WAGENINGEN
UNIVERSITY & RESEARCH

Biometris
Quantitative methods brought to life

ApoE4 dose effects on serum metabolites in Alzheimer's Disease

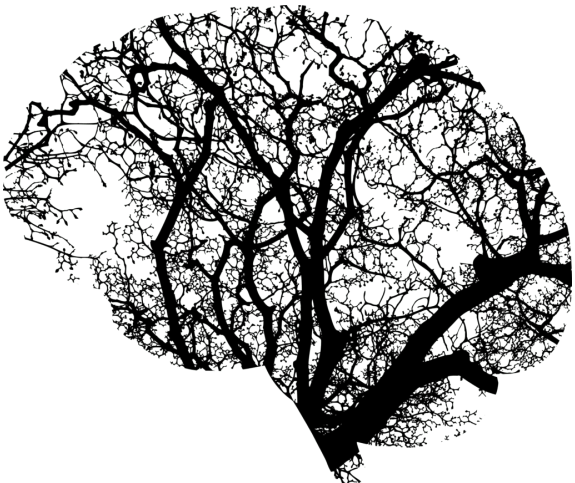
a Data Science approach

George Miliarakis

MSc Thesis
Wageningen University & Research
Wageningen, March 2024

Supervisor
C.F.W. Peeters
Mathematical & Statistical Methods (Biometris)
Wageningen University & Research

Supervisor
Yannick Vermeiren
Nutrition, Brain and Cognitive Ageing
Wageningen University & Research



Foreword

This is where you type your foreword. In the Foreword the student states clearly his/her contribution that originates from the master thesis project, and, if applicable, what contribution(s) possibly followed from the student's internship on the same topic. It is also stated in the foreword what data sources were used, and whether the data have a degree of confidentiality. In addition, possible acknowledgements may be made to people who have contributed to (certain parts of) the thesis.

Abstract

This is where you type your abstract. The abstract must be communicable to a broad audience and contain, next to a summary text, an outreach item such as an infographic or a link to a video/website/web application. Such as, for example, the nice figure below.

Abbreviations

ABCA1: ATP-binding cassette transporter A1. 3, 4
ApoE: Apo-lipoprotein E. 3
APP: amyloid precursor protein. 4
ATP: Adenosine tri-phosphate. 3
AUC: Area Under the ROC curve. 11, 15

BBB: blood-brain barrier. 3, 4

CSF: cerebrospinal fluid. 2

FA: Factor Analysis. 11
FAD: familial Alzheimer's Disease. 1

GGM: Gaussian graphical model. 12
GSK-3β: Glycogen synthase kinase-3β. 5

HDL: high density lipoprotein. 3

IDL: intermediate density lipoprotein. 3
IL: interleukin. 5
ISF: interstitial fluid. 4

LASSO: Least Absolute Shrinkage and Selection Operator. 11
LDLR: low-density lipoprotein receptor. 4
LOAD: late-onset Alzheimer's Disease. 1
LRP1: low density lipoprotein receptor-related protein 1. 4

ML: Maximum Likelihood. 11
MNL: Multinomial Logistic Regression. 10

NFkB: nuclear factor kappa B. 3

ROS: reactive oxygen species. 5

SAD: sporadic Alzheimer's Disease. 1
SMOTE: Synthetic Minority Over-Sampling Technique. 10

TLR4: toll-like receptor 4. 3
TNF-α: tumor necrosis factor α. 5

VCS: Version Control System. 8
VLDL: very low density lipoprotein. 3

XGBoost: eXtreme Gradient Boosting. 7, 11

Contents

1. Introduction

1.1. **Alzheimer's Disease.** Alzheimer's Disease (AD) is a complex, progressive neurodegenerative disorder and the most common form of dementia [1]. It was considered the 6th leading cause of death in the US in 2019, with an overall increase of 145% in mortality from 2000 to 2019 [2]. The impact AD has on patients, their caregivers, and healthcare systems is detrimental. Hence, a considerable amount of research has been performed in an effort to understand, prevent, impede or cure it. Nonetheless, important aspects of its systemic manifestations remain unknown.

Age is the main risk factor for AD, while several genetic and lifestyle risk factors, as well as biochemical pathways contribute to its development [1]. AD ~~occurs~~ in various histopathological phenotypes and presents a broad spectrum of clinical signs and symptoms [3, 4]. The AD continuum starts with subjective cognitive decline (SCD), followed by mild cognitive impairment (MCI) [5], and continues with progressive loss of global cognition, of which particularly memory, processing speed and executive functioning, spanning a total period of 10-15 years [6].

Sporadic or late-onset AD (SAD, LOAD; 95% of cases) is the most frequent phenotype, typically appearing after 65 years of age [7]. A rarer phenotype is early-onset familial AD (FAD), usually starting at ages 30–65 and passed in an autosomal dominant fashion [8]. Even though FAD mutations explain only a small percentage of AD cases, they have a great impact on AD research given their appealing genotype-phenotype links.

1.1.1. *Amyloid cascade hypothesis.* The amyloid ~~plaque~~ hypothesis has dominated the scientific discussion on the pathogenesis of AD. In historical terms, its impact was profound, as it helped distinguish and identify AD as a single disease that may be studied for treatment [9]. It suggests that chronic neuroinflammation promotes protein misfolding and accumulation in the brain, forming plaques (consisting of oligomerized amyloid Aβ_{42}) and tangles (consisting of hyperphosphorylated tau protein) [4]. Nevertheless, it does not necessarily cover all AD cases; clinical trials ~~of~~ Aβ anti-bodies as treatment ~~prove the amyloid cascade hypothesis as insufficient~~ [10, 11]. Evidence suggests that oxidative stress, metabolic abnormalities, atherosclerosis, cardiovascular effects, imbalances of intra-neuronal calcium and ~~other~~ metal ions contribute to the development of AD [10]. Kepp *et al.* propose a more complex and holistic view of AD pathology, by integrating (epi-)genetic, environmental, vascular, neuro-inflammatory and metabolic factors in predictive models [10].

1.2. **Human apolipoprotein-E gene.** LOAD is consistently shown associated with the gene that encodes apolipoprotein-E (ApoE) [12]. The structure and function of ApoE, as well as how the first defines the latter is described in Section 1.3.

1.2.1. *ApoE4 and Alzheimer's ~~D~~isease.* Humans present three variants of the ApoE gene: ε_2, ε_3 and ε_4 [13, 14], resulting in six genotypes [Fig. 1]. In caucasian populations, the most abundant allele is ApoE3 (rs7412 C/rs429358 T), with a frequency of 78%, and is considered neutral regarding the risk of AD [15]. ApoE4 (rs7412 C/rs429358 C) has a frequency of 14% and represents the strongest genetic risk factor for LOAD, with gene dose effects [16]. Conversely, ApoE2 (rs7412 T/rs429358 T) is found in 8% of Caucasian populations, and carriers of this allele exhibit a reduced risk of LOAD [15]. The various ApoE alleles are strongly linked to the primary pathological features of LOAD, namely amyloid-~~A~~β and phosphorylated tau [17]. While the association between ApoE alleles and LOAD risk or protection is observed across diverse ancestral backgrounds, the strength of this association varies [18, 19], see Section 1.2.2.

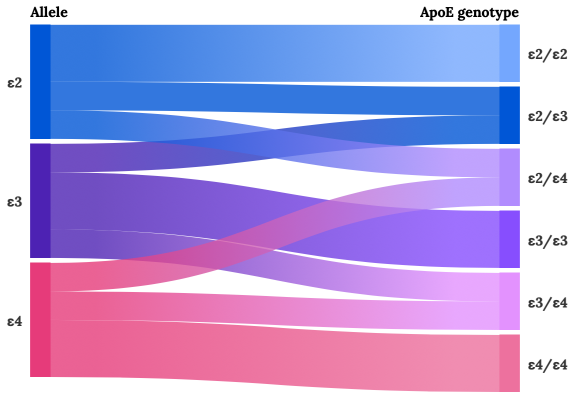


FIGURE 1. Sankey chart showing the allele distribution among the 6 ApoE genotypes.

1.2.2. *ApoE and ancestry in Alzheimer's ~~D~~isease.* The majority of studies exploring the relationship between ApoE alleles and the genetics of LOAD have primarily focused on Northern European populations[14]. However, smaller studies involving diverse ancestral backgrounds show variations in ApoE4 allele [14]. While ApoE4 is present in 14% of Caucasian Americans, its prevalence increases to 40% among African Americans, 37% in Oceania, and 26% in Australia. Southern Asia and Europe exhibit ApoE4 allelic frequencies of less than 10%, compared to Northern Europe where it rises to 25% [18, 20–22].

The epidemiological impact of ApoE alleles also differs among populations. In Korea, Japan, and Japanese-American communities, ApoE4 confers a higher risk of LOAD compared to Caucasians [19]. Conversely, for Native Americans, Hispanic Americans, African Americans, and those of African descent, ApoE4 is associated with a lower risk of LOAD than in Caucasian-American populations [19, 23–26]. A recent study in a Chinese population found that ApoE3 is more protective than ApoE2 [27]. Some of these population-specific effects are attributed to the ApoE haplotype [23, 25]. A recent discovery of a novel locus (19q13.31) could contribute to attenuating ApoE4-mediated AD risk in African Americans [28].

1.2.3. *ApoE and sex synergy in Alzheimer's ~~D~~isease.* Notably, sex (60% females) and ApoE4 allelic composition (50% has at least one ε_4 allele) are the strongest genetic risk factors for SAD [29]. In this regard, it is shown that the ApoE4 genotype has a larger impact on females, as they present greater impairment of mitochondrial energy production, compared to males [29, 30].

1.2.4. *Evolution of ApoE over time.* Interestingly, humans are the only species exhibiting polymorphism in the ApoE gene [30]. All other animal species have one ApoE variant, which resembles the human ApoE3 allele [31]. ApoE4 is the oldest human allele, followed by ApoE3 and ApoE2 in age [30]. ApoE4 may be adaptive, reducing mortality in highly infectious environments, with food scarcity and shorter lifespans [32]. However, as human environments became less septic, with food abundance and longer life expectancy, ApoE4 started to burden the arteries and brain, increasing the risk of diseases related to ageing [30]. The emergence of ApoE3 from ApoE4 putatively reflects the shift in human diet from a plant-based one to a meat-rich one, where genes adaptive to high meat consumption were and still are vital to regulate increased cholesterol levels [33].

1.2.5. *Tissue expression.* The principal producers of ApoE are hepatocytes in the liver [34]. In the CNS, ApoE is primarily expressed in glia, ~~astrocytes~~, cells of which modulate metabolic homeostasis and neuronal communication, followed by microglia, the brain immune cells [35]. Each genotype is linked to different expression levels [13]. ApoE2 carriers seem to have higher cerebrospinal fluid (CSF) levels of ApoE, compared to ApoE4 carriers [36, 37].

1.3. **Apo-lipoprotein E.**

1.3.1. *Structure and function.* ApoE is a brain-specific lipid-binding glycoprotein of 299 amino-acids (34 kDa) that comprises several types of lipoproteins, i.e., chilomicra, intermediate density lipoprotein and very low density lipoprotein [13]. Its main function in the brain is the transport of lipids (mainly cholesterol) through membrane receptors [14]. Moreover, its isoforms have an effect on diverse cellular functions, e.g. synaptic integrity, glucose metabolism, Aβ clearance, blood-brain barrier integrity and mitochondrial regulation [13]. How these relate to AD pathology will be elaborated in Section 1.3.4.

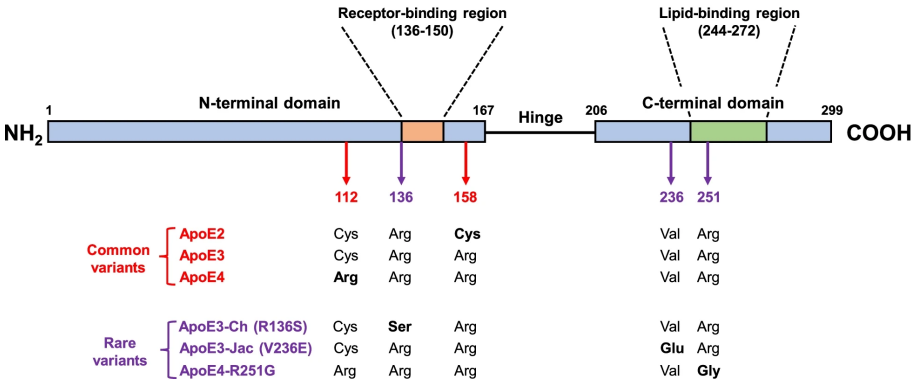


FIGURE 2. Linear representation of the ApoE protein. Three structural domains are highlighted: N-terminal, hinge and C-terminal domains. The different aminoacids at positions 112 and 158 are shown per common alleles and aminoacids at positions 136, 236 and 251 coded by rarer alleles. Source: "APOE targeting strategy in Alzheimer's disease: lessons learned from protective variants", Bu (2022) [38]

1.3.2. *ApoE isoforms.* The nuances in the amino acid composition of ApoE, specifically the presence of cysteine or arginine at positions 112 and 158, significantly impact its binding with lipids and receptors [30]. The most prevalent isoform, ApoE3, features cysteine at position 112 and arginine at position 158 [30], as shown in Fig. 2. ApoE2 has two cysteines, while ApoE4 has two arginines at these positions [30]. The C-terminal domain of ApoE (positions 273–299) is crucial for its lipidation specificity and efficiency [39].

