Title

Metabolic dysfunctions predict the development of Alzheimer's disease: Statistical and machine learning analysis of EMR data.

Authors

Liu, Rex

Durbin-Johnson, Blythe

Paciotti, Brian

et al.

Publication Date

2024-08-14

DOI

10.1002/alz.14101

Received: 8 February 2024 | Revised: 4 June 2024 | Accepted: 6 June 2024

DOI: 10.1002/alz.14101

RESEARCH ARTICLE

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

Metabolic dysfunctions predict the development of Alzheimer's disease: Statistical and machine learning analysis of EMR data

Rex Liu[1] | Blythe Durbin-Johnson[2] | Brian Paciotti[3] | Albert T. Liu[4] | Alyssa Weakley[5] | Xin Liu[1] | Yu-Jui Yvonne Wan[6]

[1]Department of Computer Science, University of California, Davis, Sacramento, California, USA

[2]Department of Public Health Sciences, University of California, Davis, Sacramento, California, USA

[3]Data Center of Excellence, University of California, Davis, Sacramento, California, USA

[4]Department of Obstetrics/Gynecology, University of California, Davis, Sacramento, California, USA

[5]Department of Neurology, University of California, Davis, Sacramento, California, USA

[6]Department of Medical Pathology and Laboratory Medicine, University of California, Davis, Sacramento, California, USA

Correspondence

Yu-Jui Yvonne Wan, Department of Pathology and Laboratory Medicine, University of California, Davis Health, Room 3400B, Research Building III, 4645 2nd Ave, Sacramento, CA 95817, USA.
Email: yjywan@ucdavis.edu

Funding information

California Department of Public Health; Chronic Disease Control Branch; Alzheimer's Disease Program, Grant/Award Numbers: 18-10925, 22-10079

Abstract

INTRODUCTION: The incidence of Alzheimer's disease (AD) and obesity rise concomitantly. This study examined whether factors affecting metabolism, race/ethnicity, and sex are associated with AD development.

METHODS: The analyses included patients $\geq$ 65 years with AD diagnosis in six University of California hospitals between January 2012 and October 2023. The controls were race/ethnicity, sex, and age matched without dementia. Data analyses used the Cox proportional hazards model and machine learning (ML).

RESULTS: Hispanic/Latino and Native Hawaiian/Pacific Islander, but not Black subjects, had increased AD risk compared to White subjects. Non-infectious hepatitis and alcohol abuse were significant hazards, and alcohol abuse had a greater impact on women than men. While underweight increased AD risk, overweight or obesity reduced risk. ML confirmed the importance of metabolic laboratory tests in predicting AD development.

DISCUSSION: The data stress the significance of metabolism in AD development and the need for racial/ethnic- and sex-specific preventive strategies.

KEYWORDS

alcohol abuse, metabolic liver disease, metabolism, non-infectious hepatitis, obesity

Highlights

- Hispanics/Latinos and Native Hawaiians/Pacific Islanders show increased hazards of Alzheimer's disease (AD) compared to White subjects.
- Underweight individuals demonstrate a significantly higher hazard ratio for AD compared to those with normal body mass index.
- The association between obesity and AD hazard differs among racial groups, with elderly Asian subjects showing increased risk compared to White subjects.
- Alcohol consumption and non-infectious hepatitis are significant hazards for AD.
- Machine learning approaches highlight the potential of metabolic panels for AD prediction.

1 | INTRODUCTION

The incidence of Alzheimer's disease (AD) has been rising, which can be due to the prevalence of obesity, longer lifespans, educational attainment, or improved diagnostic tools.[1,2] Regarding lifespan, the trends might vary by population.[3] For example, historically, non-Hispanic Whites had higher life expectancies compared to non-Hispanic Blacks, but the differences have been narrowing over time.[4] Nevertheless, AD still has an uneven burden on aged Black Americans. Thus, the differences and risks for AD for different races remain to be addressed. Although there has been increased attention to AD racial disparity, there are still many challenges, which include issues with recruitment, lack of biological data, and uncertainties in diagnostic criteria.[5–7]

Alarmingly, the worldwide prevalence of obesity has nearly tripled since 1975.[8] In the United States, almost three quarters of adults ages $\geq$ 20 are either overweight or obese.[9] Obesity is associated with a range of health risks, including metabolic syndromes or liver disease, cardiovascular diseases (CVD), and type 2 diabetes mellitus (T2DM).[10,11] All of these, which are comorbidities of obesity, might be risks for AD.[12] Possibly as a more direct cause, obesity at midlife is an established risk for AD.[13] However, late-life high body mass index (BMI) is associated with lower amyloid beta (Aβ) load, higher brain volumes, and slower cognitive decline. Thus, late-life obesity can be a protective factor for AD.[14] This paradox should be validated in the context of racial/ethnic groups. California is known for its high racial/ethnic diversity and has large Hispanic and Latino populations. Moreover, California is also home to large and diverse Asian populations.[15] Overall, California provides a unique opportunity to study risks for AD in ethnically diverse populations.

The liver is the most important metabolic organ. Western diet intake and aging, which both stress metabolism and induce chronic inflammation, can contribute to the development of AD.[16,17] The current study tests a hypothesis that metabolic dysfunction is an AD hazard in a race- or sex-specific manner. The long-term goal is to develop population-based preventive strategies. We studied electronic medical records (EMR) data from six University of California (UC) health systems (Davis, San Francisco, Los Angeles, Irvine, Riverside, and San Diego), covering both northern and southern California. The studied patients had late-onset AD diagnoses; a condition likely influenced by lifestyle. Demographic data, laboratory test data, and diseases that might affect metabolic functions were included in statistical and machine learning (ML) analyses.

