frontiers in Radiology

METHODS
published: 21 January 2022
doi: 10.3389/fradi.2021.777030

Integrating Transcriptomics, Genomics, and Imaging in Alzheimer's Disease: A Federated Model

Jianfeng Wu[1], *Yanxi Chen*[1], *Panwen Wang*[2], *Richard J. Caselli*[3], *Paul M. Thompson*[4], *Junwen Wang*[2*†] *and Yalin Wang*[1*†] *for the Alzheimer's Disease Neuroimaging Initiative*

[1] School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, United States, [2] Department of Health Sciences Research and Center for Individualized Medicine, Mayo Clinic Arizona, Scottsdale, AZ, United States, [3] Department of Neurology, Mayo Clinic Arizona, Scottsdale, AZ, United States, [4] Imaging Genetics Center, Stevens Institute for Neuroimaging and Informatics, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

Alzheimer's disease (AD) affects more than 1 in 9 people age 65 and older and becomes an urgent public health concern as the global population ages. In clinical practice, structural magnetic resonance imaging (sMRI) is the most accessible and widely used diagnostic imaging modality. Additionally, genome-wide association studies (GWAS) and transcriptomics—the study of gene expression—also play an important role in understanding AD etiology and progression. Sophisticated imaging genetics systems have been developed to discover genetic factors that consistently affect brain function and structure. However, most studies to date focused on the relationships between brain sMRI and GWAS or brain sMRI and transcriptomics. To our knowledge, few methods have been developed to discover and infer multimodal relationships among sMRI, GWAS, and transcriptomics. To address this, we propose a novel federated model, Genotype-Expression-Imaging Data Integration (GEIDI), to identify genetic and transcriptomic influences on brain sMRI measures. The relationships between brain imaging measures and gene expression are allowed to depend on a person's genotype at the single-nucleotide polymorphism (SNP) level, making the inferences adaptive and personalized. We performed extensive experiments on publicly available Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset. Experimental results demonstrated our proposed method outperformed state-of-the-art expression quantitative trait loci (eQTL) methods for detecting genetic and transcriptomic factors related to AD and has stable performance when data are integrated from multiple sites. Our GEIDI approach may offer novel insights into the relationship among image biomarkers, genotypes, and gene expression and help discover novel genetic targets for potential AD drug treatments.

Keywords: Alzheimer's disease, brain imaging, GWAS, transcriptomics, chow test, federated learning

OPEN ACCESS

Edited by:
Li Shen,
University of Pennsylvania,
United States

Reviewed by:
Jingwen Yan,
Purdue University Indianapolis,
United States
Eun Jeong Min,
Catholic University of Korea,
South Korea

***Correspondence:**
Junwen Wang
wang.junwen@mayo.edu
Yalin Wang
ylwang@asu.edu

[†]These authors have contributed equally to this work

Specialty section:
This article was submitted to Artificial Intelligence in Radiology, a section of the journal Frontiers in Radiology

Received: *14 September 2021*
Accepted: *21 December 2021*
Published: *21 January 2022*

Citation:
Wu J, Chen Y, Wang P, Caselli RJ, Thompson PM, Wang J and Wang Y (2022) Integrating Transcriptomics, Genomics, and Imaging in Alzheimer's Disease: A Federated Model. Front. Radiol. 1:777030. doi: 10.3389/fradi.2021.777030

INTRODUCTION

Alzheimer's disease (AD) is a major public health concern, with the number of affected individuals expected to triple, reaching 13.8 million, by the year 2050 in the U.S. alone (1). Current therapeutic failures in patients with dementia due to AD may be due to interventions that are too late or targets that are secondary effects and less relevant to disease initiation and early progression (2). Mounting evidence suggests that germline mutations, e.g., DNA single nucleotide polymorphisms (SNPs), play an important role in AD etiology and progression (3, 4). Among various genetic risk factors, Apolipoprotein E (*APOE*) has the strongest association to late-onset AD, and the e4 allele is associated with increased risk, whereas the e2 allele is associated with decreased risk (5). Known genetic risk variants could be used to identify presymptomatic individuals at risk for AD and support diagnostic assessment of symptomatic subjects. By taking into account patients' genetic risk factors, at-risk individuals could be more readily identified, diagnostic precision could be improved, and targetable disease mechanisms for new drug development may be discovered (6–9). By enabling each patient to receive earlier diagnoses, risk assessments, and optimal treatments, personalized or precision medicine holds promise for improving early AD intervention while also lowering costs (10).

Recent clinical trials targeting single molecular mechanisms have failed (11, 12). Rather, it might be necessary to tackle the problem from a holistic or multimodality perspective (13, 14). Indeed, the NIH and the scientific community realized this problem a while ago and have already started to produce multi-omics data. For example, the Alzheimer's Disease Sequencing Project (ADSP) data repository contains genomic level data derived from genome-wide association studies (GWAS) (4, 15), whole-exome sequencing (WES) (16, 17), and whole-genome sequencing (WGS), and RNA level data including mRNA, miRNA, and long non-coding RNA profiling from either microarray or RNA-Seq (18). And transcriptome-wide association studies (TWASs) provides a way to use eQTLs and expression data to guide GWAS of AD (19). Brain imaging has played a significant role in the study of Alzheimer's disease (20). Integrating imaging data and omics data is becoming an emerging data science field known as imaging genomics (21). The major task of this field is to perform integrated analysis of imaging and omics data, often combined with other biomarkers, as well as clinical and environmental data. The ultimate goal is to gain new insights into the underlying mechanisms of human health and disease, to better inform the development of new diagnostic, therapeutic, and preventative approaches.

Various imaging genetics methods have been developed to integrate imaging and genetic data. However, most studies have focused on imaging, imaging combined with GWAS data (22–24), imaging with transcriptomics (25), or GWAS with transcriptomics (26). For example, imaging genetics methods have been used to link SNPs with image features (27), and expression quantitative trait loci (eQTL) have been used to discover *APOE*-related genes (28). However, relatively few methods have been developed to integrate GWAS/WES/WGS, imaging, and transcriptomic data to infer multimodal relationships. For instance, Liu et al.(29) use a brain-wide gene expression profile available in the Allen Human Brain Atlas (AHBA) as a 2-D prior to guide the brain imaging genetics association analysis. Their transcriptome-guided SCCA (TG-SCCA) framework incorporates the gene expression information into the traditional SCCA model. Such a multimodal approach may give us a more holistic view of the evidence from multiple sources to provide novel insights on the molecular mechanisms of AD pathogenesis and prognosis. Besides, both gene expression and imaging features are dynamic and change with time and throughout the disease, whereas germline SNPs are unchanged over an individual's lifetime. We need a better model for studying SNP-image-gene expression relationships to consider both the dynamic changes in imaging and gene expression features and understand how they are affected by an individual's SNPs. Such knowledge will provide novel insights into the relationship among image biomarkers, genotypes and gene expression, and may help discover novel genetic targets for pharmaceutical interventions.

AD is a complex multifactorial disorder that involves many biological processes. The launch of the Alzheimer's Precision Medicine Initiative (APMI) and its associated cohort program in 2016—facilitated by the academic core coordinating center run by the Sorbonne University Clinical Research Group in Alzheimer's Precision Medicine—is intended to improve clinical diagnostics and drug development research in Alzheimer's disease (30). Hampel et al. (30) indicate the challenges for precision medicine, including secure data access accompanied by rigorous privacy protection and the availability of data to qualified researchers who may use them to exercise their creative thinking with an *a posteriori* approach or to test their *a priori* hypotheses. Integrating data from multiple sites and sources is common practice to achieve larger sample sizes and increase the statistical power. Unprecedentedly large amounts of biomedical data now exist across hospitals and research institutions. However, different institutions may not be readily able to share biomedical research data due to patient privacy concerns, data restrictions based on patient consent or institutional review board (IRB) regulations, and legal complexities; this can present a major obstacle to pooling large scale datasets to discover and understand AD-related genetic information. To remedy this distributed problem, a large-scale collaborative network, ENIGMA consortium, was built (31). Federated learning is an important direction of interest in multi-site neuroimaging research; the use of distributed computing offers an approach to learn from data spread across multiple sites without having to share the raw data directly or to centralize it in any one location. Even so, most ENIGMA and other GWAS studies currently focus on the influence of genetic variants on human brain structures (22, 32–34) or functional measures (35) and relatively few have studied the relationships among image biomarkers, genotypes, and gene expression.

In this paper, we propose a novel Federated Genotype-Expression-Image Data Integration model (GEIDI) based on the Chow test (36). The intuition behind our multi-omics framework is illustrated in **Figure 1**. Some important image-expression relationships (correlations) may be diluted when the population

is mixed together. Still, when we stratify the population based on their genotypes (a gene like *APOE* or a SNP like *rs942439*), we can observe strong correlations (AA and BB groups) across subgroups. Accordingly, as shown in **Figure 2**, our model is designed to detect if the relationships between X (imaging biomarker) and Y (gene expression) are different among the subgroups. The *p*-value of the model is then used to prioritize the trios (genotype-expression-image).