1.3.3. *Lipidation nuances of ApoE isoforms.* For ApoE to exert its effects, it needs to be lipidated [13]. ApoE undergoes lipidation via ATP-binding cassette transporter A1 (ABCA1), a lipid efflux protein [40, 41]. The lipidation degree varies among ApoE isoforms, with ApoE4 exhibiting the least efficient lipidation [39, 42]. This discrepancy in lipidation has been linked to alterations in lipoprotein size and type, in that ApoE4 "prefers" large triglyceride-rich VLDL, while ApoE2 and ApoE3 have a higher affinity for phospholipid-rich HDL particles [43]. The lower affinity of ApoE4 for HDL particles leads to increased levels of unlipidated ApoE, resulting in its aggregation [44]. Moreover, ApoE4 fibrils are more neurotoxic than those of ApoE2 and ApoE3 [45].

Poor lipidation leads to poor ApoE recycling [30]. The latter favors the entrapment of ABCA1 in endosomes, away from the cell membrane, thereby pooling cholesterol in the cell membrane rather than attaching it to HDL particles [46]. This increased cholesterol content in the cell membrane amplifies toll-like receptor 4 (TLR4) signaling in macrophages, activating NFkB and inducing an inflammatory gene response [30]. ApoE4 accumulation also sequestrates insulin receptor (IR) in endosomes, impacting cellular energy preferences [47]. This leads to a decrease in glucose utilization for ATP production and an increase in fatty acid oxidation [48].

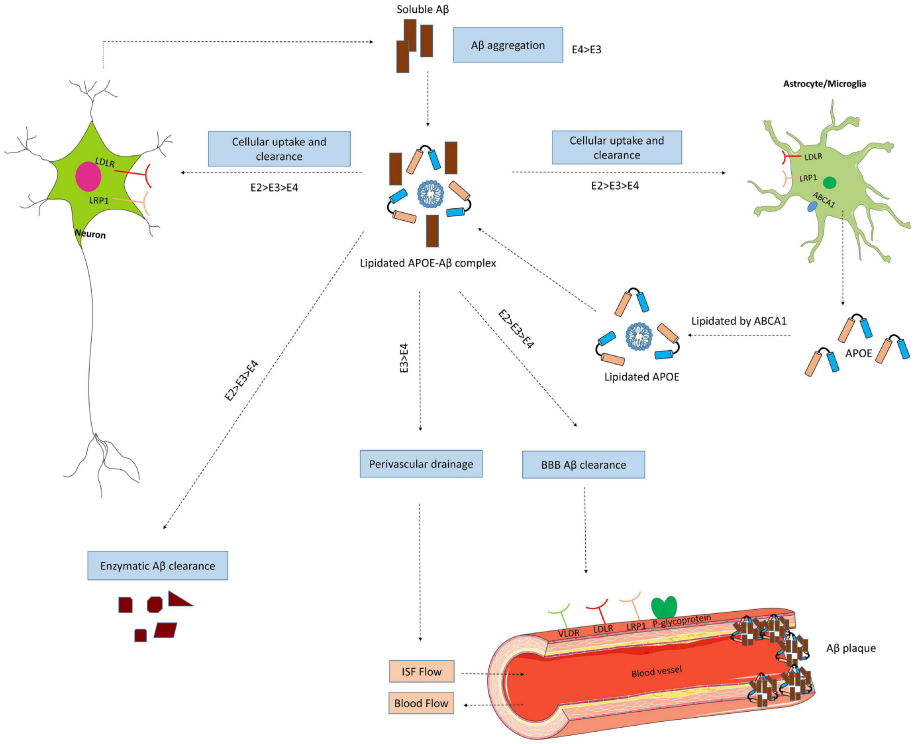


FIGURE 3. Effects of ApoE isoforms on the metabolism and removal of Aβ. Aβ is primarily cleaved from amyloid precursor protein (APP). In the brain, ApoE, mainly expressed in astrocytes and microglia, undergoes lipidation by ATP-binding cassette transporter A1 (ABCA1) to create lipoprotein particles. ApoE increases the accumulation and buildup of Aβ, or promotes cellular uptake of Aβ by astrocytes or microglia via endocytosis of the lipidated ApoE-Aβ in an isoform-specific manner. This process involves several receptors, such as low-density lipoprotein receptor (LDLR) and low density lipoprotein receptor-related protein 1 (LRP1). ApoE also facilitates isoform-specific breakdown of Aβ outside the cells. At the blood-brain barrier, soluble Aβ is predominantly transported from the interstitial fluid (ISF) into the bloodstream via LRP1 and P-glycoproteins. Additionally, ApoE plays a role in the peri-vascular drainage of Aβ. Insufficient clearance of Aβ can lead to its accumulation in the brain tissue, contributing to the formation of Aβ oligomers and amyloid plaques. Source: "APOE and Alzheimer's Disease: From Lipid Transport to Physiopathology and Therapeutics", Husain *et al.* (2021) [13]

1.3.4. *Interplay between ApoE lipidation and Alzheimer's ~~D~~isease.* As mentioned earlier, ApoE isoforms have differential pleotropic effects on various cellular functions. ApoE4 induces pro-inflammatory response, leading to the dysfunction of the blood-brain barrier, which in turn impairs cognitive functions [49–51]. Moreover, ApoE modulates the primary neuropathological hallmarks of AD: neuroinflammation, Aβ plaques and tau tangles [13]. Evidence from human and transgenic mice studies reveals increased brain Aβ and amyloid plaque loads in ApoE4 carriers, compared to ApoE3; with the lowest levels in ApoE2 carriers [52–54]. Higher plaque loads are associated with ApoE4 due to its higher affinity for Aβ and poor clearance capacity [51]. Additionally, high levels of ApoE4 in neurons remarkably increase tau protein phosphorylation, while high concentrations of ApoE3 ~~don't~~ seem to have an effect [55–58]. Notably, ApoE

directly inhibits phosphorylation of tau by GSK-3β [59]. An overview of the Aβ-independent effects of ApoE is shown in Fig. 3.

1.4. **ApoE4-mediated metabolic changes in Alzheimer's ~~D~~isease.** Metabolism entails the repertoire of chemical reactions that keep living organisms alive. Metabolites –especially lipid [60–62]– , perceived as functional intermediates of AD development, are rigorously studied for bio-markers or targets for treatment [63].

1.4.1. *Measured in post-mortem brain tissue.* A metabolomic profiling of brain tissue, obtained *post-mortem* from AD patients and healthy controls showed pronounced impairments in sterol and sphingolipid levels in ApoE4 carriers with AD [64]. However, another *post-mortem* metabolomic analysis ~~didn't~~ reveal nuances significantly correlated with ApoE4 [65], although they showed trends in increased cholesterol esters, unsaturated lipids, and sphingomyelin species.

1.4.2. *Measured in blood.* Transcriptomic and lipidomic analyses in humanized ApoE mice associated ApoE4 with decreased free fatty acid levels, many increased tricarboxylic acid (TCA) cycle metabolites, as well as changes in plasma levels of phosphatidylcholines and unsaturated fatty acids [66, 67]. A recent

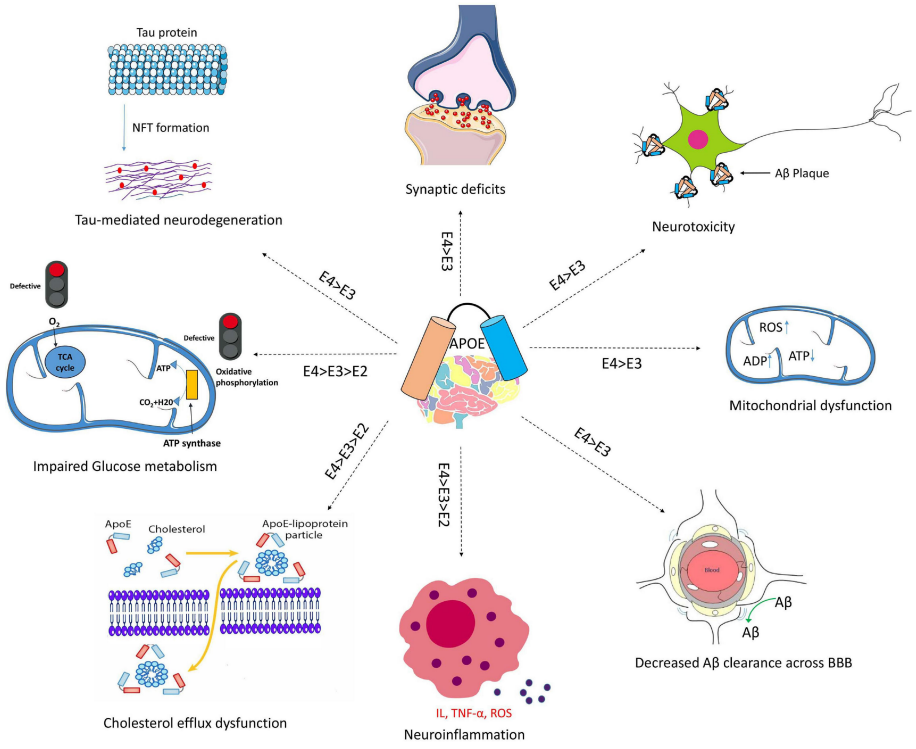


FIGURE 4. Schematic overview of Aβ-independent effects of ApoE in AD pathology. ApoE4 increases the phoshorylation of tau proteins, leading to the creation of tangles, inducing neurodegeneration, synaptic deficits and neurotoxicity. Moreover, ApoE4 is associated with decreased cholesterol efflux and neuroinflammation mediated by interleukin (IL), tumor necrosis factor α (TNF-α), reactive oxygen species (ROS). In the mitochondria, ApoE4 impairs glucose metabolism and ATP production. Source: "APOE and Alzheimer's Disease: From Lipid Transport to Physiopathology and Therapeutics", Husain *et al.* (2021) [13]

study on 58 individuals found six downregulated plasma metabolites (including lysophospholipids and cardiolipin) in ApoE4 carriers [68]. Further, the plasma metabolome of the latter reveals a preference for aerobic glycolysis [69]. Significant correlations of ApoE genotype and sex with metabolites were observed, i.e. several phosphatidylcholines were found in a large study of more than 1500 individuals [29].

Perturbed serum metabolites associated with AD are aminoacids, amines [70, 71], cholesteryl esters [62], sphingolipids [60, 63, 71–73], fatty acids [61, 70], glycerophospholipids [72, 74–76], phosphatidylcholines [77] and lipid peroxidation compounds [61]. These molecules are usually identified via high-throughput metabolomic pipelines (coupled with Mass Spectrometry (MS) detectors) that trace all compounds in a sample and result in high-dimensional data [78]. The latter often require advanced statistical methods e.g. projection to latent structures [76, 79] or graphical models [80] in order to extract putatively meaningful information. With such techniques, de Leeuw et al. discovered distinct serum metabolic signatures among AD patients-controls and those carrying at least one ApoE ε_4 allele [70], as they appear in Fig. 2. The different metabolic profiles, however, among ApoE4 non-carriers remain obscure. The present data science approach shows ApoE4-mediated differentially expressed metabolites, potentially unveiling distinct pathways of metabolic deregulation in AD.

