Deep learning methods improve polygenic risk analysis and prediction for Alzheimer's disease

Xiaopu Zhou

Hong Kong University of Science and Technology

Yu Chen

Shenzhen Institutes of Advanced Technology

Fanny Ip

Hong Kong University of Science and Technology

Yuanbing Jiang

Hong Kong University of Science and Technology

Han Cao

The Hong Kong University of Science and Technology

Huan Zhong

The Hong Kong University of Science and Technology

Jiahang Chen

The Hong Kong University of Science and Technology

Tao Ye

Hong Kong University of Science and Technology <https://orcid.org/0000-0002-8075-7323>

Yuewen Chen

Shenzhen Institutes of Advanced Technology

Yulin Zhang

The Hong Kong University of Science and Technology

Shuangshuang Ma

The Hong Kong University of Science and Technology

Ronnie Lo

The Hong Kong University of Science and Technology

Estella Tong

Hong Kong University of Science and Technology

Alzheimer's Disease Neuroimaging Initiative Alzheimer's Disease Neuroimaging Initiative

Alzheimer's Disease Neuroimaging Initiative

Vincent Mok

The Chinese University of Hong Kong <https://orcid.org/0000-0002-8102-8835>

Timothy Kwok

The Chinese University of Hong Kong <https://orcid.org/0000-0001-9253-3549>

Qihao Guo

Huashan Hospital, Shanghai Medical College, Fudan University

Kin Mok

University College London

Maryam Shoai

UCL Institute of Neurology

John Hardy

University College London

Lei Chen

The Hong Kong University of Science and Technology

Amy Fu

Hong Kong University of Science and Technology

Nancy Ip (✉ boip@ust.hk)

Hong Kong University of Science and Technology <https://orcid.org/0000-0002-2763-8907>

Article

Keywords: Alzheimer's disease, risk prediction, polygenic score, Lasso, deep learning, neural network, artificial intelligence

Posted Date: September 7th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-818364/v1>

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

[Read Full License](#)

Version of Record: A version of this preprint was published at Communications Medicine on April 6th, 2023. See the published version at <https://doi.org/10.1038/s43856-023-00269-x>.

Abstract

Recent advances in genetic sequencing have enabled comprehensive genetic analyses of human diseases, resulting in the identification of numerous genetic risk factors for heritable disorders including Alzheimer's disease (AD). Such analyses enable AD risk prediction well before disease onset, which is critical for early interventions. However, current analytical approaches have limited ability to accurately estimate the risk effects of genetic variants owing to epistatic effects, which have been overlooked in most previous studies, resulting in unsatisfactory disease risk prediction. Herein, we modeled AD polygenic risk using deep learning methods, which outperformed existing models (i.e., weighted polygenic risk score and lasso models) for classifying disease risk. Moreover, by examining the associations between the outcomes from deep learning methods and multi-omics data obtained from our in-house Chinese AD cohorts, we identified the pathways that are potentially regulated by AD polygenic risk, including immune-associated signaling pathways. Thus, our results demonstrate the utility of deep learning methods for modeling the genetic risks of human diseases, which can facilitate both disease risk classification and the study of disease mechanisms.

Introduction

Recent biotechnological advances have expanded the study of human diseases from descriptive to quantitative analyses. In particular, large-scale “omics” studies on human tissues and fluids have enabled comprehensive investigations of disease mechanisms, thus facilitating diagnostics and the development of intervention strategies^{1,2}. Genomic variations are a major category of risk factors that contribute to numerous human diseases³. Accordingly, genetic studies of human diseases have revealed key causal mutations, leading to the identification of key biological pathways—for example, the identification of a role for the amyloid-processing pathway in Alzheimer's disease (AD) as evidenced by coding mutations in the *APP* and *PSEN* genes—and providing critical information for the development of intervention strategies^{4–6}. In addition, the identification of disease-causing variants enables early risk assessment and prediction; for example, the detection of *APP*, *PSEN1*, and *PSEN2* mutations, may aid the risk assessment for both early- and late-onset AD⁶. Importantly, as most genetic variants are present in the human genome, they are a reliable source of information for disease risk prediction throughout life⁷. Therefore, profiling genetic variations in the preclinical stages of a disease enables disease risk prediction in individuals before disease onset, which is especially valuable for clinical investigations and the development of intervention strategies for age-related diseases such as AD.

AD is one of the most common neurodegenerative diseases and is highly prevalent in elderly populations (~10% among people aged ≥ 65 years)⁸. As early prediction of AD risk aids its management⁹, effective genetic tools that can predict AD risk are vital for disease intervention. Notably, epidemiological studies show that genetic factors play a pivotal role in the pathogenesis of AD, further supporting the utility of genetic information for AD risk prediction^{10,11}. However, recent genome-wide association studies (GWASs) have revealed that AD is polygenic in nature. *APOE*-e4, a coding mutation that alters the

function of ApoE protein, is one of the most widely accepted AD genetic risk factors¹². Specifically, compared to noncarriers, people harboring 1 or 2 copies of the *APOE*-ε4 allele have a 3- and 15-fold greater risk of developing AD, respectively¹³. Although there are other common gene variants that are associated with AD (e.g., variants in the *BIN1*, *CD33*, and *SORL1* loci), they exert small to medium effect sizes on AD risk^{14–24}. Of note, although the U.S. Food and Drug Administration (FDA) approved the detection of *APOE* genotypes (i.e., *APOE*-ε4) for disease risk screening²⁵, the presence or absence of this genotype only estimates an individual's risk of AD and is not a definitive screening test. Moreover, the genotype detection cannot indicate the risk of AD for most people, because the prevalence of *APOE*-ε4 is only approximately 20% in populations of European descent and 10% in Asian populations. Therefore, appropriate models that encompass multiple informative genetic variants are required to determine an individual's risk of developing AD.

There have been tremendous efforts to develop polygenic models to estimate disease risk on the basis of genetic information^{26,27}. Thus, polygenic risk score (PRS) models are widely used to estimate individuals' polygenic risk of diseases with multiple genetic risk variants. A PRS model (or weighted PRS model), which is typically constructed by multiplying the weighted sum of risk allele dosages by their corresponding effect sizes, generates scores that can indicate the risk of a given disease²⁸. Numerous studies have investigated the utility of the PRS model for classifying human patients with various diseases such as schizophrenia²⁹, Parkinson's disease³⁰, cardiovascular diseases³¹, and AD^{12,32}. Notably, one study achieved relatively high accuracy for classifying AD when applying a PRS model to patients with pathologically confirmed AD [area under the receiver operating characteristic curve (auROC) = 0.84], highlighting the potential utility of this model for predicting disease risk³³.

Importantly, the currently used PRS model assumes that the risk effects of the studied variants are independent. However, this assumption is likely incorrect given that epistatic effects are observed in AD and other diseases^{34–36}, which suggests that the polygenic score model could be optimized. Therefore, it is of interest to investigate whether other statistical learning and deep learning methods can enhance polygenic risk analysis and disease classification. These methods include “least absolute shrinkage and selection operator” (i.e., lasso) models and neural network models—both of which can ascertain the contributions of individual variants to a given disease and construct a model for disease classification. Specifically, we previously showed that the lasso model might classify diseases more accurately than the currently used PRS model³⁷. Furthermore, other studies indicate that the neural network model might better model the interrelationships among genetic risk factors for disease prediction and classification^{38,39}. Nevertheless, the performance of these models (i.e., lasso and neural network models) for AD polygenic risk analysis has not been systematically evaluated.

Therefore, this study systematically evaluated the performance of polygenic score models based on weighted PRS, lasso, and neural network models to predict AD risk. Furthermore, by combining the results from these models with other AD-associated endophenotypic data, we identified potential pathological

mechanisms that contribute to AD polygenic risk. Taken together, our results suggest that deep learning methods can be used to predict AD risk and identify disease mechanisms.