Our data revealed the disproportionate AD burden on Hispanics/Latinos and Native Hawaiians/Pacific Islanders, surprisingly, but not in Black subjects in California. Moreover, being underweight in late life was a hazard for AD. Alcohol abuse or dependence, as well as non-infectious hepatitis, were hazards in race/ethnicity- or sex-specific

RESEARCH IN CONTEXT

1. **Systematic review**: Previous works have illustrated the complex interplay between the development of Alzheimer's disease (AD) and metabolic health, including conditions such as obesity, diabetes, and alcohol abuse. However, many of these works have focused primarily on homogeneous populations, often overlooking the potential impact of race/ethnicity and sex on AD risk. Furthermore, while some studies have explored the relationship between metabolic factors and AD risk in diverse populations, there remains a gap in understanding the specific associations.
2. **Interpretation**: The present study investigates the relationship between metabolic-related health issues and AD risk in a diverse population. By analyzing electronic medical records data from six University of California health systems, this study identified several key findings. First, Hispanics/Latinos and Native Hawaiians/Pacific Islanders exhibited an increased hazard of AD compared to White subjects, highlighting the importance of considering racial/ethnic disparities in AD risk. Additionally, the study revealed that alcohol abuse and dependence were significant hazards for AD, particularly among women. Moreover, the presence of non-infectious hepatitis was positively associated with higher AD incidence, underscoring the need to consider comorbidities in AD risk assessment. Interestingly, while overweight or obesity was associated with a reduced risk of AD in aged populations, underweight individuals had an increased AD risk. Furthermore, the impact of weight status on AD risk varied across racial/ethnic groups, with obese conditions significantly increasing AD hazard among Asian subjects. Machine learning techniques further supported the importance of metabolic laboratory tests in predicting AD development, highlighting the potential utility of biomarkers in AD risk assessment and early detection.
3. **Future directions**: Building upon the findings of this study, future research should aim to elucidate the underlying mechanisms driving the observed associations between metabolic factors and AD risk in diverse populations. Longitudinal studies are needed to understand the temporal relationship between metabolic health and AD development, as well as the potential moderating effects of genetic and environmental factors. Additionally, intervention studies should explore the efficacy of targeted preventive strategies tailored to specific racial/ethnic and sex groups to reduce AD. Furthermore, efforts to integrate metabolic biomarkers into existing AD risk prediction models may help improve the accuracy of early detection and the development of personalized preventive interventions. Overall, addressing the racial/ethnic and sex disparities in AD risk requires a multifaceted approach that considers the interplay among metabolic health, lifestyle factors, and sociodemographic determinants.

manners. These findings stress the importance of having metabolic health to prevent AD.

2 | METHODS

2.1 | Study cohort

This study used Health Insurance Portability and Accountability Act (HIPAA)–compliant health data from the University of California Health Datawarehouse (UCHDW), which included data from six health systems. The Center for Data-driven Insights and Innovation built, maintained, extracted, and harmonized the data using the Observational Medical Outcomes Partnership common data model (OMOP) version 5.1. With OMOP, disparate clinical data can be compiled into one UCHDW using standard data structures and medical vocabularies such as Logical Observation Identifiers, Names, and Codes; Prescription Norms; and Systematized Nomenclature of Medicine. Currently, the OMOP database at UCHDW comprises 942,483 patients who have had at least one visit after January 1, 2012. Table S1 in supporting information shows the process of mapping that standardizes medical terminologies. Table S2 in supporting information shows the demographic information of the UCHDW cohort.

2.2 | Inclusion and exclusion criteria

Patients ≥ 65 years with a diagnosis of AD after January 1, 2012, were identified based on AD-related International Classification of Diseases (ICD)-9/10 codes (Table S3 in supporting information). Patients aged ≥ 89 had their age set as 89 by HIPAA regulations. These patients were labeled as "AD" (cases). AD patients who had a history of intracranial injury, genetic susceptibility, or early-onset AD were excluded (Table S4 in supporting information). The study encompassed 23,182 AD patients (aged ≥ 65), and 62.29% were female. The mean age of these patients was 85.14 ± 5.33 years.

The controls (166,931 patients aged ≥ 65) were randomly selected with matched race/ethnicity (Table S5 in supporting information). The control patients did not have AD, mild cognitive impairment, intracranial injury, or genetic susceptibility. Table S5 summarizes the demographic information of studied AD and control cohorts. Figure 1 and Figure 2 provide a detailed overview of the inclusion and exclusion criteria.

2.3 | Features and variables

Sixty variables classified into three categories were studied to assess their impact on AD development (Table S6 in supporting information). The first category includes demographic information: age, sex, race/ethnicity, Area Deprivation Index (ADI), UC institution, and BMI. Specifically, ADI aggregates various socioeconomic data at the census block group level (e.g., income, education, employment, housing quality) to measure socioeconomic disadvantage that occurs at the neighborhood level.[18] A rank of 1 means very low disadvantage, and 10 is the highest level of socioeconomic disadvantage in California. The secondary category includes diseases that affect metabolic function, like non-alcoholic steatohepatitis or autoimmune hepatitis, alcohol drinking, metabolic diseases, and T2DM, as well as CVD. The last category is lab tests, encompassing 37 laboratory test data (Table S6). Variables are referred to as features in the context of ML.

