Performance Battery (SPPB), Activities of Daily Living (ADL), Instrumental Activities of Daily Living (IADL), and the 5-item Geriatric Depression Scale (GDS-5). This analysis identified five distinct aging phenotypes: non-frailty, multimorbid frailty, metabolic frailty, cognitive frailty, and functional frailty. For each phenotype, we determined the underlying diseases and conditions and assessed survival rates. The calculator outputs the probability of an individual belonging to each of the five classes, with the final assigned phenotype being the class with the highest probability. In summary, the healthy (non-frail) aging phenotype was defined by, among other parameters, the absence of frailty and high functional capacity, as indicated by higher scores on the Lawton IADL and Barthel scales.

DNA extraction and sequencing

DNA was extracted from whole blood samples using the QIAamp DNA Mini Kit (*Qiagen*, Germany). Whole-genome sequencing libraries were prepared using the Nextera DNA Flex Kit (*Illumina*, USA). The samples were sequenced to 150 bp reads using the NovaSeq 6000 S4 Reagent Kit (300 cycles) (*Illumina*, USA).

FASTQ files were obtained by demultiplexing the sequencing data in BCL format using bcl2fastq2 Conversion Software v2.20 (*Illumina*) (Retrieved from: bcl2fastq and bcl2fastq2 Conversion Software Downloads. Available at: https://emea.support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software/downloads.html. Accessed 21 Nov 2022). Flow cell-wide sequencing quality was assessed

using Sequencing Analysis Viewer v2.4.7 (*Illumina*) (Retrieved from: Sequencing Analysis Viewer Support. Available at: https://support.illumina.com/sequencing/sequencing_software/sequencing_analysis_viewer_sav.html. Accessed 21 Nov 2022). The quality of the FASTQ files was evaluated using FastQC v0.11.9 (Retrieved from: Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Available at: https://www.bioinformatics.babraham.ac.uk/projects/fastqc. Accessed 21 Nov 2022). The reads were aligned to the reference genome (GRCh38.d1.vd1) (Retrieved from: GDC Reference Files | NCI Genomic Data Commons. Available at: https://gdc.cancer.gov/about-data/gdc-data-processing/gdc-reference-files.

Accessed 21 Nov 2022) using the Illumina DRAGEN Bio-IT Platform (v07.021.510.3.5.7) (Retrieved from: Accurate and efficient calling of small and large variants from popgen datasets using the DRAGEN Bio-IT Platform. Available at: https://www.illumina.com/content/illumina-marketing/amr/en_US/science/genomics-research/articles/popgen-variant-calling-with-dragen.html. Accessed 28 Oct 2024). The alignment quality of the BAM files was checked using DRAGEN FastQC v0.11.9, SAMtools v1.13 [12], and mosdepth v0.3.1 [13]. All samples were checked for duplicates, unmapped reads, and other quality metrics. A detailed description of the filtering process is provided in the Supplementary Materials section. The mean sequencing coverage was 30x for all samples.

Small variant calling of up to 50 bp was performed using Strelka2 v2.9.10 (*Illumina*, USA) [14] with default settings. Identification of duplicate samples (i.e., samples obtained from the same participant in the study) was carried out using the PICARD CrosscheckFingerprints v2.26.11 program (Retrieved from: Picard Toolkit (2020). Available at: https://broadinstitute.github.io/picard/. Accessed 21 Nov 2022). The PICARD built-in map file "hg38_chr.map", as described by Javed et al. [15], was used as a set of haplotypes.

The entire bioinformatic analysis—from raw BCL data to final VCF files—was automated within Docker containers. All software and versions cited in the study are fixed within the corresponding Docker images built for the analysis.

Accounting for population structure and ethnic diversity

To account for population structure and ethnic diversity, a principal component analysis (PCA) was carried out, and the first 10 principal components were used as covariates in the GWAS.

Genome-wide Association Study (GWAS)

Ninety percent of the sample data were used in the GWAS, while the remaining 10% were used in the validation of the polygenic score models.

The GWAS was performed in two variations:

1) Longevity GWAS: the longevity phenotype (case group, n = 3,309) vs. the population sample (control group, n = 20,142). Sex and the first 10 principal components were used as covariates;
2) Healthy longevity GWAS: the healthy longevity phenotype (case group, n = 345) vs. other aging phenotypes (control group, n = 2,827). Sex, age, and the first 10 principal components were used as covariates.

For each GWAS, cross-validation was performed by randomly selecting 7/9 of the sample in nine iterations. Each resulting subsample obtained was studied using the algorithm described below.

The following variants were removed:

1) Variants violating the Hardy–Weinberg equilibrium in the control group (p-value<10^{-6});
2) Variants with an allele frequency (AF) greater than 0.99;
3) Multiallelic variants;
4) Variants with a minor allele frequency (MAF) of less than 0.01.
5) Variants with a proportion of genotyped samples less than 95%

Genome-wide associations were tested using the following logistic regression equation:

$$log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_c \cdot C + \beta_g \cdot G,$$

where:

β_0=constant;

βc=covariate effect vector;

C=covariate vector;

βg=genotype effect vector;

G=genotype vector.

Genotyping (GT) rate, or the percentage of samples successfully genotyped, for each variant was assessed based on the PASS call status (with standard settings), genotyping quality (GQ) ≥ 10, and allele depth (AD) ≥ 10 calculated as the sum of reads supporting each allele at a specific position (sumAD). Variants with GT below 95% were excluded from subsequent analysis. Samples of variants with GT ≥ 95 that did not meet specific quality thresholds were also excluded. This stringent filtering reduced the sample size for specific variants but enhanced the overall quality of the dataset.

