



OPEN

Genetic and non-genetic factors associated with the phenotype of exceptional longevity & normal cognition

Bin Han¹, Huashuai Chen^{2,3}, Yao Yao⁷, Xiaomin Liu^{4,5}, Chao Nie^{4,5}, Junxia Min⁶, Yi Zeng^{2,7}✉ & Michael W. Lutz⁸✉

In this study, we split 2156 individuals from the Chinese Longitudinal Healthy Longevity Survey (CLHLS) data into two groups, establishing a phenotype of exceptional longevity & normal cognition versus cognitive impairment. We conducted a genome-wide association study (GWAS) to identify significant genetic variants and biological pathways that are associated with cognitive impairment and used these results to construct polygenic risk scores. We elucidated the important and robust factors, both genetic and non-genetic, in predicting the phenotype, using several machine learning models. The GWAS identified 28 significant SNPs at p -value $< 3 \times 10^{-5}$ significance level and we pinpointed four genes, *ESR1*, *PHB*, *RYS3*, *GRIK2*, that are associated with the phenotype through immunological systems, brain function, metabolic pathways, inflammation and diet in the CLHLS cohort. Using both genetic and non-genetic factors, four machine learning models have close prediction results for the phenotype measured in Area Under the Curve: random forest (0.782), XGBoost (0.781), support vector machine with linear kernel (0.780), and ℓ_2 penalized logistic regression (0.780). The top four important and congruent features in predicting the phenotype identified by these four models are: polygenic risk score, sex, age, and education.

Cognitive Impairment (CI) is defined as the loss of ability in cognitive functions, such as remembering, learning, and concentrating, which negatively impacts affected individuals' daily activities¹. In the stage of mild cognitive impairment (MCI), affected individuals start to experience memory issues without seriously hindering their abilities to execute daily activities. In the stage of severe cognitive impairment, which is referred as dementia², individuals tend to lose basic functionalities of comprehending, memorizing, or even talking and writing. Many diseases are associated with the development of CI, such as Alzheimer's Disease (AD), Vascular Dementia, Parkinson's Disease (PD), Progressive Supranuclear Palsy, and Lewy Body Disease³.

Aging is one of the major risk factors for the development and onset of CI^{4,5}. As forecast in the world population aging report⁶, the population aged 60 years or above is expected to grow by 56% between 2015 and 2030. By 2050, that population is projected to be 2.1 billion globally. The increasing aging population size, with correspondingly increasing prevalence of CI, imposes great burdens at the levels of individuals, families, and communities⁷⁻⁹. Thus, it is vitally important to study factors that are associated with CI and to investigate potential therapeutic or lifestyle interventions, for the sake of improving the quality of life or delaying the onset of CI and reducing economic costs.

There have been many previous studies on the pertinent factors associated with cognitive impairment. For example, using discovery and multiple replication cohorts, Davies et al.⁵ identified several significant genetic loci that are associated with CI, such as rs2075650 and rs115566 located in *TOMM40* and rs429358 located in the *APOE* region. Lv et al.¹⁰ probed the association between the rate of cognitive decline and the mortality rate.

¹Department of Statistical Science, Duke University, Durham, NC, USA. ²Center for the Study of Aging and Human Development, Medical School of Duke University, Durham, NC, USA. ³Business School of Xiangtan University, Xiangtan, China. ⁴BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China. ⁵BGI-Shenzhen, Shenzhen, China. ⁶The First Affiliated Hospital, Institute of Translational Medicine, School of Medicine, Zhejiang University, Hangzhou, China. ⁷Center for Healthy Aging and Development Studies, National School of Development, Raissun Institute for Advanced Studies, Peking University, Beijing, China. ⁸Department of Neurology, Duke University School of Medicine, Durham, NC, USA. ✉email: zengyi@duke.edu; Michael.Lutz@duke.edu

They concluded that faster cognitive decline rate is associated with higher mortality rate, specifically among individuals aged between 65–79 years old and cognitively normal individuals, regardless of their initial cognitive abilities. With respect to the analysis on non-genetic factors, Casanova et al.¹¹ constructed a random forest model to investigate important predictors for cognitive trajectories, identifying education, age, and gender, as top predictors. Many other studies have also analyzed the genetic and non-genetic contributors to CI in community-based cohorts^{4,12,13}.

Nevertheless, there are few studies that analyze the effects from both genetic and non-genetic factors on CI to our knowledge. Consequently, the primary goal of our research is to investigate which factors, both genetic and non-genetic, are significantly associated with cognitive impairment, in contrast to intact cognition in late life. Specifically, we approached the problem from the following two original perspectives:

- We conducted GWAS using the approach of splitting individuals into two groups, considering a phenotype of exceptional longevity & normal cognition. The work by Perez-Gracia et al.¹⁴ demonstrated that the selection of individuals with exceptionally characteristic, clinically-related phenotypes, referred to as extreme phenotypes, can promote efficiency in identifying important factors at the molecular level. Additionally, Estep¹⁵ proposed a similar definition of long-living individuals with intact cognition as the extreme phenotype. Consequently, we split the samples into two groups: (1) participants with the exceptional longevity & normal cognition phenotype—aged 90 or above with normal cognition. (2) individuals that were cognitively impaired without age restrictions. The goal here is trying to raise the contrast between the two groups and help identify significant protective effects from the group with exceptional longevity & normal cognition and risk effects from the cognitively impaired group. Note that the definition of exceptional longevity (age ≥ 90 years) can differ in different population and social contexts.
- Machine learning has been widely used for prediction tasks. There has been previous work on applying machine learning techniques to identify the important predictors of cognitive trajectories by Casanova et al.¹¹ as mentioned above. In our analysis, we considered several machine learning models to predict CI and identify significant factors using feature importance plots. Different from Casanova et al.'s work where they only worked with random forest models, we considered multiple machine learning models to reduce potential bias from using just one model as model choices have potential impacts on the results. We also introduced genetic effects in the form of Polygenic Risk Scores (PRS) and considered more non-genetic variables.

Results

Summary statistics. The non-genetic factors from the CLHLS survey include socioeconomic characteristics, health status, and living habits referring to previous studies^{4,10,16}. Specifically, socioeconomic factors include: age, sex, education, occupation, co-residence, and marital status. Living-habit factors include: staple food (major food source), fruits intake, vegetables intake, current smoker, former smoker, current drinker, former drinker, and exercise currently. Health status factors include whether participants have hypertension, diabetes, heart diseases, cardiovascular diseases, and respiratory diseases. Except for age, education, occupation, marital status, and staple food, other factors are binary features taking on yes/no answers. Please refer to Table 1 for summary statistics of non-genetic factors. In terms of our study sample, Table 2 displayed the counts of individuals stratified by sex and age groups (5-year intervals), conditioned on cognitive status. We noticed that there are more female participants than male counterparts and there are more cognitively impaired participants than cognitively normal ones. Almost half of the individuals are within the 100–105 age group. Additionally, we ran univariate tests (T-test and Chi-square tests of Independence) on variables to check if there are differences in the mean values or if there is any association existing between the two cognitive groups and the variables. Please refer to Table 2 in the Supplemental material.

GWAS result. We present the set of significant SNPs in Table 3 using 3×10^{-5} as the p-value threshold. The table contains 28 significant and independent SNPs associated with CI, with 14 SNPs having odds ratios greater than one and 14 less than one. Out of all the significant SNPs, rs13198061 in gene *ESR1*, rs56368572 in *CTNND2*, rs954303 near *RNU4-58P* and rs939432 in gene *RYR3* have p-values less than 1×10^{-5} and odds ratio less than 1, indicating that those SNPs are associated with potential protective effects of preventing cognitive impairment. The SNPs in *ESR1* and *RYR3* are of specific interest as they have been extensively studied and found to be involved with immunologic processes and brain functions, which are previously reported to be associated with cognitive impairment and decline. Regarding to the significant SNPs that have odds ratios greater than 1, indicating that they are related with progressive effects towards cognitive impairment, we found out that rs935129 in *RP11-81K2.1* and *PHB*, rs6726046 in *DGKD* and *AC0129221.4*, and rs13028996 in *SAG* are significant at the level less than 1×10^{-5} . We also completed the GWAS including years of education as a covariate and report the results in Table 1 in the Supplemental Material.

Polygenic risk score. As mentioned in the introduction, our primary focus is on the identification of statistically significant factors, both genetic and non-genetic, that are associated with CI. However, correction for multiple testing in GWAS might undermine the power of association analysis to identify small effect-size variants that have biological or clinical importance. Additionally, genetically complex diseases typically involve numerous small effect-size genetic factors. Therefore, to select significant SNPs from GWAS and construct PRS, we tested three different p-values, from a relatively stringent one to a comparably relaxed one. Then we used PRS to predict cognitive status (binary response variable—impaired or intact cognition) using Area Under the Curve (AUC) as the measurement. The three p-values are 1×10^{-5} , 2×10^{-5} , and 3×10^{-5} and the corresponding number of SNPs are 7, 19, and 28 respectively.

Feature	Male (N = 533)	Female (N = 1623)	Total (N = 2156)
Age (years), mean (SD)	100.0 (3.55)	101.7 (3.54)	101.3 (3.62)
Education (years), mean (SD)	2.9 (3.9)	0.3 (1.3)	0.9 (2.49)
Occupation			
White-Collar	64 (0.12)	16 (0.01)	80 (0.04)
Other	469 (0.88)	1607 (0.99)	2076 (0.96)
Marital status			
Single	450 (0.84)	1598 (0.98)	2048 (0.95)
Partnered	83 (0.16)	25 (0.02)	108 (0.05)
Staple food			
Corn	13 (0.02)	54 (0.03)	67 (0.03)
Rice	295 (0.55)	859 (0.53)	1154 (0.53)
Wheat	129 (0.24)	467 (0.28)	596 (0.28)
Other	96 (0.18)	243 (0.15)	339 (0.16)
Co-residence	449 (0.85)	1370 (0.84)	1819 (0.84)
Fruit intake	73 (0.14)	188 (0.12)	261 (0.12)
Vegetables intake	433 (0.81)	1324 (0.82)	1757 (0.81)
Current smoker	93 (0.17)	78 (0.05)	171 (0.08)
Former smoker	221 (0.41)	159 (0.10)	380 (0.18)
Current drinker	103 (0.19)	133 (0.08)	236 (0.11)
Former drinker	205 (0.38)	211 (0.13)	416 (0.19)
Exercise currently	141 (0.26)	185 (0.11)	326 (0.15)
Hypertension	88 (0.17)	286 (0.18)	374 (0.17)
Diabetes	6 (0.01)	12 (0.01)	18 (0.01)
Heart	53 (0.1)	131 (0.08)	184 (0.09)
Cardiovascular disease	29 (0.05)	79 (0.05)	108 (0.05)
Respiratory	77 (0.14)	140 (0.09)	217 (0.10)

Table 1. Summary statistics of non-genetic factors. Data are provided as count (percentage), unless specified in the feature column. From variable “Co-residence” to “Respiratory”, all the features take binary values of either Yes or No. Their count values sum individuals with Yes response.