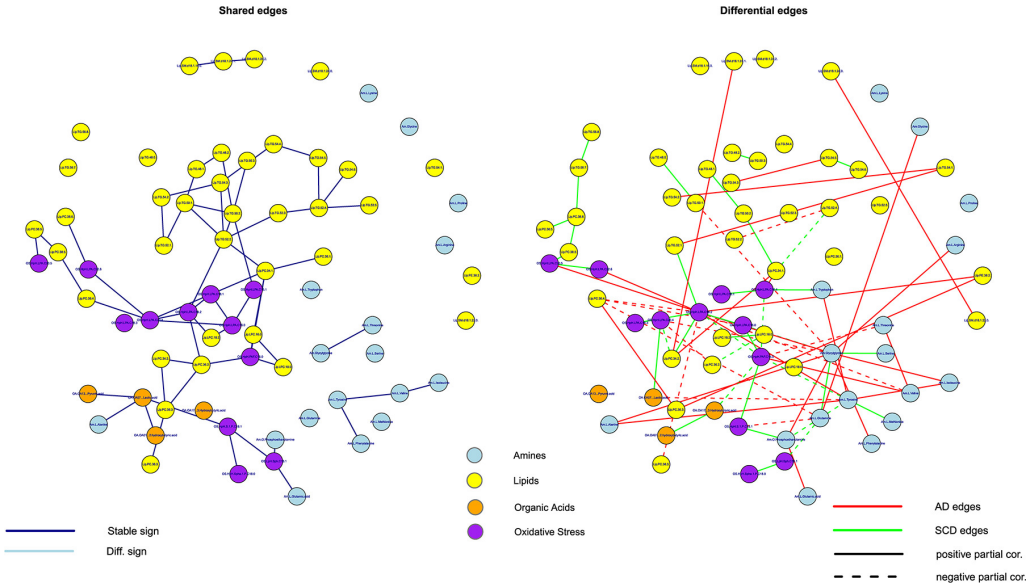


FIGURE 5. Mutual (left-hand panel) and distinct (right-hand panel) metabolic network topologies between ApoE4 carriers and non-carriers in AD, as published by de Leeuw et al.. Red edges represent links that are present exclusively in ApoE4 carriers with AD. Green edges represent connections found in the SCD group. Solid edges represent positive partial correlations, while dashed edges represent negative partial correlations. Abbreviations: SCD, subjective cognitive decline. Source: "Blood-based metabolic signatures in Alzheimer's disease", De Leeuw *et al.* (2017) [70]

1.5. **Research Questions.** ApoE4 carriers –particularly females– experience metabolic disturbances and are at increased risk of SAD. The mechanistic links, however, between ApoE4 dose, metabolism and AD development are not entirely known [61]. Tracking ApoE4 dose effects on serum metabolites can reveal metabolic perturbations preceding or co-existing with AD. Hence, in an effort to elucidate ApoE4-mediated changes in serum metabolites in AD and SCD, one could state the following research questions (RQ):

Are there mechanistic links between ApoE4 dose and serum metabolome in AD?

(1) Are there ApoE4 dose effects on serum metabolite levels in AD?
(2) How discriminatory are serum metabolites of ApoE4 status and/or AD?
(3) How do the covariance network topologies of metabolites differ between ApoE4 carriers and non-carriers?

1.6. **Approach and Overview.** An introduction to the data used in this study is found in Section 2.1. A general overview of the software is found in Section 2.2 and 2.5, while the R `sessionInfo` is found in Appendix B. To facilitate statistical analysis, two new features were created for the first two research questions: ApoE4dose (0, 1, 2) and ApoE4AD (4 possible phenotype: AD without ApoE4, AD with at least 1 ApoE4, SCD without ApoE4, SCD with at least 1 ApoE4), respectively, as described in Section 2.3.

The statistical methods applied to answer the research questions were adapted from De Leeuw *et al.*'s "Blood-based metabolic signatures in Alzheimer's disease" and are described in Section 2.4. To screen for ApoE4 dose effects on serum metabolites in AD, two approaches were taken: a global test (correcting only for sex) and nested linear model comparison using ANCOVA F-tests (correcting for several factors: Table 1, except CSF markers).

To test the (added) classification potential of serum metabolites against ApoE4AD, several multi-class classification models were fitted. First, a benchmark multinomial logistic regression model was fitted using only clinical background data as predictors. Second, the full metabolomic panel was addded on top of the clinical data in the same model. Third, the metabolites were projected to a latent orthogonal space, where 6 meta-metabolites (accounting for around 30% of the variance) replaced the original metabolites. Finally, the meta-metabolites were fitted on top of the clinical data in a multinomial logistic regression model, a decision tree and an eXtreme Gradient Boosting (XGBoost) model. The discrimatory performance of the aforementioned models was then holistically evaluated and compared.

To visualise and compare the covariance network topologies of metabolites among ApoE4 carriers and non-carriers, the high precision matrices were first sparsified using Ridge and then prunned controlling the False Discrovery Rate. Network topologies were plotted and their statistics were calculated and compared between the ApoE4 carriers and non-carriers using Wilcoxon Signed Rank test.

The results of the analysis are presented in Section 3 and discussed in Section 4.

The study is concluded in Section 5

2. Methods

2.1. **Subjects.** The data were collected from 126 AD patients and 121 SCD (n = 247) individuals ~~in the context of the~~ Amsterdam Dementia Cohort [70, 81]. In this study two data sets were used: a targeted metabolomics panel and clinical background data, i.e. age at diagnosis, sex, smoking status, alcohol consumption, hypertension (and medication), hyperlipidemia (and medication), anticoagulant medication, anti-depressants, mean arterial pressure (MAP) and body mass index (BMI) (see Table 1). The metabolomics set contains p = 230 metabolites (amines, organic acids, lipids and oxidative stress compounds). The methodology for the metabolomic analysis and ApoE genotyping can be found at de Leeuw et al.'s Blood-based metabolic signatures in Alzheimer's Disease [70]: SMT1 . The data was cleaned as described in the same article. The resulting data-sets for AD and SCD are high-dimensional, in the sense that they contain more variables than observations ($p > n$). Another particularity of the data is the covariance and collinearity of the variables. Therefore, appropriate measures need to be taken to prevent model over-fitting –the algorithm being unable to distinguish signal from noise and fitting the latter– and to correct for spurious correlations.

Table 1. Clinical background characteristics used as control variables of ApoE4 (dose) effects. The molecules measured in CSF were used only in the Multi-class classification.

	Clinical feature	**Type**
Anthropometric	Age	Discrete
	Sex	Binary
Intoxications	Smoking (past, current, no)	Nominal
	Alcohol	Binary
Commorbidities	Hypertension	Binary
	Diabetes mellitus	Binary
	Hypercholesterolemia	Binary
Medication	Cholesterol lowering	Numeric
	Antidepressants	Binary
	Antiplatelet	Binary
CSF	$A\beta_{42}$	Numeric
	tau	Numeric
	p-tau	Numeric

2.2. **Data management.** The FAIR principles for data management and stewardship in science were published by Wilkinson *et al.* in 2016 [82]. FAIR stands for Findable, Accessible, Interactive, and Reusable data; the intention is to create and use data that are well-documented and reproducible. These principles were considered at every step of the study and implemented when applicable. The statistical analysis was performed in R (version 4.3.2), and the current report was written in LaTeX(Tex Live version 2023). All files are stored in a private Github repository –with git as Version Control System (VCS).

2.3. **Feature Engineering.** The ApoE allele and genotype frequencies in the population are reflected in the AD and SCD data, as shown in Table 2. That is, the most abundant allele is $\varepsilon 3$, followed by $\varepsilon 4$ and $\varepsilon 2$. The most common genotype was $\varepsilon 3\varepsilon 3$, followed by $\varepsilon 3\varepsilon 4$ and $\varepsilon 4\varepsilon 4$.

The ApoE genotypes is valuable information and might be interesting to screen for metabolic nuances between them. However, the genotypes are not equally represented in the data and hence, testing for differences among them would not be reliable. The following two sections describe the features created to focus on ApoE4 status (0 or at least 1 allele) and dose (0, 1, or 2 alleles) to balance the data when studying only the AD group and both AD and SCD.

TABLE 2. ApoE genotype counts (top part) ApoE4 counts: features created to balance the genotype counts (bottom part)

		AD	SCD	Total
ApoE genotypes	$\varepsilon2\varepsilon2$	2	0	2
	$\varepsilon2\varepsilon3$	15	3	18
	$\varepsilon2\varepsilon4$	5	2	7
	$\varepsilon3\varepsilon3$	69	37	106
	$\varepsilon3\varepsilon4$	26	59	85
	$\varepsilon4\varepsilon4$	3	26	29
ApoE4 alleles	1x	31	61	92
	2x	3	26	29
	≥ 1x	34	87	121
	No	86	40	126

2.3.1. *AD group.* [70] divided the subjects into two classes: those carrying at least one $\varepsilon4$ allele, and $\varepsilon4$ non-carriers. However, in order to study the $\varepsilon4$ *dose* effects, the genotypes can be ~~binned~~ into groups, based on the number of $\varepsilon4$ alleles they carry: 0, 1 or 2. The number of ApoE4 allele doses are shown in the bottom part of Table 2 (first two and last row).

2.3.2. *SCD and AD.* In order to incorporate ApoE4 status, as well as the diagnosis of AD, a four-level feature (AD without ApoE4, AD with at least 1 ApoE4, SCD without ApoE4, SCD with at least 1 ApoE4, "ApoE4AD") was created, as shown in the last two rows of Table 2.

2.4. **Statistical Analysis.**

2.4.1. *ApoE4 dose effects on serum metabolite levels in AD.* To test if the number of ApoE4 alleles have an effect on mean metabolite levels in AD, two methods were applied: a global test and nested linear model comparison with ANCOVA.

Global Test The concept of a global test was first introduced by Simon *et al.*, and proposed an approach based on permutations to cater to the high dimensionality of gene expression data. The R package globaltest [84–86] developed by Goeman *et al.* features a multinomial logistic regression model, fitting genes to predict clinical or biological group membership (number of ApoE4 alleles in this case). This method is also appropriate for other types of -omics data, such as metabolomics in this study [86]. The null hypothesis is that metabolite levels are independent of the ApoE4 dose X, i.e. $H_0 : P(Y|X) = P(Y)$, where $X \in \mathbb{R}^{nxp}$. The test statistic under H_0 follows, asymptotically, a normal distribution. The `gt` function of the package was used to screen for nuances in metabolite levels on the number of ApoE4 alleles, correcting for sex. To asses ApoE4 dose effects correcting for clinical data nested linear model comparison was performed, as described in the next section.

Nested linear models To test for ApoE4 dose effects on each metabolite, $p = 230$ model comparisons were needed, hence a function was created to iterate over the metabolites, fit the nested models, compare them using ANCOVA F-tests, store and adjust the p-values, filter those below .05, then consolidate the coefficients of the meaningful full models and the p-values of their t-tests in a table and display it. To decrease run time, as well as harness the power of multiple cores, furrr's `future_map` function was used for parallel iterations.

The dependent variable in each model was a metabolite; the nested model (1), has only clinical variables (Table 1 except CSF markers) while the full model (2), features the clinical variables, plus the number of

ApoE4 alleles (0, 1 or 2 -nominal) as explanatory variables. Let y_j represent the j-th metabolite, x_k the k-th clinical variable, and $D_{\varepsilon 4}$ the ApoE4 dose; the nested models then are:

$$y_j = \beta_0 + \sum_{k=1}^{m} \beta_k x_k + \epsilon \quad (1)$$

$$y_j = \beta_0 + \sum_{k=1}^{m} \beta_k x_k + \beta_{m+1} D_{\varepsilon 4} + \epsilon \quad (2)$$

For every metabolite j = 1,...,p the hypothesis test is H_0: $\beta_{m+1} = 0$ vs H_α: $\beta_{m+1} \neq 0$. Under H_0 the test statistic, F follows an $F_{1,n-(m+2)}$ distribution. This implies p hypothesis tests, which qualifies as multiple testing. A method to treat the latter is controlling False Discovery Rate [87], that is controlling the expected ratio of incorrectly rejected H_0 hypotheses, globally e.g. at an α of 0.05. First, the p-values are sorted in ascending order, and for every j p-value each α is multiplied by j divided by the total number of p-values [87] m = 230 in this study. In other words, after adjustment, a null hypothesis j may be rejected only if its associated F-test's p-value is less than a fraction (j/m) of α.

Algorithm 1: Benjamini–Hochberg's procedure to control FDR

1 Specify α, the level at which to control the FDR.
2 Compute p-values, $p_1, ..., p_m$, for the m null hypotheses $H_{01}, ..., H_{0m}$.
3 Order the m p-values so that $p(1) \leq p(2) \leq ... \leq p(m)$.
4 Define

$$L = \max\{j : p(j) \leq \frac{j}{m}\alpha\}$$

5 Reject all null hypotheses H_{0j} for which $p_j \leq p(L)$.

2.4.2. *Classification of ApoE4 status and/or AD.* In machine learning, a class denotes a group of objects that share common characteristics, such as having AD or ApoE4 [88]. Classification, in this context, denotes training a (supervised) learning algorithm on labeled data (containing their class) [88]. The classifier learns patterns in the data and is then able to predict class membership for unknown data [88].

Bias-Variance trade-off The degree to which a user can understand and interpret the prediction or decisions made by a statistical model is defined as *interpretability* [89]. It is of interest in this study to find the optimal balance between the performance of a model with its interpretability. The *bias-variance trade-off* was formally introduced by Geman *et al.* and refers to the trade-off between the accuracy (opposite of bias) and precision (opposite of variance) of a prediction. It also refers to the trade-off between model flexibility (or complexity) and interpretability [90]. One may consider this trade-off during model and evaluation method selection, as some impose more bias or variance than others.