Results

Superior performance of the neural network model for Alzheimer's disease risk prediction

To systematically evaluate the performance of different polygenic score models for AD classification, we obtained genotype and phenotype data from 3 AD cohorts: the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort⁴⁰, the Late Onset Alzheimer's Disease (LOAD) Family Study cohort⁴¹, and the National Institute on Aging Alzheimer's Disease Centers (ADC) cohort^{42,43} ($N = 11,352$ including 6,681 patients with AD and 4,671 normal controls [NCs]; **Supplementary Table 1**). We selected the AD-associated variants from the AD GWAS summary statistics reported by Jansen et al.¹⁸ with 3 p -value thresholds— $1E-4$, $1E-6$, and $1E-8$ —which yielded 8,100, 2,959, and 1,799 single nucleotide polymorphisms (SNPs), respectively, for model construction. A schematic flow diagram of the study is shown in **Figure 1**.

To determine whether the neural network model can predict AD risk more accurately than existing models, we constructed weighted PRS, lasso, and 5-layer neural network models using the selected AD-associated SNPs mentioned above and tested their performance for disease risk classification. The effect sizes (i.e., weights) for the weighted PRS model were obtained from the AD GWAS summary statistics reported by Jansen et al.¹⁸. Meanwhile, the lasso and neural network models required additional datasets to cross-validate their performance and were constructed using the data from 2 of the 3 AD cohorts (**Supplementary Table 1**): the data of 70% of individuals from the LOAD ($n = 2,995$ of 4,278) and ADC ($n = 3,984$ of 5,692) cohorts were used for model training, and the data of the remaining 30% ($n = 2,991$ of 9,970 total) were used to evaluate model accuracy at the end of each epoch. The data from the ADNI cohort ($N = 1,382$ including 689 patients with AD and 693 NCs) were used as a cross-evaluation dataset to assess the AD classification performance of the lasso and neural network models.

The performance of each model was evaluated by calculating the auROC and area under the precision-recall curve (auPRC), with higher values indicating more accurate AD classification. For all 3 models, higher classification accuracy was achieved when more SNPs were included in model construction; for example, the auROCs for the neural network model with 1,799, 2,959, and 8,100 SNPs were 0.82, 0.83, and 0.84 respectively, when applied to the entire ADC cohort (**Figures 2A, B; Table 1**). Moreover, when the same numbers of SNPs were used for model construction, the neural network model outperformed the weighted PRS and lasso models for AD classification; for example, when 8,100 SNPs were included in model construction, the auROCs for the weighted PRS, lasso, and neural network models in the entire ADC cohort were 0.64, 0.81, and 0.84, respectively; the auPRCs were 0.80, 0.82, and 0.92, respectively (**Figures 2A–D; Table 1**).

To further assess the performance of those models, we conducted an additional analysis by inclusion of a parallel weighted PRS analysis (denoted as wPRS2) using summary statistics from the IGAP 2019 Rare Variant Analysis stage 1 data (21,982 AD cases and 41,944 NC)¹⁷, which provides a more accurate estimation on the effect size of individual variants as only clinical-confirmed cases and normal controls were included in the analysis. Meanwhile, we removed the potential duplicated samples ($n = 16, 348$, and 51 for ADC, LOAD and ADNI, respectively) inferred by the identity-by-descent (IBD) analysis ($PI_{HAT} > 0.90$) to avoid overestimation of the model performance in the validation dataset. The results suggested that the neural network model significantly outperformed both the two weighted PRS models with any of the 3 selected variant pools in all the three AD cohorts ($p < 0.05$; **Supplementary Figure 2; Supplementary Tables 2 and 3**). Meanwhile, the neural network model also exerted a significant better performance in AD classification when compared to the Lasso models in ADC and LOAD cohorts, and in ADNI cohort when 2,959 variants (obtained from p -value threshold of $1E-4$) were used for model construction (**Supplementary Figure 2; Supplementary Tables 2 and 3**). Hence, our results demonstrate the possible superiority of the neural network model for AD classification.

Selection of variants to construct a polygenic score model for Alzheimer's disease in the Chinese population

To evaluate the general applicability of polygenic score models for AD classification, we examined their performance for AD classification in another ethnic population: the Chinese population. Accordingly, we first identified the common variants known to be associated with AD from various representative AD GWASs conducted in populations of various ethnic backgrounds^{14,16–19,21,44–46}, which yielded 216 AD GWAS hits (**Supplementary Tables 4 and 5**). We then examined whether these variants were associated with AD in 2 in-house Chinese AD cohorts for which whole-genome sequencing (WGS) data were available: Chinese AD WGS cohort 1 ($N = 2,340$ comprising 1,116 patients with AD, 309 patients with mild cognitive impairment [MCI], and 915 NCs)²⁴ and Chinese WGS cohort 2 ($N = 1,009$ comprising 356 patients with AD and 653 NCs) (**Supplementary Table 1**)³⁷. Notably, logic regression analysis with inclusion of age, sex, and genomic structure (represented by the top 5 principal components) as covariates showed that 38 of the 216 SNPs were significantly associated with AD in 1 of the 2 cohorts (**Supplementary Tables 6 and 7**). Meta-analysis of both cohorts further identified that among these 38 SNPs, 33 were significantly associated with AD (meta- $p < 0.05$; **Table 2; Supplementary Table 8**). In addition to these 33 SNPs, 4 SNPs (i.e., rs16824536, rs9271058, rs61732533, and rs111278892) were significantly associated with AD in 1 of the 2 cohorts and exhibited concordant risk trends in the other cohort (**Supplementary Table 9**).

Accordingly, we used the 37 AD-associated SNPs to calculate polygenic scores using the lasso and neural network models (see **Methods; Supplementary Table 9**). The lasso model that included the 37 SNPs for AD classification yielded an auROC of 0.71 (**Figure 3A; Supplementary Table 10**). This suggests that the abovementioned variants can indeed be used to classify people at risk of AD in the Chinese population. Meanwhile, we conducted subanalyses by segregating the 37 SNPs into 2 groups: a model including the

12 variants residing in the *APOE* locus (Lasso_ *APOE* model) and another including the 25 variants outside the locus (Lasso_non *APOE* model). Interestingly, while the Lasso_ *APOE* model could distinguish patients with AD from NCs (auROC = 0.65) as expected, the Lasso_non *APOE* model could also distinguish these two groups (auROC = 0.61; **Supplementary Figure 3; Supplementary Table 10**). Thus, corroborating the findings of other polygenic score studies of AD^{47,48}, these results suggest that SNPs in the non-*APOE* region also contribute to the pathogenesis of AD.

Next, we compared the AD classification performance of the polygenic score systems using the neural network model and lasso model. Notably, in Chinese WGS cohort 1, the neural network model distinguished patients with AD and MCI from NCs more accurately than the lasso regression model ($p < 0.01$; AD vs. NC: auROC = 0.77 and 0.71, respectively; MCI vs. NC: auROC = 0.62 and 0.60, respectively; **Figures 3A, B; Supplementary Figures 3 and 5; Supplementary Tables 10 and 11**). However, in Chinese WGS cohort 2, the neural network and lasso regression models exhibited comparable performance for AD classification ($p = 0.64$; auROC = 0.63 and 0.63 for the lasso and neural network models, respectively; **Supplementary Figures 4 and 5; Supplementary Tables 10 and 11**). Hence, our results collectively support the application of a neural network model that uses genetic information to model AD risk.

Performance of the neural network model for Alzheimer's disease risk classification in the Chinese population

Given the superior performance of the neural network model constructed above, we further examined whether it could classify individuals according to their relative disease risk. Accordingly, the scores of individuals in the Chinese cohorts calculated using the neural network model ("neural network risk score" hereafter) showed a clear distinction between the individuals with the lowest and highest scores. Specifically, patients with AD tended to have higher scores (**Figure 3C**). Therefore, we applied a multiple Gaussian fitting model to the risk scores obtained from the neural network model to stratify individuals into low-, medium-, and high-risk groups and subsequently examined the association of risk scores with disease status (see **Methods** and **Figure 3C**). In both Chinese WGS cohorts, larger proportions of patients with AD fell within the medium- and high-risk groups than within the low-risk group; for example, in Chinese WGS cohort 1, 70.9%, 49.4%, and 22.2% of patients with AD were classified into the high-, medium-, and low-risk groups, respectively. (**Figure 3D; Supplementary Table 12**). Furthermore, individuals in the high-risk group had greater risks of AD and MCI than those in the low- or medium-risk group ($p < 1E-5$ and $1E-3$ for Chinese WGS cohorts 1 and 2, respectively; **Supplementary Table 12**). Moreover, in Chinese WGS cohort 1, individuals in the medium-risk group had higher risks of AD ($p < 2E-16$) and MCI ($p = 1.25E-2$) than those in the low-risk group (**Supplementary Table 12**). Therefore, the PRS derived from the neural network model can predict the relative risk of AD in individuals from the Chinese population.