2.4 | Statistical analysis

Our analysis showed that patients with severe liver diseases died 6 to 7 years earlier than age 77, which was the median age of AD diagnosis in our study cohort. The mean age of death for patients who had liver cancer, cirrhosis, or toxic liver disease was 73, 72, and 74, respectively. To mitigate the impact of mortality on analysis and have sufficient lab data and time for follow-up visits, we decided to set the age cutoff at 70. Statistical analyses were done using 34,277 patients who had at least one visit to a UC institution before age 69 and who remained alive without AD at age 70. Within this group, 866 patients later had AD diagnoses after age 70.

The data characteristics based on the final follow-up status are presented in Table 1. Time to AD (age 70 as time 0) was modeled using the Cox proportional hazards model,[19] treating loss to follow-up or death as censoring events. The Cox proportional hazards model was used to investigate the time to AD and to address the presence of control subjects who might have eventually developed AD given longer follow-ups. Using models without covariate adjustments, the impact of demographics (sex, race/ethnicity) and ADI on AD diagnosis was studied. When assessing the effect of BMI on the timing of AD diagnosis, adjustments were made for sex, race/ethnicity, ADI, and UC institution. The minimum and maximum pre-age 70 BMI for each subject were used in separate analyses. Differences between ethnicity/race and sex in the effects of a specific variable on AD were examined through models incorporating two-way interaction effects between the variable in question and race/ethnicity and sex. When evaluating the impact of medical diagnoses on the time to AD, the models considered the following covariates: sex, race/ethnicity, ADI, and UC institution. The resulting *p* values for diagnoses were adjusted using the Bonferroni correction.

A multivariable Cox proportional hazards model of time to AD by subject characteristics was fitted using variables that were significant (adjusted or raw $p < 0.05$, as applicable) in univariable analysis. The model included alcohol, non-infectious hepatitis, maximum BMI (race-based), sex, race, UC site, and ADI. The minimum race-based BMI and non–race-based BMI measures were not included due to high collinearity with maximum race-based BMI.

To study the impact of laboratory test variables (Table S6) on the timing to AD, the models used the pre-age 70 minimum or maximum values of each lab measurement (in separate models). Models included sex, race/ethnicity, ADI, and UC institution as covariates and the inverse probability of missingness weighted observations. The

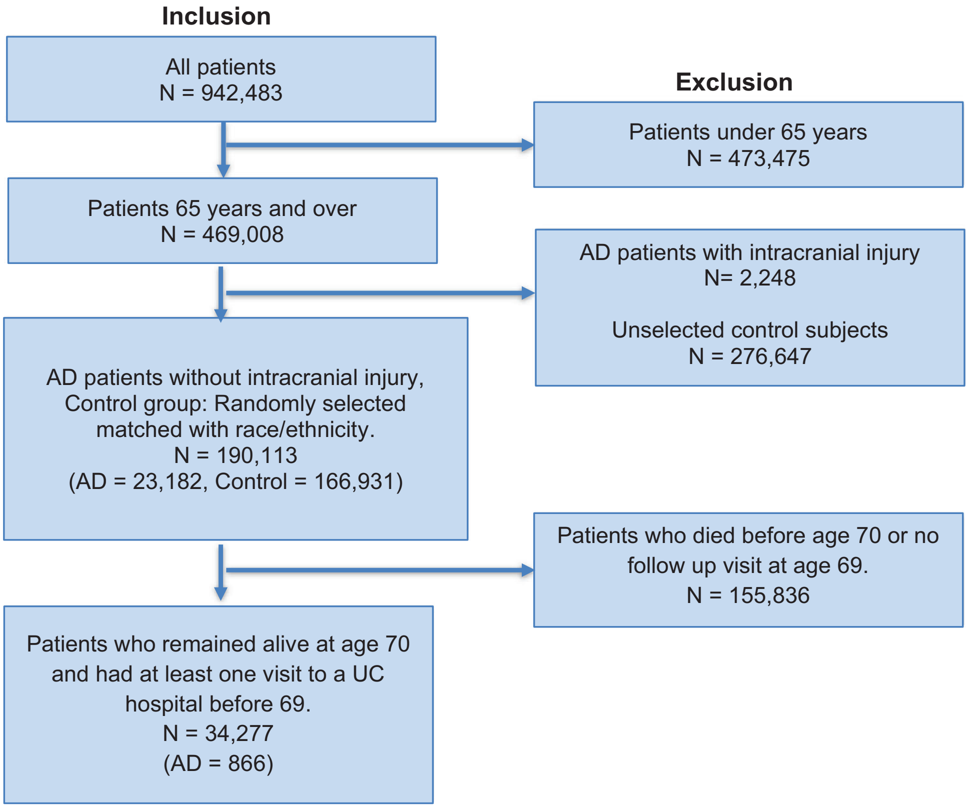
Inclusion
All patients
N = 942,483
Exclusion
Patients under 65 years
N = 473,475
Patients 65 years and over
N = 469,008
AD patients with intracranial injury
N= 2,248
Unselected control subjects
N = 276,647
AD patients without intracranial injury,
Control group: Randomly selected
matched with race/ethnicity.
N = 190,113
(AD = 23,182, Control = 166,931)
Patients who died before age 70 or no
follow up visit at age 69.
N = 155,836
Patients who remained alive at age 70
and had at least one visit to a UC
hospital before 69.
N = 34,277
(AD = 866)

FIGURE 1 Inclusion/exclusion criteria for statistical analysis. The starting point includes all patients in the UCHDW cohort. Numbers in each box correspond to the number of patients included/excluded, AD patients, and control patients. AD, Alzheimer's disease; UC, University of California; UCHDW, University of California Health Datawarehouse.