All calculations were performed using Statsmodels v0.12.2 (Python) and Spark Cluster parallelization. Variants with p-values less than 5.0×10^{-8} were considered

significant. The results were visualized using the LocusZoom Javascript library.

The results of the nine studies were averaged by calculating the geometric mean for p-values and the arithmetic mean for the regression coefficients. Polymorphisms with an average p-value of less than 5×10^{-8} were considered to have genome-wide significance.

Statistical analysis

Statistical analysis was carried out in Python 3.9.12. To compare results between the GWASs, pairwise scatter plots were generated for all variants (SNPs), where the coordinates represent their p-values from the corresponding studies (Fig. 4B).

Polygenic score models

Polygenic score (PGS) models were constructed using polymorphisms that reached a significance threshold below 5×10^{-6} in at least one stage of the cross-validated GWAS. Sex was used as a covariate for both models; age was used as a covariate for the PGS model for healthy longevity.

Before modeling, the dataset was randomly split into training and test sets in a 4:1 ratio and stratified by the target variable. All PGS models were constructed using two machine learning methods: logistic regression (LR) and random forest (RF).

For the LR-based model, features were selected using ANOVA test results and L2 regularization (Python v3.10, scikit-learn v1.6.1 library). The following equation was used for LR:

$$loss = \sum_{i=1}^{n}(-y_i * \log\frac{\exp{(X_i * \beta + \beta_0)}}{1+\exp{(X_i * \beta + \beta_0)}} - (1-y_i) * (1 - \log\frac{\exp(X_i * \beta + \beta_0)}{1+\exp(X_i * \beta + \beta_0)})) + \frac{1}{2}\lambda\sum_{j=1}^{k}\beta_j^2$$

After hyperparameter tuning, the optimal values were λ = 37 and k = 27.

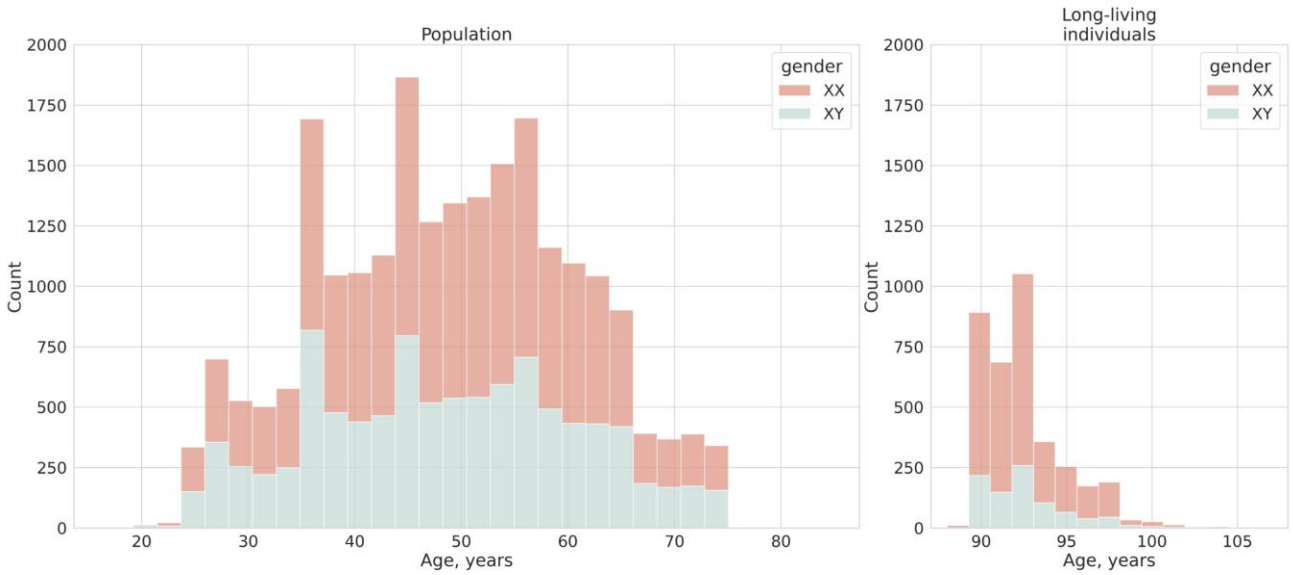

Figure 1. Age and sex distribution in the examined cohorts.

The RF-based models were constructed using RandomForestClassifier (Python v3.10, scikit-learn v1.6.1 library). Hyperparameters were selected via tenfold cross-validation, with stratification by the target variable on the training set across a grid with n_estimators: {100, 150, 200, 230, 250, 270, 300, 350, 400, 450, 1100, 1300, 1500, 2000, 2100, 2300, 2500}, max_depth: {4, …, 128}. The percentage of features to be split ranged from 10% to 100% of the total number of features. The optimal set of hyperparameters was selected by maximizing the F1 score with n_estimators=2100, max_depth=16, and max_features=0.2 for the model for the population vs. long-living adults; n_estimators=350, max_depth=4, and max_features=0.3 for the model for healthy long-living adults vs. non-healthy long-living adults. Node splitting was based on the Gini impurity index. Permutation feature importance was assessed using the built-in tools of scikit-learn v1.6.1.