We further design various experiments on publicly available data from the Alzheimer's Disease Neuroimaging Initiative (ADNI, adni.loni.usc.edu) to demonstrate that our model may detect the genetic factors most related to AD better than the state-of-the-art Matrix eQTL. The overall intent of the work is to detect relationships that inform the design or repurposing of drugs to target these subgroups to achieve precision medicine. We first use a hypergeometric analysis and an AD-related gene list from alzgene.org to evaluate the ability of our federated GEIDI model to discover AD-related gene expression. To further aid in the discovery of genes that may be potential AD drug targets, we also use Pearson correlations analyses to demonstrate the divergence in stratified populations. Additionally, we design experiments to show that our model can discover more AD-related SNPs, based on tests with 1,217 known AD-associated SNPs and 1,217 randomly selected SNPs. Finally, we evaluate the stability of our model under different multi-site conditions. With the ADNI dataset, we set off to test our hypothesis that the proposed federated GEIDI model may be an effective federated model that can provide novel insights into the relationship among image biomarkers, genotypes, and gene expressions and the discovery of novel genes for potential AD drug targets.

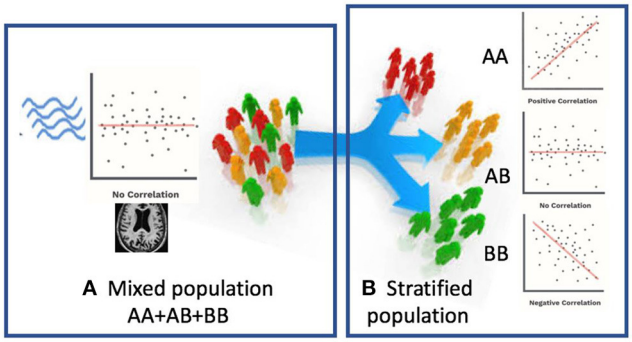

FIGURE 1 | Schematic view of our multi-omics approach. **(A)** When patients are mixed together, image-expression correlation may be low. **(B)** When a certain genotype stratifies patients, some subsets (AA, BB) have a high correlation.

DATA AND METHODS

Data Preprocessing

The data in this work are from the Alzheimer's Disease Neuroimaging Initiative (ADNI, adni.loni.usc.edu) and the TADPOLE challenge (tadpole.grand-challenge.org) (37). The ADNI was launched in 2003 as a public-private partnership led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of MCI and early AD. The genome-wide association study of ADNI is designed to provide researchers with the opportunity to combine genetics with imaging and clinical data to help investigate the mechanisms of the disease. For up-to-date information, see adni.loni.usc.edu/data-samples/data-types/genetic-data/. From the ADNI GWAS, we analyzed data from 697 subjects, including AD patients, people with mild cognitive impairment

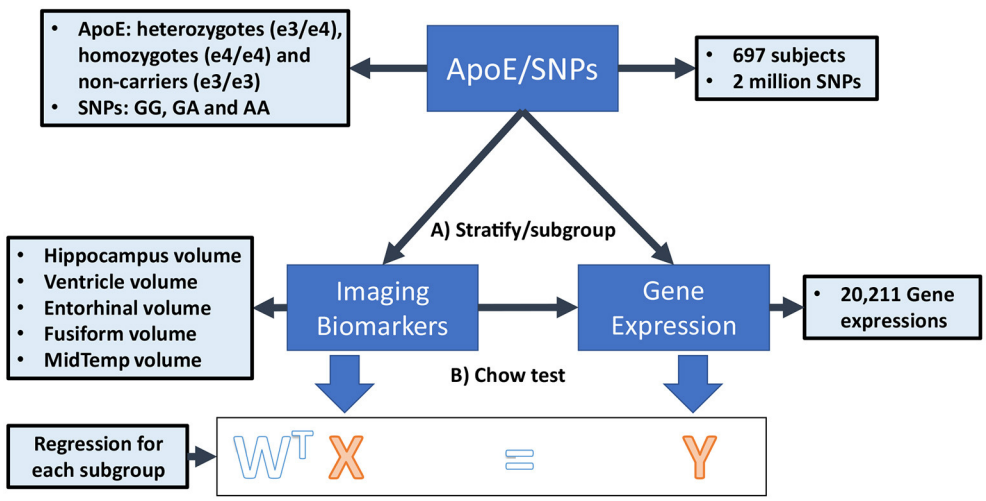

FIGURE 2 | The federated GEIDI model on ADNI data. **(A)** Stratify samples into subgroups with different genotypes of a gene (e.g., *APOE*) or at a specific SNP locus (e.g., *rs942439*) **(B)** Federated GEIDI is used to detect if the relationships between X (imaging biomarker) and Y (gene expression) are different among the subgroups. The *p*-value of federated GEIDI will then be used to prioritize the trios (genotype-expression-image).

(MCI), and cognitively unimpaired (CU) subjects, for whom the demographic information is shown in **Table 1**. Each sample has three types of modalities of data: genotypes of known AD risk genes (e.g., *APOE*) and SNPs from genome-wide association studies (GWAS), gene expression measurements (for 20,211 genes) from microarray-based transcriptomic profiling of samples' blood, and imaging biomarkers from structural magnetic resonance imaging (sMRI) data of subjects' brains. We use *plink* to perform a quality check of the genotype data. The SNPs in the normal group that deviate significantly from Hardy-Weinberg equilibrium are removed (38). The LINNORM package (39) was adopted to perform data transformation on the expression data for normality and homoscedasticity. Recent evaluations (40, 41) show that LINNORM typically performs better than current DEG analysis methods for both single-cell and bulk RNA-Seq, such as Seurat (42) and DESeq2 (43).

Eventually, we get 2,059,586 SNPs, *APOE* genotype, and expression data for 20,211 genes for each sample. Besides, from the TADPOLE challenge, we obtained two brain imaging biomarkers for each subject calculated using FreeSurfer (44) with sMRI, including the volume of the hippocampus and middle temporal gyrus (MidTemp). To adjust for individual differences in head size, the volume of each region is adjusted by the intracranial vault volume (ICV) of each subject (volume/ICV). The difference between the dates for gene expression collection and MRI scan is <5 months.

Federated Genotype-Expression-Image Data Integration Framework

Econometrician Gregory Chow first proposed the Chow test in 1960 (36) to determine whether correlation coefficients estimated in two subgroups are significantly different. In econometrics, it is most commonly used in time series analysis to test for the presence of a structural break at a period that can be assumed to be known as *a priori* (for instance, a significant historical event such as a war). For example, we can model the data as $y = wX+\epsilon$. Then, the data can be broken into two groups according to some event and fitted to the regression model as, $y_1 = w_1x_1 + \epsilon$ and $y_2 = w_2x_2 + \epsilon$. The null hypothesis of the Chow test asserts that $w_1 = w_2$ and the model errors ϵ are independent and identically distributed from a normal distribution with unknown variance. Let S_C, S_1, and S_2 be the sum of squared residuals for the three regression models, respectively, N_1 and N_2 are the number of observations in each group, and k is the number of parameters. The Chow test statistic is $F = \frac{(S_C-(S_1+S_2))/k}{(S_1+S_2)/(N_1+N_2-2k)}$, which follows the F-distribution with k and $N_1 + N_2 - 2k$ degrees of freedom.

TABLE 1 | Demographic information for the subjects we study from the ADNI.

Group	Sex (M/F)	Age	MMSE
AD ($n = 96$)	59/37	74.8 ± 7.5	21.8 ± 4.1
MCI ($n = 366$)	209/157	72.0 ± 7.5	28.0 ± 1.7
CU ($n = 235$)	115/120	74.4 ± 5.8	29.1 ± 1.2

Values are mean ± standard deviation, where applicable.

Although the Chow test is commonly used in the financial industry, it is seldom used in the biomedical field (45). In this work, we first generalize the Chow test model to estimate the multi-subgroup condition and further introduce a federated learning technique to the model. We apply the proposed model to the ADNI dataset to detect the significant trios among genotype, gene expression, and imaging biomarkers and discover the dominant genetic and transcriptomics factors for brain structures.

Standardization

We simulate the multi-site condition by separating all the samples into I hypothetical institutions ($I = 5$) on Apache Spark (spark.apache.org), a state-of-the-art distributed computing platform (Although the ADNI data can be centralized, such a federated analysis would allow the method to be scaled up to much larger datasets, including genomic data that is difficult to centralize for logistic or regulatory reasons). As illustrated in **Figure 2**, the samples in each institution can be further partitioned into at most three subgroups $\left(g = 1, 2, 3\right)$ according to the subject's genotype at certain SNP loci (e.g., GG, GA, AA) or a gene (e.g., stratified by the three *APOE* genotypes considered in this study). Accordingly, X_i^g and y_i^g, respectively, represent the image biomarkers and gene expression values in the gth group of the ith institution. The data from the gth group in all I institutions will be fitted into a regression model in a federated strategy.