Age groups	Cognitively impaired		Cognitively normal	
	Female	Male	Female	Male
90–95	9	6	12	21
95–100	191	78	156	122
100–105	615	115	316	142
105–110	189	26	95	18
110+	26	2	14	3

Table 2. Study sample stratified by age groups and sex, conditioned on cognitive status. The age groups are inclusive on the right bounds and exclusive on the left bounds.

As shown in Fig. 1, the average AUC from fivefold cross validation noticeably increases as p -value threshold increases. Specifically, the PRS from two relatively smaller thresholds (1×10^{-5} , 2×10^{-5}) have average AUC equal to 0.631 and 0.706 respectively. The PRS from the largest p -value threshold (3×10^{-5}) has the best performance, with average AUC of 0.742. It demonstrates that an overly stringent p -value threshold might overlook some genetic effects, which are small at SNP level but large at aggregated individual level and could enhance the predictive power. Therefore, for the machine learning models, we included the PRS based on SNPs using 3×10^{-5} as the threshold, as it has the best performance.

The association between the PRS and the phenotypes in the replication cohort was not significant ($p = 0.49$). However, higher scores of the PRS were associated with cognitive impairment relative to the extreme phenotype (age ≥ 90 years and no cognitive impairment) (OR = 1.05, 95% CI 0.92–1.20). Female sex was significantly associated with the extreme phenotype relative to cognitive impairment ($p = 0.0023$, OR = 0.61, 95% CI 0.45–0.84).

Machine learning models. Using both genetic and non-genetic factors, four out of five machine learning models we examined— ℓ_2 penalized logistic regression, support vector machine (SVM) with linear kernel, random forest, and XGBoost have similar performances, with average AUCs from fivefold cross-validation around

SNP	Chr.	Position	Nearest gene	A1	A2	MAF	p	Odds ratio	Lower-95CI	Upper-95CI
Odds ratio < 1										
rs13198061	6	152,306,894	ESR1*	T	C	0.051	1.4×10^{-6}	0.49	0.37	0.66
rs939432	15	33,986,294	RYR3*	C	A	0.274	2.1×10^{-6}	0.71	0.61	0.82
rs954303	16	59,581,776	RNU4-58P (7606)	A	G	0.155	2.2×10^{-6}	0.66	0.55	0.78
rs56368572	5	11,300,912	CTNND2*	T	C	0.094	8.7×10^{-6}	0.62	0.50	0.77
rs4816332	21	30,201,706	N6AMT1 (42807)	C	T	0.400	1.0×10^{-5}	0.74	0.65	0.85
rs1030695	4	130,318,150	RP11-419L4.1 (91973)	T	A	0.299	1.1×10^{-5}	0.73	0.64	0.84
rs1293144	20	52,917,208	PFDN4 (72617)	T	G	0.371	1.3×10^{-5}	0.75	0.66	0.85
rs62001981	15	25,279,909	RP11-701H24.10* & PWAR6*	T	C	0.196	1.8×10^{-5}	0.70	0.60	0.82
rs9404070	6	101,463,320	GRIK2 (383344)	G	A	0.415	2.0×10^{-5}	0.76	0.67	0.86
rs76299633	13	40,727,639	LINC00332 (28307)	G	A	0.119	2.0×10^{-5}	0.66	0.55	0.80
rs9676032	18	48,297,450	MRO (27124)	T	A	0.131	2.1×10^{-5}	0.67	0.56	0.81
rs28673399	4	71,371,765	AMTN (12492)	G	A	0.448	2.2×10^{-5}	0.76	0.67	0.86
rs10500293	19	46,431,638	NOVA2 (5354)	G	A	0.440	2.7×10^{-5}	0.77	0.68	0.87
rs72627042	3	23,906,287	UBE2E1*	T	C	0.058	2.8×10^{-5}	0.55	0.41	0.72
Odds ratio > 1										
rs13028996	2	234,246,225	SAG*	C	T	0.466	1.1×10^{-6}	1.38	1.21	1.57
rs6726046	2	234,287,221	DGKD* & AC019221.4*	A	G	0.375	1.5×10^{-6}	1.37	1.21	1.56
rs935129	17	47,486,016	RP11-81K2.1* & PHB*	A	G	0.387	8.3×10^{-6}	1.35	1.18	1.53
rs2792251	1	164,541,977	PBX1*	G	A	0.136	1.1×10^{-5}	1.53	1.27	1.85
rs6547617	2	85,655,402	SH2D6*	A	T	0.406	1.1×10^{-5}	1.35	1.18	1.55
rs10037430	5	180,569,007	OR2V2 (12936)	C	T	0.069	1.2×10^{-5}	1.93	1.44	2.59
rs7710849	5	82,220,225	RP11-78C3.1 (3287)	T	A	0.052	1.3×10^{-5}	2.04	1.48	2.80
rs79669991	22	43,936,861	-	A	G	0.253	1.5×10^{-5}	1.39	1.20	1.61
rs2418761	10	107,295,345	RNU6-463P (822)	C	T	0.110	1.5×10^{-5}	1.57	1.28	1.93
rs7927292	11	44,730,158	RP11-45A12.2 (10784)	A	C	0.082	2.0×10^{-5}	1.71	1.34	2.18
rs741171	16	6,652,854	RP11-420N3.2* & RBFOX1*	G	A	0.206	2.3×10^{-5}	1.41	1.21	1.66
rs57164734	11	44,773,258	TSPAN18*	G	C	0.089	2.3×10^{-5}	1.66	1.31	2.11
rs2528812	7	22,446,110	STEAP1B (12953)	C	T	0.420	2.4×10^{-5}	1.32	1.16	1.50
rs4934715	10	35,364,992	CUL2*	T	G	0.306	2.8×10^{-5}	1.35	1.17	1.55

Table 3. Information of significant SNPs from GWAS using 3×10^{-5} as the p -value threshold. The nearest genes are either the genes that contain the variants (overlapping) or the nearest upstream/downstream gene to the variants. * indicates overlapping gene. Distances to the nearest upstream/downstream genes are listed in the parenthesis, measured in bp distance.

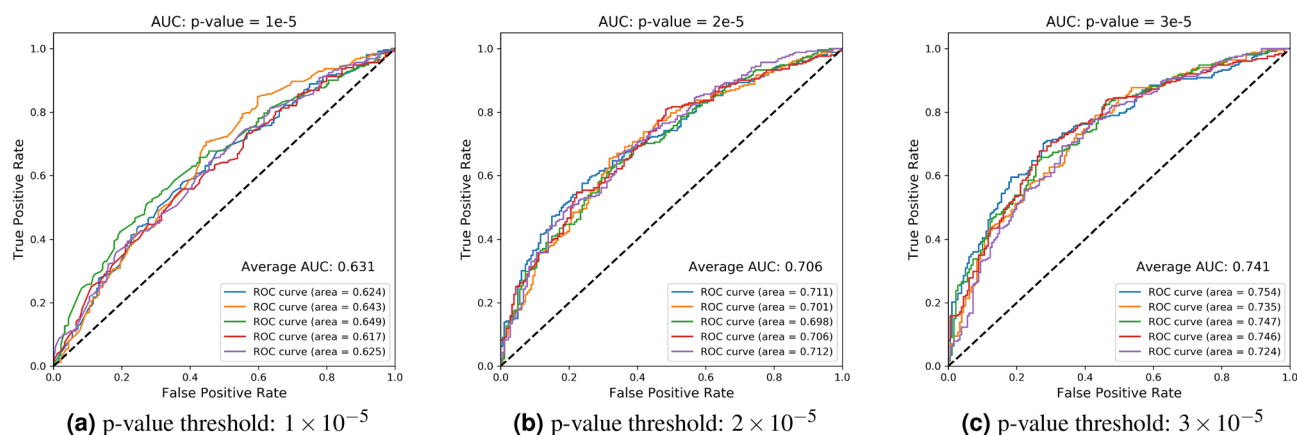


Figure 1. Average AUC from fivefold cross validation using PRS to predict the cognitive impairment. The p -value threshold increases from 1×10^{-5} in panel (a) to 2×10^{-5} in panel (b) and to 3×10^{-5} in panel (c).

Model	Performance (AUC)	Top six important features (descending in importance)
ℓ_2 Penalized Logistic Regression	0.780	PRS, education, age, sex, vegetables intake, former smoker
SVM—linear kernel	0.780	PRS, education, age, sex, co-residence, vegetables intake
Random forest	0.782	PRS, age, education, staple food, sex, exercise currently
XGBoost	0.781	PRS, education, staple food, sex, age, exercise currently

Table 4. Model performances and important features from the best four predictive models. Model performances are almost identical and four importance factors are congruent among the models—PRS, education, age, and sex.

Category	GeneSet	N Genes	N Overlap	P	adjP	Genes	Link
Immunologic signatures	GSE12392_WT_VS_IFNB_KO_CD8A_POS_SPLEEN_DC_DN	200	4	4.59×10^{-6}	0.022	RYR3:PHB:GRIK2:ESR1	Link1
Transcription factor targets	AACTTT_UNKNOWN	1928	7	6.09×10^{-5}	0.037	PBX1:RYR3:RBF1:DGKD:UBE2E1:CTNND2:ESR1	Link2

Table 5. Results of the gene to function pathway analysis. Link 1: http://www.gsea-msigdb.org/gsea/msigdb/cards/GSE12392_WT_VS_IFNB_KO_CD8A_POS_SPLEEN_DC_DN Link 2: http://www.gsea-msigdb.org/gsea/msigdb/cards/AACTTT_UNKNOWN “adjP” refers to adjustment for multiple comparisons.

0.780. CART has relatively lower prediction performance, with average AUC equal to 0.738. We generated the feature importance plots from the four best performing models and the result are listed in Table 4. Interestingly, all models identified PRS (based on SNPs extracted from GWAS using 3×10^{-5} as the p -value threshold), education, age, and sex to be important factors, with slightly different orders. It demonstrates that those factors are important and robust in predicting CI. Additionally, simple linear models, including logistic regression and SVM with linear kernel, are consistent on the factor “Vegetables Intake”, while complex models, including random forest and XGBoost, considered intricate interactions among features and agreed on the factors “Staple Food” and “Exercise Currently”.