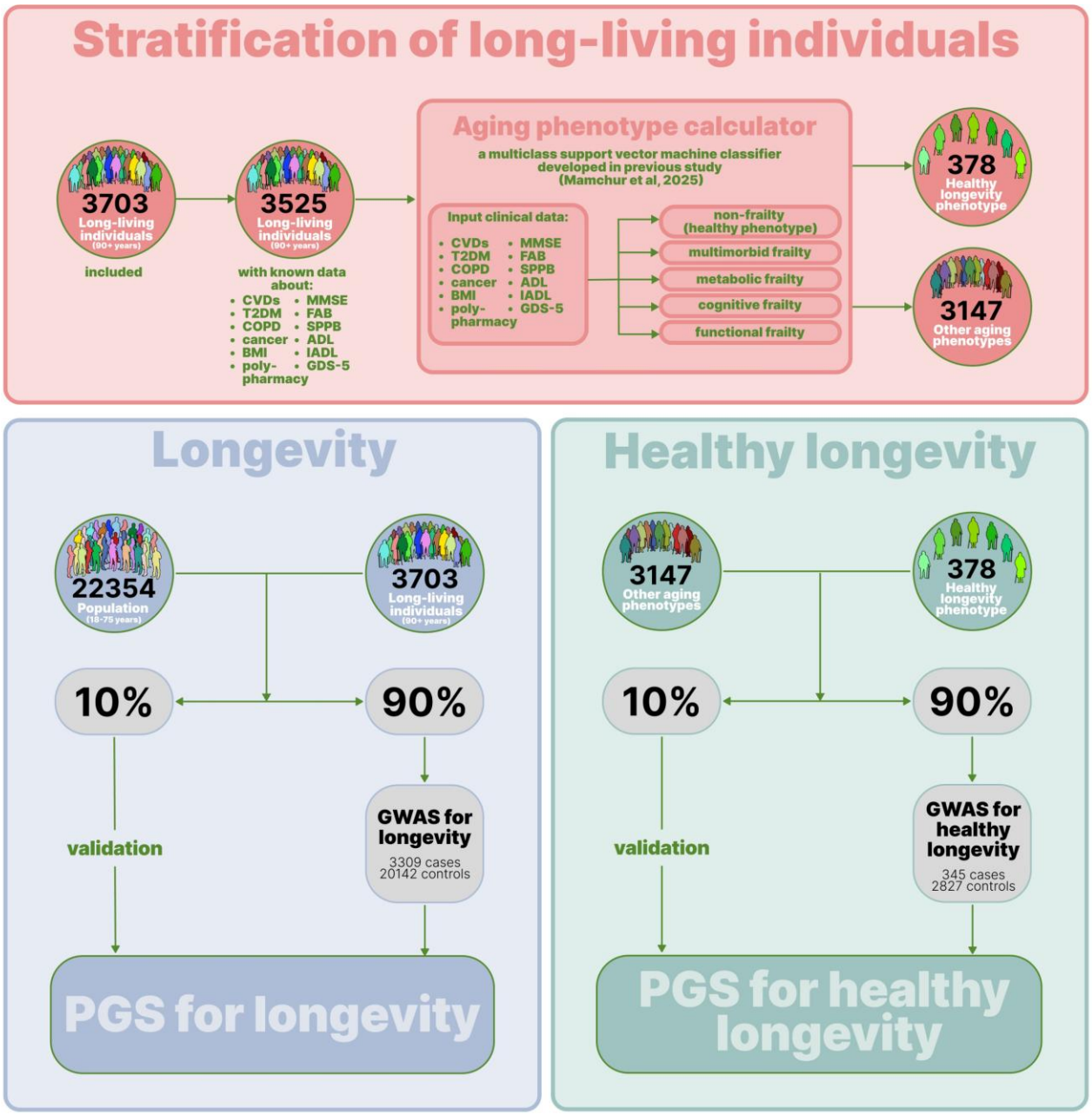

Figure 2. Flow diagram of the study design.

RESULTS

Participants

The study analyzed data from a sample of 26,057 individuals comprising a cohort of 3,703 long-living adults aged ≥90 years and a representative population sample of 22,354 adults aged 18–75 years. Figure 1 shows the age and sex distribution in the cohorts.

Ninety percent of the sample (n = 23,451) was randomly selected for the GWAS for longevity, while the remaining 10% (n = 2,606) was reserved for PGS model validation.

The GWAS for healthy longevity only included data from 3,525 long-living adults, for whom sufficient information was available to determine the aging phenotype. Similarly, data from 90% of the long-living adults were selected for the GWAS (n=3,172), and the remaining 10% were used to validate the PGS models. In addition to sex, age was introduced as a covariate. Figure 2 shows a flow diagram illustrating the study design.

GWAS for longevity

After quality filtration, 9,829,859 variants were tested. After cross-validation, the GWAS results for long-living adults and the population sample revealed a single polymorphism significantly associated with longevity—rs429358 in the *APOE* gene (Table 1; Fig. 3A). This finding confirms the association between this variant and longevity in the Russian population.