Federated Chow Test Analysis

Using federated linear regression, we can calculate four linear models for all the I institutions, including three models for three subgroups and one for all samples in the three subgroups. $\overline{w^{(1)}}$, $\overline{w^{(2)}}$, $\overline{w^{(3)}}$ and $\overline{w^{(C)}}$ are their optimal coefficient vectors. The Chow test assumes that the errors ϵ are independent and identically distributed from a normal distribution by an unknown variance. The null hypothesis of the Chow test asserts that $\overline{w^{(1)}}$, $\overline{w^{(2)}}$, and $\overline{w^{(3)}}$ are equal. The predictive test suggested by Chow is then:

$$F = \frac{\left(S^{(C)} - \left(S^{(1)} + S^{(2)} + S^{(3)}\right)\right) / \left(2k\right)}{\left(S^{(1)} + S^{(2)} + S^{(3)}\right) / \left(N^{(1)} + N^{(2)} + N^{(3)} - 3k\right)}, \quad (1)$$

where $S^{(C)}$ is the sum of squared residuals from the combined data from the three subgroups, $S^{(1)}$ is the sum of squared residuals from the first group, and so on for $S^{(2)}$ and $S^{(3)}$. $N^{(1)}$, $N^{(2)}$, and $N^{(3)}$ are the number of samples in each subgroup, and k is the number of parameters. Under the null hypothesis, the test statistic follows the F-distribution with $2k$ and $N^{(1)} + N^{(2)} + N^{(3)} - 3k$ degrees of freedom. The global center will calculate F by gathering all the least square losses and the number of subjects for each subgroup and combined data from each institution. For example, for the first subgroup, the global least-square loss is $S^{(1)} = \sum_{i=1}^{I} S_i^{(1)}$ and the global subject number is $N^{(1)} = \sum_{i=1}^{I} N_i^{(1)}$. Eventually, the p-value will be calculated at the global coordinating center and assigned to each institution.

Federated Linear Regression

Many regression models may be selected for the Chow test model, such as linear regression (46), polynomial regression (47), ridge regression (48), and so on. In this study, we focus on studying the differences in the relationships between imaging biomarkers and gene expression among different groups. Complex regression models, like polynomial regression, may lead to over-fitting and meaningless results. Also, sparse or penalized regression methods, such as ridge regression, require an appropriate regularization parameter. Therefore, in this work, linear regression would be the most rational choice.

Since the federated regression models for each subgroup are the same, we omit the group superscripts here. For the data in one subgroup of all the I institutions, we can calculate the linear regression equation as: $y = Xw + \epsilon$, where $X \in R^{N\times k}$ represents the independent variables, $y \in R^N$ is a vector of the observations on a dependent variable, $w \in R^k$ is a coefficient vector, and $\epsilon \in R^N$ is the disturbance vector. N is the number of observations in the group, and k is the number of parameters. Then, the coefficient vector w can be estimated by minimizing the least squared function, $S(w) = \frac{1}{2} \left\|Xw - y\right\|_2^2$.

To avoid centralizing the data, (X_i, y_i), from each institution, we first rewrite the minimization problem as, $\min \sum_{i=1}^{I} S_i\left(w; X_i, y_i\right) = \frac{1}{2}\sum_{i=1}^{I}\left\|X_i w - y_i\right\|_2^2$. Then, the global gradient can be calculated as, $\nabla S(w) = X^T\left(Xw - y\right) = \sum_{i=1}^{I} X_i^T\left(X_i w - y_i\right) = \sum_{i=1}^{I}\nabla S_i(w)$. Therefore, instead of centralizing the data, the global center only needs to gather the partial gradient, $\nabla S_i(w)$, which is calculated with (X_i, y_i) at each local institution. After computing the global gradient, $\nabla S(w)$, the global center will send it back to ith local institution. Finally, w will be updated at each institution by gradient descent with the same learning rate, $w \leftarrow w - \eta\nabla S(w)$. The reason for not updating w at the global center is to avoid possible data reconstruction. When w is zero, the local gradient sent to the center is $-X_i^T y_i$. Then, the global center can easily acquire $X_i^T X_i w$ and X_i might be reconstructed if w is known to the center. Consequently, our optimization strategy is able to preserve data privacy for all institutions. The whole framework of our federated Genotype-Expression-Image Integration model is summarized in **Algorithm 1**. And the code can be downloaded at our website, https://github.com/JianfengWu1993/GEIDI.

Algorithm 1 Federated Genotype-Expression-Image Data Integration Model.

Input: Data pairs of the I institutions, (X_1, y_1), ..., (X_i, y_i), ..., (X_I, y_I) and the sample numbers of each group, $\left(N_1^{(1)}, N_1^{(2)}, N_1^{(3)}\right)$, ..., $\left(N_i^{(1)}, N_i^{(2)}, N_i^{(3)}\right)$, ..., $\left(N_I^{(1)}, N_I^{(2)}, N_I^{(3)}\right)$

Output: p-value of the studying Genotype-Expression-Image trio

Initialize: $w^{(1)}, w^{(2)}, w^{(3)}, w^{(C)} = \mathbf{0}$

1: **for** $g = \{1, 2, 3, C\}$ **do**
2: **while** *convergence and maximum number of iterations are not reached* **do**
3: Get an image patch $\mathbf{x_i}$ from $\mathbf{X}$.
4: Each institution computes the gradient: $\nabla S_i^{(g)}\left(w^{(g)}\right) = [X_i^{(g)}]^T\left(X_i^{(g)} w^{(g)} - y_i^{g(g)}\right)$.
5: Global center computes and sends global gradient to each institution: $\nabla S^{(g)}\left(w^{(g)}\right) = \sum_{i=1}^{I}\nabla S_i^{(g)}\left(w^{(g)}\right)$.
6: Each institution updates the coefficient with the global gradient: $w^{(g)} \leftarrow w^{(g)} - \eta\nabla S^{(g)}\left(w^{(g)}\right)$.
7: **end while**
8: Each institution calculates the sum of squared residual: $S_i^{(g)}\left(w^{(g)}; X_i^{(g)}, y_i^{(g)}\right)$.
9: Global center gathers the global sum of squared residual: $S^{(g)} = \sum_{i=1}^{I} S_i^{(g)}$.
10: Global center gathers the global sample numbers: $N^{(g)} = \sum_{i=1}^{I} N_i^{(g)}$.
11: **end for**
12: Global center calculates F value with equation (1) and then computes and sends p-value to all institutions.

Performance Evaluation Protocol

We firstly use our model to identify AD-related gene expression. From the publicly available database, alzgene.org, and the GWAS results from International Genomics of Alzheimer's Project (IGAP) (6), we select 632 known AD-related genes. When we fix the genotype and imaging biomarker, we can calculate a p-value for each of the 20,211 gene expressions. We rank the 20,211 p-values and identify which of the known AD-related genes are featured in the top N gene expressions. In section Discovering AD-Related Gene Expressions, the top N gene expressions are the ones with a p-value < 0.05. In section Discovering AD-Related SNPs, N is a fixed number (100 and 200). We introduce hypergeometric analysis (49) to evaluate the model's performance to detect the known AD-related genes. The probability mass function of hypergeometric analysis is defined as,

$$p(k, M, n, N) = \frac{\binom{n}{k}\binom{M-n}{N-k}}{\binom{M}{N}} \tag{2}$$

In our case, the number of population (M) is 20,211, the sample size (n) is 632, the number of samples drawn from the population (n) is the selected top N gene expressions, and the number of the observed successes (k) is the number of overlapping genes between 632 known AD-related genes and the top N gene expressions.

Secondly, with different genotypes, the pattern of hypergeometric enrichment will vary. The AD-related genotypes should, in general, have a more significant hypergeometric enrichment. From alzgene.org, we also obtain 1217 known AD-related SNPs. And we randomly select 1217 SNPs from the ADNI database as non-AD-related SNPs. After ranking the SNPs

with the p-value based on hypergeometric analysis, we compute the number of AD-related SNPs found in the top m SNPs as true positive rate (TPR) and evaluate the performance of the models with TPR.

Finally, to prove the stability of our federated GEIDI, we compare the residuals of the federated linear regression model under different multi-site conditions. If the residuals are the same under different conditions, the F value and p-value will stay unchanged.

RESULTS

Discovering AD-Related Gene Expressions

APOE Related Gene Expressions

APOE genotype is a well-known genetic biomarker for predicting subjects' risk for AD. We stratify 697 subjects into three subgroups based on their *APOE* genotype status: non-carriers (e3/e3), heterozygotes (e3/e4), and homozygotes (e4/e4). Federated GEIDI is then adopted to discover genes correlated with hippocampus volume differentially across the three subgroups. We first run federated GEIDI with the volume of both sides of the hippocampus and the expression measures for 20,211 genes. Next, 1,625 gene expression measures are selected with $p < 0.05$. We evaluate the enrichment of these genes and the 632 AD-related genes annotated on alzgene.org and find 73 overlapping genes, yielding a hypergeometric enrichment $p = 0.00039$. Among the 73 overlapping genes, the top ten gene expressions are those measured for *CAST, CST3, GSTO1, LSS, MS4A4A, NPC1, PMVK, PPM1H, PPP2R2B,* and *SORCS2*. Besides, the top ten genes in the 1,625 gene expressions are *IK, BRPF3, BTN3A2, LOC101929275, TDRG1, PAFAH1B1, SERINC3, ALKBH6, VPS45, and LGALS1*. We also perform the false discovery rate (FDR) (50) test on the 20,211 p-values but none of the corrected p-values are significant. The list for these selected gene expressions is attached in **Supplementary Material (table 1.csv)**.

Additionally, we perform the same experiments on the volume of the middle temporal gyrus (MidTemp); the results are shown in **Table 2**. 2,415 gene expressions are significant and 92 of them overlap with the 632 AD-related genes - with a hypergeometric enrichment $p = 0.00624$. The top ten gene expressions are those measured for *ABCA2, COL11A1, CST3, GNA11, HMOX1, HSPA1B, MAOA, MS4A4A, PRKAB2,* and *SORCS2*. And the top ten genes in the 2,415 gene expressions are *GLRA3, CAMK2N2, MCOLN2, BPIFA1, KIT, CST3, SLC20A2, LGALS4, TNFSF8,* and *LCOR*. After performing FDR on the 20,211 genes, three gene expressions are significant, including *GLRA3*, *CAMK2N2*, and *BPIFA1*. The list for these selected gene expressions is attached in **Supplementary Material (table 2.csv)**.