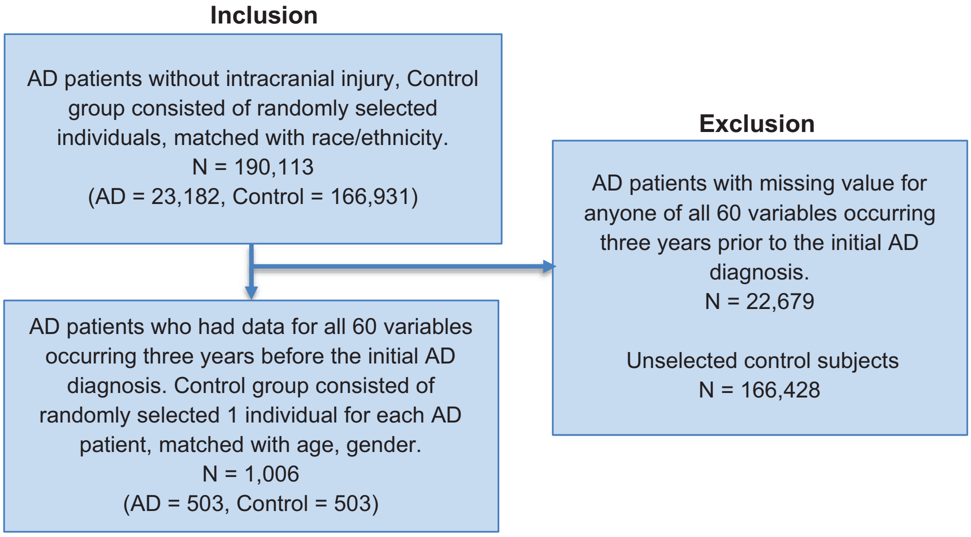
Inclusion
AD patients without intracranial injury, Control
group consisted of randomly selected
individuals, matched with race/ethnicity.
N = 190,113
(AD = 23,182, Control = 166,931)
Exclusion
AD patients with missing value for
anyone of all 60 variables occurring
three years prior to the initial AD
diagnosis.
N = 22,679
Unselected control subjects
N = 166,428
AD patients who had data for all 60 variables
occurring three years before the initial AD
diagnosis. Control group consisted of
randomly selected 1 individual for each AD
patient, matched with age, gender.
N = 1,006
(AD = 503, Control = 503)

FIGURE 2 Inclusion/exclusion criteria for machine learning. The starting point includes all patients in the Studied AD and Control cohort selected from the UCHDW cohort. Numbers in each box correspond to the number of patients included/excluded, AD patients, and control patients. AD, Alzheimer's disease; UCHDW, University of California Health Datawarehouse.

TABLE 1 Demographic information of patients from the matched case–control dataset at age 70 and had at least one visit to a UC institution before age 69.

	All patients	No AD at last follow-up	AD diagnosed
Number	34,277	33,411	866
Age at last follow-up			
Mean, years (SD)	72.5 (3.92)	72.4 (3.92)	75.8 (2.04)
Median [min, max]	73 [65, 89]	73 [65, 89]	76 [71, 80]
Sex			
Female	20,732 (60.5%)	20,201 (60.5%)	531 (61.3%)
Male	13,545 (39.5%)	13,210 (39.5%)	335 (38.7%)
Race			
White	24321 (71.0%)	22571 (71.0%)	613 (70.8%)
Black	2040 (6.0%)	1868 (5.9%)	39 (4.5%)
Asian	3571 (10.4%)	3348 (10.5%)	79 (9.1%)
Hispanic/Latino	3520 (10.3%)	3242 (10.2%)	102 (11.8%)
American Indian/Alaska Native	62 (0.2%)	60 (0.2%)	<10
Native Hawaiian/Other Pacific Islander	170 (0.5%)	160 (0.5%)	10 (1.2%)
Multi-race	593 (1.7%)	572 (1.7%)	21 (2.4%)
ADI			
Mean (SD)	4.18 (2.76)	4.19 (2.76)	3.95 (2.60)
Median [min, max]	4.00 [1, 10]	4.00 [1, 10]	3.00 [1, 10]

Abbreviations: AD, Alzheimer's disease; ADI , Area Deprivation Index; SD , standard deviation; UC, University of California.

Bonferroni correction adjusted *p* values for multiple tests across different laboratory variables.

All statistical analyses were conducted using R version 4.2.2 within the Spark analytical platform Databricks.

2.5 | Machine learning

Unlike statistical models, which primarily examine the impact of a single variable on time to AD, ML analysis incorporated 60 variables/features in ML. To study the impact of different features/variables, lab tests were classified into five groups: (1) metabolic panel, (2) blood counts, (3) serum lipids, (4) sugar, and (5) heart function (Table S6). Each group was removed to evaluate its impact on the ML model's performance.