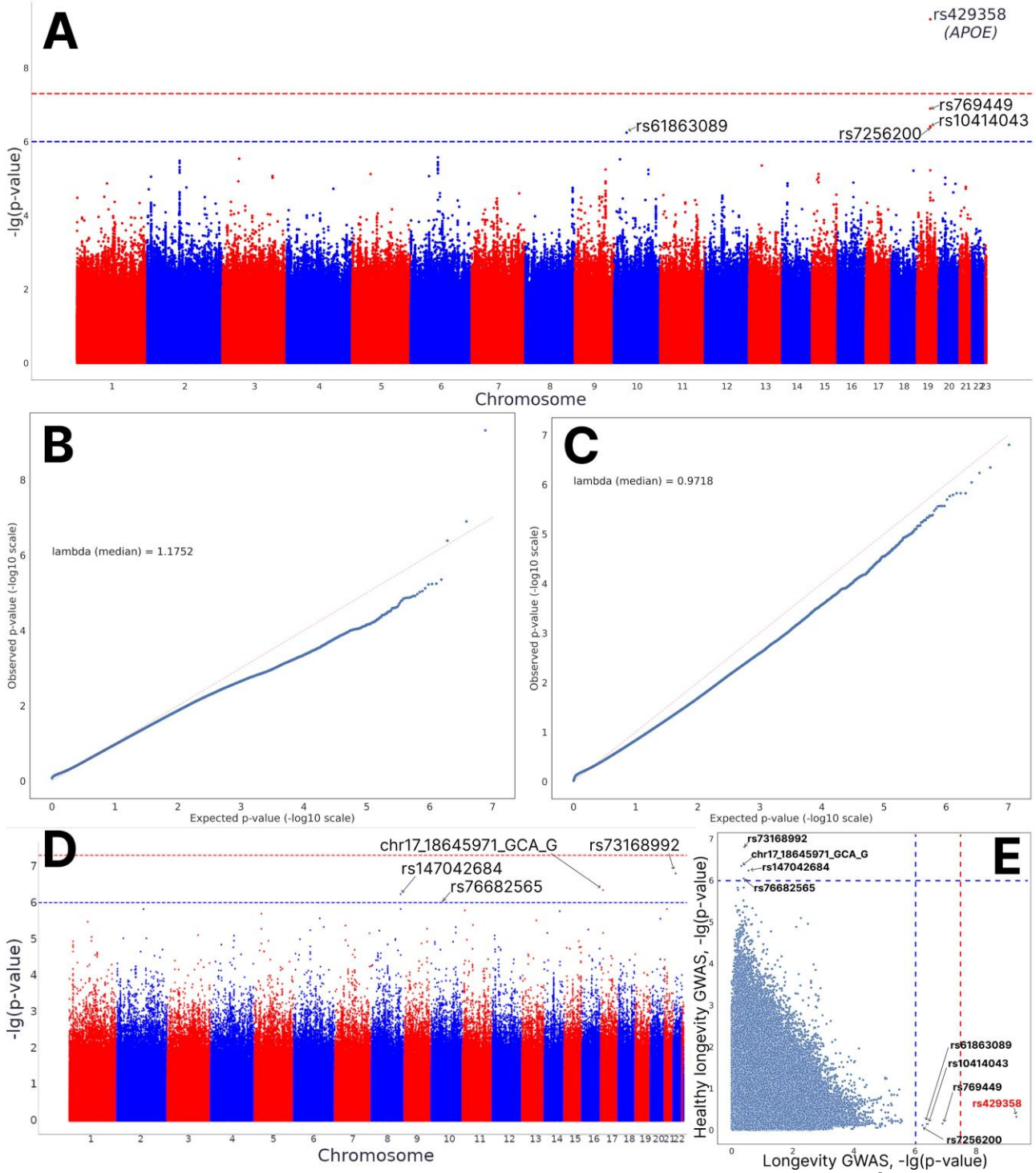

Figure 3. GWAS results. A. Results of longevity GWAS (Manhattan plot). Red dotted line: p-value = $5\times10^{-8;}$ blue dotted line: p-value = 10^{-6}. B. QQ-plot for longevity GWAS. C. QQ-plot for healthy longevity GWAS. D. Results of healthy longevity GWAS. Red dotted line: p-value = 5×10^{-8}; blue dotted line: p-value = 10^{-6}. E. Comparison of two GWASs: p-value for SNPs in longevity GWAS vs. p-value for SNPs in healthy longevity GWAS.

Table 1. Results of the GWAS for longevity results: variants showing the most significant associations with the trait.

Chromo-some	Position	Reference allele	Alternative allele	Regression p-value	Odds ratio	dbSNP ID	Gene name	Variant type
19	44908684	T	C	4.82×10^{-10}	0.67	rs429358	*APOE*	missense
19	44906745	G	A	1.27×10^{-7}	0.68	rs769449	*APOE*	intronic
19	44912456	G	A	3.86×10^{-7}	0.70	rs10414043	*APOC1*	2KB Upstream Variant
19	44912678	G	T	4.12×10^{-7}	0.73	rs7256200	*APOC1*	2KB Upstream Variant
10	132913337	G	T	5.73×10^{-7}	0.82	rs61863089	*CFAP46*	intronic

Note. dbSNP: the Single Nucleotide Polymorphism database.

Additionally, four polymorphisms—rs769449 in *APOE*, rs10414043 and rs7256200 in *APOC1*, and rs61863089 in *CFAP46*—did not meet the genome-wide significance threshold but did surpass the 10^{-6} threshold, suggesting a potential association that could be confirmed with a larger sample size (Table 1). Importantly, all of these polymorphisms exhibited a negative association with longevity, indicated by an odds ratio <1, and were associated with an increased risk of aging-related diseases. Consequently, these variants were significantly less frequent in the cohort of long-living adults than in the general population sample.

However, the presence of the significant variant rs429358 is not a definitive determinant of longevity. Despite its persistent negative association with longevity, a substantial number of long-living adults were carriers of this allele. Moreover, this cohort was phenotypically heterogeneous. Therefore, we assessed the frequency of this variant across the aging phenotypes using the calculator presented in our previous publication [10]. Table 2 presents the results of this assessment.