Pathway analysis. Results of the gene to function pathway analysis are shown in Table 5. Two pathways were significant at the $p \leq 0.05$ level after adjustment for multiple comparisons, one is an immunological signature, the other a signature based on a highly conserved transcription factor motif. Two genes are common to these pathways: *ESR1* and *RYR3*.

Discussion

Our study investigated the effects of genetic and non-genetic factors on cognitive impairment. We approached the problem with two novel aspects compared with prior studies. First, we adopted the approach of stratifying individuals with a phenotype of exceptional longevity & normal cognition and contrasted them with cognitively impaired patients using GWAS, aiming to identify significant genetic variants that have progressive effects towards or protective effects from CI. Considering that our data has the world’s largest sample size of Chinese centenarians and considering the historical background of China, this approach would help us identify stronger biological mechanisms. Our study identified numerous significant SNPs in GWAS and pinpointed the corresponding genes that have coherent and inter-connected biological interpretations through immunological systems, brain functions, metabolic pathways, and diets. Second, we used machine learning techniques to predict cognitive impairment using both genetic and non-genetic factors, which have not been well characterized by prior studies. The four best performing models consistently identified four important factors for predicting cognitive impairment: PRS, sex, age, and education. In addition, the two simple models agreed on factor “Vegetables Intake”. The two complex models in addition agreed on two factors: “Exercise Currently” and “Staple Food”, considering convoluted interactive effects among factors.

The PRS is identified as the most important factor from all four models. It is not surprising to see the results because of two reasons. First, genetic pre-disposition affects the onset and severity of CI. Second, as mentioned in section M5, our study considered individuals with the phenotype of exceptional longevity & normal cognition, contrasted with the cognitively impaired individuals, which could potentially identify moderate or strong genetic effects. Age is expected to be significant as ageing is strongly associated with CI. Although the PRS was not significantly associated with the phenotype in the replication cohort, the direction of the effect indicated that this score reflects an increased risk for CI as described in the analysis of the CLHLS data. The effect size of the PRS was also consistent with the effect sizes reported for Alzheimer’s disease genetic associations^{17,18}. Replication in an independent dataset is critical for genetic studies including development and testing of PRS. The CLHLS is a unique resource in terms of the number of centenarians and long-lived individuals which makes identification of a suitable replication cohort challenging. The advantages of the ROSMAP cohort for replication are the long follow up period with repeated cognitive testing, up to 24 years (mean = 5.6 years, SD = 4.9), consistent definition of the phenotype with the CLHLS study and the large number of individuals with the extreme phenotype (579).

Reasons for lack of replication of the PRS association with the phenotype are likely associated with the lack of power to detect a statistically-significant effect. However, it is also possible that adding additional SNPs to the PRS would improve the prediction performance, especially for coverage of individuals with different genetic ancestry. Replication in other cohorts to confirm the usage of the PRS to predict CI relative to aging to 90+ years with intact cognition is warranted.

Prior studies have extensively and consistently shown that sex is an important factor related with longevity, cognitive impairment/decline, and the development of Alzheimer Disease. Au et al.¹⁹ conducted meta-analysis including 56 studies and showed that there was a higher prevalence, but not incidence, of non-amnesic MCI among women than among men. A similar conclusion has been arrived by Li and Singh²⁰: elderly women displayed faster rate of cognitive decline and more severe cognitive deterioration than elderly men. Additionally, closely related with our study population, An et al.²¹ inspected the cognitive patterns among the middle-aged and elder Chinese people, specifically considering sex differentiation. They reported that females tend to have verbal memory advantages over their male counterparts independent of age. On the other hand, males tend to have more intact cognition in general, with better functions in attention, execution, and processing speed. The disparities in cognition exhibited between the two sex groups have not been fully disclosed, but are considered to be the aftermath of differences in years of educations among that generation of the elderly Chinese population^{19,21}.

Several studies have shown that education is an important factor in affecting the development of CI^{4,12,22}. These studies consistently concluded that higher education (or longer years in education) helps maintain the cognitive functions through cognitive practice, thus reducing the risks of developing CI. Additionally, education has also been extensively studied and widely accepted^{23–25} as an important factor associated with Alzheimer Disease (AD). Similar to the conclusion in CI literature, higher educational levels are associated with lower risk of AD.

As shown in Table 4, simple linear models agreed on “Vegetables Intake” as an important factor. In addition, random forest and XGBoost identified “Staple Food” and “Exercise Currently” as important factors, considering intricate interactions among factors. “Vegetables” and “Staple Food” can be related with an individual’s type of diet. The finding that the staple food is an important factor is interesting, specifically considering the environment of China where our study is based. China is generally divided into Southern and Northern areas according to some geographical division. Consequently, the living habits, including the main source of food, are usually different between the two regions, which might have an impact on the starch and diet-related biology. In general, there are some studies that have shown that cognitive decline/cognitive aging is associated with the individual’s diet. For example, two studies^{26,27} have shown that Dietary Approach to Stop Hypertension (DASH) and Mediterranean diet is associated with slower rates of cognitive aging. The study by Samieri et al.²⁸ concluded that diets involving fish in-take may slow down the rate of cognitive aging as well. Another study²⁹ demonstrated that different types of nutrition in the diet can modify the potential risk of developing cognitive impairment in the future.

In addition to the general studies of the relationship between diet and cognitive decline, there are some specific works that investigate the association between cognitive abilities and diet among Chinese elderly people, which is more relevant to our study. Two research studies^{30,31} have consistently shown that a healthier diet, such as diets with more nuts, vegetables, and fruits, could help decrease the risks of cognitive impairment. Similarly, the study by Wang et al.³² arrived at the conclusion that Chinese diets that lack of legume and animal oil might increase the prevalence of mild cognitive impairment. Additionally, it is recognized that Chinese diets are abundant in carbohydrates, mainly starch and sucrose from sources such as rice. The study by Qin et al.³³ has shown that a wheat-based diverse diet that consists of similar components as the Mediterranean diet could slower the rate of cognitive decline.

Active physical exercise is also found to be important in predicting CI from the two complex models. However, prior human and animal studies investigating the relationship between physical exercise and cognitive functions are not always consistent³⁴. Some displayed strong positive associations while other showed minimum or no relationships. For example, Baker et al. studied the effects of aerobic exercises on cognition using randomized and clinically controlled trials. They demonstrated the sex-specific effects that aerobic exercises help improve executive control process for older women³⁵. The study from Laurin et al. also showed that physical exercises are associated with lower risk of CI, dementia and Alzheimer Disease, compared with individuals with no exercises³⁶. Similarly, Geda et al. conducted a population-based case-control study and concluded that moderate exercise in midlife or later life, no matter what frequency of the exercise, is associated with reduction in odds of developing mild cognitive impairment (MCI)³⁷. However, there are studies that did not identify significant relationships between CI and physical exercise. Young et al. concluded that with their collected randomized controlled trials (RCTs), there is no evidence of cognitive benefits among cognitively healthy elderly people from cardiorespiratory exercises³⁸.

Four genes identified in the pathway analysis (*ESR1*, *PHB*, *RYR3*, *GRIK2*) constitute part of an immunologic signature in the MSigDB database³⁹ that was identified as significantly associated with the cognitive impairment phenotype. Both the individual genes that comprise this signature, and innate and adaptive immunity as biological processes have been associated with cognitive impairment and cognitive aging^{40–46}. Polymorphisms in estrogen receptor genes (*ESR1* and *ESR2*) have been associated with risk of developing cognitive impairment and in turn may play a role in cognitive aging^{42,47}. These polymorphisms have been demonstrated to have an impact in both men⁴² and women^{42,48}. Moreover, interactions between *ESR1* and the *APOE* gene for AD risk have also been reported^{49,50}. Some of the *ESR1* polymorphisms associated with Alzheimer’s disease and mild cognitive impairment are low-frequency (MAF < 2%)⁵¹. The specific *ESR1* SNP detected as associated with cognitive impairment in our study is not in linkage disequilibrium with the SNPs reported in Yaffe et al. ($r^2 \leq 0.014$) and are greater than 100 kB distant from these SNPs, therefore likely represent a different association signal. Genetic alterations in the estrogen metabolic pathway have been reported to be associated with risk of Alzheimer’s Disease in a study of a southern Chinese population⁵² and association of *ESR1* with one-year cognitive decline in healthy oldest-old individuals has also been cited⁵³.

The *RYR3* gene codes for the ryanodine receptor that functions to release calcium from intracellular storage for use in cellular signaling and biochemical processes. *RYR3* is an isoform of the ryanodine receptor that is expressed in specific regions of mammalian brain that are involved in the development of cognitive dysfunction and Alzheimer's disease, e.g. the hippocampus⁵⁴. Sustained calcium dysregulation contributes to neurodegeneration and cognitive impairment^{55,56} and ryanodine receptor levels change during the lifespan^{57,58} and are altered in mild cognitive impairment and Alzheimer's disease⁵⁹.

The other two genes in the signature, *GRIK2* and *PHB* have literature support for roles in cognitive aging and longevity. *GRIK2* belongs to the kainate family of glutamate receptors that function as ligand-activated ion channels. This gene is highly expressed in brain and mutations in the gene have been associated with cortical development, autism and schizophrenia^{60–62}. Interestingly, a large (6036 cases) genome-wide association study identified SNPs near *GRIK2* as showing suggestive levels of association with longevity (OR = 1.2, $p = 5.09e^{-8}$)⁶³. *PHB*, or prohibitin, is characterized as playing a role in human cellular senescence and tumor suppression and in model organisms, as a modulator of longevity⁶⁴. Prohibitins modulate mitochondrial fusion and have a role in forming protein and lipid scaffolds^{65,66}. *Caenorhabditis elegans* studies have shown that prohibitins moderate fat metabolism and energy production and in turn influence aging⁶⁴. In a mouse model, mutations in *PHB2* (one of the two homologous *PHB* proteins) triggered massive neurodegeneration with accumulation of abnormal mitochondria and hyperphosphorylated tau⁶⁷. Interestingly, in a *C. elegans* model, mutations in *PHB* were strongly associated with genotype-dependent responses to dietary restriction⁶⁸.