To account for variations in testing frequencies and visit periods among patients, the values for each feature in AD patients were based on their last lab test results 3 years before the AD diagnosis date. For the control subjects, the values for each feature corresponded to their last lab test data 3 years before their final visit. A 3-year timeframe was selected to enable AD prediction while maintaining a sufficient sample size for analysis. Based on this criterion, the study included 503 AD patients; each had data for all 60 features. To build a balanced dataset, control patients included 503 age- and sex-matched randomly selected patients. Table 2 shows demographic information of AD and control cohorts used for ML.

TABLE 2 Demographic information of balanced ML dataset after using case–control age–sex-matching strategy.

Characteristics	AD	Controls
Number	503	503
Mean age, years (SD)	82.9 (5.2)	82.9 (5.20)
Sex, number (%)		
Female	323 (64.2%)	323 (64.2%)
Male	180 (35.8%)	180 (35.8%)
Race, number (%)		
White	337 (67%)	329 (65.4%)
Black	40 (8%)	47 (9.4%)
Asian	60 (11.9%)	81 (16.1%)
Hispanic/Latino	60 (11.9%)	45 (8.9%)
American Indian/Alaska Native	<10	<10
Native Hawaiian/Other Pacific Islander	<10	<10

Abbreviations: AD, Alzheimer's disease; ML, machine learning; SD, standard deviation.

In the data processing step, one-hot encoding was used for categorical variables, converting each category into a binary vector representation.[20] This data processing method ensured categorical data were effectively incorporated into the analysis and avoided ordinal assumptions. For numerical variables, min–max normalization was applied, scaling the values to a defined range, that is, 0 and 1, to ensure a consistent treatment in the model. All models were conducted using Scikit-learn version 0.20 in Databricks.

Ten-fold cross-validation was implemented, a reliable estimate of model performance compared to a train/test split to reduce bias and variance. The dataset was randomly shuffled and split into ten folds. In each of the ten folds, the model was trained on nine of these folds and tested on the remaining one. This process was repeated with a different fold as the test was set in each iteration until all ten folds were used for testing. The mean classification accuracy and standard deviation (SD) of all iterations were calculated to determine the performance of each algorithm.

3 | RESULTS

3.1 | Demographic overview for statistical and ML analysis

The study population for statistical analysis comprised 34,277 patients, with 33,411 having no AD diagnosis at the last follow-up and 866 diagnosed with AD. The mean age was 72.5 years (SD = 3.92). The mean age of AD patients was 75.8 (SD = 2.04). Females were 60.5% of the overall population, with similar distributions in controls (60.5%) and AD groups (61.3%). For ML analyses, 1006 subjects were evenly split into AD patients and controls (503 each), and both cohorts had a mean age of 82.9 years (SD = 5.2). Sex distribution was identical in AD and control groups, of which 64.2% were females.

3.2 | Racial disparity of AD

Different racial/ethnic groups have distinct dietary habits and lifestyles that may contribute to AD development. Table 3 shows the univariable Cox proportional hazard analyses of sex, race/ethnicity, and ADI to the timing of AD diagnosis. Compared to White subjects, Hispanics/Latinos ($p = 0.013$) and Native Hawaiians/Pacific Islanders had increased hazards of AD ($p = 0.001$). However, neither sex nor ADI was associated with time to AD (Table 3). Table S7 in supporting information shows the results of multivariable Cox proportional hazard models of time to AD by demographic information, as well as alcohol and non-infectious hepatitis. After adjusting variables in the model, the Hispanic or Latino race and Native Hawaiian or Pacific Islanders had a higher hazard of AD relative to White subjects.

3.3 | The impact of late-life BMI on AD

We studied both the maximum and minimum to assess the impact of BMI (based on the Centers for Disease Control [CDC] guideline) because BMI values fluctuated. After adjusting for sex, race, ADI, and UC site, Cox proportional hazards models revealed that underweight for at least one assessment had a significantly higher hazard ratio (HR) than those with normal BMI ($p \leq 0.001$). In contrast, irrespective of using either the maximum or minimum, overweight or obese had reduced HR, indicating a protective effect (Table 3). However, Asian subjects showed a significantly increased hazard of AD with obesity versus overweight, which differed significantly from the small protective effect found in White subjects when modeling by using the maximum BMI ($p = 0.021$, Table 4). Conversely, there was no significant increase in AD hazard between races when using the minimum BMI in modeling (Table S8 in supporting information).

Adjusting for all other variables in the multivariable Cox proportional hazard model, obese or overweight (maximum BMI) was associated with a significantly lower hazard of AD relative to normal BMI. Moreover, high ADI increases AD hazard (Table S7).

3.4 | Alcohol drinking and non-infectious hepatitis are AD hazards

We studied whether liver dysfunctions can be hazards for AD. Additionally, other metabolic-related health issues, including T2DM and CVD, as well as alcohol drinking, were examined (Table 5). Both non-infectious hepatitis (HR = 5.181) and alcohol drinking (HR = 2.595) were significant hazards by using the univariable Cox proportional hazard models of time to AD (adjusted p value < 0.001, Table 5). Although toxic liver disease and T2DM had elevated HR, they did not reach statistical significance. The ICD-10 codes used to diagnose alcohol drinking and non-infectious hepatitis are listed in Table S9A,B in supporting information, respectively.

Based on the result of the multivariable Cox proportional hazards model of time to AD by demographic information, alcohol and non-infectious hepatitis were associated with a significantly higher hazard of AD relative to reference groups (Table S7).