Multi-class classification models The R package caret streamlines the training and comparison of classification and regression models, offering broad parametrization options [91]. The functions `trainControl` and `train` were used to fit the models. A sampling method used to deal with the unbalanced classes was SMOTE (Synthetic Minority Over-Sampling Technique), as implemented by the package DMwR2 [92].

Considering interpretability, Multinomial Logistic Regression (MNL) is inherently interpretable. Let $y = k$, with $k \in N, [1, 4]$ representing the k-th class of ApoE4AD and β_{kj} its set of coefficients, β_{lj} the coefficients of the rest of classes for j-th metabolite, then a MNL model calculated the probability

$$\Pr(y = k|X = x) = \frac{e^{\sum_{j=1}^{p} \beta_{kj} x_j}}{\sum_{l=1}^{4} \sum_{j=1}^{p} e^{\beta_{lj} x_j}}$$

When $p > n$, the coefficient estimation method has low bias and high variance, in that small changes in the training data can result in very different coefficient estimates [93]. Regularization trades off a small increase in bias for a great decrease in variance, by shrinking the unimportant coefficients towards zero. LASSO (Least Absolute Shrinkage and Selection Operator) [94], also called L1-regularization shrinks the MNL coefficients to 0, thus weeding out spurious correlations and reducing the number of predictors [94]. It does so by introducing the term

$$\lambda \sum_{j=1}^{p} |\beta_j|$$

where $\lambda \geq 0$ is a tuning parameter that balances the coefficient shrinking effect. The package nnet, as implemented in caret's `train` function was used in this case.

A method to treat multi-collinearity and high dimensionality is a 2-stage Maximum Likelihood (ML) Factor Analysis (FA), such as the one the package FMradio [79] performs. In the 1st stage, a L2-regularised ML estimation is used to filter out redundant features from the data matrix. In the second stage, ML FA projects the aforementioned matrix to an orthogonal space where the features are replaced by -fewer- factors that explain their covariance. One can then use the produced factor scores as predictors in MNL.

Decision Trees (DT) are inherently interpretable, non-parametric models, that fit well large and complicated data sets. They have a tree-like structure that splits the data into branches and leaves(nodes) [95]. The `rpart2` [96] function of rpart was used.

Boosting models, are ensemble models that fit several weak learners (such as linear/logistic regression or DTs) sequentially, reweighing the data, and take their weighted majority vote [97, 98]. Despite Boosting tends to outperform DTs, it often operates as a *black box* and is poorly interpretable. The state-of-the-art eXtreme Gradient Boosting (`xgbTree`) of the package xgboost [99] was used.

Algorithm 2: Multi-class classification of ApoE4 and AD status pipeline

1. Fit clinical data only in MNL
2. Fit clinical data + 230 serum metabolites in MNL
3. Fit clinical data + 6 meta-metabolites in MNL
4. Fit clinical data + 6 meta-metabolites in a Decision Tree
5. Fit clinical data + 6 meta-metabolites in an XGBoost
6. Evaluate performance and compare

First, a benchmark model was created fitting the clinical background data to predict ApoE4AD in a penalised MNL model. Second, the 230-metabolite panel was fitted on top of the clinical background data in a penalised MNL model. Then the full metabolite panel was projected into 6 latent factors (meta-metabolites), explaining around 30% of their variance. The 6-factor metabolite projection was then fitted on top of the clinical data in a penalised MNL, a DT and an XGBoost model. The aforementioned models were hyper-parameter tuned over a grid of values and the best was selected using repeated (100 times) 10-fold Cross-Validation (CV).

Model performance was holistically assessed and compared, with the repeated 10-fold CV-obtained ROC curves and their respective Area Under the Curve AUC using the pROC package [100] and other metrics such as Accuracy, Precision, Recall and F1-score from caret's `confusionMatrix`.

2.4.3. *Metabolite Covariance Network Analysis.* Network science offers a unifying framework for data and system representation, applicable to any domain [101]. A network, in an abstract sense, consists of nodes connected with links, also referred to as edges. In data science, a network whose nodes represent random features, whose joint probability distribution is defined by the ensemble of their edges is called *graphical model* [80]. A metabolomic covariance correlation network represents the ensemble of metabolites based on their covariance, showing nuances among the samples that non-graphical statistical methods on individual

metabolites may fail to detect [102]. It may provide insights into correlated metabolites that don't belong in the same metabolic pathway [102].

A *Gaussian graphical model* (GGM) is an undirected graph that represents the conditional independence properties of the features [103]. The statistic employed by GGMs is the partial correlation which also adjusts for indirect correlation, i.e. two metabolites are correlated with a third one and are shown correlated with each other [104]. For instance, let $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be a GGM consisting of a set $\mathcal{V}$ of p vertices, corresponding to random features $Y_1, ..., Y_p$ with joint probability distribution $P \sim N_p(\mathbf{0}, \mathbf{\Sigma})$, and set of edges $\mathcal{E}$, such that for all pairs $\{Y_i, Y_j\}$ with $i \neq j$:

$$\mathbf{\Sigma}^{-1}_{ij} = (\mathbf{\Omega}_{ij}) = 0 \iff Y_i \perp\!\!\!\perp Y_j \mid \{Y_k : k \neq i, j\} \iff (i, j) \notin \mathcal{E}$$

In natural language, a zero value in the inverse covariance matrix (usually referred to as precision matrix $\mathbf{\Omega}$) implies that the respective random features are independent, given the rest of features, and they are not connected by an undirected edge ($(i, j) \notin \mathcal{E}$) [80].

In this study, the package ragsridges [80] was used to generate the feature covariance matrix, as well as the precision matrix, regularise it and represent it in a GGM –as shown in Fig. 2.

2.5. **R packages.**

caret [91]
dplyr [105]
DMwR2 [92]
globaltest [84–86]
heatmaply [106]
FMradio [79]
nnet [107]
pROC [100]
rags2ridges [80]
rpart [96]
xgboost [99]
furrr [108]

3. Results

3.1. **ApoE4 dose effects on serum metabolite levels in AD.**

3.1.1. *Global Test.* Testing for ApoE4 dose-effect on serum metabolites of AD patients, correcting for sex (Ho: ApoE4 dose has no effect on mean metabolite levels, Ha: it has an effect), showed a significant global difference in metabolites (p = 0.017). The most significantly affected metabolites are triglycerides and diglycerides (FDR-adjusted p-value <0.05) See Table 3 and Fig. 6.

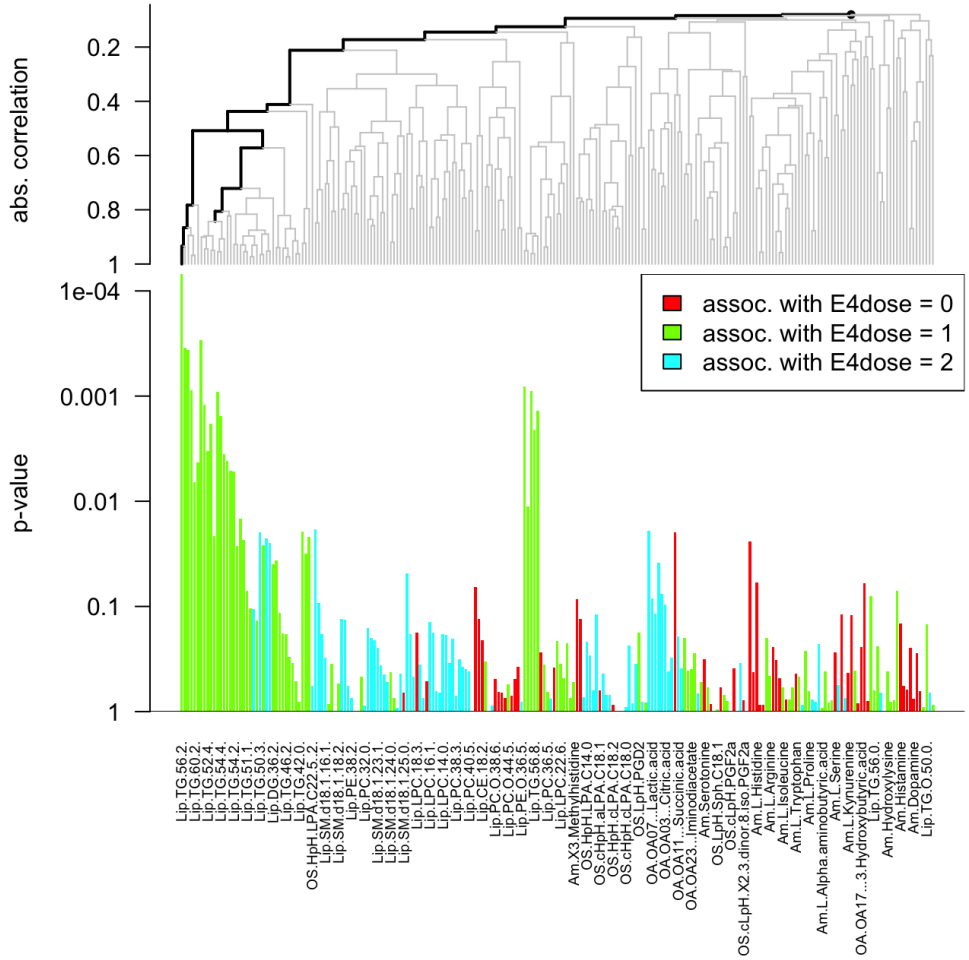


FIGURE 6. Covariates plot showing the metabolites affected by the number of ApoE4 alleles

TABLE 3. Metabolites affected by ApoE4 dose, correcting for sex as per globaltest [86]

	Inheritance	Assoc. with	p-value	FDR
Lip.TG.56.2.	0.046	1 ApoE4	<0.001	0.016
Lip.TG.58.1.	0.117	1 ApoE4	<0.001	0.021
Lip.DG.36.3.	0.19	1 ApoE4	<0.001	0.021
Lip.TG.56.3.	0.244	1 ApoE4	<0.001	0.021
Lip.TG.52.3.	0.424	1 ApoE4	0.001	0.031
Lip.TG.54.5.	0.48	1 ApoE4	0.001	0.027
Lip.TG.58.2.	0.722	1 ApoE4	0.001	0.027
Lip.TG.54.4.	0.761	1 ApoE4	0.002	0.033
Lip.TG.58.9.	1	1 ApoE4	0.001	0.027
Lip.TG.56.7.	1	1 ApoE4	0.001	0.027
Lip.TG.58.8.	1	1 ApoE4	0.001	0.032
Lip.TG.54.6.	1	1 ApoE4	0.002	0.035
Lip.TG.56.8.	1	1 ApoE4	0.002	0.037
Lip.TG.52.4.	1	1 ApoE4	0.003	0.055
Lip.TG.54.3.	1	1 ApoE4	0.004	0.055
Lip.TG.56.6.	1	1 ApoE4	0.004	0.059
Lip.TG.56.1.	1	1 ApoE4	0.004	0.059
Lip.TG.52.2.	1	1 ApoE4	0.005	0.063
Lip.TG.54.2.	1	1 ApoE4	0.005	0.063
Lip.TG.60.2.	1	1 ApoE4	0.007	0.077
Lip.TG.58.10.	1	1 ApoE4	0.011	0.124
Lip.TG.51.3.	1	1 ApoE4	0.015	0.154
Lip.SM.d18.1.18.1.	1	2 ApoE4	0.018	0.169
OA.2.ketoglutaric.acid	1	2 ApoE4	0.019	0.169
Lip.TG.52.0.	1	1 ApoE4	0.019	0.169
Lip.TG.50.3.	1	2 ApoE4	0.02	0.169
OA.Uracil	1	no ApoE4	0.02	0.169
Lip.TG.52.5.	1	1 ApoE4	0.021	0.174
Lip.TG.54.0.	1	1 ApoE4	0.022	0.174
Lip.TG.50.2.	1	2 ApoE4	0.022	0.174
Lip.TG.51.2.	1	1 ApoE4	0.023	0.174
Am.L.Glutamine	1	no ApoE4	0.024	0.174
Lip.TG.50.1.	1	2 ApoE4	0.025	0.174
Lip.TG.50.4.	1	1 ApoE4	0.026	0.176
Lip.TG.52.1.	1	1 ApoE4	0.027	0.176
Lip.TG.54.1.	1	1 ApoE4	0.031	0.203
Lip.TG.50.0.	1	1 ApoE4	0.037	0.229
OA.Malic.acid	1	2 ApoE4	0.038	0.234
Lip.DG.36.2.	1	1 ApoE4	0.039	0.235
OS.HpH.PAF.C16.0	1	2 ApoE4	0.049	0.283

3.1.2. *Nested Linear Models.* Several metabolites from all classes were affected by ApoE4 dose, both in AD and SCD, however, after correcting for multiple testing by controlling FDR, none of the effects ~~were~~ significant at $\alpha = 0.05$. Among AD patients, ApoE4 dose seems to have positive effect on several triglycerides, diglycerides, putrescine, 2-ketoglutraric acid, lysophosphatidylcholin, HpH.PAF.-C16.0 and -C18.0 (Table 4). Among individuals with SCD, lipid metabolites were not ~~affected~~ as much as in the AD group, with only two sphingomyelin species showing a difference. Aminoacids L-serine, tryptophan, glycine,trytptophan, L-homoserine, putrescine were affected in this group. L-Tryptophan is negatively associated with ApoE4 dose (at 1x and 2x ApoE4), while L-serine, glycine and L-homoserine are negativelly associated only with ApoE4 homozygotes. (Table 4).