To determine the relevance of the resultant neural network risk scores to clinical outcomes, we examined the association between individuals' scores and their cognitive functioning. Given the data available for

each cohort, cognitive function was assessed according to the Mini-Mental State Examination (MMSE) in Chinese WGS cohort 1 and the Montreal Cognitive Assessment (MoCA) in Chinese WGS cohort 2. Notably, in Chinese WGS cohort 1, neural network risk scores were significantly associated with cognitive performance among all participants ($p < 2E-16$; **Figure 3E**), non-AD participants ($p = 3.10E-04$; **Figure 3F**), *APOE*- $\epsilon 3$ homozygous participants ($p < 2E-16$; **Figure 3G**), and *APOE*- $\epsilon 4$ carriers ($p = 2.18E-07$; **Figure 3H**; **Supplementary Table 13**). In addition, the neural network risk scores were significantly associated with the MoCA scores among all participants in Chinese WGS cohort 2 (**Supplementary Figure 4D**). Hence, the neural network risk scores calculated herein can predict cognitive function in the Chinese population.

Determination of Alzheimer's disease pathological mechanisms according to polygenic risk

To investigate the mechanisms whereby the identified variants (i.e., SNPs) modulate disease risk, we evaluated the associations between biological indices and polygenic risk in Chinese WGS cohort 2, which contained comprehensive data on AD endophenotypes³⁷. First, we determined the associations between polygenic scores and levels of amyloid-beta ($A\beta$), tau/p-tau181, and neurofilament light polypeptide (NfL), which are collectively termed the “ATN biomarkers” and reflect the progression and severity of AD⁴⁹. Accordingly, polygenic scores were significantly associated with plasma ATN biomarker levels among all participants (**Figure 4A**; **Supplementary Table 14**). However, detailed analysis showed that the associations between polygenic scores and plasma biomarker levels were significant in NCs but not in patients with AD, suggesting that the identified risk variants modulate AD-associated pathways independent of disease state (**Figures 4A–D**; **Supplementary Table 14**). Moreover, among all participants, polygenic scores were significantly associated with changes in brain structure, namely changes in the volumes of specific brain regions/tissues during AD progression^{50,51}, including the amygdala ($p = 6.53E-03$), grey matter ($p = 1.21E-02$), and the hippocampus ($p = 4.92E-02$) (**Supplementary Figure 5**; **Supplementary Table 15**). Moreover, polygenic scores were significantly associated with white matter hyperintensity, which is a marker of demyelination and axonal loss in the brain⁵² ($p = 2.69E-02$; **Supplementary Figure 5**; **Supplementary Table 15**). Meanwhile, there was only a weak association between polygenic scores and pons volume among patients with AD ($p = 3.97E-02$; **Supplementary Figure 5**; **Supplementary Table 15**). Hence, our results suggest that AD polygenic risk is associated with known AD biomarkers, particularly in people who have not developed AD.

To better understand how AD polygenic risk is associated with endophenotypic changes in human regardless of disease state, we comprehensively analyzed the associations between polygenic scores and the changes in the levels of 1,160 plasma proteins that potentially reflect changes in multiple biological pathways in NCs from Chinese WGS cohort 2. We found that polygenic scores were significantly associated with levels of 80 of the 1,160 assayed proteins; the plasma proteins that exhibited the strongest associations were PLTP (phospholipid transfer protein; $p = 2.67E-03$), which is involved in cholesterol metabolism⁵³, and CCL19 (chemokine ligand 19; $p = 6.65E-07$), a cytokine involved in inflammation⁵⁴ (**Figures 4E–G**; **Supplementary Table 16**). In addition, Gene Ontology (GO) enrichment analysis suggested that polygenic scores were associated with plasma proteins involved in

TNF- α - and cytokine-related pathways that are closely related to the immune system⁵⁵ (**Figure 4H; Supplementary Tables 17 and 18**). Furthermore, protein–protein interaction network analysis of those plasma proteins involved in cytokine-related pathways again suggested their enriched interaction, particularly TNF- α (enrichment $p < 1E-16$; **Figure 4I; Supplementary Table 19**). Together, these results suggest that AD polygenic risk variants are associated with the modulation of immune-associated signaling pathways in the human body before AD onset and independent of disease state.

Utility of neural network models for studying disease mechanisms

Polygenic risks, which integrate the impacts of multiple genetic risk factors, may reflect differences in the activities of distinct biological pathways within a given cell type or among different cell types. Although such differences cannot be inferred from the single score derived from the PRS, lasso, or neural network models, they may be embedded in the intermediate outputs of the hidden layers in neural network models. Hence, neural network models can potentially be used to determine how distinct biological processes contribute to AD polygenic risk.

As the penultimate layer of our neural network model summarizes the polygenic effects of the 37 studied SNPs into 5 modules (**Figure 5A**), we hypothesized that these modules represent distinct biological processes that inform AD polygenic risk. Accordingly, we retrieved the outputs of the 5 nodes in the penultimate layer of our neural network model and found that the outcomes for the 5 modules were not perfectly correlated (**Figure 5B**), which suggests that the 5 modules contain more information than the final polygenic score. Therefore, to understand the biological effects represented by the individual modules, we applied a multivariate model that simultaneously incorporates the outcomes of the 5 modules to determine their associations with individuals' endophenotypes (i.e., plasma protein levels). Notably, the levels of 336 plasma proteins were significantly associated with the outcomes of the 5 modules (**Supplementary Table 20**). Furthermore, GO enrichment analysis showed that the plasma proteins associated with the 5 modules could be classified into 4 clusters (designated "C1" to "C4") with distinct biological functions (**Figure 5C**). For instance, the plasma proteins classified into C1, C3, and C4 are associated with immune pathways; those in C3 and C4 are associated with cell communication; and those in C1 and C4 are associated with TNF- α -related signaling (**Figure 5C**).

Accordingly, we hypothesized that the involvement of specific risk variants in gene expression regulation—possibly in specific cellular contexts—underlies the observed associations between polygenic risk and plasma protein levels⁵⁶. Therefore, to determine whether specific plasma proteins associated with genetic variants were predominantly expressed in specific blood cell types, we conducted a cell-type enrichment analysis of the affected plasma proteins in each cluster. Interestingly, the plasma proteins in C1 and C4 were mainly expressed by B cells, those in C2 by erythroblasts and megakaryocytes, and those in C4 by dendritic cells and eosinophils (**Figure 5C; Supplementary Tables 21 and 22**). Furthermore, protein–protein interaction network analysis revealed that the affected proteins expressed in B cells were closely interconnected (enrichment $p = 1E-12$; **Figure 5D; Supplementary Tables 23 and 24**), suggesting that the polygenic risk variants might affect the cellular state of B cells. Specifically, the levels of the plasma

protein TCL1 family AKT coactivator A (TCL1A), which is uniquely expressed by B cells and associated with B-cell maturation⁵⁷, varied with polygenic risk; furthermore, its plasma level was altered in patients with AD compared to that in NCs (**Figures 5E, F**). Hence, these results reveal the possible effects of AD polygenic risk variants on specific biological pathways and the states of blood cell subtypes.