By race, the impact of alcohol drinking diagnosis on time to AD significantly affected White subjects but not others (Table S10 in supporting information). Due to the limited number of Native Hawaiian/Pacific Islander subjects diagnosed with alcohol-related issues, individuals from this demographic were excluded from the model involving AD by alcohol with race and sex interactions. Furthermore, the effect of alcohol drinking on time to AD was worse for women than men (interaction effect $p = 0.034$, Table S10). It is intriguing to note that alcoholic-related hepatitis was not an AD hazard after adjusting for race, sex, ADI, and UC location (adjusted $p = 0.95$, Table 5).

Standard lab tests frequently used in an annual routine health checkup were studied, and none was linked to AD. Table S11 in supporting information presents the results of the Cox proportional hazards models by analyzing both the minimum and maximum of 37 laboratory tests. Thus, the individual laboratory test did not show a significant hazard to AD.

3.5 | Metabolic panel as a novel data source for AD prediction

Five ML approaches, namely logistic regression, linear support vector machine, decision tree, random forest, and gradient boosting decision trees (GBDT) were used to determine whether the studied variables had predicting power for AD diagnosis.[21–25] Table S12 in supporting information shows the mean classification accuracy and

TABLE 3 Univariable Cox proportional hazard models of time to AD by demographic information.

Variable	Hazard ratio	95% CI	*p*-value
Sex			
Female vs. male	0.992	(0.865, 1.137)	0.907
Race			
American Indian or Alaska Native vs. White	1.365	(0.341, 5.471)	0.660
Asian vs. White	0.848	(0.671, 1.071)	0.166
Black vs. White	0.877	(0.634, 1.212)	0.428
Hispanic/Latino vs. White	1.302	(1.055, 1.605)	0.013*
Native Hawaiian/Pacific Islander vs. White	2.708	(1.449, 5.059)	0.001*
ADI	1.000	(0.975, 1.025)	0.985
BMI			
Minimum before age 70			
Underweight vs. normal	2.018	(1.627, 2.501)	< 0.001**
Overweight vs. normal	0.599	(0.505, 0.709)	< 0.001**
Obese vs. normal	0.409	(0.316, 0.529)	< 0.001**
Maximum before age 70			
Underweight vs. normal	0.864	(0.407, 1.833)	0.703
Overweight vs. normal	0.748	(0.629, 0.889)	0.001*
Obese vs. normal	0.718	(0.601, 0.858)	< 0.001**

Note: Models of the effect of sex, race, and ADI on time to AD were not covariate-adjusted. Other models included sex, race/ethnicity, area deprivation index, and UC institution as covariates.
Abbreviations: AD, Alzheimer's disease; ADI, Area Deprivation Index; BMI, body mass index; CI, confidence interval; HR, hazard ratio.
$^*p < 0.05$, $^{**}p < 0.001$.

SDs for the five algorithms evaluated on the balanced ML dataset using 10-fold cross-validation. The GBDT algorithm achieved the highest accuracy of 62.43% for AD prediction when all 60 variables were included, surpassing the 50% accuracy expected by random guess. However, when only demographic information was included (age, sex, race/ethnicity, ADI, UC institution, and BMI), all five algorithms yielded an accuracy of ≈ 50% for AD classification. This illustrated the effectiveness of the case–control age–sex-matching strategy in mitigating the demographic-related impact on AD classification, thereby increasing the potential to identify other relevant risk factors like diagnosis and laboratory tests.

Interestingly, comparing the exclusion of disease diagnoses versus lab test data, the predicting power tended to be lower when lab test data were excluded. This finding suggested the relative importance of lab test data in predicting AD. We, therefore, further analyzed the lab tests by classifying them into five categories (Table S6). GBDT was used because it provided better prediction power than others. The results revealed that excluding the metabolic panel lab test data resulted in a mean accuracy of 54.88%, with a drop of ≈ 7.55%, compared to the accuracy of 62.43% when all variables were included (Table 6). Subsequent *t* test results produced a *p* value of 0.0003, surpassing the predefined significance threshold of 0.05, indicating a statistically significant enhancement with the inclusion of the metabolic panel lab data. This finding suggests the potential of using the metabolic panel as a novel source for AD prediction.

4 | DISCUSSION

The study unveiled a compelling relationship between BMI and AD hazards. Subjects categorized as underweight, according to CDC guidelines, had a higher HR for AD compared to those with normal BMI. Conversely, obese/overweight patients had reduced HR, revealing protective effects. These findings are particularly pertinent for subjects who had at least one visit to a UC institution before age 69 and who remained alive without AD at age 70. Our results aligned with prior research conducted on close to 2 million people in the UK, which showed that being underweight in middle and old age had an increased risk of dementia.[13] Another study showed that higher BMI was associated with an increased risk of dementia when weight was measured > 20 years before dementia diagnosis, but such association was reversed when BMI was assessed < 10 years before dementia diagnosis.[26] Our statistical analysis focused specifically on older ages (> 69), and the data revealed that underweight was a risk of developing AD, consistent with the conclusions of both data above in the aged group. Moreover, our novel findings showed that BMI as a risk for AD was race/ethnic different. It is important to note that in elderly Asians, obesity was significantly associated with an increased hazard of AD compared to being overweight ($p = 0.022$, shown in Table 4).