The immunologic signature based on gene set enrichment for the genes in the predictive risk score represents a small subset (4 out of 200) of genes that are downregulated in CD8A+ splenic dendritic cells in a mouse model where interferon beta 1 (*IFNB1*) is knocked out⁶⁹. *IFNB1*, along with other type 1 interferons link the innate and adaptive arms of the immune system. The involvement of genes representing several metabolic and signaling functions and, specifically, the role for dendritic cells that comprise the biochemical basis for the immunologic signature, contribute to a possible role for association with the cognitive aging phenotype utilized in our study. Dendritic cells capture antigens, which are transported to the lymphatic system. Mechanistically, identification of this signature in the context of our study, that examined cognitive resilience in nonagenarians and centenarians, is of interest because of the likely involvement of the immune system and immune response to exogenous antigens. In the mouse model associated with the signature identified in our study, the absence of *IFNB1* constitutes an “exceptional phenotype” for a model where a critical factor needed for T cell stimulation is removed. Over the course of a lifetime, an individual is exposed to numerous bacterial, viral and parasitic infections. Of particular relevance to our study from the perspective of examining environmental and genetic factors that contribute to long-term cognitive resilience, low but constitutive production of *IFNB1* was shown to be necessary to maintain dendritic cells in a state that is responsive to antigens associated with these infections⁶⁹. Dendritic cells that can capture and process antigens under noninflammatory conditions are considered to acquire tolerogenic properties (e.g. induction of tolerance) which may have strong relevance for understanding the genetic backgrounds associated with cognitive resilience⁶⁹. The overall immunologic gene signature is complex with components including members of gene families of transcription factors, cell differentiation markers and homeodomain proteins. Biochemical linkages between the 4 genes in the signature identified in our study likely represent several underlying physiological processes involved in immune function and not a single biochemical pathway that involves signaling between these genes.

The design of our study is to compare individuals who live to an advanced age, 90+ years with individuals who develop cognitive impairment in middle to late life. Cognitive impairment can result from several pathological processes including Alzheimer's disease, vascular dementia, cerebrovascular disease, Lewy body dementia, frontotemporal dementia and mixed dementia. The CLHLS survey does not include clinical diagnoses that differentiate these causes and therefore cognitive impairment is equivalent to all cause dementia for the purposes of our study. Alzheimer's disease pathology is likely a major cause of many of the dementia cases in our study, accounting for 60 to 80% of dementia cases⁷⁰.

Genome-wide association studies (GWAS) and development of predictive risk scores for AD are active areas of research. The study by Kunkel et al.⁷¹ identified 25 genetic loci associated with AD with many loci supported by earlier GWAS studies^{72,73}. Mapping genes to the SNPs identified in the GWAS study enumerated a list of genes that have been investigated through pathway analysis, fine mapping, gene-based association analysis. AD risk-genes include *APOE*, *PICALM*, *BIN1*, *CRI* and *TREM2*; these risk genes and loci are often presented as a Manhattan plot⁷¹. Our study did not identify any of the AD risk genes as associated with the cognitive phenotype, however the phenotypes in the studies are different in terms of cognitive impairment instead of diagnosed AD and a comparison group of individuals resistant to cognitive impairment until late life (> 90 years) in comparison to cognitively normal individuals, typically aged 50–80. Interestingly, there was considerable overlap between the biochemical pathways identified in our study and the biochemical pathways associated with AD, notably immune system and inflammation^{71,74–77}. Calcium regulation is biological process involved with AD^{78–80} and the *RYR3* gene identified in our study is a key molecule involved in this process⁸¹. Lipid metabolism is a key pathway identified through pathway analysis of AD GWAS⁷¹ and *ESR1* polymorphisms have been shown to impact this pathway⁸², including lipid metabolism in the brain^{83,84}.

The predictive risk score (PRS) developed for this study predicted cognitive impairment with AUCs approximately 0.74. This level is similar to a risk score used to predict MCI vs. controls that had an AUC of 0.67. For pathologically-confirmed AD cases, as compared with clinical diagnosis, a PRS score demonstrated an AUC of 0.84⁸⁵, among the best performance for predictive biomarkers for AD that included imaging, biofluid and cerebrospinal fluid based measures.

Studying genetic and non-genetic factors that contribute to the absence or delayed onset of cognitive impairment in resilient individuals may inform lifestyle and therapeutic opportunities for intervention. The biological mechanisms involved in cognitive resilience are complex and our study supports roles for several key metabolic

pathways and genes. Strengths of this study include a large cohort comprised of elderly individuals where a substantial proportion demonstrated resilience to cognitive impairment with ages greater than 90 years and consideration of both genetic and environmental factors. Our study also has some limitations. Although the CLHLS cohort contains a large number of centenarians, the sample size is still relatively small for GWAS. Moreover, replication in an independent cohort would be essential to increase the statistical rigor of the genetic association results. Unfortunately, there is no equivalent dataset of individuals of similar ancestry where the dataset is enriched for centenarians and long-lived individuals, essential for our study design. In recognition of these limitations, we have interpreted the biological and gene-specific results in context of prior studies and also used a polygenic risk score to provide a single composite measure of risk from numerous small effect size genetic variants. These steps improve the likelihood for replication. The study represents one of few studies conducted in Han Chinese in contrast to numerous genetic studies of individuals of European ancestry. Future studies are needed to examine whether the results of our study will translate to individuals of other ancestries.

Methods

M1. Chinese longitudinal healthy longevity survey (CLHLS). We worked with the Chinese Longitudinal Healthy Longevity Survey (CLHLS) Series, which are publicly available from National Archive of Computerized Data on Aging (NACDA)⁸⁶. The survey data provides information about physical and mental health status, socioeconomic status, demographics of participants aged 65 and above from 22 provinces in China. The first wave of survey started in 1998, followed with six more waves conducted in 2000, 2002, 2005, 2008–2009, 2011–2012, and 2014. Genotype data is available for 4477 participants with coverage over 7,000,000 SNPs after imputation.

The CLHLS has the world's largest sample size of centenarians to date⁸⁷. Unlike the population structure of western countries, consisting of people with different ancestries, the population structure in China is comparably homogeneous, with fewer immigrants from other parts of the world. Additionally, even though China has 55 minority groups, Han Chinese (the Han group) accounts for 92% of the total population. To obtain homogeneous genetic information, the surveys only included Han Chinese. A study by Xu and Jin⁸⁸ has shown that the average differences in genetics among European population ($F_{st} = 0.009$) was much higher than that among the Han Chinese population ($F_{st} = 0.002$), based on Human Genome Diversity Panel data. Consequently, the nature of the CLHLS data is less likely to be affected by the population stratification than western cohorts.

We utilized the information from the last survey that each individual participated in before he/she was lost-to-follow-up or deceased. For example, if the participant was deceased in the 2005 survey wave, then we used the survey information from 2002 wave. As the final step, we matched participants from the survey with the 4477 individuals who have available genotypes. Since not all participants in the survey have DNA information and not all individuals with genotypes were engaged in the survey, out of 4477 individuals, we ended up having 2243 individuals as our sample. Notice that the final sample has 2165 individuals due to the selection of individuals with a phenotype of exceptional longevity & normal cognition. The composition of the study sample is shown in Fig. 2.

M2. Genetic data and quality controls. The participants were genotyped using the Illumina HumanOmniZhongHua-8 BeadChips, which was built by selecting optimized tag SNP content from the 1000 Genomes Project and the three HapMap phases⁸⁷. To further increase genome coverage, imputation analysis was performed to infer the genotypes of all SNPs ($MAF \geq 0.01$) using IMPUTE software version 2⁸⁹ and the 1000 Genomes Project integrated phase 1 release as reference panel. SNPs with a quality score (Rsq) of < 0.9 were discarded before analysis. Gene dosages from SNPs that were directly genotyped were used when available, otherwise imputed dosages were used. Therefore, missing values were possible for the directly genotyped SNPs. For SNPs with a substantial frequency of missing data, we used a proxy SNP in the PRS as detailed in M8. More details of the genotyping are articulated in the study by Zeng et al.⁸⁷.

Genotype data prior to quality control consists of 4477 individuals and 7,443,066 SNPs. We imposed quality controls on the set of genetic data using PLINK (v1.9)⁹⁰. The maximum per-SNP missing (`-geno`), Hardy-Weinberg disequilibrium p -value (`-hwe`), and maximum per-person missing (`-mind`) are set to default values, which are 0.1, 0.0001, and 0.05 respectively. The minimum minor allele frequency (`-maf`) is set to 0.05. Post quality control, 4,611,702 SNPs were available for analysis. All the subsequent analyses, such as regional stratification using PCA, frequency calculation, and GWAS, are all based on the genetic data that passed the quality controls.

M3. Non-genetic data imputation and outcome assessment. We used Mini-Mental State Exam (MMSE) score to measure the outcome—individuals' cognitive functions. The questions used to calculate MMSE in the Chinese version were carefully designed based on international standards⁹¹. Both internal⁹¹ and external⁹² assessments have been conducted and ensured the quality and validity of the questionnaire. The maximum MMSE score is 30. If a person scores 18 or above, he/she is scored as cognitively normal; otherwise, cognitively impaired. We chose 18 as the cutoff value based on several prior studies^{4,12} on cognitive impairment. For each participant in the survey, we calculated MMSE from each survey wave that he/she participated in, until the participant was lost-to-follow-up or passed away. If participants were not able to answer the survey questions—except for the question about number of years of education, we treated the answers as “Wrong” with corresponding 0 score.

There are missing values in some non-genetic factors. Count of missing values are listed in the Table 6, from which we could see that the missing rates are generally low. We imputed the missing values using the Multi-variate Imputation by Chained Equations (MICE) Package⁹³ in R (v3.6.2)⁹⁴. The imputations were done using CART (Classification and Regression Tree) to account for different types of the factors and potentially complex

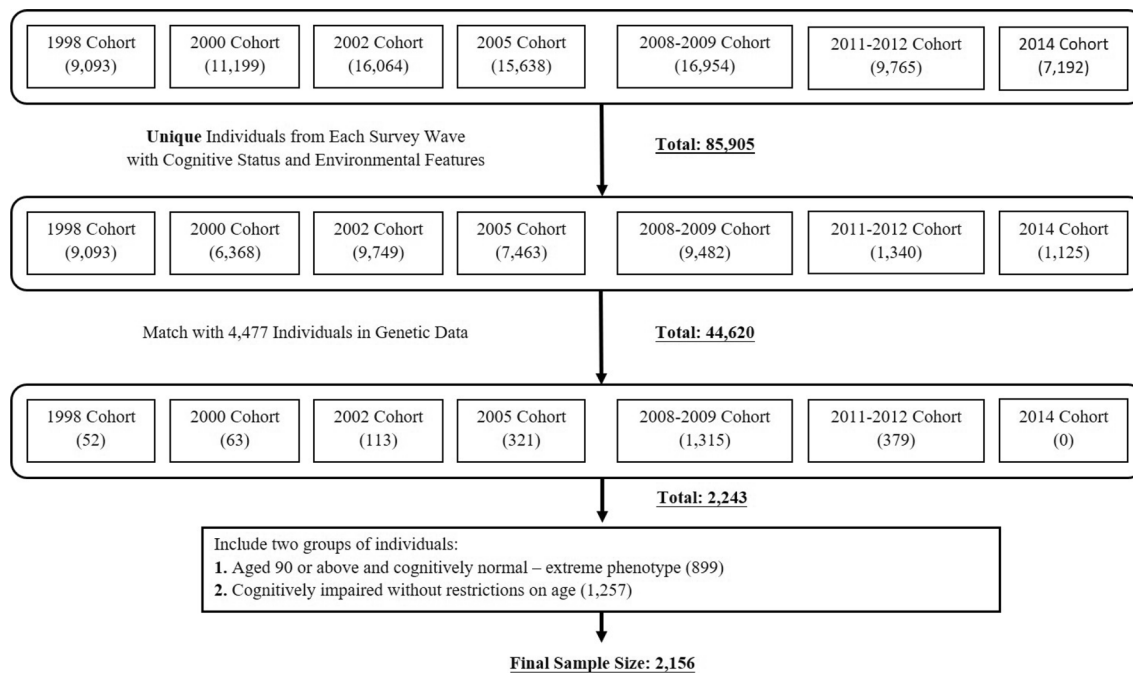


Figure 2. Composition of the study sample. The numbers in the parenthesis are the sample size for each survey cohort.