Table 2. Frequency of the rs429358 allele in long-living adults with different aging phenotypes.

Aging phenotype*	Group in GWAS for healthy longevity	Frequency of the rs429358 allele	P-value (chi-square)	Post-hoc analysis (pairwise chi-square); only significant pairs
Non-frailty	case	0.056	2.44×10^{-7}	Non-frailty vs. Cognitive frailty (5×10^{-6}), Non-frailty vs. Functional frailty (8.2×10^{-4}), Metabolic frailty VS Cognitive frailty (1.6×10^{-4}), Metabolic frailty vs. Functional frailty (0.009), Multimorbid frailty vs. Cognitive frailty (5.2×10^{-6}), Multimorbid frailty vs. Functional frailty (0.004)
Metabolic frailty	control	0.059		
Multimorbid frailty		0.067		
Cognitive frailty		0.12		
Functional frailty		0.097		

**Note: Aging phenotypes were determined using the calculator presented in our previous publication [10].*

No significant differences in the frequency of rs429358 were observed between the non-frailty, metabolic frailty, and multimorbid frailty groups. However, differences were detected between the functional and cognitive frailty groups. Carriers of this variant were at a greater risk of falling into one of the two most frail groups, which are associated with a high one-year mortality rate [10]. However, the presence of rs429358 does reliably differentiate between healthy and non-healthy phenotypes.

Although the absence of the rs429358 variant is associated with a higher probability of surviving to age 90, it is not a definitive factor for longevity or sustained functional capacity in this age group. This finding leads to the central question of our study: which variants are specifically associated with healthy longevity?

GWAS for healthy longevity

To identify genetic variants associated with healthy longevity, we conducted an additional comparison, in which 345 long-living adults with the healthy aging phenotype served as the case group, and 2,827 long-living adults with other aging phenotypes served as controls (Table 2). After quality filtration, 10,312,508 variants were tested. No significant results were obtained (Fig. 3D). The absence of significant SNPs was further supported by the QQ plot (Fig. 3C). Table 3 shows polymorphisms with a p-value < 10^{-6}, which may reach significance with a larger sample size.

We compared the two studies (Fig. 3E) and found no correlation between their results. This indicates that genetic factors specifically associated with healthy longevity are distinct from those associated with longevity defined as survival to age 90.

Table 3. Results of the GWAS for healthy longevity: variants showing the most significant associations with the healthy phenotype.

Chromo-some	Position	Reference allele	Alternative allele	Regression p-value	Odds ratio	dbSNP ID	Gene name	Variant type
22	25831221	C	T	1.57×10^{-7}	11.2	rs73168992	*MYO18B*	intronic
17	18645971	GCA	G	4.52×10^{-7}	10.4	-	*TBC1D28*	2KB Upstream Variant
8	83290733	G	A	5.82×10^{-7}	8.9	rs147042684	-	intergenic
10	23586606	T	C	9.02×10^{-7}	10.1	rs76682565	*LOC105376454*	500B Downstream Variant

Note. dbSNP: the Single Nucleotide Polymorphism database.

Polygenic score models

Based on the results of both GWASs, predictive PGS models were constructed for longevity (PGS for longevity) and healthy longevity (PGS for healthy longevity). The models were built using both logistic regression (LR) and random forest (RF) methods. Performance (ROC AUC) was evaluated using three datasets: training (80%) and internal test (20%) sets from the GWAS sample and external test sets (10%) from the initial sample reserved for validation. The ROC AUC of the LR-based PGS model for longevity was 0.68 in the training set, 0.685 in the internal test set, and 0.66 in the external test set (Fig. 4A).