Matrix eQTL (51) is a state-of-the-art software to study the association between genotype and gene expression. We also leverage the linear model and the ANOVA model in Matrix eQTL to evaluate the *APOE* genotype and the measured expression levels of the 20,211 genes. For the linear model, there are 2,657 significant gene expressions and 98 overlapping genes, leading to a hypergeometric enrichment $p = 9.76E-03$. For the ANOVA model, 3,234 gene expressions are selected, and 110 known genes are found, which leads to a p-value $= 2.665E-02$. The results show that our federated GEIDI can detect the most gene candidates that are significantly enriched for known AD genes. As the volume of hippocampus has the best performance in detecting AD-related genes, we use it as the imaging biomarker for all the remaining experiments.

SNP Related Gene Expressions

In this experiment, we stratify the subjects into three subgroups based on their SNP status. We choose *rs942439*, as this SNP was reported in alzgene.org, and also one of the top hits in our experiment of discovering AD-related SNPs (the details about selecting AD-related SNPs will be introduced in section Discovering AD-Related SNPs). And we use the volume of both sides of hippocampus as the imaging biomarker because of its superior performance in the first experiment. Federated GEIDI is used to detect any known AD gene whose expression is differentially associated with hippocampus volume in the subgroups stratified by the genotype at *rs942439* locus.

As shown in **Table 3**, 1,587 gene expressions are significant and 60 of them are reported in alzgene.org and IGAP GWAS results, leading to a hypergeometric enrichment $p = 0.017$. Of these 60 gene expression measures, the top ten genes are *ADRB1, ALOX5, ATXN1, CBS, FGF1, FLOT1, HSPA1A, RFTN1, SORL1,* and *XRCC1*. Besides, the top ten genes in the 1,587 gene expressions are *AIF1L, KRT23, CA2, C2ORF88, HSPA1A, LRGUK, LGALS3BP, IFT46, DDX23,* and *FAM166B*. After performing the FDR test on the 20,211 genes, none of the gene expression is significant. The list for these selected gene expressions is attached in **Supplementary Material (table 3.csv)**.

We also perform eQTL analysis on the SNP, *rs942439*. For a linear regression model, 1,794 gene expressions are selected, and, of these, 66 genes are reported in alzgene.org and IGAP GWAS results, yielding a hypergeometric enrichment $p = 0.021$. For the ANOVA model, 1,347 gene expression values are significant and,

TABLE 2 | Hypergeometric statistics for *APOE*.

Structures	Selected genes	Overlapping genes	*P*-value
Hippocampus	1,625	73	0.00039
MidTemp	2,415	92	0.00624
Linear regression	2,657	98	0.00976
ANOVA	3,234	110	0.02665

TABLE 3 | Hypergeometric statistics for *rs942439*.

Structures	Selected genes	Overlapping genes	*P*-value
Hippocampus	1,587	60	0.017
Linear regression	1,794	66	0.021
ANOVA	1,347	49	0.033

of these, 49 genes are reported in alzgene.org and IGAP; in this case, the hypergeometric enrichment was $p = 0.033$.

In the experiment, one of the most significant gene expression measures is for *XRCC1*, for which the *p*-value is $4.332E-03$. *XRCC1* is a gene coding for the X-ray repair cross-complementing protein; it was previously reported to be weakly associated with AD in a Turkish population (52).

As shown in **Figure 3**, we further adopt Pearson's correlation to evaluate the relationship between the hippocampal volume (*x*-axis) (adjusted for ICV) and *XRCC1* gene expression (*y*-axis) of each subgroup. **Figure 3A** illustrates the distribution for all the samples. **Figures 3B–D** show the distribution for the samples with "GG", "GA" and "AA" genotype, respectively. Above each subfigure, *R* and *p* are the Pearson correlation coefficient and *p*-value, and N is the number of subjects. Even so, there is always some missing information in the genotype data. Hence, before we run federated GEIDI as well as the Pearson correlation statistics, we remove the subjects without the specific genotype. Because of this, the total number N in **Figure 3A** is 579 instead of 697. We find samples with an "AA" genotype had hippocampal volume negatively correlated with expression levels of *XRCC1* ($N = 37$, $R = 0.37$, $p = 0.022$). In contrast, the analysis in all samples (**Figure 3A**) or subjects with either "GG" or "GA" genotype (**Figures 3B,C**) showed that the Pearson correlation coefficients were not significant in the overall, pooled sample. This result indicates that our method can establish associations among SNP, imaging, and gene expression data that include known AD risk factors.

We further apply the above procedure to discover genes that have never been reported to be associated with AD. As shown in **Figure 4D**, *SEC14L2* gene expression is negatively associated with hippocampal volume only in the subpopulation with "AA" genotype at rs942439 locus ($N = 37$, $R = -0.47$, $p = 0.003$). Interestingly, the opposite correlation is found in a subpopulation with "GA" genotype (**Figure 4C**, $N = 208$, $R = 0.15$, $P = 0.03$), and when applied to all pooled subjects, the total population does not show significant correlations (**Figure 4A**, $N = 579$, $R = 0.07$, $p = 0.09$) and the subpopulation with "GG" genotype doesn't show any significant correlations (**Figure 4B**, $N = 334$, $R = 0.062$, $p = 0.258$). The *SEC14L2* gene encodes a protein that stimulates squalene monooxygenase, a downstream enzyme in the cholesterol biosynthesis pathway.

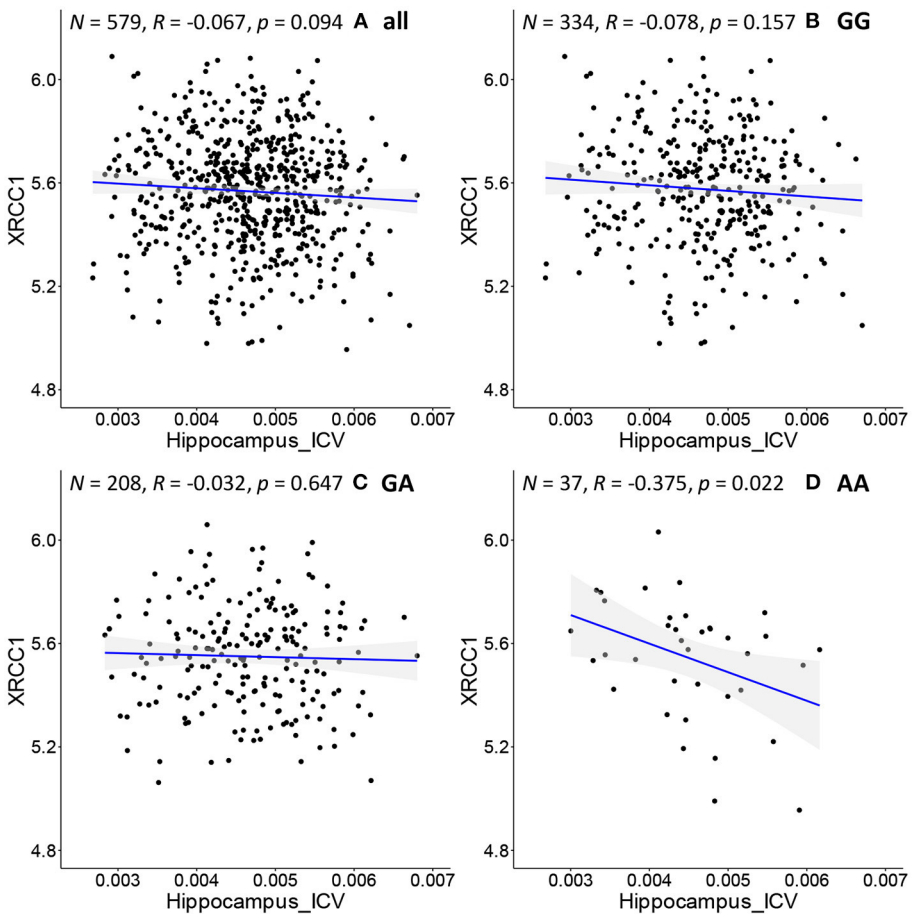

FIGURE 3 | Correlation of image biomarkers and *XRCC1* gene expression in subpopulations stratified by the sample's genotype at *rs942439*. **(A)** all samples **(B)** individuals with "GG" genotype; **(C)** those with "GA" genotype **(D)** those with "AA" genotype.