Discussion

Recent technological advances have enabled the relatively inexpensive and accurate profiling of individual genomes for use in disease risk prediction. However, as commonly used weighted PRS models fail to account for epistatic effects, they are suboptimal for modeling polygenic risk effects. Accordingly, here we show that an AD polygenic risk score generated by a neural network model can classify AD more accurately than other currently applied models. Specifically, our neural network model is comparably accurate to a weighted PRS model based on pathologically confirmed AD patients³³, yielding an AD classification accuracy (auROC = 0.84) close to the upper limit of that using genetic information alone³³. Furthermore, our neural network model requires far fewer variants than the weighted PRS model (8,100 vs. 29,017 SNPs, respectively) to achieve similar AD classification accuracy. Of particular interest, we evaluated the ability of this model with varying numbers of SNPs to predict AD risk in people of different ethnic backgrounds. Regardless of inputs, the neural network model could classify AD risk and distinguish patients with AD more accurately than the other models. Hence, our results demonstrate that neural network models might be superior to weighted PRS models for modeling polygenic risk based on genetic information. It is of interest to further investigate the utility of neural network models for modeling the risks of other heritable diseases for early risk prediction at a population scale, opening a new avenue for their use in conjunction with genomics studies.

Most previous polygenic risk studies focus on predicting disease risk but do not discuss how the identified polygenic risks contribute to the disease in question. Notably, by determining the associations between the risk scores (specifically the outputs from the hidden layers) obtained from deep learning analysis and information about specific biomarkers in the brain and circulatory system, we identified how AD polygenic risk is associated with possible pathophysiological changes in individuals; in particular, we identified changes in immune-associated pathways and in specific blood cells (e.g., B cells), supporting the role of inflammation in AD pathogenesis⁵⁸. Hence, our results also support the utility of deep learning methods for investigating the underlying disease mechanisms of AD, which may facilitate the development of precision medicine for early intervention in parallel to early risk prediction⁵⁹.

The performance of the neural network model for polygenic risk analysis, however, can be further improved. First, incorporating more variants into the classification model may better capture the genetic signatures that contribute to the pathogenesis of the disease in question, increasing the accuracy of disease classification. Meanwhile, misdiagnoses and misclassification of patients (or NCs) in the training dataset can impact model accuracy; in the case of AD, this can be corrected by including disease biomarkers that reflect the disease states, including brain amyloid load and levels of fluid AD biomarkers

(e.g., A β and p-tau181). Thus, adopting deep learning methods for a cohort with more accurate diagnostic data may further improve the prediction accuracy of the neural network model. Specifically, the recent development of AD biomarkers including positron emission tomography scan- and blood-based biomarkers (e.g., p-tau181) might enable the rapid identification of a “clean” cohort with more accurate diagnostic data for polygenic risk analysis, which may also improve model accuracy^{60,61}. In addition, there was a noticeable difference in the classification accuracy obtained in the training and validation datasets in the current study. Such differences may be derived from substantial differences among the study cohorts (e.g., recruitment criteria or cohort characteristics), or potential bias in the currently applied models. Therefore, further investigation with larger sample size is required to optimize the model performance before its implementation in clinical practice.

Taken together, our results suggest the utility of deep learning methods for classifying disease risk as well as their potential applications in unrevealing disease mechanisms. Further studies are required to explore their utility for forecasting disease risk at a population scale, as well as potential applications in disease mechanism studies and therapeutics development.

Methods

Study data

To investigate the performance of the polygenic score models for AD classification, we included the array data from the ADC cohort (phs000372.v2.p1), LOAD cohort (phs000168.v2.p2), and ADNI cohort (<http://adni.loni.usc.edu/>) in the analysis. The demographic data for these cohorts are presented in **Supplementary Table 1**. The details of the quality control and imputation processes are presented in the **Supplementary Methods**.

We included 2 Chinese WGS cohorts to study polygenic score models. The data for Chinese WGS cohort 1 ($N = 2,340$ comprising 1,116 patients with AD, 309 patients with MCI, and age- and sex-matched 915 NCs) have been published elsewhere²⁴. After quality control, data from 2,340 participants were retained for analysis. Data for Chinese WGS cohort 2 ($N = 1,009$ comprising 356 patients with AD and 653 age- and sex-matched NCs) have also been published elsewhere³⁷. The phenotypic data of the participants analyzed in this study were based on the most recent diagnostic records (as of December 2019).

Variant selection for model construction

We selected variants to evaluate polygenic score models on the basis of the AD GWAS summary statistics reported by Jansen et al.¹⁸. We applied 3 different p -value thresholds (i.e., $<1E-8$, $<1E-6$, and $<1E-4$) to the resultant variants for model construction. We retained the SNPs detected by all imputed array data from the ADC, LOAD, and ADNI cohorts that also fell into the corresponding p -value ranges for model construction and comparison (selected according to SNP ID).

Construction and testing of the weighted polygenic risk score model

We constructed the weighted PRS model by elementwise multiplication of allele dosage and the corresponding effect sizes from the AD GWAS summary statistics¹⁸ for the variants selected according to the p -value thresholds. We applied the same calculation method to evaluate the reported model for the Chinese population⁶². We calculated weighted PRSs in *RStudio* (v1.3.1056) using *R* programming (v4.0.2). We further used the effect sizes of the selected variants from another AD GWAS summary statistics¹⁷ to generate a parallel weighted PRS model. More than 96% of the variants selected in the first GWAS were captured by the second GWAS for model construction.

Construction and testing of the lasso model

We applied logistic regression to the ADC and LOAD datasets to regress out potential confounding effects from age, sex, and genomic structure (represented by the top 5 principal components). We then randomly selected 70% of the participants from the ADC and LOAD cohorts ($n = 4,198$ patients with AD and 2,780 NCs) as the training dataset to establish the model. We constructed the logistic lasso regression model using the *cv.glmnet()* function in the *glmnet* package, with 5-fold cross-validation ($\alpha = 1$, $\text{type.measure} = "mse"$, $\text{nfolds} = 5$) for the variants selected according to the p -value thresholds³⁷. We selected the λ -value that retained the most variants for the risk score calculation. We subsequently applied the model to the entire ADC, LOAD, and ADNI datasets using the *predict()* function with the appropriate λ -value as described above. Regarding the Chinese WGS datasets, we used Chinese WGS cohort 1 for the training dataset; we applied the same approach for model construction using the information from the 37 variants selected by the association test. We subsequently used Chinese WGS cohort 2 to evaluate the resultant model. We performed all analyses of the lasso polygenic score model in *RStudio* (v1.3.1056) using *R* programming (v4.0.2). The value of the random seed was fixed to the same constant value prior to all analyses.

Construction and testing of the neural network model

The training dataset for the neural network model was the same as that used for the lasso model above (i.e., 70% of the ADC and LOAD datasets), with the remaining 30% used for validation ($n = 1,794$ patients with AD and 1,198 NCs) during the model training step. We constructed the neural network model using the *Sequential()* function from the *Keras* package, an API for TensorFlow. Before the analysis, we annotated the selected SNPs for their associated loci using *ANNOVAR*⁶³ (77, 141, and 696 loci for p -value thresholds $<1E-8$, $<1E-6$, and $<1E-4$, respectively). We designed a 7-layer model, with the first and third layers being dropout layers ($\text{dropout rate} = 0.2$ or 0.3). We designed the number of nodes as follows: 3 * number of loci (based on the assumption of a maximum of 3 different haplotypes associated with AD in each locus), 1 * number of loci (corresponding to the locus number), 22 (corresponding to the chromosome number), 5 (arbitrary number corresponding to the potential number of pathways that affect the disease in parallel), and 1 (corresponding to the risk score). We applied exponential decay using the *ExponentialDecay()* function to accelerate the analysis ($\text{learning rate} = 0.3$, $\text{decay steps} = 100000$, $\text{decay rate} = 0.96$, $\text{staircase} = \text{True}$). During model training for the Chinese WGS datasets, we used Chinese WGS cohorts 1 and 2 as the training and validation datasets, respectively. Accordingly, we designed a 7-

layer model for the study, with the second and fourth layers being dropout layers (*dropout rate = 0.3*); the numbers of nodes in each layer were 50, 30, 10, 5, and 1. For both models, there were 300–1,000 epochs, with a batch size of 256 for model training. We applied the *binary crossentropy loss* function and evaluated model accuracy using the *accuracy* metrics. Moreover, we used *backend.function()* to extract the outputs from the nodes from specific layers for further analyses. The value of the random seed was fixed to the same constant value prior to all analyses.