TABLE 4. ApoE4 dose effects on metabolites in AD and SCD: significant F tests (α = 0.05) from nested linear model comparison. Reduced model: (1), Full model: (2). Am: Aminoacid, Lip: Lipid, DG: Diglyceride, LPC: Lysophosphatidylcholine, SM: Sphingomyelin, TG: Triglyceride, OA: Organic Acid, OS: Oxidative Stress compound

	Metabolite	P(>F)	FDR	No ε4		1x ε4		2x ε4	
				Coef.	P(>t)	Coef.	P(>t)	Coef.	P(>t)
AD	Lip.TG.52.3.	0.001	0.156	-1.4	0.818	2.6	0.004	3.7	0.001
	Lip.TG.52.4.	0.003	0.156	0.6	0.888	1.6	0.008	2.4	0.002
	Lip.DG.36.3.	0.002	0.156	0.0	0.631	0.0	0.012	0.0	0.001
	OS.HpH.PAF.C16.0	0.002	0.156	34.8	<0.001	2.6	0.017	4.6	0.001
	Lip.TG.52.2.	0.006	0.255	-5.1	0.469	2.1	0.03	3.7	0.002
	Lip.TG.54.5.	0.01	0.373	1.9	0.496	0.9	0.016	1.3	0.006
	Lip.TG.50.2.	0.012	0.398	-4.5	0.297	0.9	0.141	2.2	0.003
	Lip.TG.50.1.	0.018	0.45	-2.3	0.539	0.7	0.179	1.8	0.005
	Lip.TG.54.4.	0.02	0.45	3.5	0.318	1.2	0.015	1.4	0.02
	Lip.TG.54.6.	0.017	0.45	-0.4	0.795	0.5	0.027	0.7	0.009
	Am.Putrescine	0.029	0.515	0.0	<0.001	0.0	0.035	0.0	0.017
	OA.2.ketoglutaric.acid	0.026	0.515	0.0	0.366	0.0	0.953	0.0	0.014
	Lip.TG.50.3.	0.028	0.515	-1.2	0.668	0.6	0.127	1.3	0.008
	Lip.TG.52.5.	0.042	0.538	0.5	0.789	0.4	0.134	0.7	0.014
	Lip.TG.56.7.	0.042	0.538	-2.4	0.095	0.5	0.015	0.4	0.105
	Lip.TG.56.8.	0.044	0.538	-1.4	0.055	0.2	0.014	0.2	0.151
	Lip.TG.58.9.	0.042	0.538	-0.4	0.068	0.1	0.013	0.0	0.41
	Lip.LPC.16.0.	0.033	0.538	4.2	0.028	0.3	0.237	0.9	0.009
	OS.HpH.LPA.C18.0	0.044	0.538	0.5	0.038	0.0	0.178	0.1	0.013
SCD	Am.L.Serine	0.002	0.285	4.7	<0.001	0.2	0.115	-1.1	0.003
	Am.L.Tryptophan	0.002	0.285	3.7	0.004	-0.3	0.047	-1.4	0.002
	Am.Glycine	0.006	0.45	4.3	0.001	0.4	0.015	-0.9	0.052
	Am.L.homoserine	0.047	0.931	0.0	0.001	0.0	0.723	0.0	0.014
	Am.Putrescine	0.045	0.931	0.0	0.969	0.0	0.013	0.0	0.927
	Lip.SM.d18.1.22.0.	0.043	0.931	3.3	0.002	0.3	0.015	0.3	0.448
	Lip.SM.d18.1.23.0.	0.029	0.931	1.2	0.01	0.1	0.012	0.2	0.277

3.2. **Classification of ApoE4 status and/or AD.** All classification models had a multi-class ROC AUC aboee 80%. The worst performing model was Multinomial Logistic Regression fitting the clinical data only (Table 1), while the best performing one was XGBoost fitting the clinical data and the 6 meta-metabolites. Adding serum metabolite information (either the full 230-metabolite matrix or its 6-factor projection) seems

to increase the discriminatory power of the models. Notably, looking at individual ROC curves, all models were able to discriminate better among certain classes (AD+E4/SCD+E4, AD+E4/SCD, AD-E4/SCD+E4 and AD-E4/SCD-E4) compared to others (AD+E4/AD-E4 and SCD+E4/SCD-E4).

TABLE 5. Performance metrics of multi-class classification of ApoE4 and AD status, obtained from 10-fold CV repeated 100 times. NIR: No Information Rate, AUC: Area Under the (ROC) Curve

	NIR	Accuracy (95/% CI)	P(NIR > Accuracy)	AUC
Clinical features only		0.4807 (0.4744, 0.4869)	$< 2.2e^{-16}$	0.814
Clinical features + 230 metabolites		0.5751 (0.5689, 0.5813)	$< 2.2e^{-16}$	0.834
Clinical features + 6 latent factors	0.3522	0.5313 (0.525, 0.5375)	$< 2.2e^{-16}$	0.818
Decision Tree		0.5082 (0.502, 0.5145)	$< 2.2e^{-16}$	0.818
XGBoost		0.5944 (0.5882, 0.6005)	$< 2.2e^{-16}$	0.836

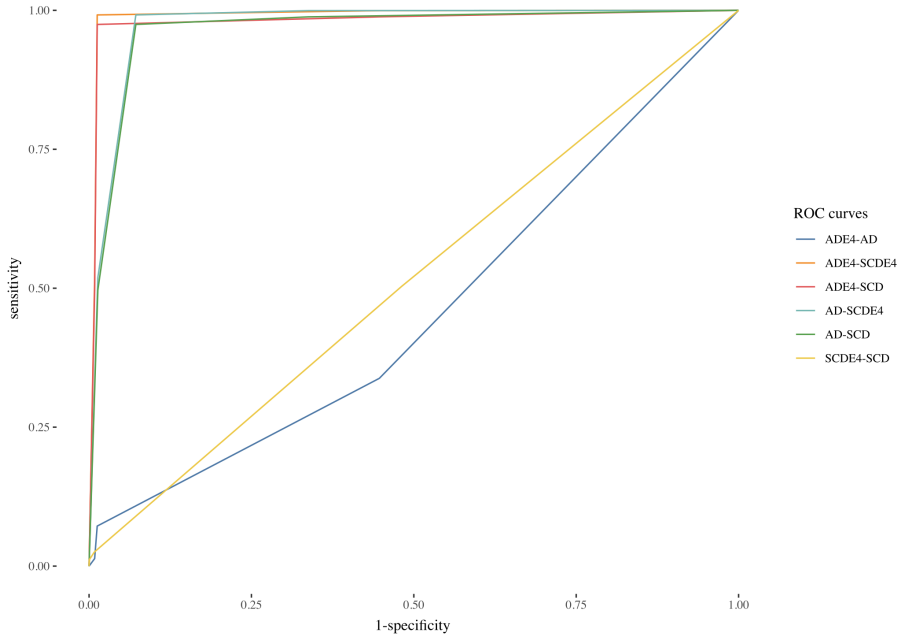


FIGURE 7. ROC curves of benchmark model: Multinomial Logistic Regression fitting the clinical background features only), obtained from repeated (100 times) 10-fold CV.

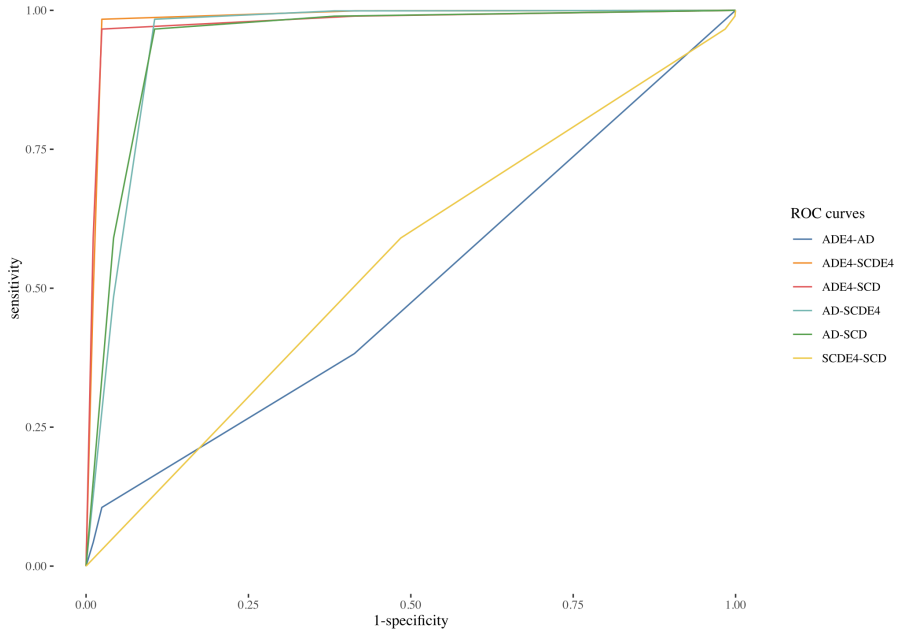


FIGURE 8. ROC curves of Multinomial Logistic Regression fitting the 230 metabolites on top of the clinical background variables, obtained from repeated (100 times) 10-fold CV

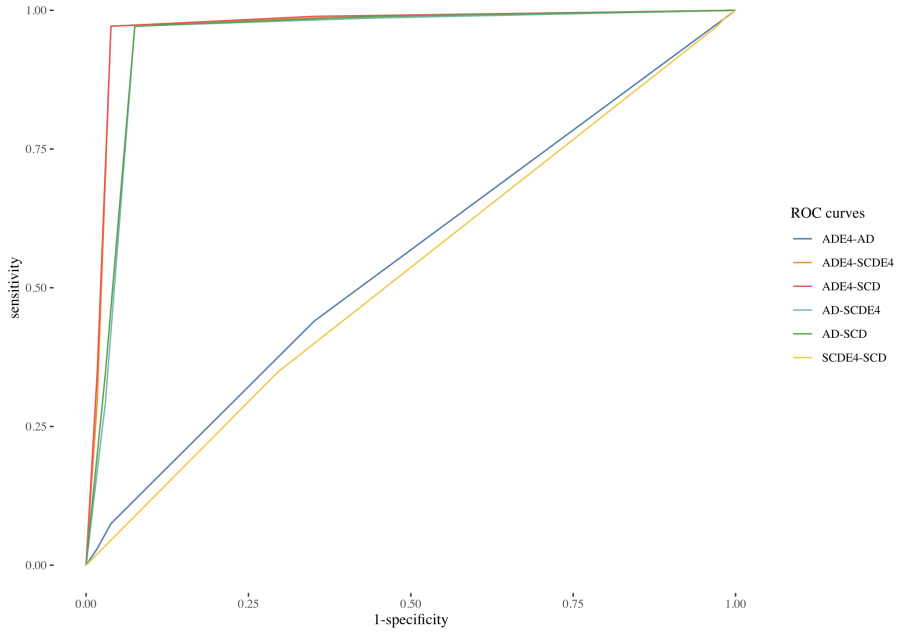


FIGURE 9. ROC curves of Multinomial Logistic Regression fitting the 6 ML-estimated latent factors on top of the clinical background variables, obtained from repeated (100 times) 10-fold CV.

3.3. **Metabolite Covariance Network Analysis.** The metabolite covariance network analysis among ApoE4 carriers and non-carriers showed distinct correlations between the metabolites. The ApoE4 carriers' network was less sparse, having less edges comparet to the non-carriers'. A Wilcoxon Signed Rank Test showed distinct degrees of centrality between ApoE4 carriers and non-carriers (p = 0.001).