Variable	Count of missing value (%)
Education	10 (0.44%)
Occupation	126 (5.61%)
Marital status	11 (0.49%)
Co-residence	9 (0.40%)
Staple food	1 (0.04%)
Fruit intake	4 (0.18%)
Vegetables intake	5 (0.22%)
Current smoker	4 (0.18%)
Former smoke	1 (0.04%)
Current drinker	10 (0.45%)
Former drinker	2 (0.09%)
Exercise currently	19 (0.85%)
Hypertension	76 (3.39%)
Diabetes	87 (3.88%)
Heart	71 (3.17%)
Cardiovascular disease	75 (3.34%)
Respiratory	63 (2.81%)

Table 6. Count of missing values of non-genetic variables. Variables that do not have missing values are not listed.

relationships among the factors. The item and individual missing rates were checked, as well as the distribution of imputed values, to ensure the quality of imputation.

M4. Regional stratification using PCA. It is traditionally thought that China has two major regions—Northern and Southern China, based on geographical division, where people maintain different living habits under the influences of climates and natural resources. Empirically, the study by Xu et al.⁹⁵ have examined and showed internal differences between the Han Chinese population in Southern and Northern areas. Zeng et al.⁸⁷ worked with the same CLHLS data sets and confirmed that the top two eigenvectors are sufficient to adjust the population stratification. Therefore, to account for the potential variations in population characteristics from two regions, we conducted Principle Component Analysis (PCA) and extracted the top two eigenvectors cor-

responding to the first and second largest eigenvalues. The two eigenvectors are further used as covariates in the GWAS.

M5. Selection of exceptional longevity & normal cognition phenotype. The 2156 individuals from the sample were split into two groups for GWAS. The first group includes individuals that were aged 90 or above, but were still cognitively intact. We designate this group of people as individuals with the phenotype of exceptional longevity & normal cognition. The other group consists of individuals that are cognitively impaired without any age constraint. The term “extreme phenotype” comes from the study by Perez-Graia et al.¹⁴. They concluded that samples with extreme phenotype have abundant genetic information about risk or protective effects from SNPs. Considering the nature of the CLHLS cohort and the goal of our research, emphasizing participants with the phenotype of exceptional longevity & normal cognition can potentially help us identify biological mechanisms with strong effects on cognition. In contrast with previous GWAS studies on individuals from western countries, the historical events that happened in China in the twentieth century are drastically different, imposing some external effects on participants from CLHLS⁹¹. Therefore, the long-lived people in China might have enriched genetic backgrounds that interact with the non-genetic factors, such as environmental factors, to provide protective effects on CI, considering the brutal and harsh environments they have suffered from and survived in the past, such as national and civil wars, revolutions, and starvation.

M6. Genome-wide association study. The GWAS analysis was performed using logistic regression using PLINK(v1.9), conditioning on sex and the two eigenvectors extracted from the PCA. Both genetic variants and environmental factors have been shown to contribute to educational attainment⁹⁶. Studies have also demonstrated that genetic variants linked to education also predict longevity⁹⁷. Since the study design centers on the GWAS identifying genetic factors associated with cognitive impairment while analysis of environmental factors, including educational attainment is considered in the second phase, we included age, sex and population stratification as covariates in the GWAS and not educational attainment. This also removed the potential multiple confounding between cognitive measures, educational attainment and longevity at the level of the GWAS and is consistent with covariate adjustment in a recent large-scale study of genetic loci influencing general cognitive function⁹⁸.

Chagnon et al.⁹⁹ discussed two general approaches of how to choose the p -value threshold in order to identify the significant SNPs: (1) If the goal is to identify SNPs with stronger statistical power, then it is suggested that a more stringent p -value threshold be selected, which usually gives less than 100 SNPs. (2) On the other hand, if the scientific question is approached as a prediction problem where strong prediction power is desired, we should not ignore less significant SNPs with larger p -values. Aggregating the small contributions from each SNP could enhance the prediction results. Under this circumstance, researchers usually proceed with a relaxed threshold which generates hundreds more SNPs than a stringent p -value.

Even though our approach is prediction-based, our primary focus is on the investigation of statistically important genetic and non-genetic factors in affecting CI. Therefore, we used 100 as the checkpoint for the number of significant SNPs from GWAS. We experimented with three p -value thresholds— 1×10^{-5} , 2×10^{-5} , and 3×10^{-5} , which gave us 15, 41, and 76 significant SNPs respectively. Using thresholds of 4×10^{-5} and 5×10^{-5} gave us 120 and 170 SNPs correspondingly, which are more than the checkpoint 100 so we stopped at 3×10^{-5} . Those are preliminary SNPs with potentially high linkage disequilibrium. We used the `-clump` command in PLINK (default setting on three parameters: `-clump-p2 0.01`, `-clump-r2 0.5`, `-clump-kb 250`) to clump them into smaller subsets of independent SNPs. The numbers of clumped SNPs are 7, 19, and 28 corresponding to the three p -values. All the analyses, including constructing the PRS and machine learning models and biological pathway analysis, were based on the clumped sets of SNPs.

M7. Missing dosage replacement. To account for substantial missing values for SNPs directly genotyped, we replaced the missing values with the dosage values from proxy SNPs using the following approach:

- For those SNPs not in the dbSNP database, such that they do not have any reference number (“rs” number), we use Kaviar (v160204-Public)¹⁰⁰ software to map the location with the corresponding reference number. We used the hg19(GRCh37) coordinate system.
- For each significant SNP, we looked up proxy SNPs using LDlink (v4.1.0)¹⁰¹. The reference panel was the population of Han Chinese in Beijing, China. All the proxy SNPs were checked to be on the same chromosome as the original SNP. Additionally, as discussed by Chagnon et al.⁹⁹, proxy SNPs with $R^2 \geq 0.8$ are considered to be very good proxies. Therefore, we only include proxies with $R^2 \geq 0.8$.

For each significant SNP, we created two reference panels. The first panel, “all_panel”, contains all the proxy SNPs for each target SNP. Those proxy SNPs are listed in descending order in terms of R^2 , so that when we replaced missing dosage values, we always checked the proxy with largest R^2 first. The second panel, “reverse_panel”, contains proxies that have different minor alleles from the target SNPs, adjusting the minor allele for the proxies accordingly. For instance, if the minor/major allele of the original SNP is A/T, while the proxy SNP has T/A, then we put this proxy SNP into the reverse panel. Note that the “reverse_panel” is a subset of the “all_panel”. When replacing the missing dosage value, if the proxy SNP is in the “reverse_panel”, instead of directly using the dosage value from the proxy SNP, we adjust the dosage to the allele count.

- For each individual and for each significant SNP where the dosage value is missing, we replaced it with the dosage value from the best (highest R^2) and available (existing in our genome data) proxy SNP. If the best available proxy is in the “reverse_panel”, then we used (2-dosage value).

- All remaining missing dosage values were set to zero. The underlying assumption for this is that we expect that most of the population have at least one copy of the major allele, which is best estimated by the 0 dosage value.

M8. PRS construction. PRS can be viewed as a simple model used to predict the risk for genetically complex diseases. PRS is calculated by aggregating SNP-level information from GWAS to “account for the phenotypic variation observed in complex traits, by assuming an additive, non-multiplicative, effect of multiple variants with variable effect sizes” (Ibanez et al. 2019)¹⁰². This aggregated information has stronger predictive power for complex diseases than individual genetic variants, notably when effect sizes are small to moderate. With the beta coefficient of each significant SNP_{*i*} and the corresponding dosage value for each individual, we constructed a PRS for each individual *j* using the formula:

$$\text{PRS}_j = \sum_{i=1}^n \beta_i * \text{dosage}_{ij}$$

where β_i is the beta coefficient for SNP_{*i*} and dosage_{*ij*} is the corresponding number of effect alleles from the individual *j*. The PRS can be viewed as the summation of genetic effects that each participant possesses, associated with the underlying probability of developing CI. Then, PRS was used as a single predictor to predict the likelihood of CI. We constructed PRS using three different *p*-value thresholds and assessed their performances correspondingly. We applied fivefold cross validation to evaluate the PRS’s performance. The entire data was split into fivefolds. We evaluated the performance of PRS five times for each of the fivefolds, using Area Under the Curve (AUC) and reported the average of the results.

The PRS derived from the CLHLS GWAS analysis was tested for association with the equivalent phenotype in an independent cohort, the Rush ROSMAP (Religious Order Study/Memory and Aging Project). Participants came from two community-based longitudinal cohort studies of diverse participants, the Religious Orders Study (ROS)¹⁸, the Rush Memory and Aging Project (MAP)¹⁰³. The studies were approved by the Institutional Review Board of Rush University Medical Center. Participants were enrolled without known dementia and each participant signed an informed consent and agreed to annual clinical evaluations. The ROS and MAP participants were predominantly white Americans. Importantly, both studies were conducted by the same team of investigators and share a large common core of testing batteries and uniform structured clinical evaluations. This makes it possible for a combined analysis. The same phenotype definition, based on survival to age 90 years with intact cognition versus cognitive impairment at any age was used, based on the MMSE scores obtained for the participants. Descriptive statistics for the cohort are included in Supplementary Table 3.

M9. Machine learning models. In addition to PRS, we considered machine learning models to predict the phenotype—binary cognitive status, using both genetic and non-genetic factors and to identify important predictive factors by assessing feature importance. We examined five machine learning models— ℓ_2 penalized logistic regression, classification and regression tree (CART), support vector machine (SVM), random forest, and boosted decision trees (XGBoost implementation). Five models were utilized so that we were able to compare their prediction performances and evaluate feature importance, which could potentially reduce the bias from simply assessing one model and help us pinpoint important factors that are robust and congruent from different models.