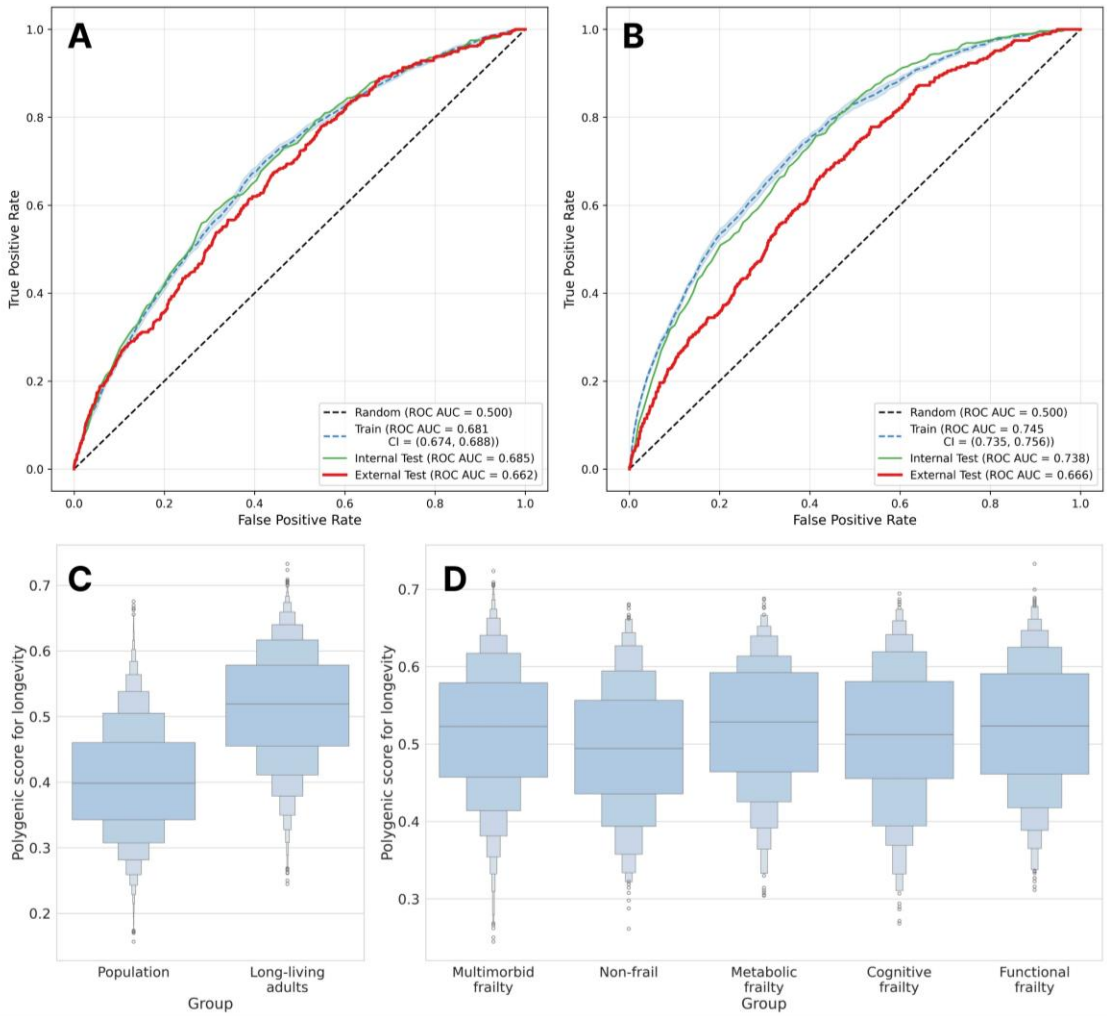

Figure 4. PGSs for longevity. (A) ROC curves for the LR-based model. **(B)** ROC curves for the RF-based model. **(C–D)** The distribution of PGSs for longevity (RF) across the study groups. *Training set: 80% of the GWAS sample; internal test set: 20% of the GWAS sample; and external test set: 10% of the initial sample reserved for validation.

The ROC AUC of the RF-based PGS model for longevity was 0.75 in the training set, 0.74 in the internal test set, and 0.67 in the external test set (Fig. 4B). Due to its superior performance, the RF-based model was selected as the primary model for further analysis. This model included 729 predictors: sex and 728 genetic variants. Table 4 shows the significance of the top ten variants included in the RF-based model as predictors. Sex was the predictor with the highest weight, followed by the rs1272044102 variant in the *AIRE* gene, rs429358 in the *APOE* gene, rs2096792627 in the *CAMTA1* gene, rs2549886183 in the *GPHN* gene, and a number of intergenic variants. The complete list of predictors and their permutation feature importance values are presented in Supplementary Table 1.

Table 4. Feature importance values for RF-based PGS models.

Feature	dbSNP ID	Gene	Permutation feature importance
PGS for longevity			
sex	-	-	0.087
chr21_44286154_G_C	rs1272044102	*AIRE*	0.016
chr5_61282896_ATACATG_A	-	-	0.014
chr19_44908684_T_C	rs429358	*APOE*	0.012
chr1_7739222_A_C	rs2096792627	*CAMTA1*	0.012
chr14_66984851_T_G	rs2549886183	*GPHN*	0.011
chr5_152479903_G_A	rs527909885	-	0.010
chr5_152479947_C_G	rs75413073	-	0.007
chr5_152479860_C_T	rs76164916	-	0.007
chr15_97195357_G_C	rs2061219	-	0.006
PGS for healthy longevity			
age	-	-	0.489
sex	-	-	0.348
chr11_20734966_T_G	rs137858196	*NELL1*	0.078
chr21_22800533_A_ATTTT	rs33923923	-	0.053
chr17_45952050_T_G	rs1395542202	*MAPT*	0.032

Figures 4C–D show the distribution of polygenic scores (PGSs) across different study groups. While a notable difference in the average PGSs is evident between long-living adults and the population sample, the overall predictive performance of the model was moderate, as indicated by the accuracy metrics. Furthermore, the PGSs for longevity did not effectively differentiate between healthy and non-healthy long-living adults. These findings support our hypothesis that longevity (reaching age 90) and healthy longevity are driven by different mechanisms.

The ROC AUC of the LR-based PGS model for healthy longevity was 0.625 in the training set, 0.62 in the internal test set, and 0.52 in the external test set (Fig. 5A). The ROC AUC of the RF-based PGS model for healthy longevity was 0.61 in the training set, 0.60 in the internal test set, and 0.61 in the external test set (Fig. 5B). The RF-based model showed higher predictive ability in the external test set and will be discussed further.

Table 4 shows the permutation feature importance of the variants included in the model as predictors. We observed a sharp decline in the model's predictive performance on the validation set, which may be explained by two factors: 1) a limited size of the validation set; 2) specificity of the associations between these genetic variants and healthy longevity to the GWAS cohort and not to the validation cohort. This finding is not unexpected, given the suggestive character of the associations between the polymorphisms and the phenotype in question. This result, however, underscores the critical importance of stratifying long-living adults in longevity studies.