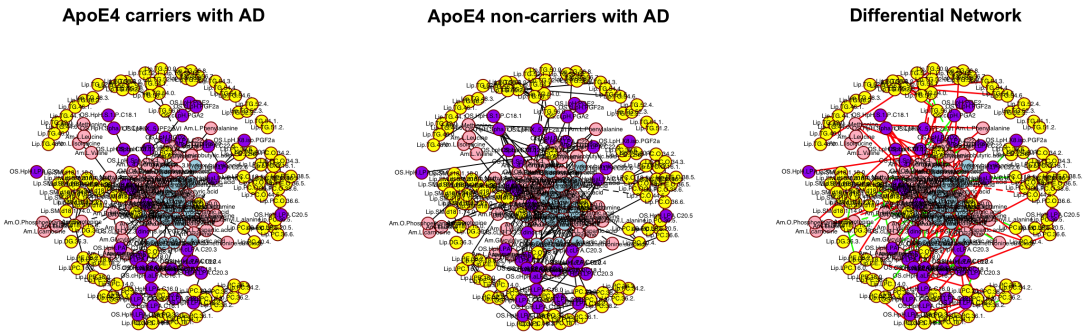


FIGURE 10. Prunned metabolite covariance network topologies among ApoE4 carriers (a) and non-carriers in AD (b); differential edge network (c). The ApoE4-positive metabolite covariance network is more sparse compared to ApoE4-negative. Red edges represent links that are present exclusively in ApoE4 carriers. Green edges represent connections found in ApoE4 non-carriers.

4. Discussion

5. Conclusion

References

1. Penke, B., Szűcs, M. & Bogár, F. New Pathways Identify Novel Drug Targets for the Prevention and Treatment of Alzheimer's Disease. *International Journal of Molecular Sciences* **24,** 5383. ISSN: 1422-0067. https://www.mdpi.com/1422-0067/24/6/5383 (Mar. 2023).
2. 2023 Alzheimer's disease facts and figures. *Alzheimer's & Dementia* **19,** 1598–1695. ISSN: 1552-5279. https://onlinelibrary.wiley.com/doi/full/10.1002/alz.13016%20https://onlinelibrary.wiley.com/doi/abs/10.1002/alz.13016%20https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.13016 (Apr. 2023).
3. Heneka, M. T. *et al.* Neuroinflammation in Alzheimer's disease. *The Lancet. Neurology* **14,** 388–405. ISSN: 1474-4465. https://pubmed.ncbi.nlm.nih.gov/25792098/ (Apr. 2015).
4. Edwards, F. A. A Unifying Hypothesis for Alzheimer's Disease: From Plaques to Neurodegeneration. *Trends in Neurosciences* **42,** 310–322. ISSN: 1878108X. http://www.cell.com/article/S016622361930027X/fulltext%20http://www.cell.com/article/S016622361930027X/abstract%20https://www.cell.com/trends/neurosciences/abstract/S0166-2236(19)30027-X (May 2019).
5. Aaldijk, E. & Vermeiren, Y. The role of serotonin within the microbiota-gut-brain axis in the development of Alzheimer's disease: A narrative review. *Ageing Research Reviews* **75,** 101556. ISSN: 1568-1637. https://www.sciencedirect.com/science/article/pii/S1568163721003032 (2022).
6. Scheltens, P. *et al.* Alzheimer's disease. *Lancet (London, England)* **388,** 505–517. ISSN: 1474-547X. https://pubmed.ncbi.nlm.nih.gov/26921134/ (July 2016).
7. Beydoun, M. A. *et al.* Epidemiologic studies of modifiable factors associated with cognition and dementia: Systematic review and meta-analysis. *BMC Public Health* **14,** 1–33. ISSN: 14712458. https://bmcpublichealth.biomedcentral.com/articles/10.1186/1471-2458-14-643 (June 2014).
8. Van Cauwenberghe, C., Van Broeckhoven, C. & Sleegers, K. The genetic landscape of Alzheimer disease: clinical implications and perspectives. *Genetics in Medicine 2016 18:5* **18,** 421–430. ISSN: 1530-0366. https://www.nature.com/articles/gim2015117 (Aug. 2015).
9. Hardy, J. Alzheimer's disease: The amyloid cascade hypothesis: An update and reappraisal. *Journal of Alzheimer's Disease* **9,** 151–153. ISSN: 1387-2877 (Jan. 2006).
10. Kepp, K. P., Robakis, N. K., Høilund-Carlsen, P. F., Sensi, S. L. & Vissel, B. The amyloid cascade hypothesis: an updated critical review. https://doi.org/10.1093/brain/awad159 (2023).
11. Kurkinen, M. *et al.* The Amyloid Cascade Hypothesis in Alzheimer's Disease: Should We Change Our Thinking? *Biomolecules 2023, Vol. 13, Page 453* **13,** 453. ISSN: 2218-273X. https://www.mdpi.com/2218-273X/13/3/453/htm%20https://www.mdpi.com/2218-273X/13/3/453 (Mar. 2023).
12. Corder, E. H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261,** 921–923. ISSN: 00368075 (Aug. 1993).
13. Husain, M. A., Laurent, B. & Plourde, M. APOE and Alzheimer's Disease: From Lipid Transport to Physiopathology and Therapeutics. *Frontiers in Neuroscience* **15,** 630502. ISSN: 1662453X (Feb. 2021).
14. Yang, L. G., March, Z. M., Stephenson, R. A. & Narayan, P. S. Apolipoprotein E in lipid metabolism and neurodegenerative disease. *Trends in Endocrinology & Metabolism* **34,** 430–445. ISSN: 1043-2760. http://www.cell.com/article/S1043276023000929/fulltext%20http://www.cell.com/article/S1043276023000929/abstract%20https://www.cell.com/trends/endocrinology-metabolism/abstract/S1043-2760(23)00092-9 (Aug. 2023).
15. Liu, C. & al. et, e. Apolipoprotein E and Alzheimer disease: risk, mechanisms, and therapy. *Nat. Rev. Neurosci.* **9,** 106–118 (2013).
16. Strittmatter, W. & al. et, e. Apolipoprotein E: high-avidity binding to β-amyloid and increased frequency of type 4 allele in late-onset familial Alzheimer disease. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 1977–1981 (1993).
17. Deming, Y. & al. et, e. Genome-wide association study identifies four novel loci associated with Alzheimer's endophenotypes and disease modifiers. *Acta Neuropathol.* **133,** 839–856 (2017).
18. Belloy, M. & al. et, e. A quarter century of APOE and Alzheimer's disease: progress to date and the path forward. *Neuron* **101,** 820–838 (2019).
19. Farrer, L. & al. et, e. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. *JAMA* **278,** 1349–1356 (1997).
20. Egert, S., Rimbach, G. & Huebbe, P. ApoE genotype: from geographic distribution to function and responsiveness to dietary factors. *Proceedings of the Nutrition Society* **71,** 410–424. ISSN: 1475-2719. https://www.cambridge.org/core/journals/proceedings-of-the-nutrition-society/article/apoe-genotype-from-geographic-distribution-to-function-and-responsiveness-to-dietary-factors/D9F35B3FFE46F48CAFE131E0E94CC62C (Aug. 2012).
21. Eisenberg, D. T., Kuzawa, C. W. & Hayes, M. G. Worldwide allele frequencies of the human apolipoprotein E gene: Climate, local adaptations, and evolutionary history. *American Journal of Physical Anthropology* **143,** 100–111. ISSN: 1096-8644. https://onlinelibrary.wiley.com/doi/full/10.1002/ajpa.21298%20https://onlinelibrary.wiley.com/doi/abs/10.1002/ajpa.21298%20https://onlinelibrary.wiley.com/doi/10.1002/ajpa.21298 (Sept. 2010).

22. Logue, M. W. *et al.* A Comprehensive Genetic Association Study of Alzheimer Disease in African Americans. *Archives of Neurology* **68,** 1569–1579. ISSN: 0003-9942. https://jamanetwork.com/journals/jamaneurology/fullarticle/1108047 (Dec. 2011).
23. Blue, E. E., Horimoto, A. R., Mukherjee, S., Wijsman, E. M. & Thornton, T. A. Local ancestry at APOE modifies Alzheimer's disease risk in Caribbean Hispanics. *Alzheimer's & Dementia* **15,** 1524–1532. ISSN: 1552-5279. https://onlinelibrary.wiley.com/doi/full/10.1016/j.jalz.2019.07.016%20https://onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2019.07.016%20https://alz-journals.onlinelibrary.wiley.com/doi/10.1016/j.jalz.2019.07.016 (Dec. 2019).
24. Suchy-Dicey, A., Howard, B., Longstreth, W. T., Reiman, E. M. & Buchwald, D. APOE genotype, hippocampus, and cognitive markers of Alzheimer's disease in American Indians: Data from the Strong Heart Study. *Alzheimer's & Dementia* **18,** 2518–2526. ISSN: 1552-5279. https://onlinelibrary.wiley.com/doi/full/10.1002/alz.12573%20https://onlinelibrary.wiley.com/doi/abs/10.1002/alz.12573%20https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.12573 (Dec. 2022).
25. Rajabli, F. *et al.* Ancestral origin of ApoE ϵ4 Alzheimer disease risk in Puerto Rican and African American populations. *PLOS Genetics* **14,** e1007791. ISSN: 1553-7404. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007791 (Dec. 2018).
26. Naslavsky, M. S. *et al.* Global and local ancestry modulate APOE association with Alzheimer's neuropathology and cognitive outcomes in an admixed sample. *Molecular Psychiatry 2022 27:11* **27,** 4800–4808. ISSN: 1476-5578. https://www.nature.com/articles/s41380-022-01729-x (Sept. 2022).
27. Chen, H.-K. & al. et, e. Apolipoprotein E4 domain interaction mediates detrimental effects on mitochondria and is a potential therapeutic target for Alzheimer disease. *J. Biol. Chem.* **286,** 5215–5221 (2011).
28. Rajabli, F. *et al.* A locus at 19q13.31 significantly reduces the ApoE ϵ4 risk for Alzheimer's Disease in African Ancestry. *PLOS Genetics* **18,** e1009977. ISSN: 1553-7404. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009977 (July 2022).
29. Arnold, M. *et al.* Sex and APOE ϵ4 genotype modify the Alzheimer's disease serum metabolome. *Nature Communications 2020 11:1* **11,** 1–12. ISSN: 2041-1723. https://www.nature.com/articles/s41467-020-14959-w (Mar. 2020).
30. Yassine, H. N. & Finch, C. E. APOE Alleles and Diet in Brain Aging and Alzheimer's Disease. *Frontiers in Aging Neuroscience* **12,** 544681. ISSN: 16634365 (June 2020).
31. Hunsberger, H. C., Pinky, P. D., Smith, W., Suppiramaniam, V. & Reed, M. N. The role of APOE4 in Alzheimer's disease: strategies for future therapeutic interventions. *Neuronal Signaling* **3,** 1–15. ISSN: 20596553. /neuronalsignal/article/3/2/NS20180203/110989/The-role-of-APOE4-in-Alzheimer-s-disease%20https://dx.doi.org/10.1042/NS20180203 (June 2019).
32. Trumble, B. C. *et al.* Apolipoprotein E4 is associated with improved cognitive function in Amazonian forager-horticulturalists with a high parasite burden. *FASEB Journal* **31,** 1508–1515. ISSN: 15306860 (Apr. 2017).
33. Finch, C. E. & Sapolsky, R. M. The evolution of Alzheimer disease, the reproductive schedule, and apoE isoforms. *Neurobiology of Aging* **20,** 407–428. ISSN: 01974580 (1999).
34. Mahley, R. Central nervous system lipoproteins: ApoE and regulation of cholesterol metabolism. *Arterioscler. Thromb. Vasc. Biol.* **36,** 1305–1315 (2016).
35. Lanfranco, M. & al. et, e. Expression and secretion of apoE isoforms in astrocytes and microglia during inflammation. *Glia* **69,** 1478–1493 (2021).
36. Castellano, J. M. *et al.* Human apoE isoforms differentially regulate brain amyloid-β peptide clearance. *Science Translational Medicine* **3.** ISSN: 19466234 (June 2011).
37. Cruchaga, C. *et al.* Cerebrospinal fluid APOE levels: An endophenotype for genetic studies for Alzheimer's disease. *Human Molecular Genetics* **21,** 4558–4571. ISSN: 09646906 (Oct. 2012).
38. Bu, G. APOE targeting strategy in Alzheimer's disease: lessons learned from protective variants. *Molecular Neurodegeneration* **17,** 1–4. ISSN: 17501326. https://molecularneurodegeneration.biomedcentral.com/articles/10.1186/s13024-022-00556-6 (Dec. 2022).
39. Hu, J. *et al.* Opposing effects of viral mediated brain expression of apolipoprotein E2 (apoE2) and apoE4 on apoE lipidation and Aβ metabolism in apoE4-targeted replacement mice. *Molecular Neurodegeneration* **10,** 1–11. ISSN: 17501326. https://molecularneurodegeneration.biomedcentral.com/articles/10.1186/s13024-015-0001-3 (Mar. 2015).
40. Flowers, S. A. & Rebeck, G. W. APOE in the normal brain. *Neurobiology of Disease* **136,** 104724. ISSN: 0969-9961 (Mar. 2020).
41. Courtney, R. & Landreth, G. E. LXR Regulation of Brain Cholesterol: From Development to Disease. *Trends in Endocrinology and Metabolism* **27,** 404–414. ISSN: 18793061. http://www.cell.com/article/S1043276016300042/fulltext%20http://www.cell.com/article/S1043276016300042/abstract%20https://www.cell.com/trends/endocrinology-metabolism/abstract/S1043-2760(16)30004-2 (June 2016).
42. Heinsinger, N. M., Gachechiladze, M. A. & Rebeck, G. W. Apolipoprotein E Genotype Affects Size of ApoE Complexes in Cerebrospinal Fluid. *Journal of Neuropathology & Experimental Neurology* **75,** 918–924. ISSN: 0022-3069. https://dx.doi.org/10.1093/jnen/nlw067 (Oct. 2016).