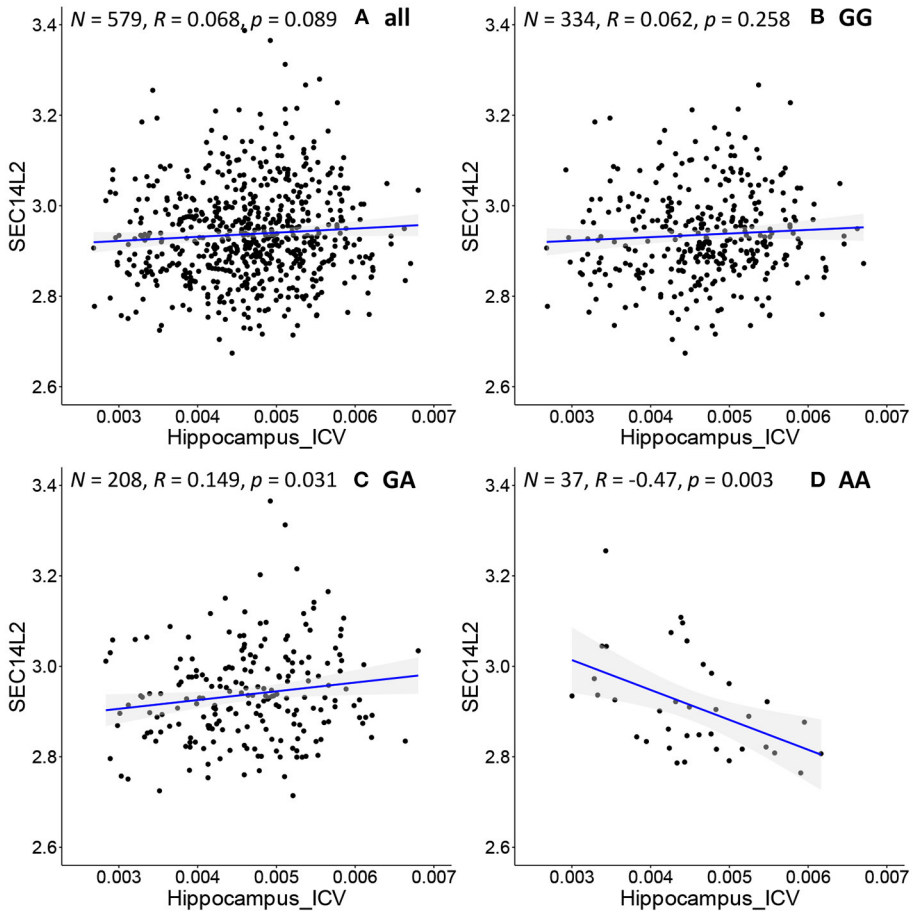

FIGURE 4 | Correlation of image biomarkers and *SEC14L2* gene expression in subpopulation stratified by the sample's genotype at *rs942439*. **(A)** all samples **(B)** those with "GG" genotype; **(C)** those with "GA" genotype; **(D)** those with "AA" genotype.

This gene has never been reported to be associated with AD, but high cholesterol levels have been linked to early-onset AD (53). This result indicates that our method can detect strong correlations in specific subpopulations that cannot be detected in the whole population. We also observe conflicting directions in different subpopulations, as shown by "GA" and "AA" subpopulations showing opposite correlations. This also highlights the importance of individualized medicine in patient management, as the same drug may have opposing effects in different groups of samples. Thus, federated GEIDI offers a new approach to discover novel genes related to AD as potential drug targets.

Discovering AD-Related SNPs

In the experiments of section Discovering AD-Related Gene Expressions, we used hypergeometric statistics to evaluate the ability of our proposed model to discover AD-related gene expressions that are differentially associated with imaging measures in populations stratified by *APOE* haplotype. In this experiment, we also use hypergeometric statistics to assess the discovery rate of known AD-related genes, in the set of genes whose expression shows different correlations with imaging markers, in samples stratified according to different genotypes. Sets that are enriched in AD-related SNPs will have a more significant *p*-value in the hypergeometric test that assesses enrichment. Since the hippocampal volume measure showed superior performance for this task, among all the imaging biomarkers in section Discovering AD-Related Gene Expressions, we adopt it as the brain imaging measure in this experiment. To illustrate the effectiveness of our GEIDI model, we perform the same experiment with the linear model in Matrix eQTL, which can evaluate the associations between SNPs and gene expression. To adjust for multiple comparisons, we will convert raw *p*-values to false discovery rate (FDR) and consider trios with FDR <0.05 as functionally important.

When we analyze each SNP with our federated GEIDI and Matrix eQTL, we will obtain a *p*-value for each of the 20,211 expressed genes. Instead of selecting the significant gene expressions with a *p*-value $<$ 0.05, we respectively rank the *p*-value of all the gene expressions calculated by the two methods and select the top N (100 and 200) gene expressions to apply the hypergeometric analysis. With the *p*-value from this

hypergeometric analysis (which assesses enrichment for known AD-associated genes), we may rank the SNPs and obtain the most AD-related ones. Then, we try to prove that our GEIDI is able to detect more AD-related SNPs. From alzgene.org, we also created a list of 1,217 AD-related SNPs, and we randomly selected another 1,217 SNPs as the non-AD-related ones. After ranking the SNPs with the *p*-value computed by the two methods, we calculate the true positive rate (TPR) for the top *m* SNPs, which measures the percentage of AD-related SNPs in the selected top *m* SNPs. For example, the last number in **Table 4** is 0.57, which means 57% of the top 500 SNPs are AD-related ones. As the results in **Table 4**, our federated GEIDI can always achieve superior performance than Matrix eQTL.

In **Figure 5**, we visualize the *p*-values of these 2,434 SNPs from hypergeometric analysis in the Manhattan plots. The top figure is the Manhattan plot for the result with the top 100 gene expressions and the bottom one is for the result of the top 200 gene expressions. The SNPs, *rs4889013* and *rs11940059*, are the top-ranked ones for both results. When we select 100 or 200 as the number of samples drawn from the population, three parameters in Equation (2) are fixed and only the number of observed successes, *k,* varies for different SNPs. Therefore, the *p*-value from different SNPs might be the same if their numbers of observed successes are the same. This explains why results of some SNPs locate at the same horizontal position.

Federated Learning Stability Analysis

In this experiment, we aim to demonstrate that the performance of our federated GEIDI model is not greatly affected by different data distribution models across institutions. In practice, it would be convenient and efficient to run association tests on data that might be distributed across multiple servers without transferring it all to a centralized location. We developed this algorithm with R language and simulated the distributed condition on a cluster with several conventional x86 nodes, of which each contains two Intel Xeon E5-2680 v4 CPUs running at 2.40 GHz. Each institution is assigned one computing node. We synthesized 1,000 samples and randomly assigned them to different independent hypothetical institutions, including one institution, three institutions, five institutions and seven institutions. We compared the residuals from each linear regression model for each condition and found the residuals remained unchanged, as shown in **Table 5**. The first column is the ground truth residual and the rest are the residuals for our federated linear model under different data distribution conditions. The residuals are the same, which means that the results of our Federated GEIDI will remain stable under different multi-site conditions. Therefore, these results demonstrate the correctness and stability of our federated GEIDI model.

TABLE 4 | True Positive Rates of AD-related SNPs in the top *m* SNPs.

SNP (*m*) / EXP (*N*)	10	50	100	200	500
Matrix eQTL: linear regression					
100	0.50	0.58	0.52	0.50	0.53
200	0.50	0.60	0.55	0.58	0.54
Matrix eQTL: ANOVA					
100	0.60	0.58	0.52	0.49	0.55
200	**0.60**	0.60	0.57	0.55	0.55
Federated GEIDI					
100	**0.60**	**0.60**	**0.60**	**0.61**	**0.60**
200	**0.60**	**0.62**	**0.61**	**0.58**	**0.57**

The SNPs are ranked with the p-value from hypergeometric analysis with the top. N gene expressions as the number of samples drawn from the population.

DISCUSSION

In this work, we propose a novel federated Genotype-Expression-Imaging Data Integration (GEIDI) model to identify the genetic and transcriptomic influences on brain sMRI measures. We performed various experiments with our model on the publicly available ADNI dataset, and we have two main findings. First, our federated GEIDI is an effective multimodal approach that provides novel insights into the relationship among image biomarkers, genotypes, and gene expression, and may be useful to discover novel genes as potential AD drug targets. It has better performance in detecting AD-related gene expressions and SNPs than the linear regression model and ANOVA model in the state-of-the-art Matrix eQTL approach. In addition, our model may not only detect known AD-associated genes as potential drug targets, such as *XRCC1*, but may also help in discovering novel genes as potential drug targets, such as *SEC14L2*. Second, compared to Matrix eQTL, our federated GEIDI provides a way to investigate extremely large datasets from different institutions without violating data privacy. The statistical power of the model will also be increased with the larger sample size. Our work may lay down a solid foundation for future multi-site large-scale imaging genetics research.

Comparison Analysis of Federated GEIDI and Matrix eQTL

Expression quantitative trait loci (eQTL) analysis (54, 55) is designed to identify the significant associations between SNPs and gene expression, which can help understand the biochemical processes occurring in living systems, discover the genetic factors that influence the onset and progression of certain diseases, and determine the pathways affected by them. There are many eQTL analysis methods, including linear regression, ANOVA models, Bayesian regression (56), and so on. Matrix eQTL (51) is the state-of-the-art software for computationally efficient eQTL analysis, and it supports additive linear and ANOVA models. It has been widely used in the study of human genetic traits and diseases. However, it has two main limitations. First, although Matrix eQTL is very computationally efficient, it cannot work on data that is distributed across different institutions. Nowadays, unprecedentedly large volumes of biomedical and genetic data have been collected by different hospitals and research institutions, and this aggregate of available data may significantly advance the study of factors influencing disease.