Whole-genome sequencing

WGS (5× coverage) was performed by Novogene. The genomic DNA libraries were sequenced on an Illumina HiSeq × Ten and NovaSeq platform (150-bp paired-end reads). We adopted the GotCloud pipeline⁶⁴ to detect variants from our low-pass WGS data. In brief, sequencing data were subjected to FastQC (www.bioinformatics.babraham.ac.uk/projects/fastqc/) to check quality and Trimmomatic⁶⁵ to trim and filter low-quality reads. Clean data were mapped to the GRCh37 reference genome containing decoy fragments using BWA-mem. Subsequent analysis was conducted by subjecting data to the GotCloud pipeline with data processing and variant detection using the default settings²⁴. We then subjected the clean genotype files to Beagle⁶⁶ for genotyping refinement. For the Hong Kong Chinese WGS cohort, we also applied Thunder⁶⁷ for genotyping refinement after Beagle processing.

Analysis of plasma protein and brain imaging data

Next, we analyzed plasma protein and brain imaging data collected from a subgroup of participants from Chinese WGS cohort 2³⁷. Specifically, we analyzed plasma A β ₄₂, A β ₄₀, tau, and NfL levels in 157 patients with AD and 125 NCs by single-molecule detection assay (Neurology 3-Plex A Advantage Kit, #101995; NF-light Advantage Kit, #103186; Simoa). Specifically, we examined plasma p-tau181 levels in 154 patients with AD and 118 NCs (pTau-181 Advantage V2 Kit, #103714, Simoa). Detection was performed at the Quanterix Accelerator Lab (Boston, MA, USA). Moreover, the plasma samples of 97 patients with AD and 69 NCs from Chinese WGS cohort 2 were further subjected to Olink Proteomics (Boston, MA, USA) to determine the abundance of 1,160 plasma proteins by proximity extension assay. The following panels were used for the analysis: Cardiometabolic (91802), Cardiovascular II (91202), Cardiovascular III (91203), Cell Regulation (91702), Development (91703), Immune Response (91701), Inflammation (91301), Metabolism (91801), Neuro Exploratory (91502), Neurology (91501), Oncology II (91402), Oncology III (91403), and Organ Damage (91901).

For brain imaging analysis, we retrieved T1-weighted magnetization-prepared rapid acquisition with gradient echo (MPRAGE) and fluid-attenuated inversion recovery (FLAIR) sequences for 78 patients with AD and 104 NCs from Prince of Wales Hospital. We deidentified the raw imaging files and sent them to BrainNow Medical Technology (Hong Kong, China) to analyze volumetric information in different brain regions and white matter hyperintensity levels. We did not perform multiple test adjustment because of the limited sample sizes of the plasma protein and brain MRI data.

Statistical analysis

We performed a meta-analysis with a fixed-effects model using *METASOFT* software (v2.0.0) for variant analysis. We evaluated the classification accuracy of the models by calculating the auROC using the *roc()* function in the *pROC* package or the auPRC using the *pr.curve()* function in the *PRROC* package. The 95% CI of the auROC were estimated using bootstrap methods from *ci.auc()* function. The *roc.test()* function using bootstrap method was applied for testing the potential difference in auROC obtained from different models. The low-, medium-, and high-risk groups were determined by fitting risk scores to a Gaussian mixture model using the *normalmixEM()* function from the *mixtools* package for patients with MCI in Chinese WGS cohort 1 ($k = 3$, $maxit = 200$, $ECM = T$). We calculated the probability of an individual being classified into the low- or high-risk group using the corresponding scores as the input for the fitted probability distributions (using values of μ and σ from the fitted Gaussian mixture model). Meanwhile, we calculated the probability of an individual being classified into the medium-risk group by calculating the difference between 1 and the sum of the probabilities of being in the low- or high-risk group.

In addition, we performed association analysis between polygenic score or risk group and disease phenotype by logistic regression using the *glm()* function from the *stats* package. We also performed association analysis between polygenic score or risk/phenotype group and cognitive performance, plasma biomarkers, and brain volume by using a robust linear regression model using the *lmrob()* function from the *robustbase* package. Regarding cognitive performance, we applied rank-based, inverse-normal transformation to the cognitive scores using the *RankNorm()* function from the *RNOmni* package. We included age, sex, and genomic structure (represented by the top 5 principal components) as covariates. Regarding the brain imaging data, we further included intracranial volume as a covariate to normalize the possible interindividual variation in brain volume.

We performed GO enrichment analysis of the UniProt IDs in the Database for Annotation, Visualization and Integrated Discovery (DAVID) (<https://david.ncifcrf.gov/>). Moreover, we performed protein–protein interaction network analysis of the UniProt IDs in the Search Tool for the Retrieval of Interacting Genes/Proteins (STRING) database (<https://string-db.org/>). For cluster analysis, we applied *k*-means clustering to separate plasma proteins into individual clusters using the *kmeans()* function from the *stats* package for absolute values of *t*-statistics obtained from association tests between the levels of plasma proteins (i.e., normalized protein expression) and the polygenic scores from the 5 different modules. We determined the optimal number of clusters by using the elbow method implemented in the *fviz_nbclust()* function from the *factoextra* package.

For the cell-type enrichment analysis, we obtained the gene expression levels measured by RNA sequencing in individual blood cell types from the BLUEPRINT database (<http://dcc.blueprint-epigenome.eu/>). We performed cell-type enrichment analysis using the *TissueEnrich* package. Briefly, cell-type-specific transcript levels measured as fragments per kilobase per million mapped fragments (FPKM) for 1,159 plasma proteins were subjected to the *teGeneRetrieval()* function to first identify genes expressed by specific groups of cells (*foldChangeThreshold* = 1.5, *expressedGeneThreshold* = 5). We then

applied *teEnrichmentCustom()* for the enrichment analysis to identify the specific cell types associated with individual gene clusters (*tissueSpecificGeneType* = 1).

Data visualization

We generated a schematic diagram of the study using Microsoft PowerPoint (v2105). We generated heatmaps of AD classification accuracy (i.e., auROC and auPRC), box plots, volcano plots, bar charts, and dot plots using GraphPad Prism (v8.3.0). We plotted ROC and PR curves using the *plot()* function in R. Moreover, we generated histograms of polygenic scores using the *ggplot()* function from the *ggplot2* package with the *geom_density_ridges_gradient()* function from the *ggridges* package. We generated a heatmap to visualize distinct protein clusters using the *heatmap.2()* function from the *gplots* package. Finally, we visualized the protein–protein network using *Cytoscape* (v3.8.2) based on the node and interaction score information obtained from the STRING database (v11.0).

Declarations

Data availability

The genotype data used in the study for variant selection can be accessed in the corresponding manuscripts: the National Institute on Aging—Late Onset Alzheimer’s Disease Family Study (LOAD) raw data can be accessed in dbGaP at [phs000168.v2.p2](#); the Alzheimer’s Disease Genetics Consortium (ADGC) Genome Wide Association Study—NIA Alzheimer’s Disease Centers Cohort (ADC) raw data can be accessed in dbGaP at [phs000372.v2.p1](#); the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset can be accessed at ADNI database (<http://adni.loni.usc.edu/>).

The genetic and AD-associated endophenotypic data analysis results are given in the **Supplementary Information**. For raw data, the consent form signed by individual participants stated that the research content will be kept private under supervision of the hospital and research team. Therefore, these data will be made available and shared only in the context of a formal collaboration; applications for data sharing and project collaboration will be processed and reviewed by a Review Panel hosted at the Hong Kong University of Science and Technology. Researchers may contact [sklneurosci@ust.hk] for further details on project collaboration and sharing of the data from this study.

Code availability

The code used in this manuscript is available upon request.