- *ℓ_2 Penalized Logistic Regression* an ℓ_2 penalty term is applied to the sum of squared coefficients from logistic regression to prevent over-fitting.
- *Classification and Regression Tree (CART)*¹⁰⁴: the algorithm constructs binary decision trees by splitting input features on certain values that maximizes information gain at each node. It stops when a criterion is satisfied. We controlled the depth of the tree to prevent over-fitting.
- *Support Vector Machine (SVM)*¹⁰⁵ the algorithm generates a hyperplane to separate binary response variables by maximizing the total distances between the hyperplane and all data points. We assessed linear kernel, polynomial kernel, and radial basis kernel. The model with linear kernel is simple but had the best performance in our case.
- *Random Forest*¹⁰⁶ random forest is an ensemble of multiple decision trees. The implementation we used takes the poll of predictions from individual trees as the final predictive result. Random forest is considered a complex model that possess strong predictive power.
- *Boosted Decision Trees*¹⁰⁷ an algorithm that combines weak tree classifiers into a strong classifier using boost approach. The implementation we applied is XGBoost¹⁰⁸. The model is considered complex.

To be consistent with the process of assessing the PRS’s performances, we applied a fivefold cross validation procedure on the entire data to select the best hyper-parameters. One fold was used as the validation set and the other four folds were involved in parameter tuning and model constructing. We repeated the process until each one of the fivefolds was used as the validation set once. Then we constructed the final model on the entire data using the best parameters and calculated the feature importance. According to the official documentation from scikit-learn package¹⁰⁹, the incorporated `feature_importance` function might inflate the importance of numerical variables using impurity-based measurement. Therefore, we adopted the suggested `permutation_importance`, though we did not notice significant differences between the two approaches. All the modeling procedures were conducted in Python (v3.7)¹¹⁰ with publicly available packages.

Note that we used the entire data as training set, without constructing a holdout test set. This step is taken to minimize the loss of power for statistical analysis considering our sample size. Since we carefully tuned hyper-parameters in the cross-validation process to protect against over-fitting, we think that the final model constructed on the entire data is generalizable.

M10. Functional mapping and biological signature analysis. Functional annotation of the GWAS results was performed using the SNPnexus (v4.0) platform^{111–114} designed for annotation and interpretation of sets of SNPs derived from GWAS. SNPnexus maps SNPs in genomic loci to genes using positional mapping based on maximum distances between SNPs and genes. Biological function of the genes was annotated based on the Genetic Association Database and through literature searches. Biological signature analysis was carried out with the GENE2FUNC (v1.3.6) algorithm implemented in Functional Mapping and Annotation of Genome-Wide Association Studies (FUMA, v1.3.5)¹¹⁵. Of the 31 input genes, 21 were identified with a unique Entrez¹¹⁶ identification number and all genes with an Entrez identification number (35,142) were used as the background gene set for the hypergeometric test. The Molecular Signatures Database database^{39,117} (v7.0, August 2019) was used for the set of potential biological signatures. The Benjamini-Hochberg method was used as a correction for multiple testing with a maximum adjusted p -value of 0.05 for gene-set enrichment tests.

Ethical approval. The Research Ethics Committees of Duke University and Peking University granted approval for the Protection of Human Subjects for the Chinese Longitudinal Healthy Longevity Survey, including collection of DNA samples used for the present study. The survey respondents who contributed their DNA samples gave informed consent before participation. All of the GWAS experiments and methods of analyses in the present study were performed in accordance with relevant guidelines and regulations. The Religious Order Study and the Rush Memory and Aging Project were approved by an Institutional Review Board of Rush University Medical Center. All subjects signed an informed consent, an Anatomic Gift Act, and a repository consent to allow their biospecimens and data to be used for ancillary studies. Conduct of the studies was performed in accordance with relevant guidelines/regulations set forth by the Rush University Medical Center.

Data availability

The Chinese Longitudinal Healthy Longevity Survey (CLHLS) data are publicly available from National Archive of Computerized Data on Aging (NACDA). The genetic data can be obtained by requesting from the corresponding author Dr. Yi Zeng. Data on the ROSMAP studies can be requested at www.radc.rush.edu.

Received: 25 May 2020; Accepted: 12 October 2020

Published online: 05 November 2020

References

1. CDC. Cognitive impairment: A call for action, now! (2011). https://www.cdc.gov/aging/pdf/cognitive_impairment/cogimp_poili_cy_final.pdf. Accessed 21 Feb 2020.
2. Hugo, J. & Ganguli, M. Dementia and cognitive impairment: epidemiology, diagnosis, and treatment. *Clin. Geriatr. Med.* **30**, 421–442. <https://doi.org/10.1016/j.cger.2014.04.001> (2014).
3. Tsolaki, M. Clinical workout for the early detection of cognitive decline and dementia. *Eur. J. Clin. Nutr.* **68**, 1186–1191. <https://doi.org/10.1038/ejcn.2014.189> (2014).
4. Gao, M. *et al.* The time trends of cognitive impairment incidence among older Chinese people in the community: based on the clhls cohorts from 1998 to 2014. *Age Ageing* **46**, 787–793. <https://doi.org/10.1093/ageing/afx038> (2017).
5. Davies, G. *et al.* A genome-wide association study implicates the apoe locus in nonpathological cognitive ageing. *Mol. Psychiatry* <https://doi.org/10.1038/mp.2012.159> (2013).
6. United Nations. World Population Ageing (2015). https://www.un.org/en/development/desa/population/publications/pdf/ageing/WPA2015_Report.pdf. Accessed 21 Feb 2020.
7. Alzheimer's Disease International. World Alzheimer Report 2015: The Global Impact of Dementia; an Analysis of Prevalence, Incidence, Cost and Trends (2015). <https://www.alz.co.uk/research/WorldAlzheimerReport2015.pdf>. Accessed 21 Feb 2020.
8. Comas-Herrera, A., Wittenberg, R., Pickard, L. & Knapp, M. Cognitive impairment in older people: future demand for long-term care services and the associated costs. *Int. J. Geriatr. Psychiatry* **22**, 1037–1045. <https://doi.org/10.1002/gps.1830> (2007).
9. Association, A. 2010 alzheimer's disease facts and figures. *Alzheimer's & Dement.* **6**, 158–194. <https://doi.org/10.1016/j.jalz.2010.01.009> (2010).
10. Lv, X. *et al.* Cognitive decline and mortality among community-dwelling Chinese older people. *BMC Med.* **17**, 63. <https://doi.org/10.1186/s12916-019-1295-8> (2019).
11. Casanova, R. *et al.* Investigating predictors of cognitive decline using machine learning. *J. Gerontol. Ser. B* **75**, gby054. <https://doi.org/10.1093/geronb/gby054> (2020).
12. Zhang, Z. Gender differential in cognitive impairment and decline of the oldest old in china. *J. Gerontol. Ser. B* **61**, 107–115. <https://doi.org/10.1093/geronb/61.2.S107> (2006).
13. Lutz, M. W. *et al.* Analysis of pleiotropic genetic effects on cognitive impairment, systemic inflammation, and plasma lipids in the health and retirement study. *Neurobiol. Aging* **80**, 173–186. <https://doi.org/10.1016/j.neurobiolaging.2018.10.028> (2019).
14. Pérez-Gracia, J. L. *et al.* Selection of extreme phenotypes: the role of clinical observation in translational research. *Clin. Transl. Oncol.* **12**, 174–180. <https://doi.org/10.1007/s12094-010-0487-7> (2010).
15. Estep, P. *The Mindspan Diet: Reduce Alzheimer's Risk, Minimize Memory Loss, and Keep Your Brain Young* (Ballantine Books, New York, 2016).
16. An, R. & Liu, G. G. Cognitive impairment and mortality among the oldest-old Chinese. *Geriatr. Psychiatry* <https://doi.org/10.1002/gps.4442> (2016).
17. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
18. Bennett, D. A., Schneider, J. A., Arvanitakis, Z. & Wilson, R. S. Overview and findings from the religious orders study. *Curr. Alzheimer Res.* **9**, 628–645 (2012).
19. Au, B., Dale-McGrath, S. & Tierney, M. C. Sex differences in the prevalence and incidence of mild cognitive impairment: a meta-analysis. *Ageing Res. Rev.* **35**, 176–199. <https://doi.org/10.1016/j.arr.2016.09.005> (2017).