43. Nguyen, D. *et al.* Molecular basis for the differences in lipid and lipoprotein binding properties of human apolipoproteins E3 and E4. *Biochemistry* **49,** 10881–10889. ISSN: 00062960. https://pubs.acs.org/doi/abs/10.1021/bi1017655 (Dec. 2010).
44. Hatters, D. M., Peters-Libeu, C. A. & Weisgraber, K. H. Apolipoprotein E structure: insights into function. *Trends in Biochemical Sciences* **31,** 445–454. ISSN: 09680004 (Aug. 2006).
45. Hatters, D. M., Zhong, N., Rutenber, E. & Weisgraber, K. H. Amino-terminal Domain Stability Mediates Apolipoprotein E Aggregation into Neurotoxic Fibrils. *Journal of Molecular Biology* **361,** 932–944. ISSN: 00222836 (Sept. 2006).
46. Rawat, V. *et al.* ApoE4 Alters ABCA1 Membrane Trafficking in Astrocytes. *Journal of Neuroscience* **39,** 9611–9622. ISSN: 0270-6474. https://www.jneurosci.org/content/39/48/9611%20https://www.jneurosci.org/content/39/48/9611.abstract (Nov. 2019).
47. Zhao, N. *et al.* Apolipoprotein E4 Impairs Neuronal Insulin Signaling by Trapping Insulin Receptor in the Endosomes. *Neuron* **96,** 115–129. ISSN: 1097-4199. https://pubmed.ncbi.nlm.nih.gov/28957663/ (Sept. 2017).
48. Svennerholm, L., Boström, K. & Jungbjer, B. Changes in weight and compositions of major membrane components of human brain during the span of adult human life of Swedes. *Acta Neuropathologica* **94,** 345–352. ISSN: 00016322. https://link.springer.com/article/10.1007/s004010050717 (Oct. 1997).
49. Marottoli, F. M. *et al.* Peripheral inflammation, apolipoprotein E4, and amyloid-β interact to induce cognitive and cerebrovascular dysfunction. *ASN Neuro* **9.** ISSN: 17590914. https://journals.sagepub.com/doi/10.1177/1759091417719201 (Aug. 2017).
50. Teng, Z. *et al.* ApoE Influences the Blood-Brain Barrier Through the NF-κB/MMP-9 Pathway After Traumatic Brain Injury. *Scientific Reports 2017 7:1* **7,** 1–8. ISSN: 2045-2322. https://www.nature.com/articles/s41598-017-06932-3 (July 2017).
51. Kloske, C. M. & Wilcock, D. M. The Important Interface Between Apolipoprotein E and Neuroinflammation in Alzheimer's Disease. *Frontiers in Immunology* **11.** ISSN: 16643224 (Apr. 2020).
52. Huang, Y. W. A., Zhou, B., Wernig, M. & Südhof, T. C. ApoE2, ApoE3, and ApoE4 Differentially Stimulate APP Transcription and Aβ Secretion. *Cell* **168,** 427–441. ISSN: 10974172 (Jan. 2017).
53. Tachibana, M. *et al.* Rescuing effects of RXR agonist bexarotene on aging-related synapse loss depend on neuronal LRP1. *Experimental Neurology* **277,** 1–9. ISSN: 0014-4886 (Mar. 2016).
54. Safieh, M., Korczyn, A. D. & Michaelson, D. M. ApoE4: an emerging therapeutic target for Alzheimer's disease. *BMC Medicine 2019 17:1* **17,** 1–17. ISSN: 1741-7015. https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1299-4 (Mar. 2019).
55. Cao, J. *et al.* ApoE4-associated phospholipid dysregulation contributes to development of Tau hyper-phosphorylation after traumatic brain injury. *Scientific Reports* **7.** ISSN: 20452322 (Dec. 2017).
56. Shi, Y. *et al.* ApoE4 markedly exacerbates tau-mediated neurodegeneration in a mouse model of tauopathy. *Nature 2017 549:7673* **549,** 523–527. ISSN: 1476-4687. https://www.nature.com/articles/nature24016 (Sept. 2017).
57. Vasilevskaya, A. *et al.* Interaction of APOE4 alleles and PET tau imaging in former contact sport athletes. *NeuroImage: Clinical* **26,** 102212. ISSN: 2213-1582 (Jan. 2020).
58. Wang, C. & al. et, e. Gain of toxic apolipoprotein E4 effects in human iPSC-derived neurons is ameliorated by a small-molecule structure corrector. *Nat. Med.* **24,** 647–657 (2018).
59. Hoe, H. S., Freeman, J. & Rebeck, G. W. Apolipoprotein e decreases tau kinases and phospho-tau levels in primary neurons. *Molecular Neurodegeneration* **1.** ISSN: 17501326 (2006).
60. Barupal, D. K. *et al.* Sets of coregulated serum lipids are associated with Alzheimer's disease pathophysiology. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **11,** 619–627. ISSN: 2352-8729 (Dec. 2019).
61. Fernández-Calle, R. *et al.* APOE in the bullseye of neurodegenerative diseases: impact of the APOE genotype in Alzheimer's disease pathology and brain diseases. *Molecular Neurodegeneration 2022 17:1* **17,** 1–47. ISSN: 1750-1326. https://link.springer.com/articles/10.1186/s13024-022-00566-4%20https://link.springer.com/article/10.1186/s13024-022-00566-4 (Sept. 2022).
62. Proitsi, P. *et al.* Association of blood lipids with Alzheimer's disease: A comprehensive lipidomics analysis. *Alzheimer's & Dementia* **13,** 140–151. ISSN: 1552-5260 (Feb. 2017).
63. Oeckl, P. *et al.* Glial Fibrillary Acidic Protein in Serum is Increased in Alzheimer's Disease and Correlates with Cognitive Impairment. *Journal of Alzheimer's disease : JAD* **67,** 481–488. ISSN: 1875-8908. https://pubmed.ncbi.nlm.nih.gov/30594925/ (2019).
64. Bandaru, V. V. R. *et al.* ApoE4 disrupts sterol and sphingolipid metabolism in Alzheimer's but not normal brain. *Neurobiology of Aging* **30,** 591–599. ISSN: 0197-4580 (Apr. 2009).
65. Novotny, B. C. *et al.* Metabolomic and lipidomic signatures in autosomal dominant and late-onset Alzheimer's disease brains. *Alzheimer's & Dementia* **19,** 1785–1799. ISSN: 1552-5279. https://onlinelibrary.wiley.com/doi/full/10.1002/alz.12800%20https://onlinelibrary.wiley.com/doi/abs/10.1002/alz.12800%20https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/alz.12800 (May 2023).

66. Area-Gomez, E. *et al.* APOE4 is Associated with Differential Regional Vulnerability to Bioenergetic Deficits in Aged APOE Mice. *Scientific Reports 2020 10:1* **10,** 1–20. ISSN: 2045-2322. https://www.nature.com/articles/s41598-020-61142-8 (Mar. 2020).
67. Zhao, N. *et al.* Alzheimer's Risk Factors Age, APOE Genotype, and Sex Drive Distinct Molecular Pathways. *Neuron* **106,** 727–742. ISSN: 10974199. http://www.cell.com/article/S0896627320301859/fulltext%20http://www.cell.com/article/S0896627320301859/abstract%20https://www.cell.com/neuron/abstract/S0896-6273(20)30185-9 (June 2020).
68. Peña-bautista carmen, c. *et al.* Metabolomics study to identify plasma biomarkers in alzheimer disease: ApoE genotype effect. *Journal of Pharmaceutical and Biomedical Analysis* **180,** 113088. ISSN: 0731-7085 (Feb. 2020).
69. Farmer, B. & al. et, e. APOE4 lowers energy expenditure in females and impairs glucose oxidation by increasing flux through aerobic glycolysis. *Mol. Neurodegener.* **16,** 62 (2021).
70. De Leeuw, F. A. *et al.* Blood-based metabolic signatures in Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **8,** 196–207. ISSN: 2352-8729 (Jan. 2017).
71. Green, R. E. *et al.* Investigating associations between blood metabolites, later life brain imaging measures, and genetic risk for Alzheimer's disease. *Alzheimer's Research and Therapy* **15,** 1–13. ISSN: 17589193. https://alzres.biomedcentral.com/articles/10.1186/s13195-023-01184-y%20http://creativecommons.org/publicdomain/zero/1.0/ (Dec. 2023).
72. Varma, V. R. *et al.* Brain and blood metabolite signatures of pathology and progression in Alzheimer disease: A targeted metabolomics study. *PLOS Medicine* **15,** e1002482. ISSN: 1549-1676. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1002482 (Jan. 2018).
73. Sun, L. *et al.* Association between Human Blood Metabolome and the Risk of Alzheimer's Disease. *Annals of Neurology* **92,** 756–767. ISSN: 1531-8249. https://onlinelibrary.wiley.com/doi/full/10.1002/ana.26464%20https://onlinelibrary.wiley.com/doi/abs/10.1002/ana.26464%20https://onlinelibrary.wiley.com/doi/10.1002/ana.26464 (Nov. 2022).
74. Jia, L. *et al.* A metabolite panel that differentiates Alzheimer's disease from other dementia types. *Alzheimer's & Dementia* **18,** 1345–1356. ISSN: 1552-5279. https://onlinelibrary.wiley.com/doi/full/10.1002/alz.12484 (July 2022).
75. Huo, Z. *et al.* Brain and blood metabolome for Alzheimer's dementia: findings from a targeted metabolomics analysis. *Neurobiology of Aging* **86,** 123–133. ISSN: 0197-4580 (Feb. 2020).
76. Weng, W. C., Huang, W. Y., Tang, H. Y., Cheng, M. L. & Chen, K. H. The Differences of Serum Metabolites Between Patients With Early-Stage Alzheimer's Disease and Mild Cognitive Impairment. *Frontiers in Neurology* **10,** 482026. ISSN: 16642295 (Nov. 2019).
77. Simpson, B. N. *et al.* Blood metabolite markers of cognitive performance and brain function in aging. *Journal of Cerebral Blood Flow and Metabolism* **36,** 1212–1223. ISSN: 15597016. https://journals.sagepub.com/doi/full/10.1177/0271678X15611678 (July 2016).
78. Oka, T., Matsuzawa, Y., Tsuneyoshi, M. & Nakamura, Y. Multiomics analysis to explore blood metabolite biomarkers in an Alzheimer's Disease Neuroimaging Initiative cohort. https://www.researchsquare.com%20https://www.researchsquare.com/article/rs-2973576/v1 (June 2023).
79. Peeters, C. F. W. *et al.* Stable prediction with radiomics data. https://arxiv.org/abs/1903.11696v1 (Mar. 2019).
80. Peeters, C. F., Bilgrau, A. E. & van Wieringen, W. N. rags2ridges: A One-Stop-ℓ2-Shop for Graphical Modeling of High-Dimensional Precision Matrices. *Journal of Statistical Software* **102,** 1–32. ISSN: 1548-7660. https://www.jstatsoft.org/index.php/jss/article/view/v102i04 (May 2022).
81. Van Der Flier, W. M. & Scheltens, P. Amsterdam Dementia Cohort: Performing Research to Optimize Care. *Journal of Alzheimer's disease : JAD* **62,** 1091–1111. ISSN: 1875-8908. https://pubmed.ncbi.nlm.nih.gov/29562540/ (2018).
82. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data 2016 3:1* **3,** 1–9. ISSN: 2052-4463. https://www.nature.com/articles/sdata201618 (Mar. 2016).
83. Simon, R. M. *et al.* Design and Analysis of DNA Microarray Investigations Series: Statistics for Biology and Health, 199 (2004).
84. Goeman, J. J., Van de Geer, S., De Kort, F. & van Houwellingen, H. C. A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics (Oxford, England)* **20,** 93–99. ISSN: 1367-4803. https://pubmed.ncbi.nlm.nih.gov/14693814/ (Jan. 2004).
85. Goeman, J. J., Van De Geer, S. A. & Van Houwelingen, H. C. Testing against a high dimensional alternative. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68,** 477–493. ISSN: 1467-9868. https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9868.2006.00551.x (June 2006).
86. Goeman, J., Oosting, J., Finos, L., Solari, A. & Edelmann, D. The Global Test and the globaltest R package (2023).
87. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57,** 289–300. ISSN: 00359246 (1995).
88. Drummond, C. in *Encyclopedia of Machine Learning* (eds Sammut, C. & Webb, G. I.) 171–171 (Springer US, Boston, MA, 2010). ISBN: 978-0-387-30164-8. https://doi.org/10.1007/978-0-387-30164-8_111.