Acknowledgments

We thank Pauline Kwan, Hazel Mok, Dr. Phillip Y.C. Chan, Choi Ying Ling, and Bonnie W. Wong for coordinating the collection of clinical data. We also thank Ka Chun Lok, Cara Wing Si Kwong, Tiffany Tze Wing Mak, Yan Ma, and Saijuan Liu for their excellent technical assistance as well as other members of the Ip Laboratory for many helpful discussions. This work was supported in part by the National Key R&D

Program of China (2018YFE0203600), the Research Grants Council of Hong Kong (the Theme-Based Research Scheme [T13-607/12R] and the Collaborative Research Fund [C6027-19GF]), the Areas of Excellence Scheme of the University Grants Committee (AoE/M-604/16), the Innovation and Technology Commission (ITCPD/17-9), the Guangdong Provincial Fund for Basic and Applied Basic Research (2019B1515130004), NSFC-RGC Joint Research Scheme (32061160472), the Guangdong Provincial Key S&T Program (2018B030336001), the Shenzhen Knowledge Innovation Program (JCYJ20180507183642005 and JCYJ20200109115631248), and the Guangdong–Hong Kong–Macao Greater Bay Area Center for Brain Science and Brain-Inspired Intelligence Fund (2019001 and 2019003). Please refer to the **Supplementary Note** for acknowledgments corresponding to the ADNI dataset, the Alzheimer’s Disease Genetics Consortium (ADGC) Genome Wide Association Study–NIA Alzheimer’s Disease Centers Cohort (ADC dataset), and the National Institute on Aging–Late Onset Alzheimer’s Disease Family Study (LOAD dataset). A portion of the data used in the preparation of this article was obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu); as such, the investigators within the ADNI contributed to the design and implementation of that initiative and/or provided data but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found in the **Supplementary Note** and at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Competing interests

The authors declare no competing interests.

References

1. Hasin, Y., Seldin, M. & Lusis, A. Multi-omics approaches to disease. *Genome Biol.* **18**, 83 (2017).
2. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nat. Rev. Genet.* **19**, 299–310 (2018).
3. Claussnitzer, M. *et al.* A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
4. Hardy, J. The Amyloid Hypothesis of Alzheimer’s Disease: Progress and Problems on the Road to Therapeutics. *Science (80-)*. **297**, 353–356 (2002).
5. Hardy, J. Amyloid, the presenilins and Alzheimer’s disease. *Trends Neurosci.* **20**, 154–159 (1997).
6. Lanoiselée, H. M. *et al.* APP, PSEN1, and PSEN2 mutations in early-onset Alzheimer disease: A genetic screening study of familial and sporadic cases. *PLoS Med.* (2017) doi:10.1371/journal.pmed.1002270.
7. Melzer, D., Pilling, L. C. & Ferrucci, L. The genetics of human ageing. *Nat. Rev. Genet.* **21**, 88–101 (2020).
8. 2020 Alzheimer’s disease facts and figures. *Alzheimer’s Dement.* **16**, 391–460 (2020).

9. Zissimopoulos, J., Crimmins, E. & St.Clair, P. The Value of Delaying Alzheimer's Disease Onset. *Forum Heal. Econ. Policy* **18**, 25–39 (2015).
10. Gatz, M. *et al.* Heritability for Alzheimer's disease: The study of dementia in Swedish twins. *Journals Gerontol. - Ser. A Biol. Sci. Med. Sci.* **52**, (1997).
11. Gatz, M. *et al.* Role of genes and environments for explaining Alzheimer disease. *Arch. Gen. Psychiatry* **63**, 168–174 (2006).
12. Zhou, X., Fu, A. K. & Ip, N. Y. APOE signaling in neurodegenerative diseases: an integrative approach targeting APOE coding and noncoding variants for disease intervention. *Current Opinion in Neurobiology* (2021) doi:10.1016/j.conb.2021.02.001.
13. Liu, C. C., Kanekiyo, T., Xu, H. & Bu, G. Apolipoprotein e and Alzheimer disease: Risk, mechanisms and therapy. *Nature Reviews Neurology* vol. 9 106–118 (2013).
14. Jun, G. R. *et al.* Transethnic genome-wide scan identifies novel Alzheimer's disease loci. *Alzheimer's Dement.* **13**, 727–738 (2017).
15. Jia, L. *et al.* Prediction of Alzheimer's disease using multi-variants from a Chinese genome-wide association study. *Brain* **8**, 99 (2020).
16. Marioni, R. E. *et al.* Genetic Stratification to Identify Risk Groups for Alzheimer's Disease. *J. Alzheimer's Dis.* **57**, 275–283 (2017).
17. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A β , tau, immunity and lipid processing. *Nat. Genet.* **51**, 414–430 (2019).
18. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
19. Kunkle, B. W. *et al.* Novel Alzheimer Disease Risk Loci and Pathways in African American Individuals Using the African Genome Resources Panel: A Meta-analysis. *JAMA Neurol.* **78**, 102–113 (2021).
20. Jia, L. *et al.* Prediction of Alzheimer's disease using multi-variants from a Chinese genome-wide association study. *Brain* (2020) doi:10.1093/brain/awaa364.
21. Bellenguez, C., Küçükali, F., Jansen, I., Andrade, V. & Moreno-grau, S. New insights on the genetic etiology of Alzheimer's and related dementia. *MedRxiv* 1–35 (2020).
22. Schwartzentruber, J. *et al.* Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat. Genet.* **53**, 392–402 (2021).
23. Wightman, D. P. *et al.* Largest GWAS (N=1,126,563) of Alzheimer's Disease Implicates Microglia and Immune Cells. *medRxiv* 2020.11.20.20235275 (2020) doi:10.1101/2020.11.20.20235275.

24. Zhou, X. *et al.* Identification of genetic risk factors in the Chinese population implicates a role of immune system in Alzheimer's disease pathogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1697–1706 (2018).
25. Arias, J. J., Tyler, A. M., Douglas, M. P. & Phillips, K. A. Private payer coverage policies for ApoE-e4 genetic testing. *Genet. Med.* (2021) doi:10.1038/s41436-020-01042-4.
26. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics* (2018) doi:10.1038/s41588-018-0183-z.
27. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nature Reviews Genetics* (2018) doi:10.1038/s41576-018-0018-x.
28. Choi, S. W., Mak, T. S. H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nature Protocols* (2020) doi:10.1038/s41596-020-0353-1.
29. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* (2009) doi:10.1038/nature08185.
30. Escott-Price, V. *et al.* Polygenic risk of Parkinson disease is correlated with disease age at onset. *Ann. Neurol.* (2015) doi:10.1002/ana.24335.
31. Sun, L. *et al.* Polygenic risk scores in cardiovascular risk prediction: A cohort study and modelling analyses. *PLoS Med.* (2021) doi:10.1371/JOURNAL.PMED.1003498.
32. Harrison, J. R., Mistry, S., Muskett, N., Escott-Price, V. & Brookes, K. From Polygenic Scores to Precision Medicine in Alzheimer's Disease: A Systematic Review. *J. Alzheimer's Dis.* **74**, 1271–1283 (2020).
33. Escott-Price, V., Myers, A. J., Huentelman, M. & Hardy, J. Polygenic risk score analysis of pathologically confirmed Alzheimer disease. *Ann. Neurol.* **82**, 311–314 (2017).
34. Sorbi, S. *et al.* Epistatic effect of APP717 mutation and apolipoprotein E genotype in familial Alzheimer's disease. *Ann. Neurol.* **38**, 124–7 (1995).
35. Combarros, O., Cortina-Borja, M., Smith, A. D. & Lehmann, D. J. Epistasis in sporadic Alzheimer's disease. *Neurobiol. Aging* **30**, 1333–49 (2009).
36. Hohman, T. J., Koran, M. E. & Thornton-Wells, T. Epistatic genetic effects among alzheimer's candidate genes. *PLoS One* (2013) doi:10.1371/journal.pone.0080839.
37. Zhou, X. *et al.* Genetic and polygenic risk score analysis for Alzheimer's disease in the Chinese population. *Alzheimer's Dement. (Amsterdam, Netherlands)* **12**, e12074 (2020).
38. Badré, A., Zhang, L., Muchero, W., Reynolds, J. C. & Pan, C. Deep neural network improves the estimation of polygenic risk scores for breast cancer. *J. Hum. Genet.* (2020) doi:10.1038/s10038-020-