DISCUSSION

Based on the findings from the GWAS for longevity, the rs429358 variant in the *APOE* gene was the only variant to achieve genome-wide significance, confirming the established association between this gene and longevity [7]. This variant defines the *APOE* ε4 allele—a missense mutation resulting in an amino acid substitution at C112R. Furthermore, this variant has been linked to Alzheimer's disease and reduced life expectancy [16]. Our previous research also suggests a connection between this variant and the onset of cognitive impairment in long-living adults, as assessed by the Mini Mental State Examination [17]. Notably, two variants, rs769449 (*APOE*) and rs10414043 (*APOC1*), which did not reach genome-wide significance, have also been associated with cognitive impairment [17]. The frequency of these variants was significantly lower in long-living adults than in the general population sample. Therefore, we hypothesize that the development of cognitive impairment, or

conversely, the preservation of cognitive function before the age of 90, is a critical determinant of surviving beyond this age. The biological mechanisms mediated by the rs429358 variant and other longevity-decreasing variants likely become activated well before the age of 90.

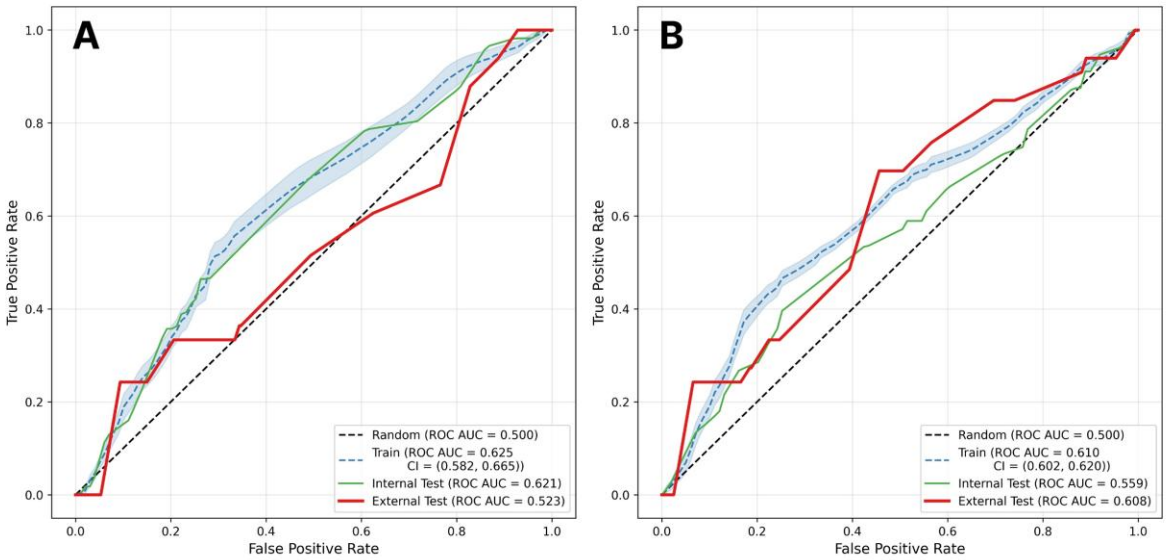

Figure 5. PGS for healthy longevity. (A) ROC curves for LR-based model. (**B)** ROC curves for the RF-based model. *Training set: 80% of the GWAS sample; internal test set: 20% of the GWAS sample; and external test set: 10% of the initial sample reserved for validation.

This hypothesis is motivated by the finding that the variants associated with longevity do not inform us about the quality of life at an older age. Long-living adults are not a homogeneous population, as modern medicine enables a longer lifespan despite chronic diseases. However, the primary interest lies in extending healthspan, not just lifespan. Using the aging phenotype calculator from our previous study [10], we demonstrated that neither the *APOE* gene nor other longevity-associated genes can reliably predict the preservation of functional capacity at the age of 90 and beyond. Although our GWAS for healthy longevity did not yield polymorphisms with genome-wide significance, it conclusively demonstrated that neither *APOE* nor *APOC1* variants have such significance for this phenotype.

Furthermore, the likelihood of longevity cannot be predicted solely based on the *APOE* gene. In our cohort, 14.7% of long-living adults were carriers of the rs429358 variant, with 0.3% being homozygotes. This frequency suggests that this genetic variant cannot be a definitive determinant of longevity. These findings highlight the polygenic nature of longevity. In developing our PGS models, we incorporated the concept of nonlinear interactions between genetic factors, proposed by Yashin et al. [9]. Consequently, we employed nonlinear machine learning models, specifically random forests.

It is important to note that, in addition to rs429358, the model also included polymorphisms located in other genes and intergenic regions. For example, the list of the top 10 predictors (Table 4) includes variants in the *AIRE, CAMTA1*, and *GPHN* genes. Interestingly, while these genes have not been directly associated with longevity, the *CAMTA1* gene is known to play an important role in the pathogenesis of age-related diseases, such as cancer and cognitive impairment [18]. This suggests that the key to longevity may lie not in individual variants but in their nonlinear interactions. Our future research will focus on long-living adults and other demographic groups carrying both rs429358 and at least one of the identified genetic variants to examine their interactions.

The model for healthy longevity included three polymorphisms in the *NELL1* and *MAPT* genes and an intergenic region. An association between longevity and the *MAPT* gene has been previously reported. As a key player in neurodegenerative processes [19], this gene may directly influence life expectancy, as observed in animal models [19, 20]. It is important to note that the specific associations identified for healthy longevity have not yet been independently replicated. Therefore, the transferability of our findings to non-Russian cohorts, as well as the functional significance of the identified polymorphisms, requires further investigation.