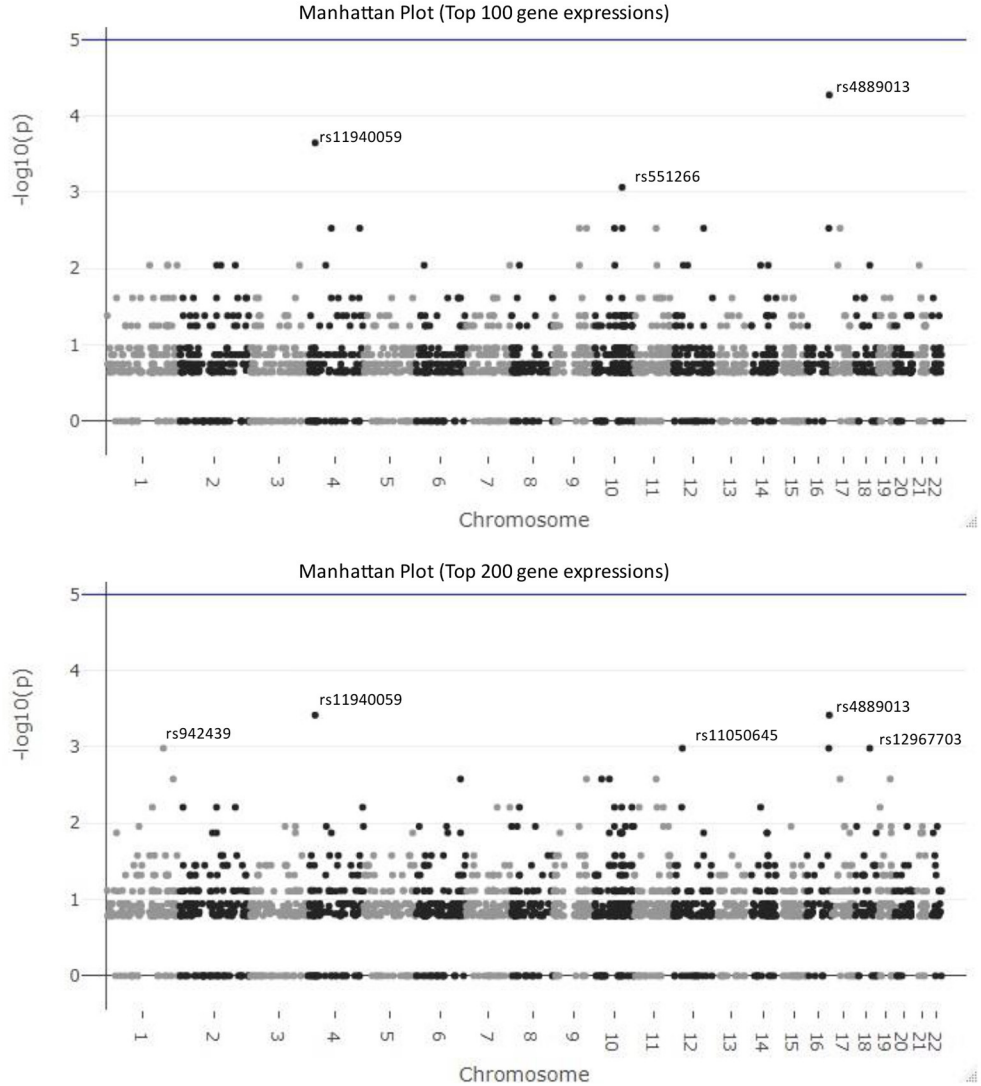

FIGURE 5 | Manhattan plots for the results of federated GEIDI. The top figure is the Manhattan plot for the results from hypergeometric analysis with the top 100 gene expressions as the number of samples drawn from the population and the bottom one is for the results with the top 200 gene expressions. The SNPs, *rs4889013* and *rs11940059*, are the top-ranked ones for both results.

TABLE 5 | Stability analysis of federated GEIDI across different institutional settings.

	Ground truth	1-institution	3-institution	5-institution	7-institution
Residual	3.9553	3.9553	3.9553	3.9553	3.9553

However, data restrictions, legal complexities, and patient privacy have all been major obstacles for researchers to obtain or share these data. Therefore, federated machine learning and distributed statistical models are becoming advantageous for current research on medical data (57, 58). Second, the models in Matrix eQTL cannot jointly consider the information from images. Changes in brain structures can play a vital role in the study and diagnosis of Alzheimer's disease, and many researchers have attempted to detect associations between genetic factors and imaging features (21, 59, 60). Therefore, introducing imaging information may greatly assist the detection of genetic factors that influence disease as an intermediate phenotype that might reflect relevant disease processes.

Our proposed federated Genotype-Expression-Imaging Data Integration model can effectively overcome these two obstacles. In the Methods section, we detailed how our model maintains each institutional data private. Additionally, our federated GEIDI model integrates GWAS data, gene expression, and imaging data. The experimental results demonstrate that our federated GEIDI model has better performance in detecting AD-related genes and SNPs. In detecting AD-related gene expression, our model achieves the strongest hypergeometric enrichment with

the volume of the hippocampus. In our tests detecting AD-related SNPs, our federated GEIDI model generally obtained a higher TPR than the linear regression model and ANOVA model. Besides, compared with existing methods, our proposed model offers novel insights into the relationship among image biomarkers, genotypes, and gene expression by considering both imaging and gene expression features—which can vary over time—and understanding how they are affected by an individual's SNPs. Compared with Matrix eQTL, the only caveat of our model is the computation time on a single computing node. Our proposed model may require more computation time compared to Matrix eQTL. For each trio, our framework has to solve four linear regression models. But Matrix eQTL only needs to calculate one correlation matrix to evaluate all the trios. Therefore, our model may require more computation time. We perform the experiments of section APOE Related Gene Expressions on a single computing node. Matrix eQTL only takes 1.4 s, while our model may require 265 s. However, due to the federated learning nature, our work may be applied to different computation nodes parallelly. It may make our work scalable to large datasets and result in comparable computation times with Matrix eQTL.

Drug Target for Precision Medicine of AD

Increasingly, a major challenge in healthcare is that many drugs are adequate for only small subgroups of patients (61). Some patients may not only suffer from adverse side effects but also waste money on ineffective drugs. Precision medicine has the potential to tailor therapy based on the best expected response and highest safety margin to ensure better patient care. By enabling each patient to receive earlier diagnoses, risk assessments, and optimal treatments, personalized medicine holds promise for improving health care while also potentially lowering costs (10). In this work, our multi-omics approach offers potential in genome-guided drug discovery. Compared to state-of-the-art methods, our model performs better in detecting AD-related genes and SNPs. Moreover, our model not only detects known genes for target drugs, like *XRCC1,* but also discovers novel potential gene expressions, like *SEC14L2*. Meanwhile, our federated framework may integrate data from multiple sources without violating the data privacy and the obtained larger sample size may help discover and understand more AD-related genetic information. Therefore, we believe our federated GEIDI model will play an important role in the study of precision medicine for AD in the future.

Limitations and Future Work

Despite the promising results of our federated GEIDI model, there are four caveats. Firstly, we only evaluated our model on data from 697 subjects from the publicly available ADNI dataset. In the future, we will add other datasets to make results more robust and reliable. For example, the Arizona APOE cohort (AZ APOE cohort) recruited 450 actively followed participants matched by age, sex, and education—including homozygous *APOE-e4* carriers and non-*e4* carriers since 1994 (62). The UK Biobank project (63) collects both large-scale genetic-genomic and phenotypic data as well as health-related information from around 500,000 volunteer participants in the UK. Assessments include biological measures, blood- and urine-based biomarkers, body, and brain imaging scans, and lifestyle parameters (64, 65). Second, the volumes of specific subcortical structures may not be ideal imaging measurements for the multiple biological processes involved in Alzheimer's disease. Surface-based morphometry analyses have achieved excellent performance for early AD detection (66–68). In recent work (69, 70), the authors created tools to generate a univariate morphometry index (UMI) for surface morphometry features on regions of interest (ROIs) that are related to beta-amyloid deposition. This induced UMI may reflect intrinsic morphological changes induced by processes of amyloid accumulation in AD and has greater signal-to-noise ratio and strong generalizability to new subjects. If we were to use such brain pathology induced UMI measures instead of volumes, our federated GEIDI model may detect additional AD-related genes whose expression is influenced by SNPs. Thirdly, all the three methods (our GEIDI model, linear regression model, and ANOVA model) can only get several significant corrected p-values on this dataset with the FDR test since only about 10 percent of the 20,211 gene expressions can get significant raw p-values. Therefore, we use hypergeometric analysis to evaluate the top findings of these methods. The hypergeometric analysis is also a classical approach to evaluate the discovery significance in genetics research (71–73). It may justify our model. In the future, we will try to apply our work to larger datasets and apply the multiple test correction models for justification. Finally, in ongoing work on blood-based biomarkers (74, 75), plasma levels of amyloid-beta (plasma Aβ) may provide an alternative but highly accurate estimate of brain amyloid positivity. In (75), plasma P-Tau181 accurately discriminated AD dementia from non-AD neurodegenerative diseases with an excellent AUC (0.94). Similarly, such plasma measures might be used in conjunction with our federated GEIDI model to better understand the effects of AD-related genotypes. We plan to analyze such datasets to further evaluate our model in the future.