It is important to note that the underrepresentation of ethnic and racial minority groups in AD research remains an issue due to a lack

TABLE 4 Estimates of the effect of the maximum BMI before age 70 on AD by race/ethnicity using the Cox proportional hazards model.

Race	Comparison	Hazard ratio	95% CI	*p*-value
White	Overweight vs. normal	0.755	(0.610, 0.934)	0.009*
	Obese vs. normal	0.724	(0.582, 0.900)	0.003*
	Obese vs. overweight	0.958	(0.784, 1.171)	0.681
Asian	Overweight vs. normal	0.593	(0.346, 1.018)	0.058
	Obese vs. normal	1.251	(0.701, 2.231)	0.447
	Obese vs. overweight	2.107	(1.111, 3.997)	0.022*
Black	Overweight vs. normal	0.608	(0.255, 1.445)	0.260
	Obese vs. normal	0.471	(0.207, 1.068)	0.071
	Obese vs. overweight	0.774	(0.3657, 1.639)	0.503
Hispanic/Latino	Overweight vs. normal	0.838	(0.478, 1.469)	0.539
	Obese vs. normal	0.660	(0.378, 1.151)	0.142
	Obese vs. overweight	0.786	(0.509, 1.213)	0.278
Comparison				
Asian—White	Overweight vs. normal	0.786	(0.444, 1.393)	0.410
	Obese vs. normal	1.728	(0.938, 3.181)	0.079
	Obese vs. overweight	2.197	(1.124, 4.297)	0.021*
Black—White	Overweight vs. normal	0.805	(0.331, 1.957)	0.633
	Obese vs. normal	0.651	(0.279, 1.512)	0.317
	Obese vs. overweight	0.807	(0.372, 1.754)	0.589
Hispanic/Latino—White	Overweight vs. normal	1.110	(0.615, 2.007)	0.728
	Obese vs. normal	0.911	(0.506, 1.641)	0.757
	Obese vs. overweight	0.821	(0.509, 1.322)	0.416

Abbreviations: BMI, body mass index; CI, confidence interval.
*$p < 0.05$.

of biological data or reliable information.[27] The current study uses EMR data that offers reliable diagnostic information. The data used were harmonized from the UCHDW, which covers 34,277 patients in both northern and southern California with enriched and diverse races and ethnicities. The findings of Hispanics/Latinos and Native Hawaiians/Pacific Islanders have an increased hazard of AD, emphasizing the importance and urgency of addressing AD health disparity issues. Surprisingly, our data did not show Black subjects had increased AD hazard. Whether this is unique for California remains to be investigated.

Light-to-moderate alcohol consumption can have detrimental effects on brain structure, including global and regional brain volume, as well as white matter integrity. These effects intensify with higher alcohol intake levels and significantly increase the risk of AD.[28] Further, excessive alcohol consumption in midlife is categorized as a modifiable risk factor for AD.[29] However, our data further revealed a significant association between alcohol drinking as a hazard to AD affecting women. Additionally, non-infectious hepatitis was a hazard to AD. This finding highlights the importance of addressing alcohol use and liver health in AD prevention strategies, particularly among women. However, our data did not show alcoholic liver disease is a hazard. It is essential to mention that our patients were formally diagnosed with alcohol abuse/dependence or intoxication. They had severe drinking problems, and the likelihood of having alcoholic hepatitis for those patients should be very high. There is a possibility that those patients with serious drinking problems did not seek medical attention for their potential liver problems. Thus, the number of patients diagnosed with alcoholic hepatitis was small.

The impacts of T2DM and CVD on AD have been studied extensively. Studies have demonstrated a higher incidence of AD in patients with T2DM or CVD compared to those without comorbidities.[30,31] However, for patients who had at least one visit to a UC institution before age 69 and who remained alive without AD at age 70, the analysis did not identify any of those diseases as hazards to AD. Studying other age ranges or increasing the sample size would validate the findings for the UC patients.

The investigation into various lab test variables shows that individual laboratory tests did not predict AD hazards. Considering these results within the context of broader hazard factors is essential. Thus, the laboratory tests were categorized into five groups for ML analysis. An exciting finding of this study is the potential of using the metabolic panels as a novel source for AD prediction. These findings highlight the crucial role of metabolic health and its broad-reaching implications for cognitive well-being. Overall, the presented

TABLE 5 Univariable Cox proportional hazard models of time to AD by diagnosis.

Variable	Hazard ratio	95% CI	*p*-value	Adjusted *p*
NAFLD: yes vs. no	0.885	(0.630, 1.242)	0.479	1.000
NASH: yes vs. no	0.805	(0.360, 1.801)	0.598	1.000
Alcoholic-related hepatitis: yes vs. no	1.031	(0.385, 2.759)	0.951	1.000
Toxic liver diseases: yes vs. no	2.458	(0.790, 7.642)	0.120	1.000
Hepatic failure: yes vs. no	0.666	(0.214, 2.071)	0.482	1.000
Non-infectious hepatitis: yes vs. no	5.181	(2.577, 10.413)	< 0.001	<0.001**
Cirrhosis: yes vs. no	0.867	(0.519, 1.448)	0.586	1.000
Inflammatory liver: yes vs. no	0.843	(0.271, 2.622)	0.767	1.000
Abscess liver: yes vs. no	1.689	(0.421, 6.786)	0.459	1.000
Autoimmune hepatitis: yes vs. no	0.635	(0.089, 4.512)	0.649	1.000
Other liver diseases: yes vs. no	0.813	(0.622, 1.063)	0.129	1.000
Liver disorders: yes vs. no	0.000	(0, Infinity)	0.986	1.000
Type 2 diabetes mellitus: yes vs. no	1.261	(1.070, 1.487)	0.005	0.118
Alcohol: yes vs. no	2.595	(1.922, 3.503)	< 0.001	<0.001**
Cardiovascular disease: yes vs. no	1.045	(0.885, 1.235)	0.602	1.000
Hot flashes: yes vs. no	0.952	(0.561, 1.617)	0.856	1.000
Hot flashes menopause: yes vs. no	0.939	(0.682, 1.293)	0.701	1.000

Note: All models included sex, race/ethnicity, Area Deprivation Index, and University of California institution as covariates. Adjusted *p* is Bonferroni corrected *p* value.