39. Yin, B. *et al.* Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype. in *Bioinformatics* (2019). doi:10.1093/bioinformatics/btz369.
40. Mueller, S. G. *et al.* The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America* vol. 15 869–877 (2005).
41. Lee, J. H., Cheng, R., Graff-Radford, N., Foroud, T. & Mayeux, R. Analyses of the national institute on aging late-onset Alzheimer's disease family study: Implication of additional loci. *Arch. Neurol.* (2008) doi:10.1001/archneur.65.11.1518.
42. Jun, G. *et al.* Meta-analysis confirms CR1, CLU, and PICALM as Alzheimer disease risk loci and reveals interactions with APOE genotypes. *Arch. Neurol.* (2010) doi:10.1001/archneurol.2010.201.
43. Naj, A. C. *et al.* Common variants at MS4A4/MS4A6E, CD2AP, CD33 and EPHA1 are associated with late-onset Alzheimer's disease. *Nat. Genet.* (2011) doi:10.1038/ng.801.
44. Wightman, D. P. *et al.* Largest GWAS (N=1,126,563) of Alzheimer's Disease Implicates Microglia and Immune Cells. *medRxiv* 2020.11.20.20235275 (2020) doi:10.1101/2020.11.20.20235275.
45. Schwartzenuber, J. *et al.* Genome-wide meta-analysis, fine-mapping, and integrative prioritization identify new Alzheimer's disease risk genes. *medRxiv* **53**, 392–402 (2020).
46. Jia, L. *et al.* Prediction of Alzheimer's disease using multi-variants from a Chinese genome-wide association study. *Brain* **8**, 99 (2020).
47. Tosto, G. *et al.* Polygenic risk scores in familial Alzheimer disease. *Neurology* **88**, 1180–1186 (2017).
48. Escott-Price, V., Myers, A., Huentelman, M., Shoai, M. & Hardy, J. Polygenic Risk Score Analysis of Alzheimer's Disease in Cases without APOE4 or APOE2 Alleles. *J. Prev. Alzheimer's Dis.* **6**, 16–19 (2019).
49. Leuzy, A., Cullen, N. C., Mattsson-Carlsson, N. & Hansson, O. Current advances in plasma and cerebrospinal fluid biomarkers in Alzheimer's disease. *Curr. Opin. Neurol.* **Publish Ah**, 266–274 (2021).
50. Poulin, S. P., Dautoff, R., Morris, J. C., Barrett, L. F. & Dickerson, B. C. Amygdala atrophy is prominent in early Alzheimer's disease and relates to symptom severity. *Psychiatry Res. - Neuroimaging* **194**, 7–13 (2011).
51. Frisoni, G. B. *et al.* Detection of grey matter loss in mild Alzheimer's disease with voxel based morphometry. *J. Neurol. Neurosurg. Psychiatry* **73**, 657–664 (2002).
52. Prins, N. D. & Scheltens, P. White matter hyperintensities, cognitive impairment and dementia: An update. *Nature Reviews Neurology* (2015) doi:10.1038/nrneurol.2015.10.

53. Huuskonen, J., Olkkonen, V. M., Jauhiainen, M. & Ehnholm, C. The impact of phospholipid transfer protein (PLTP) on HDL metabolism. *Atherosclerosis* **155**, 269–281 (2001).
54. Marsland, B. J. *et al.* CCL19 and CCL21 induce a potent proinflammatory differentiation program in licensed dendritic cells. *Immunity* **22**, 493–505 (2005).
55. Zheng, C., Zhou, X. W. & Wang, J. Z. The dual roles of cytokines in Alzheimer's disease: Update on interleukins, TNF- α , TGF- β and IFN- γ . *Translational Neurodegeneration* vol. 5 1–15 (2016).
56. Van Der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
57. Herling, M. *et al.* High TCL1 levels are a marker of B-cell receptor pathway responsiveness and adverse outcome in chronic lymphocytic leukemia. *Blood* **114**, 4675–4686 (2009).
58. Heppner, F. L., Ransohoff, R. M. & Becher, B. Immune attack: The role of inflammation in Alzheimer disease. *Nature Reviews Neuroscience* (2015) doi:10.1038/nrn3880.
59. Zhavoronkov, A., Bischof, E. & Lee, K.-F. Artificial intelligence in longevity medicine. *Nat. Aging* **1**, 5–7 (2021).
60. Porter, T. *et al.* A Polygenic Risk Score Derived From Episodic Memory Weighted Genetic Variants Is Associated With Cognitive Decline in Preclinical Alzheimer's Disease. *Front. Aging Neurosci.* **10**, 1–11 (2018).
61. Karikari, T. K. *et al.* Diagnostic performance and prediction of clinical progression of plasma phospho-tau181 in the Alzheimer's Disease Neuroimaging Initiative. *Mol. Psychiatry* (2021) doi:10.1038/s41380-020-00923-z.
62. Jia, L. *et al.* Prediction of Alzheimer's disease using multi-variants from a Chinese genome-wide association study. *Brain* (2020) doi:10.1093/brain/awaa364.
63. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: Functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* (2010) doi:10.1093/nar/gkq603.
64. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* (2015) doi:10.1101/gr.176552.114.
65. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* (2014) doi:10.1093/bioinformatics/btu170.
66. Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.*

(2007) doi:10.1086/521987.

67. Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: Implications for design of complex trait association studies. *Genome Res.* (2011) doi:10.1101/gr.117259.110.

Tables

Table 1. Model performance for Alzheimer’s disease classification

<i>p</i> -value cutoff (number of variants)	Cohort	auROC			auPRC		
		wPRS	Lasso	NN	wPRS	Lasso	NN
<1.00E-08 (<i>n</i> = 1,799)	ADC	0.6711	0.7894	0.8156	0.8205	0.8820	0.9030
	LOAD	0.6519	0.7615	0.7995	0.6247	0.7376	0.7909
	ADNI	0.6632	0.6793	0.6853	0.6638	0.6577	0.6909
<1.00E-06 (<i>n</i> = 2,959)	ADC	0.6651	0.8159	0.8273	0.8149	0.8929	0.9095
	LOAD	0.6457	0.7874	0.8162	0.6221	0.7565	0.8054
	ADNI	0.6591	0.6555	0.6892	0.6569	0.6340	0.6832
<1.00E-04 (<i>n</i> = 8,100)	ADC	0.6414	0.8074	0.8438	0.7991	0.8205	0.9155
	LOAD	0.6330	0.7939	0.8315	0.6150	0.7696	0.8147
	ADNI	0.6293	0.6947	0.6956	0.6226	0.7014	0.7016

ADC, National Institute on Aging Alzheimer’s Disease Centers cohort; ADNI, Alzheimer’s Disease Neuroimaging Initiative cohort; auPRC, area under the precision–recall curve; auROC, area under the receiver operating characteristic curve; LOAD, Late Onset Alzheimer’s Disease Family Study cohort; NN, neural network; wPRS, weighted polygenic risk score.

Table 2. Variants significantly associated with Alzheimer’s disease in the 2 Chinese AD WGS cohorts