CONCLUSION

We propose a novel federated Genotype-Expression-Image Data Integration model. Compared to similar studies, this work achieves state-of-the-art performance in discovering downstream effects of AD-related genes and SNPs. Besides, the model provides novel insights into the relationship among image biomarkers, genotypes, and gene expression and could discover novel drug targets for precision medicine. In the future, we will further validate our model with more datasets and more advanced imaging biomarkers. Specifically, we will introduce blood-based biomarkers into our model when such data are available.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: http://adni.loni.usc.edu/data-samples/data-types/genetic-data/.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Alzheimer's Disease Neuroimaging Initiative. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

ALZHEIMER'S DISEASE NEUROIMAGING INITIATIVE

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DoD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI was funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer's Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

AUTHOR CONTRIBUTIONS

JWu: methodology, investigation, formal analysis, and writing—original draft. YC: investigation. PW: methodology and conceptualization. RC: review and editing. PT: methodology and review and editing. JWa: conceptualization, supervision, funding acquisition, and writing—review and editing. YW: conceptualization, investigation, supervision, funding acquisition, and writing—review and editing. All authors contributed to the article and approved the submitted version.

FUNDING

Algorithm development and image analysis for this study were partially supported by the ASU/Mayo Seed Grant Program, the National Institute on Aging (RF1AG051710, R21AG065942, U01AG068057, R01AG069453, and P30AG072980), the National Library of Medicine (R01LM013438), the National Institute of Biomedical Imaging and Bioengineering (R01EB025032), the National Eye Institute (R01EY032125), and the Arizona Alzheimer's Consortium.

ACKNOWLEDGMENTS

Data used in preparing this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, many investigators within the ADNI contributed to the design and implementation of ADNI 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 at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fradi.2021.777030/full#supplementary-material

REFERENCES

1. Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. *Alzheimer Dementia.* (2007) 3:186–91. doi: 10.1016/j.jalz.2007.04.381
2. Hyman BT. Amyloid-dependent and amyloid-independent stages of Alzheimer disease. *Arch Neurol.* (2011) 68:1062–4. doi: 10.1001/archneurol.2011.70
3. Jill M, Martin F, Bernardino G, Benson MD. A mutation in the amyloid precursor protein associated with hereditary Alzheimer's disease. *Science.* (1991) 254:97–9. doi: 10.1126/science.1925564
4. Kunkle BW, Grenier-Boley B, Sims R, Bis JC, Damotte V, Naj AC, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet.* (2019) 51:414–30. doi: 10.1038/s,41588-019-0358-2
5. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. *Nat Genet.* (2007) 39:17–23. doi: 10.1038/ng1934
6. Lambert, J.-C., Ibrahim-Verbaas CA, Harold D, Naj AC, Sims R, et al. Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat Genet.* (2013) 45:1452–8. doi: 10.1038/ng.2802
7. Mormino EC, Sperling RA, Holmes AJ, Buckner RL, de Jager PL, Smoller JW, et al. Polygenic risk of Alzheimer disease is associated with early- and late-life processes. *Neurology.* (2016) 87:481–8. doi: 10.1212/WNL.0000000000002922
8. Singanamalli A, Wang H, Madabhushi A, Initiative ADN. Cascaded multi-view canonical correlation (CaMCCo) for early diagnosis of Alzheimer's disease via fusion of clinical, imaging and omic features. *Sci Rep.* (2017) 7:8137. doi: 10.1038/s41598-017-03925-0
9. Freudenberg-Hua Y, Li W, Davies P. The role of genetics in advancing precision medicine for alzheimer's disease—a narrative review. *Front Med.* (2018) 5:108. doi: 10.3389/fmed.2018.00108
10. Vogenberg FR, Isaacson Barash C, Pursel M. Personalized medicine: part 1: evolution and development into theranostics. *PT.* (2010) 35:560–76.
11. Cummings JL, Morstorf T, Zhong K. Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimers Res Ther.* (2014) 6:37. doi: 10.1186/alzrt269

12. Mehta D, Jackson R, Paul G, Shi J, Sabbagh M. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010-2015. *Exp Opin Invest Drugs.* (2017) 26:735–9. doi: 10.1080/13543784.2017.1323868
13. Pimplikar SW. Multi-omics and Alzheimer's disease: a slower but surer path to an efficacious therapy? *Am J Physiol Cell Physiol.* (2017) 313:C1–2. doi: 10.1152/ajpcell.00109.2017
14. Xicota L, Ichou F, Lejeune F-X, Colsch B, Tenenhaus A, et al. Multi-omics signature of brain amyloid deposition in asymptomatic individuals at-risk for Alzheimer's disease: The INSIGHT-preAD study. *EBio Med.* (2019) 47:518–28. doi: 10.1016/j.ebiom.2019.08.051
15. Saykin AJ, Shen L, Yao X, Kim S, Nho K, Risacher SL, et al. Genetic studies of quantitative MCI and AD phenotypes in ADNI: progress, opportunities, and plans. *Alzheimers Dementia.* (2015) 11:792–814. doi: 10.1016/j.jalz.2015.05.009
16. Simino J, Wang Z, Bressler J, Chouraki V, Yang Q, Younkin SG, et al. Whole exome sequence-based association analyses of plasma amyloid-β in African and European Americans; the atherosclerosis risk in communities-neurocognitive study. *PLoS ONE.* (2017) 12:e0180046–0046. doi: 10.1371/journal.pone.0180046
17. Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, et al. Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation. *Mol Psychiatry.* (2020) 25:1859–75. doi: 10.1038/s41380-018-0112-7
18. Piras IS, Krate J, Schrauwen I, Corneveaux JJ, Serrano GE, Sue L, et al. Whole transcriptome profiling of the human hippocampus suggests an involvement of the KIBRA rs17070145 polymorphism in differential activation of the MAPK signaling pathway. *Hippocampus.* (2017) 27:784–93. doi: 10.1002/hipo.22731
19. Luningham JM, Chen J, Tang S, de Jager PL, Bennett DA, Buchman AS, et al. Bayesian genome-wide TWAS method to leverage both cis- and trans-eQTL information through summary statistics. *Am J Human Genet.* (2020) 107:714–26. doi: 10.1016/j.ajhg.2020.08.022
20. Johnson KA, Fox NC, Sperling RA, Klunk WE. Brain imaging in Alzheimer disease. *Cold Spring Harbor Perspect Med.* (2012) 2:a006213. doi: 10.1101/cshperspect.a006213
21. Shen L, Thompson PM. Brain imaging genomics: integrated analysis and machine learning. *Proc IEEE Inst Electr Electron Eng.* (2020) 108:125–62. doi: 10.1109/JPROC.2019.2947272
22. Chauhan G, Adams HHH, Bis JC, Weinstein G, Yu L, Töglhofer AM, et al. Association of Alzheimer's disease GWAS loci with MRI markers of brain aging. *Neurobiol Aging.* (2015) 36:1765.e7–16. doi: 10.1016/j.neurobiolaging.2014.12.028
23. Li J-Q, Wang H-F, Zhu X-C, Sun F-R, Tan M-S, Tan C-C, et al. GWAS-linked loci and neuroimaging measures in Alzheimer's disease. *Mol Neurobiol.* (2017) 54:146–53. doi: 10.1007/s12035-015-9669-1
24. Grasby KL, Jahanshad N, Painter JN, Colodro-Conde L, Bralten J, Hibar DP, et al. The genetic architecture of the human cerebral cortex. *Science.* (2021) 367:eaay6690. doi: 10.1126/science.aay6690
25. Ritchie J, Pantazatos SP, French L. Transcriptomic characterization of MRI contrast with focus on the T1-w/T2-w ratio in the cerebral cortex. *NeuroImage.* (2018) 174:504–17. doi: 10.1016/j.neuroimage.2018.03.027
26. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet.* (2015) 16:197–212. doi: 10.1038/nrg3891
27. Stein JL, Hua X, Lee S, Ho AJ, Leow AD, Toga AW, et al. Voxelwise genome-wide association study (vGWAS). *NeuroImage.* (2010) 53:1160–74. doi: 10.1016/j.neuroimage.2010.02.032
28. Zhang A, Zhao Q, Xu D, Jiang S. Brain APOE expression quantitative trait loci-based association study identified one susceptibility locus for Alzheimer's disease by interacting with APOE ε4. *Sci Rep.* (2018) 8:8068. doi: 10.1038/s41598-018-26398-1
29. Liu K, Yao X, Yan J, Chasioti D, Risacher S, Nho K, et al. Transcriptome-guided imaging genetic analysis via a novel sparse CCA algorithm. *Graphs Biomed Image Anal Comput Anat Imaging Genet.* (2017) 10551:220–9. doi: 10.1007/978-3-319-67675-3_20
30. Hampel H, Vergallo A, Perry G, Lista S. The Alzheimer precision medicine initiative. *J Alzheimers Dis.* (2019) 68:1–24. doi: 10.3233/JAD-181121
31. Thompson PM, Jahanshad N, Ching CRK, Salminen LE, Thomopoulos SI, Bright J, et al. ENIGMA and global neuroscience: a decade of large-scale studies of the brain in health and disease across more than 40 countries. *Transl Psychiatry.* (2020) 10:100. doi: 10.1038/s41398-020-0705-1
32. Hibar DP, Stein JL, Renteria ME, Arias-Vasquez A, Desrivières S, Jahanshad N, et al. Common genetic variants influence human subcortical brain structures. *Nature.* (2015) 520:224–9. doi: 10.1038/nature14101
33. Satizabal CL, Adams HHH, Hibar DP, White CC, Knol MJ, Stein JL, et al. Genetic architecture of subcortical brain structures in 38,851 individuals. *Nat Genet.* (2019) 51:1624–36. doi: 10.1038/s41588-019-0511-y
34. Zhao B, Li T, Yang Y, Wang X, Luo T, Shan Y, et al. Common genetic variation influencing human white matter microstructure. *Science.* (2021) 372:eabf3736. doi: 10.1126/science.abf3736
35. Smit DJA, Wright MJ, Meyers JL, Martin NG, Ho YYW, Malone SM, et al. Genome-wide association analysis links multiple psychiatric liability genes to oscillatory brain activity. *Human Brain Mapping.* (2018) 39:4183–95. doi: 10.1002/hbm.24238
36. Chow GC. Tests of equality between sets of coefficients in two linear regressions. *Econometrica.* (1960) 28:591–605. doi: 10.2307/1910133
37. Marinescu RV, Oxtoby NP, Young AL, Bron EE, Toga AW, et al. *The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge: Results after 1 Year Follow-up*. (2020). Available online at: http://arxiv.org/abs/2002.03419 (accessed December 26, 2021).
38. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Human Genet.* (2007) 81:559–75. doi: 10.1086/519795
39. Yip SH, Wang P, Kocher J-PA, Sham PC, Wang J. Linnorm: improved statistical analysis for single cell RNA-seq expression data. *Nucleic Acids Res.* (2017) 45:e179. doi: 10.1093/nar/gkx1189
40. Huang D, Yi X, Zhang S, Zheng Z, Wang P, Xuan C, et al. GWAS4D: multidimensional analysis of context-specific regulatory variant for human complex diseases and traits. *Nucleic Acids Res.* (2018) 46:W114–20. doi: 10.1093/nar/gky407
41. Yip SH, Sham PC, Wang J. Evaluation of tools for highly variable gene discovery from single-cell RNA-seq data. *Brief Bioinform.* (2019) 20:1583–9. doi: 10.1093/bib/bby011
42. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol.* (2015) 33:495–502. doi: 10.1038/nbt.3192
43. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* (2014) 15:550. doi: 10.1186/s13059-014-0550-8
44. Fischl B, Sereno MI, Dale AM. Cortical surface-based analysis: ii: inflation, flattening, and a surface-based coordinate system. *NeuroImage.* (1999) 9:195–207. doi: 10.1006/nimg.1998.0396
45. Lee B, Wang J, Shen L. Identifying precision AD biomarkers with varying prognosis effects in genetics driven subpopulations. *AAIC'21: Alzheimer's Association Int. Conf. on Alzheimer's Disease, Denver* San Diego (2021).
46. Barbur VA, Montgomery DC, Peck EA. Introduction to linear regression analysis. *Statistician.* (1994) 43:339–41. doi: 10.2307/2348362
47. Rawlings JO, Pantula SG, Dickey DA. *Applied Regression Analysis: A Research Tool.* 2nd ed. New York, NY: Springer (1998).
48. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics.* (1970) 12:55–67. doi: 10.1080/00401706.1970.10488634
49. Berkopec A. HyperQuick algorithm for discrete hypergeometric distribution. *J Discrete Algorithms.* (2007) 5:341–7. doi: 10.1016/j.jda.2006.01.001
50. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Stat Soc Series B.* (1995) 57:289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x
51. Shabalin AA. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics.* (2012) 28:1353–8. doi: 10.1093/bioinformatics/bts163
52. Dogru-Abbasoglu S, Aykaç-Toker G, Hanagasi HA, Gürvit H, Emre M, Uysal M. The Arg194Trp polymorphism in DNA repair gene XRCC1 and the risk for sporadic late-onset Alzheimer's disease. *Neurol Sci.* (2007) 28:31–4. doi: 10.1007/s10072-007-0744-x