20. Li, R. & Singh, M. Sex differences in cognitive impairment and alzheimer's disease. *Front. Neuroendocrinol.* **35**, 385–403. <https://doi.org/10.1016/j.yfrne.2014.01.002> (2014). (**Sex Differences in Neurological and Psychiatric Disorders**)
21. An, Y. *et al.* Patterns of cognitive function in middle-aged and elderly Chinese adults—findings from the emcoa study. *Alzheimer's Res. Therapy* <https://doi.org/10.1186/s13195-018-0421-8> (2018).
22. Panza, F. *et al.* Alcohol consumption in mild cognitive impairment and dementia: harmful or neuroprotective?. *Geriatr. Psychiatry* <https://doi.org/10.1002/gps.3772> (2012).
23. Scarmeas, N. *et al.* Physical activity, diet, and risk of alzheimer disease. *JAMA* **302**, 627–637. <https://doi.org/10.1001/jama.2009.1144> (2012).
24. Eid, A., Mhatre, I. & Richardson, J. R. Gene-environment interactions in alzheimer's disease: a potential path to precision medicine. *Pharmacol. Therap.* **199**, 173–187. <https://doi.org/10.1016/j.pharmthera.2019.03.005> (2019).
25. Larsson, S. C. *et al.* Modifiable pathways in alzheimer's disease: mendelian randomisation analysis. *BMJ* <https://doi.org/10.1136/bmj.j5375> (2017).
26. Tangney, C. C. *et al.* Relation of dash- and mediterranean-like dietary patterns to cognitive decline in older persons. *Neurology* **83**, 1410–1416. <https://doi.org/10.1212/WNL.0000000000000884> (2014).
27. Morris, M. C. *et al.* Mind diet slows cognitive decline with aging. *Alzheimer's Dement.* **11**, 1015–1022. <https://doi.org/10.1016/j.jalz.2015.04.011> (2015).
28. Samieri, C. *et al.* Fish intake, genetic predisposition to alzheimer disease, and decline in global cognition and memory in 5 cohorts of older persons. *Am. J. Epidemiol.* <https://doi.org/10.1093/aje/kwx330> (2018).
29. Scarmeas, N., Anastasiou, C. & Yannakoulia, M. Nutrition and prevention of cognitive impairment. *Lancet Neurol.* [https://doi.org/10.1016/S1474-4422\(18\)30338-7](https://doi.org/10.1016/S1474-4422(18)30338-7) (2018).
30. Dong, L. *et al.* Diet, lifestyle and cognitive function in old Chinese adults. *Arch. Gerontol. Geriatr.* **63**, 36–42. <https://doi.org/10.1016/j.archger.2015.12.003> (2016).
31. Wu, J. *et al.* Dietary pattern in midlife and cognitive impairment in late life: a prospective study in Chinese adults. *Am. J. Clin. Nutr.* <https://doi.org/10.1093/ajcn/nqz150> (2019).
32. Wang, Z. *et al.* Is there an association between mild cognitive impairment and dietary pattern in Chinese elderly? Results from a cross-sectional population study. *BMC Public Health* <https://doi.org/10.1186/1471-2458-10-595> (2010).
33. Qin, B. *et al.* Dietary patterns and cognitive decline among Chinese older adults. *Epidemiology* <https://doi.org/10.1097/EDE.0000000000000338> (2015).
34. Kirk-Sanchez, N. J. & McGough, E. L. Physical exercise and cognitive performance in the elderly: current perspectives. *Clin. Interv. Aging* **1**, 2. <https://doi.org/10.2147/CLIA.S39506> (2014).
35. Baker, L. D. *et al.* Effects of aerobic exercise on mild cognitive impairment. *Arch. Neurol.* <https://doi.org/10.1001/archneurol.2009.307> (2010).
36. Laurin, D., Verreault, R. & Lindsay, J. Physical activity and risk of cognitive impairment and dementia in elderly persons. *Arch. Neurol.* <https://doi.org/10.1001/archneur.58.3.498> (2001).
37. Geda, Y. E. *et al.* Physical exercise, aging, and mild cognitive impairment a population-based study. *Arch. Neurol.* <https://doi.org/10.1001/archneurol.2009.297> (2010).
38. Young, J., Angevaren, M., Rusted, J. & Tabet, N. Aerobic exercise to improve cognitive function in older people without known cognitive impairment. *Cochrane Database Syst. Rev.* <https://doi.org/10.1002/14651858.CD005381.pub4> (2015).
39. Liberzon, A. *et al.* Molecular signatures database (msigdb) 3.0. *Bioinformatics* **27**, 1739–40. <https://doi.org/10.1093/bioinformatics/btr260> (2011).
40. Felsky, D. *et al.* Polygenic analysis of inflammatory disease variants and effects on microglia in the aging brain. *Mol. Neurodegener.* **13**, 38. <https://doi.org/10.1186/s13024-018-0272-6> (2018).
41. Filiano, A. J., Gadani, S. P. & Kipnis, J. Interactions of innate and adaptive immunity in brain development and function. *Brain Res.* **1617**, 18–27. <https://doi.org/10.1016/j.brainres.2014.07.050> (2015).
42. Yaffe, K. *et al.* Estrogen receptor genotype and risk of cognitive impairment in elders: findings from the health abc study. *Neurobiol. Aging* **30**, 607–14. <https://doi.org/10.1016/j.neurobiolaging.2007.08.003> (2009).
43. Gutierrez, A. & Vitorica, J. Toward a new concept of alzheimer's disease models: a perspective from neuroinflammation. *J. Alzheimer's Dis.* **64**, S329–S338. <https://doi.org/10.3233/JAD-179914> (2018).
44. Gironi, M. *et al.* A global immune deficit in alzheimer's disease and mild cognitive impairment disclosed by a novel data mining process. *J. Alzheimer's Dis.* **43**, 1199–213. <https://doi.org/10.3233/JAD-141116> (2015).
45. Di Benedetto, S., Muller, L., Wenger, E., Duzel, S. & Pawelec, G. Contribution of neuroinflammation and immunity to brain aging and the mitigating effects of physical and cognitive interventions. *Neurosci. Biobehav. Rev.* **75**, 114–128. <https://doi.org/10.1016/j.neubiorev.2017.01.044> (2017).
46. Patterson, S. L. Immune dysregulation and cognitive vulnerability in the aging brain: interactions of microglia, il-1beta, bdnf and synaptic plasticity. *Neuropharmacology* **96**, 11–8. <https://doi.org/10.1016/j.neuropharm.2014.12.020> (2015).
47. Brandi, M. L. *et al.* Association of the estrogen receptor alpha gene polymorphisms with sporadic alzheimer's disease. *Biochem. Biophys. Res. Commun.* **265**, 335–338. <https://doi.org/10.1006/bbrc.1999.1665> (1999).
48. Yaffe, K., Lui, L. Y., Grady, D., Stone, K. & Morin, P. Estrogen receptor 1 polymorphisms and risk of cognitive impairment in older women. *Biol. Psychiatry* **51**, 677–82. [https://doi.org/10.1016/s0006-3223\(01\)01289-6](https://doi.org/10.1016/s0006-3223(01)01289-6) (2002).
49. Xing, Y., Jia, J. P., Ji, X. J. & Tian, T. Estrogen associated gene polymorphisms and their interactions in the progress of alzheimer's disease. *Prog. Neurobiol.* **111**, 53–74. <https://doi.org/10.1016/j.pneurobio.2013.09.006> (2013).
50. Mattila, K. M. *et al.* Interaction between estrogen receptor 1 and the epsilon4 allele of apolipoprotein e increases the risk of familial alzheimer's disease in women. *Neurosci. Lett.* **282**, 45–48. [https://doi.org/10.1016/s0304-3940\(00\)00849-1](https://doi.org/10.1016/s0304-3940(00)00849-1) (2000).
51. Li, X. *et al.* The etiological effect of a new low-frequency esr1 variant on mild cognitive impairment and alzheimer's disease: a population-based study. *Aging (Albany NY)* **10**, 2316–2337 (2018).
52. Chen, L. H. *et al.* Genetic polymorphisms in estrogen metabolic pathway associated with risks of alzheimer's disease: evidence from a southern Chinese population. *J. Am. Geriatr. Soc.* **65**, 332–339. <https://doi.org/10.1111/jgs.14537> (2017).
53. Chaves, A. C. *et al.* Estrogen receptor-alpha gene xba i > g polymorphism influences short-term cognitive decline in healthy oldest-old individuals. *Arq. Neuropsiquiatr.* **75**, 172–175. <https://doi.org/10.1590/0004-282X20170018> (2017).
54. Hakamata, Y., Nakai, J., Takeshima, H. & Imoto, K. Primary structure and distribution of a novel ryanodine receptor/calcium release channel from rabbit brain. *FEBS Lett.* **312**, 229–35. [https://doi.org/10.1016/0014-5793\(92\)80941-9](https://doi.org/10.1016/0014-5793(92)80941-9) (1992).
55. Khachaturian, Z. S. Calcium hypothesis of alzheimer's disease and brain aging. *Ann. N. Y. Acad. Sci.* **747**, 1–11. <https://doi.org/10.1111/j.1749-6632.1994.tb44398.x> (1994).
56. Stutzmann, G. E. The pathogenesis of alzheimer's disease is it a lifelong "calciumopathy"?. *Neuroscientist* **13**, 546–559. <https://doi.org/10.1177/1073858407299730> (2007).
57. Chakraborty, S., Goussakov, I., Miller, M. B. & Stutzmann, G. E. Deviant ryanodine receptor-mediated calcium release resets synaptic homeostasis in presymptomatic 3xtg-ad mice. *J. Neurosci.* **29**, 9458–70. <https://doi.org/10.1523/JNEUROSCI.2047-09.2009> (2009).
58. Zhang, H., Sun, S., Herreman, A., De Strooper, B. & Bezprozvanny, I. Role of presenilins in neuronal calcium homeostasis. *J. Neurosci.* **30**, 8566–80. <https://doi.org/10.1523/JNEUROSCI.1554-10.2010> (2010).