The models presented in this study require further development, including the incorporation of lifestyle covariates known to influence both longevity and healthy longevity. Moreover, given the feature importance of sex in the PGS for longevity, conducting separate, sex-specific studies would be advantageous. Consequently, the assessment of the predictive power of the current models must be considered preliminary, awaiting future refinement.

Limitations

Although not impossible, longevity in the control group, comprising younger individuals, is unlikely, making it appropriate for comparison purposes.
This research focused on identifying associations between the examined variants and phenotype. For causal inference, its findings require additional experimental validation.

Conclusion

The findings of this study not only confirmed the significance of the association between the *APOE* gene variants and longevity in the Russian population but also suggested a new approach to narrowing the phenotype. Moreover, this study could be expanded to include more non-genetic factors, which would enhance the accuracy of the model in predicting the likelihood of both surviving to age 90 and maintaining health.

Supplementary Materials

The Supplementary data can be found online at: www.aginganddisease.org/EN/10.14336/AD.2025.0828.

References

[1] Passarino G, De Rango F, Montesanto A. (2016). Human longevity: Genetics or Lifestyle? It takes two to tango. Immun Ageing,13: 12.

[2] Wang J, Chen C, Zhou J, Ye L, Li Y, Xu L, et al. (2023) Healthy lifestyle in late-life, longevity genes, and life expectancy among older adults: a 20-year, population-based, prospective cohort study. Lancet Healthy Longev, 4(9): e535–e543.

[3] Ni X, Wang Z, Gao D, Yuan H, Sun L, Zhu X, et al. (2021) A description of the relationship in healthy longevity and aging-related disease: from gene to protein. Immun Ageing, 18(1): 30.

[4] Brooks-Wilson AR. (2013). Genetics of healthy aging and longevity. Hum Genet, 132(12): 1323–1338.

[5] Bin-Jumah MN, Nadeem MS, Gilani SJ, Al-Abbasi FA, Ullah I, Alzarea SI, et al. (2022) Genes and Longevity of Lifespan. Int J Mol Sci. 23(3):1499.

[6] Budovsky A, Craig T, Wang J, Tacutu R, Csordas A, Lourenço J, et al. (2013) LongevityMap: a database of human genetic variants associated with longevity. Trends Genet, 29(10): 559–560.

[7] Chung W-H, Dao R-L, Chen L-K, Hung S-I (2010). The role of genetic variants in human longevity. Ageing Res Rev, 9 Suppl 1: S67–78.

[8] Revelas M, Thalamuthu A, Oldmeadow C, Evans T-J, Armstrong NJ, Kwok JB, et al (2018). Review and meta-analysis of genetic polymorphisms associated with exceptional human longevity. Mech Ageing Dev,175: 24–34.

[9] Yashin AI, Wu D, Arbeev KG, Ukraintseva SV (2010). Joint influence of small-effect genetic variants on human longevity. Aging, 2(9): 612–620.

[10] Mamchur AA, Sharashkina NV, Erema VV, Kashtanova DA, Ivanov MV, Bruttan MV, et al. (2025). Machine learning-based decision-making in geriatrics: aging phenotype calculator and survival prognosis. Aging Dis, 16(1): 565-577.

[11] Kashtanova DA, Taraskina AN, Erema VV, Akopyan AA, Ivanov MV, Strazhesko ID, et al. (2022). Analyzing Successful Aging and Longevity: Risk Factors and Health Promoters in 2020 Older Adults. Int J Environ Res Public Health, 19(13):8178.

[12] Danecek P. et al (2921). Twelve years of SAMtools and BCFtools. GigaScience, 10(2): giab008.

[13] Pedersen BS, Quinlan AR (2018). Mosdepth: quick coverage calculation for genomes and exomes. Bioinformatics, 34(5): 867–868.

[14] Kim S, Scheffler K, Halpern AL, Bekritsky MA, Noh E, Källberg M, et al (2018). Strelka2: fast and accurate calling of germline and somatic variants. Nat Methods, 15(8): 591–594.

[15] Javed N, Farjoun Y, Fennell TJ, Epstein CB, Bernstein BE, Shoresh N. (2020). Detecting sample swaps in diverse NGS data types using linkage disequilibrium. Nat. Commun, 11(1): 3697.

[16] Deelen J, Evans DS, Arking DE, Tesi N, Nygaard M, Liu X, et al (2019). A meta-analysis of genome-wide association studies identifies multiple longevity genes. Nat Commun, 10(1): 3669.

[17] Kashtanova DA, Mamchur AA, Dzhumaniyazova IH, Ivanov MV, Erema VV, Zelenova EA, et al. (2023). Cognitive impairment in long-living adults: a genome-wide association study, polygenic risk score model and molecular modeling of the APOE protein. Front Aging Neurosci, 15:1273825.

[18] Alves G, Ornellas MH, Liehr T (2022). The role of Calmodulin Binding Transcription Activator 1 (CAMTA1) gene and its putative genetic partners in the human nervous system. Psychogeriatrics, 22(6): 869-878.

[19] Pir GJ, Choudhary B, Mandelkow E, Mandelkow EM (2016). Tau mutant A152T, a risk factor for FTD/PSP, induces neuronal dysfunction and reduced lifespan independently of aggregation in a C. elegans Tauopathy model. Mol Neurodegener, 11:33.

[20] Higham JP, Malik BR, Buhl E, Dawson JM, Ogier AS, Lunnon K, et al. (2019). Alzheimer's Disease Associated

Genes Ankyrin and Tau Cause Shortened Lifespan and Memory Loss in *Drosophila*. Front Cell Neurosci, 13:260.