53. Wingo TS, Cutler DJ, Wingo AP, Le N-A, Rabinovici GD, et al. Association of early-onset alzheimer disease with elevated low-density lipoprotein cholesterol levels and rare genetic coding variants of APOB. *JAMA Neurol.* (2019) 76:809–17. doi: 10.1001/jamaneurol.2019.0648
54. Rockman MV, Kruglyak L. Genetics of global gene expression. *Nat Rev Genet.* (2006) 7:862–72. doi: 10.1038/nrg1964
55. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. Philosophical transactions of the royal society of london. Series B. *Biol Sci.* (2013) 368:20120362. doi: 10.1098/rstb.2012.0362
56. Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* (2007) 3:e114. doi: 10.1371/journal.pgen.0030114
57. Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Sci Rep.* (2020) 10:12598. doi: 10.1038/s41598-020-69250-1
58. Ng D, Lan X, Yao MM-S, Chan WP, Feng M. Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets. *Quant Imaging Med Surgery.* (2021) 11:852–7. doi: 10.21037/qims-20-595
59. Dong Q, Zhang W, Wu J, Li B, Schron EH, McMahon T, et al. Applying surface-based hippocampal morphometry to study APOE-E4 allele dose effects in cognitively unimpaired subjects. *NeuroImage Clin.* (2019) 22:101744. doi: 10.1016/j.nicl.2019.101744
60. Yan J, Raja VV, Huang Z, Amico E, Nho K, et al. Brain-wide structural connectivity alterations under the control of Alzheimer risk genes. *Int J Comput Biol Drug Design.* (2020) 13:58–70. doi: 10.1504/IJCBDD.2020.105098
61. Schork NJ. Personalized medicine: time for one-person trials. *Nature.* (2015) 520:609–11. doi: 10.1038/520609a
62. Caselli RJ, Dueck AC, Osborne D, Sabbagh MN, Connor DJ, Ahern GL, et al. Longitudinal modeling of age-related memory decline and the APOE ε4 effect. *N Engl J Med.* (2009) 361:255–63. doi: 10.1056/NEJMoa0809437
63. Cox N. UK biobank shares the promise of big data. *Nature.* (2018) 562:194–5. doi: 10.1038/d41586-018-06948-3
64. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK biobank resource with deep phenotyping and genomic data. *Nature.* (2018) 562:203–9. doi: 10.1038/s41586-018-0579-z
65. Elliott LT, Sharp K, Alfaro-Almagro F, Shi S, Miller KL, Douaud G, et al. Genome-wide association studies of brain imaging phenotypes in UK biobank. *Nature.* (2018) 562:210–6. doi: 10.1038/s41586-018-0571-7
66. Wang Y, Zhang J, Gutman B, Chan TF, Becker JT, Aizenstein HJ, et al. Multivariate tensor-based morphometry on surfaces: application to mapping ventricular abnormalities in HIV/AIDS. *NeuroImage.* (2010) 49:2141–57. doi: 10.1016/j.neuroimage.2009.10.086
67. Zhang J, Li Q, Caselli RJ, Thompson PM, Ye J, Wang Y. Multi-source multi-target dictionary learning for prediction of cognitive decline. *Inf Process Med Imaging.* (2017) 10265:184–97. doi: 10.1007/978-3-319-59050-9_15
68. Wu J, Zhang J, Shi J, Chen K, Caselli RJ, Reiman EM, et al. Hippocampus morphometry study on pathology-confirmed alzheimer's disease patients with surface multivariate morphometry statistics. *Proc IEEE Int Symp BiomedImaging.* (2018) 2018:1555–9. doi: 10.1109/ISBI.2018.8363870
69. Wang G, Dong Q, Wu J, Su Y, Chen K, Su Q, et al. Developing univariate neurodegeneration biomarkers with low-rank and sparse subspace decomposition. *Med Image Anal.* (2020) 67:1361–8415. doi: 10.1016/j.media.2020.101877
70. Wu J, Dong Q, Zhang J, Su Y, Wu T, Caselli RJ, et al. Federated morphometry feature selection for hippocampal morphometry associated beta-amyloid and tau pathology. *Front Neurosci.* (2021) 15:1585. doi: 10.3389/fnins.2021.762458
71. Fury W, Batliwalla F, Gregersen PK, Li W. Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. In: *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*. New York (2006). p. 5531–4.
72. Falcon S, Gentleman R. Hypergeometric testing used for gene set enrichment analysis. In: Hahne F, Huber W, Gentleman R, Falcon S, editors. *Bioconductor Case Studies.* New York, NY: Springer New York (2008). p. 207–20.
73. Plaisier SB, Taschereau R, Wong JA, Graeber TG. Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures. *Nucleic Acids Res.* (2010) 38:e169–69. doi: 10.1093/nar/gkq636
74. Bateman RJ, Blennow K, Doody R, Hendrix S, Lovestone S, Salloway S, et al. Plasma biomarkers of AD emerging as essential tools for drug development: an EU/US CTAD task force report. *J Prev Alzheimers Dis.* (2019) 6:169–73. doi: 10.14283/jpad.2019.21
75. Janelidze S, Mattsson N, Palmqvist S, Smith R, Beach TG, Serrano GE, et al. Plasma P-tau181 in Alzheimer's disease: relationship to other biomarkers, differential diagnosis, neuropathology and longitudinal progression to Alzheimer's dementia. *Nat Med.* (2020) 26:379–86. doi: 10.1038/s41591-020-0755-1

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.