59. Bruno, A. M. *et al.* Altered ryanodine receptor expression in mild cognitive impairment and alzheimer's disease. *Neurobiol. Aging* **33**(1001), e1–e6. <https://doi.org/10.1016/j.neurobiolaging.2011.03.011> (2012).
60. Guzman, Y. F. *et al.* A gain-of-function mutation in the grk2 gene causes neurodevelopmental deficits. *Neurol. Genet.* **3**, e129. <https://doi.org/10.1212/NXG.000000000000129> (2017).
61. Choi, K. H., Zepp, M. E., Higgs, B. W., Weickert, C. S. & Webster, M. J. Expression profiles of schizophrenia susceptibility genes during human prefrontal cortical development. *J. Psychiatry Neurosci.* **34**, 450–8 (2009).
62. Jamain, S. *et al.* Linkage and association of the glutamate receptor 6 gene with autism. *Mol. Psychiatry* **7**, 302–10. <https://doi.org/10.1038/sj.mp.4000979> (2002).
63. Broer, L. *et al.* Gwas of longevity in charge consortium confirms apoe and foxo3 candidacy. *J. Gerontol. A Biol. Sci. Med. Sci.* **70**, 110–8. <https://doi.org/10.1093/gerona/glu166> (2015).
64. Lourenco, A. B., Munoz-Jimenez, C., Venegas-Caleron, M. & Artal-Sanz, M. Analysis of the effect of the mitochondrial prohibitin complex, a context-dependent modulator of longevity, on the *c. elegans* metabolome. *Biochim. Biophys. Acta* **1847**, 1457–68. <https://doi.org/10.1016/j.bbabi.2015.06.003> (2015).
65. Merkwirth, C. *et al.* Prohibitins control cell proliferation and apoptosis by regulating opa1-dependent cristae morphogenesis in mitochondria. *Genes Dev.* **22**, 476–88. <https://doi.org/10.1101/gad.460708> (2008).
66. Kasashima, K., Sumitani, M., Satoh, M. & Endo, H. Human prohibitin 1 maintains the organization and stability of the mitochondrial nucleoids. *Exp. Cell Res.* **314**, 988–96. <https://doi.org/10.1016/j.yexcr.2008.01.005> (2008).
67. Merkwirth, C. *et al.* Loss of prohibitin membrane scaffolds impairs mitochondrial architecture and leads to tau hyperphosphorylation and neurodegeneration. *PLoS Genet.* **8**, e1003021. <https://doi.org/10.1371/journal.pgen.1003021> (2012).
68. Schleit, J. *et al.* Molecular mechanisms underlying genotype-dependent responses to dietary restriction. *Aging Cell* **12**, 1050–61. <https://doi.org/10.1111/accel.12130> (2013).
69. Zietara, N. *et al.* Absence of ifn-beta impairs antigen presentation capacity of splenic dendritic cells via down-regulation of heat shock protein 70. *J. Immunol.* **183**, 1099–109. <https://doi.org/10.4049/jimmunol.0803214> (2009).
70. Alzheimer Association. What is alzheimer's disease? <https://www.alz.org/alzheimers-dementia/what-is-alzheimers>. Accessed 21 Feb 2020.
71. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates abeta, tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430. <https://doi.org/10.1038/s41588-019-0358-2> (2019).
72. Hollingworth, P. *et al.* Common variants at abca7, ms4a6a/ms4a4e, epha1, cd33 and cd2ap are associated with alzheimer's disease. *Nat. Genet.* **43**, 429–35. <https://doi.org/10.1038/ng.803> (2011).
73. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease. *Nat. Genet.* **45**, 1452–1458. <https://doi.org/10.1038/ng.2802> (2013).
74. Allen, M. *et al.* Novel late-onset alzheimer disease loci variants associate with brain gene expression. *Neurology* **79**, 221–8. <https://doi.org/10.1212/WNL.0b013e3182605801> (2012).
75. Karch, C. M., Cruchaga, C. & Goate, A. M. Alzheimer's disease genetics: from the bench to the clinic. *Neuron* **83**, 11–26. <https://doi.org/10.1016/j.neuron.2014.05.041> (2014).
76. Novikova, G. *et al.* Integration of alzheimer's disease genetics and myeloid cell genomics identifies novel causal variants, regulatory elements, genes and pathways. *bioRxiv* 694281. <https://doi.org/10.1101/694281> (2019).
77. Pimenova, A. A., Raj, T. & Goate, A. M. Untangling genetic risk for alzheimer's disease. *Biol. Psychiatry* **83**, 300–310. <https://doi.org/10.1016/j.biopsych.2017.05.014> (2018).
78. Alzheimer's Association Calcium Hypothesis, W. Calcium hypothesis of alzheimer's disease and brain aging: a framework for integrating new evidence into a comprehensive theory of pathogenesis. *Alzheimer's Dement* **13**, 178–182. <https://doi.org/10.1016/j.jalz.2016.12.006> (2017).
79. Gibson, G. E. & Thakkar, A. Interactions of mitochondria/metabolism and calcium regulation in alzheimer's disease: a calcein point of view. *Neurochem. Res.* **42**, 1636–1648. <https://doi.org/10.1007/s11064-017-2182-3> (2017).
80. Tong, B. C., Wu, A. J., Li, M. & Cheung, K. H. Calcium signaling in alzheimer's disease and therapies. *Biochim. Biophys. Acta Mol. Cell Res.* **1865**, 1745–1760. <https://doi.org/10.1016/j.bbamcr.2018.07.018> (2018).
81. Kushnir, A., Wajsborg, B. & Marks, A. R. Ryanodine receptor dysfunction in human disorders. *Biochim. Biophys. Acta Mol. Cell Res.* **1687–1697**, 2018. <https://doi.org/10.1016/j.bbamcr.2018.07.011> (1865).
82. Gomes-Rochette, N. F. *et al.* Association of pvuii and xbai polymorphisms on estrogen receptor alpha (esr1) gene to changes into serum lipid profile of post-menopausal women: effects of aging, body mass index and breast cancer incidence. *PLoS ONE* **12**, e0169266. <https://doi.org/10.1371/journal.pone.0169266> (2017).
83. Morselli, E. *et al.* Impact of estrogens and estrogen receptor-alpha in brain lipid metabolism. *Am. J. Physiol. Endocrinol. Metab.* **315**, E7–E14. <https://doi.org/10.1152/ajpendo.00473.2017> (2018).
84. Barros, R. P. & Gustafsson, J. A. Estrogen receptors and the metabolic network. *Cell Metab.* **14**, 289–99. <https://doi.org/10.1016/j.cmet.2011.08.005> (2011).
85. Escott-Price, V., Myers, A. J., Huentelman, M. & Hardy, J. Polygenic risk score analysis of pathologically confirmed alzheimer disease. *Ann. Neurol.* **82**, 311–314. <https://doi.org/10.1002/ana.24999> (2017).
86. Zeng, Y., Vaupel, J., Xiao, Z., Liu, Y. & Zhang, C. Chinese longitudinal healthy longevity survey (chls). *Inter-Univ. Consort. Polit. Soc. Res.* <https://doi.org/10.3886/ICPSR36692.v1> (2017).
87. Zeng, Y. *et al.* Novel loci and pathway significantly associated with longevity. *Sci. Rep.* **6**, 21243. <https://doi.org/10.1038/srep21243> (2016).
88. Xu, S. & Jin, L. A genome-wide analysis of admixture in uyghurs and a high-density admixture map for disease-gene discovery. *Am. J. Hum. Genet.* **83**, 322–336. <https://doi.org/10.1016/j.ajhg.2008.08.001> (2008).
89. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529. <https://doi.org/10.1371/journal.pgen.1000529> (2009).
90. Purcell, S. *et al.* Plink: a toolset for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.* <https://doi.org/10.1086/519795> (2007).
91. Zeng, Y. Towards deeper research and better policy for healthy aging—using the unique data of Chinese longitudinal healthy longevity survey. *China Econ. J.* **5**, 131–149. <https://doi.org/10.1080/17538963.2013.764677> (2012).
92. Goodkind, D. Review on the book healthy longevity in China: demographic, socioeconomic, and psychological dimensions. *Popul Stud.* **63**, 1–7. <https://doi.org/10.1080/00324720903216903> (2009).
93. van Buuren, S. & Groothuis-Oudshoorn, K. mice: multivariate imputation by chained equations in r. *J. Stat. Softw.* **45**, 1–67 (2011).
94. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2017).
95. Xu, S. *et al.* Genomic dissection of population substructure of han Chinese and its implication in association studies. *Am. J. Hum. Genet.* **85**, 762–774. <https://doi.org/10.1016/j.ajhg.2009.10.015> (2009).
96. Ding, X., Barban, N., Tropf, F. C. & Mills, M. C. The relationship between cognitive decline and a genetic predictor of educational attainment. *Soc. Sci. Med.* **239**, 112549. <https://doi.org/10.1016/j.socscimed.2019.112549> (2019).
97. Marioni, R. E. *et al.* Genetic variants linked to education predict longevity. *Proc. Natl. Acad. Sci. USA* **113**, 13366–13371. <https://doi.org/10.1073/pnas.1605334113> (2016).

98. Davies, G. *et al.* Study of 300,486 individuals identifies 148 independent genetic loci influencing general cognitive function. *Nat. Commun.* **9**, 2098. <https://doi.org/10.1038/s41467-018-04362-x> (2018).
99. Chagnon, M., O'Loughlin, J., Engert, J. C., Karp, I. & Sylvestre, M.-P. Missing single nucleotide polymorphisms in genetic risk scores: a simulation study. *PLoS ONE* **13**(7), e0200630. <https://doi.org/10.1371/journal.pone.0200630> (2018).
100. Glusman, G., Caballero, J., Mauldin, D. E., Hood, L. & Roach, J. C. Kaviar: an accessible system for testing snv novelty. *Bioinformatics* **27**, 3216–3217. <https://doi.org/10.1093/bioinformatics/btr540> (2011).
101. Machiela, M. J. & Chanock, S. J. Ldlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).
102. Ibanez, L., Farias, F., Dube, U., Mihindukulasuriya, K. & Harari, O. Polygenic risk scores in neurodegenerative diseases: a review. *Curr. Genet. Med. Rep.* **7**, 22–29. <https://doi.org/10.1007/s40142-019-0158-0> (2019).
103. Bennett, D. A. *et al.* Overview and findings from the rush memory and aging project. *Curr. Alzheimer Res.* **9**, 646–63 (2012).
104. Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. *Classification and Regression Trees* (CRC Press, Boca Raton, 1984).
105. Vapnik, V. & Chervonenkis, A. A note on one class of perceptrons. *Autom. Remote Control* **25**, 103–109 (1964).
106. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
107. Freund, Y. & Schapire, R. E. A short introduction to boosting. *J. Jpn. Soc. Artif. Intell.* **14**, 771–780 (1999).
108. Chen, T. & Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, KDD '16, 785–794. <https://doi.org/10.1145/2939672.2939785>. (Association for Computing Machinery, New York, NY, USA, 2016).
109. Permutation importance vs random forest feature importance (mdi). https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance.html#sphx-glr-auto-examples-inspection-plot-permutation-importance-py. Accessed 21 Feb 2020.
110. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual* (Scotts Valley, CA, 2009).
111. Chelala, C., Khan, A. & Lemoine, N. R. Snpnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* **25**, 655–61. <https://doi.org/10.1093/bioinformatics/btn653> (2009).
112. Dayem Ullah, A. Z., Lemoine, N. R. & Chelala, C. Snpnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucl. Acids Res.* **40**(W65–70), 2012. <https://doi.org/10.1093/nar/gks364> (2012).
113. Dayem Ullah, A. Z., Lemoine, N. R. & Chelala, C. A practical guide for the functional annotation of genetic variations using snpnexus. *Brief Bioinform.* **14**, 437–47. <https://doi.org/10.1093/bib/bbt004> (2013).
114. Dayem Ullah, A. Z. *et al.* Snpnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucl. Acids Res.* **46**, W109–W113. <https://doi.org/10.1093/nar/gky399> (2018).
115. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with fuma. *Nat. Commun.* **8**, 1826. <https://doi.org/10.1038/s41467-017-01261-5> (2017).
116. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez gene: gene-centered information at ncbi. *Nucl. Acids Res.* **39**, D52–7. <https://doi.org/10.1093/nar/gkq1237> (2011).
117. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA* **102**, 15545–50. <https://doi.org/10.1073/pnas.0506580102> (2005).

Acknowledgements

This work was funded in part by the National Institutes of Health/National Institute of Aging (NIH/NIA) P01 Grant, AG031719. The results published here are in part based on data obtained from the AD Knowledge Portal (<https://adknowledgeportal.synapse.org>). Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA grants P30AG10161 (ROS), R01AG15819 (ROSMAP; genomics), R01AG17917 (MAP), U01AG32984, U01AG46152 (ROSMAP), the Illinois Department of Public Health (ROSMAP), and the Translational Genomics Research Institute.

Author contributions

H.C. and Y.Z. provided the study data, both genetic and non-genetic. B.H. and M.L. designed the study approaches. B.H. contributed to the data management and statistical analysis, including running GWAS and constructing machine learning models. M.L. conducted the pathway analysis. X.L., C.N. and J.M. contributed in genotyping and genetic analysis. All authors revised and gave final approval of this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at <https://doi.org/10.1038/s41598-020-75446-2>.

Correspondence and requests for materials should be addressed to Y.Z. or M.W.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2020