	Variant	EA	Gene	Effect Size (SE)	Meta <i>p</i> -value	ρ	EAF (AD/NC)
Populations of European descent							
1	rs4663105	C	<i>BIN1</i>	0.190 (0.058)	1.12E-03	0	0.506/0.463
2	rs6733839	T	<i>BIN1</i>	0.169 (0.058)	3.82E-03	0	0.458/0.418
3	rs4504245	A	<i>CLNK</i>	0.161 (0.078)	3.80E-02	0	0.177/0.157
4	rs6605556	G	<i>HLA-DQA1</i>	-0.180 (0.088)	4.03E-02	22.49	0.107/0.127
5	rs9381040	T	<i>TREML2</i>	0.190 (0.062)	2.13E-03	0	0.356/0.333
6	rs4236673	A	<i>CLU/PTK2B</i>	-0.147 (0.074)	4.56E-02	0	0.190/0.216
7	rs11787077	T	<i>CLU</i>	-0.157 (0.072)	2.85E-02	0	0.198/0.225
8	rs1532278	T	<i>CLU</i>	-0.162 (0.071)	2.37E-02	0	0.197/0.226
9	rs9331896	C	<i>CLU</i>	-0.143 (0.072)	4.60E-02	0	0.198/0.225
10	rs867611	G	<i>PICALM</i>	-0.164 (0.060)	6.14E-03	51.73	0.353/0.394
11	rs561655	G	<i>PICALM</i>	-0.151 (0.058)	9.85E-03	0	0.448/0.489
12	rs3844143	C	<i>PICALM</i>	-0.142 (0.060)	1.80E-02	34.68	0.372/0.410
13	rs10792832	A	<i>PICALM</i>	-0.139 (0.060)	2.01E-02	13.18	0.364/0.404
14	rs3851179	T	<i>PICALM</i>	-0.139 (0.060)	2.01E-02	13.18	0.364/0.404
15	rs74685827	G	<i>SORL1</i>	0.188 (0.094)	4.57E-02	0	0.115/0.097
16	rs11218343	C	<i>SORL1</i>	-0.205 (0.065)	1.61E-03	0	0.251/0.291
17	rs3752231	T	<i>ABCA7</i>	0.122 (0.059)	3.72E-02	90.07	0.439/0.415
18	rs12151021	G	<i>ABCA7</i>	-0.118 (0.058)	4.30E-02	85.93	0.475/0.503
19	rs3752246	G	<i>ABCA7</i>	0.128 (0.061)	3.43E-02	87.24	0.369/0.345
20	rs4147929	A	<i>ABCA7</i>	0.129 (0.061)	3.28E-02	88.42	0.368/0.343
21	rs769450	A	<i>APOE</i>	-0.224 (0.074)	2.42E-03	0	0.178/0.216
22	rs429358	C	<i>APOE</i>	1.182 (0.086)	2.72E-43	0	0.256/0.090
23	rs75627662	T	<i>APOE</i>	0.663 (0.069)	8.24E-22	0	0.304/0.186
24	rs9304690	T	<i>SIGLEC11</i>	0.137 (0.070)	4.89E-02	0	0.244/0.233
Asian populations							
25	rs11668861	G	<i>NECTIN2</i>	0.527 (0.066)	1.71E-15	22.76	0.321/0.220

26	rs6859	A	<i>NECTIN2</i>	0.455 (0.062)	1.53E-13	0	0.408/0.304
27	rs3852860	C	<i>NECTIN2</i>	0.513 (0.065)	3.60E-15	52.36	0.332/0.229
28	rs71352238	C	<i>TOMM40</i>	1.045 (0.086)	9.66E-34	0	0.230/0.091
29	rs157580	A	<i>TOMM40</i>	0.334 (0.058)	7.58E-09	44.75	0.528/0.448
30	rs2075650	G	<i>TOMM40</i>	1.050 (0.086)	4.11E-34	0	0.229/0.090
31	rs157582	T	<i>TOMM40</i>	0.685 (0.069)	3.07E-23	0	0.315/0.192
32	rs439401	C	<i>APOE</i>	0.422 (0.059)	6.16E-13	62.76	0.523/0.431
33	rs4420638	G	<i>APOC1</i>	1.076 (0.080)	6.41E-41	0	0.272/0.110

AD, Alzheimer's disease; EA, effect allele; EAF, effect allele frequency; NC, normal control; SE, standard error; WGS, whole-genome sequencing. The meta-analysis was conducted using a fixed-effects model.

Figures

Figure 1

Study schematic AD, Alzheimer's disease; ATN, amyloid-beta, tau, and neurofilament light polypeptide; auROC, area under the receiver operating characteristic curve; NC, normal control; NN, neural network; PR, precision–recall; PRS, polygenic risk score; ROC, receiver operating characteristic; TNF, tumor necrosis factor; WGS, whole-genome sequencing; wPRS, weighted PRS model.

Figure 2

Application of the weighted PRS, lasso, and neural network models for Alzheimer's disease classification (A, B) Performance of the wPRS, lasso, and NN models for the classification of patients with Alzheimer's disease as indicated by (A) auROCs and (B) auPRCs. The variant pools used for model construction were selected according to the p-value cutoffs shown on the left of each panel. (C, D) Representative plots showing the Alzheimer's disease classification accuracy of different models constructed using variants with $p < 1E-4$ in individual cohorts. (C) ROC curves and (D) precision-recall curves showing Alzheimer's disease classification accuracy in different cohorts. ADC, National Institute on Aging Alzheimer's Disease Centers cohort; ADNI, Alzheimer's Disease Neuroimaging Initiative cohort; auPRC, area under the precision-recall curve; auROC, area under the receiver operating characteristic curve; LOAD, Late Onset

Alzheimer's Disease Family Study cohort; NN, neural network; ROC, receiver operating characteristic; wPRS, weighted polygenic risk score.

Figure 3

Polygenic risk analysis for Alzheimer's disease in Chinese AD (A, B) ROC curves of the polygenic score classification of patients with AD (A) and MCI (B) in Chinese WGS cohort 1. (C) Distribution of polygenic risk scores derived from the NN model for each phenotype group. The definitions of the low-, medium-, and high-risk groups are shown in the upper panel. (D) Percentages of each phenotype group in the low-,

medium-, and high-risk groups. (E–H) Associations between polygenic risk score and MMSE score in (E) all participants, (F) non-AD participants (i.e., NCs plus patients with MCI), (G) APOE-ε3 homozygous carriers, and (H) APOE-ε4 carriers. Data are presented as box-and-whisker plots. The numbers of individuals in the corresponding group are shown at the bottom of each plot. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$; robust linear regression model. AD, Alzheimer’s disease; auROC, area under the receiver operating characteristic curve; MCI, mild cognitive impairment; MMSE, Mini-Mental State Examination; NC, normal control; NN, neural network; ROC, receiver operating characteristic; WGS, whole-genome sequencing.

Figure 4

Modulatory effects of polygenic risk for Alzheimer's disease on plasma protein biomarkers in normal controls (A) Associations between the polygenic risk scores derived from the corresponding models and the levels of plasma ATN biomarkers (i.e., A β 42, A β 40, A β 42/A β 40 ratio, tau, p-tau181, and NfL) in all participants, NCs, and patients with AD. (B–D) Plasma A β 42 level (B), A β 42/A β 40 ratio (C), and p-tau181 level (D) in NCs stratified according to polygenic risk score group. (E) Volcano plots showing the associations between polygenic risk scores and plasma protein levels obtained from the high-throughput assay. (F, G) Levels of the candidate plasma proteins (F) PLTP and (G) CCL19 in NCs stratified according to polygenic risk score group. (H) Overrepresented Gene Ontology terms for plasma proteins associated with polygenic risk scores ($p < 0.05$). (I) Protein–protein interaction network of cytokines associated with polygenic risk scores. The gray nodes are the 5 proteins most strongly associated with the other nodes. Line color and thickness indicate the interaction strength of the connected nodes (darker and thicker lines denote stronger interactions). (B–D, F, G) Numbers of individuals in the corresponding group are shown at the bottom of each plot. *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$; robust linear regression model. (E, H, I) Colors denote plasma proteins or results derived from proteins that were positively (red) or negatively (blue) correlated with polygenic risk scores. A β , amyloid-beta; AD, Alzheimer's disease; CCL19, chemokine ligand 19; MCI, mild cognitive impairment; NC, normal control; NfL, neurofilament light polypeptide; NN, neural network; PLTP, phospholipid transfer protein; TNF, tumor necrosis factor; WGS, whole-genome sequencing.

respect to the absolute values of t-statistics. Representative Gene Ontology terms and cell-type enrichment analysis results are displayed in the center and right panels, respectively. (D) Protein–protein interaction network of proteins expressed by B cells. Colors denote proteins from C1 (red) and C4 (blue). (E) Cell-type-specific expression of TCL1A. (F) Plasma levels of TCL1A protein in NCs and patients with AD. **p < 0.01. AD, Alzheimer’s disease; FDR, false discovery rate; FPKM, fragments per kilobase per million mapped fragments; NC, normal control; NK, natural killer; NN, neural network; TCL1A, TCL1 family AKT coactivator A; TNF, tumor necrosis factor.

Supplementary Files

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

- [Zhousupplementary.pdf](#)
- [lpNNP.zip](#)