89. Elshawi, R., Al-Mallah, M. H. & Sakr, S. On the interpretability of machine learning-based model for predicting hypertension. *BMC Medical Informatics and Decision Making* **19,** 1–32. ISSN: 14726947. https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-019-0874-0 (July 2019).
90. Geman, S., Bienenstock, E. & Doursat, R. Neural Networks and the Bias/Variance Dilemma. *Neural Computation* **4,** 1–58. ISSN: 0899-7667. https://dx.doi.org/10.1162/neco.1992.4.1.1 (Jan. 1992).
91. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28,** 1–26. ISSN: 1548-7660. https://www.jstatsoft.org/index.php/jss/article/view/v028i05 (Nov. 2008).
92. Torgo, L. *Data Mining with R, learning with case studies, 2nd edition* http://ltorgo.github.io/DMwR2 (Chapman and Hall/CRC, 2016).
93. James, G., Witten, D., Hastie, T. & Tibshirani, R. An Introduction to Statistical Learning with Applications in R Second Edition (2023).
94. Tibshiranit, R. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* **58,** 267–288. ISSN: 2517-6161. https://onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1996.tb02080.x%20https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.x%20https://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1996.tb02080.x (Jan. 1996).
95. Song, Y. Y. & Lu, Y. Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry* **27,** 130. ISSN: 10020829. /pmc/articles/PMC4466856/%20/pmc/articles/PMC4466856/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4466856/ (Apr. 2015).
96. Therneau, T. M., Atkinson, B., Ripley, B., *et al.* rpart: Recursive partitioning. *R package version* **3** (2010).
97. Friedman, J., Hastie, T. & Tibshirani, R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The Annals of Statistics* **28,** 337–407. https://doi.org/10.1214/aos/1016218223 (2000).
98. Friedman, J. H. Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics* **29,** 1189–1232. ISSN: 00905364. http://www.jstor.org/stable/2699986 (2024) (2001).
99. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13-17-August-2016,** 785–794. https://arxiv.org/abs/1603.02754v3 (Mar. 2016).
100. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12,** 77 (2011).
101. Barabási, A.-L. *Network Science* http://networksciencebook.com/ (Cambridge University Press, 2015).
102. Perez De Souza, L., Alseekh, S., Brotman, Y. & Fernie, A. R. Network-based strategies in metabolomics data analysis and interpretation: from molecular networking to biological interpretation. *Expert Review of Proteomics* **17,** 243–255. ISSN: 17448387 (Apr. 2020).
103. Koller, D. & Friedman, N. Probabilistic Graphical Models: Principles and Techniques.
104. Amara, A. *et al.* Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. *Frontiers in Molecular Biosciences* **9,** 841373. ISSN: 2296889X (Mar. 2022).
105. Wickham, H., François, R., Henry, L., Müller, K. & Vaughan, D. *dplyr: A Grammar of Data Manipulation* R package version 1.1.4, https://github.com/tidyverse/dplyr (2023). https://dplyr.tidyverse.org.
106. Galili *et al.* heatmaply: an R package for creating interactive cluster heatmaps for online publishing. *Bioinformatics.* https://academic.oup.com/bioinformatics/article-pdf/doi/10.1093/bioinformatics/btx657/21358327/btx657.pdf (2017).
107. Venables, W. N. & Ripley, B. D. *Modern Applied Statistics with S* Fourth. ISBN 0-387-95457-0. https://www.stats.ox.ac.uk/pub/MASS4/ (Springer, New York, 2002).
108. Vaughan, D. & Dancho, M. *furrr: Apply Mapping Functions in Parallel using Futures* https://github.com/DavisVaughan/furrr, https://furrr.futureverse.org/ (2022).

Appendix A. Multiclass ROC curves

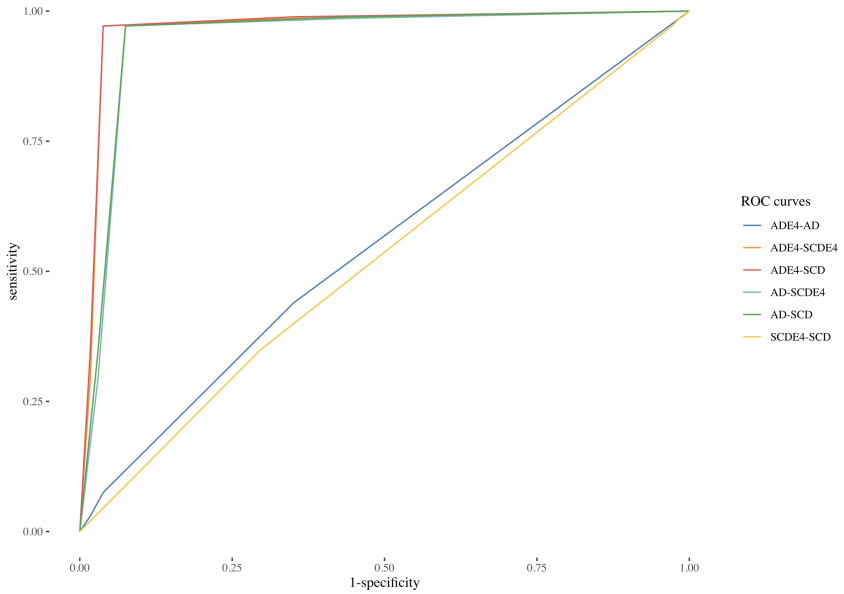


Figure 11. ROC curves of Decision Tree fitting the 6 ML-estimated latent factors on top of the clinical background variables, obtained from repeated (100 times) 10-fold CV

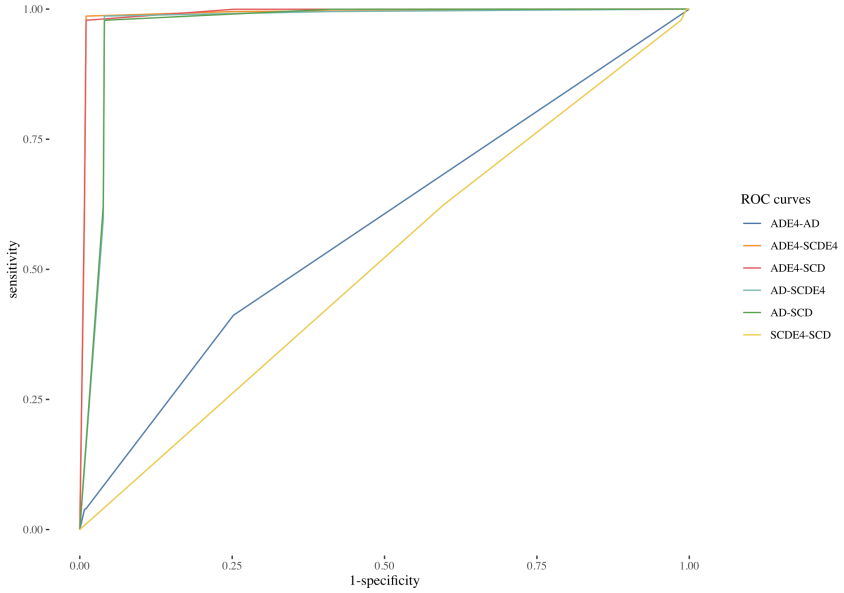


Figure 12. ROC curves of Xtreme Gradient Booster fitting the 6 ML-estimated latent factors on top of the clinical background variables, obtained from repeated (100 times) 10-fold CV

Appendix B. R Session Information

```
R version 4.3.2 (2023-10-31)
Platform: aarch64-apple-darwin20 (64-bit)
Running under: macOS Sonoma 14.2.1

Matrix products: default
BLAS:   .../vecLib.framework/Versions/A/libBLAS.dylib
LAPACK: LAPACK version 3.11.0

locale:
[1] en_US.UTF-8

time zone: Europe/Amsterdam
tzcode source: internal

attached base packages:
[1] stats     graphics  grDevices datasets  utils     methods   base

other attached packages:
 [1] rags2ridges_2.2.7       nnet_7.3-19             DMwR2_0.0.2
 [4] dplyr_1.1.4             caret_6.0-94            lattice_0.21-9
 [7] globaltest_5.56.0       survival_3.5-7          heatmaply_1.5.0
[10] xgboost_1.7.6.1         pROC_1.18.5             FMradio_1.1.1
[13] future_1.33.0           ggthemes_5.0.0          corpcor_1.6.10
[16] viridisLite_0.4.2       plotly_4.10.3           ggplot2_3.4.4
[19] rpart_4.1.21            furrr_0.3.1             viridis_0.6.4
[22] e1071_1.7-14

loaded via a namespace (and not attached):
  [1] splines_4.3.2           bitops_1.0-7            tibble_3.2.1
  [4] XML_3.99-0.16           lifecycle_1.0.4         globals_0.16.2
  [7] magrittr_2.0.3          Hmisc_5.1-1             rmarkdown_2.25
 [10] lubridate_1.9.3         zlibbioc_1.48.0         sfsmisc_1.1-16
 [13] RCurl_1.98-1.13         ipred_0.9-14            lava_1.7.3
 [16] S4Vectors_0.40.2        listenv_0.9.0           gRbase_2.0.1
 [19] codetools_0.2-19        tidyselect_1.2.0        TSP_1.2-4
 [22] jsonlite_1.8.8          Formula_1.2-5           iterators_1.0.14
 [25] snowfall_1.84-6.3       Rcpp_1.0.11             glue_1.6.2
 [28] tufte_0.13              TTR_0.24.4              GenomeInfoDb_1.38.2
 [31] fastmap_1.1.1           fansi_1.0.6             digest_0.6.33
 [34] RSQLite_2.3.4           utf8_1.2.4              tidyr_1.3.0
 [37] recipes_1.0.9           class_7.3-22            httr_1.4.7
 [40] gtable_0.3.4            timeDate_4032.109       blob_1.2.4
 [43] RBGL_1.78.0             GSEABase_1.64.0         scales_1.3.0
 [46] knitr_1.45              rstudioapi_0.15.0       tzdb_0.4.0
 [49] curl_5.2.0              proxy_0.4-27            cachem_1.0.8
 [52] foreign_0.8-85          AnnotationDbi_1.64.1    pillar_1.9.0
 [55] VGAM_1.1-9              xtable_1.8-4            cluster_2.1.4
```

```
 [58] cli_3.6.2              compiler_4.3.2         rlang_1.1.2
 [61] plyr_1.8.9             stringi_1.8.3          assertthat_0.2.1
 [64] Matrix_1.6-1.1         hms_1.1.3              bit64_4.0.5
 [67] quantmod_0.4.25        bit_4.0.5              hardhat_1.3.0
 [70] graph_1.80.0           xts_0.13.1             MASS_7.3-60
 [73] dendextend_1.17.1      backports_1.4.1        yaml_2.3.8
 [76] DBI_1.1.3              RColorBrewer_1.1-3     expm_0.999-8
 [79] purrr_1.0.2            BiocGenerics_0.48.1    seriation_1.5.4
 [82] IRanges_2.36.0         RSpectra_0.16-1        GenomeInfoDbData_1.2.11
 [85] annotate_1.80.0        parallelly_1.36.0      stats4_4.3.2
 [88] foreach_1.5.2          tools_4.3.2            snow_0.4-4
 [91] prodlim_2023.08.28     gridExtra_2.3          xfun_0.41
 [94] ca_0.71.1              withr_2.5.2            BiocManager_1.30.22
 [97] timechange_0.2.0       R6_2.5.1               colorspace_2.1-0
[100] generics_0.1.3         renv_1.0.3             data.table_1.14.10
[103] htmlwidgets_1.6.4      ModelMetrics_1.2.2.2   pkgconfig_2.0.3
[106] registry_0.5-1         XVector_0.42.0         htmltools_0.5.7
[109] Biobase_2.62.0         png_0.1-8              gower_1.0.1
[112] reshape2_1.4.4         checkmate_2.3.1        nlme_3.1-163
[115] zoo_1.8-12             stringr_1.5.1          parallel_4.3.2
[118] grid_4.3.2             reshape_0.8.9          vctrs_0.6.5
[121] htmlTable_2.4.2        evaluate_0.23          readr_2.1.4
[124] crayon_1.5.2           future.apply_1.11.0    fdrtool_1.2.17
[127] munsell_0.5.0          Biostrings_2.70.1      lazyeval_0.2.2
[130] KEGGREST_1.42.0        igraph_1.6.0           memoise_2.0.1
[133] base64enc_0.1-3        webshot_0.5.5
```