Abbreviations: AD, Alzheimer's disease; CI, confidence interval; NAFLD, non-alcoholic fatty livery disease; NASH, non-alcoholic steatohepatitis.

$^{**}p < 0.001$.

TABLE 6 The mean accuracy and standard deviations of GBDT algorithms with different variable groups.

	All variables	Exclude metabolic-related variables	Exclude blood-related variables	Exclude lipid-related variables	Exclude sugar-related variables	Exclude heart-related variables
Mean accuracy	62.43%	54.88%	59.44%	64.01%	62.63%	61.93%
Standard deviations	± 4.63%	± 4.18%	± 4.78%	± 4.57%	± 3.84%	± 4.15%

Abbreviation: GBDT, gradient boosting decision trees.

data stress the necessity for the development of preventive strategies that account for sex- and race/ethnicity-specific differences in AD risk.

Although the UCHDW database consists of patients with diverse racial and ethnic groups, the number of minority populations remains low compared to Whites. In addition, the study design required prior established disease diagnosis and lab test results, and patients without follow-ups were excluded, which limited the number of studied patients. Increasing the sample size and including more minority populations would further validate and enhance the robustness of our findings. Because the study focused on patients who visited UC institutions before age 69 and remained alive without AD at age 70, but had AD diagnosis after 70, this study design might limit the applicability of the findings to broader age ranges and in other health-care settings. Moreover, while our study identified significant associations between metabolic issues with AD development based on race/ethnicity and sex, the study design did not allow us to study the underlying biological mechanisms.

Regarding mechanisms, emerging evidence revealed the significance of the diet–liver–brain axis.[17,32,33] Moreover, hepatic encephalopathy can be an excellent clinical example of how liver function affects cognition.[34] However, encephalopathy is an extreme condition, and early stages of liver disease, like metabolic dysfunction–associated fatty liver disease, can affect general cognition.[35] Using mice, our published data show that Western diet-fed mice have systemic inflammation, microglia activation, and reduced learning ability and memory.[36] Additionally, shaping the gut microbiota using prebiotics improves diet-associated cognitive decline demonstrated in mouse models.[17] Thus, uncovering molecular biomarkers within the gut–liver axis based on sex and race might be essential for early prediction, diagnosis, and intervention.

ACKNOWLEDGMENTS

The authors thank the Center for Data-driven Insights and Innovation at UC Health (https://www.ucop.edu/uc-health/departments/center-for-data-driven-insights-and-innovations-cdi2.html) for its analytical and technical support related to use of the UC Health Data Warehouse and related data assets, including the UC COVID Research Data Set. This study is supported by grants funded by the California Department of Public Health, Chronic Disease Control Branch, Alzheimer's Disease Program 18-10925 and 22-10079.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial interests. The findings and conclusions in this report are those of the authors and do not necessarily represent the views or opinions of the California Department of Public Health or the California Health and Human Services Agency. Author disclosures are available in the supporting information.

DATA AVAILABILITY STATEMENT

The data within the UCHDW Database are patient data and are thus protected by laws. All Python scripts used, as well as additional information related to this paper, can be requested.

CONSENT STATEMENT

Not applicable.

ORCID

Rex Liu https://orcid.org/0000-0001-9642-8156
Yu-Jui Yvonne Wan https://orcid.org/0000-0003-2243-7759

REFERENCES

1. Zhu D, Montagne A, Zhao Z. Alzheimer's pathogenic mechanisms and underlying sex difference. *Cell Mol Life Sci*. 2021;78(11):4907-4920.
2. Soh Y, Whitmer RA, Mayeda ER, et al. Timing and level of educational attainment and late-life cognition in the KHANDLE study. *Alzheimer's Dement*. 2024;20(1):593-600.
3. Matthews KA, Xu W, Gaglioti AH, et al. Racial and ethnic estimates of Alzheimer's disease and related dementias in the United States (2015-2060) in adults aged ≥65 years. *Alzheimer's Dement*. 2019;15(1):17-24.
4. Rogers RG, Lawrence EM, Hummer RA, Tilstra AM. Racial/ethnic differences in early-life mortality in the United States. *Biodemography Soc Biol*. 2017;63(3):189-205.
5. Barnes LL. Alzheimer's disease in African American individuals: increased incidence or not enough data? *Nat Rev Neurol*. 2022;18(1):56-62.
6. Kawas CH, Corrada MM, Whitmer RA. Diversity and disparities in dementia diagnosis and care: a challenge for all of us. *JAMA Neurol*. 2021;78(6):650-652.
7. Mitchell UA, Shaw BA, Torres JM, Brown LL, Barnes LL. The effects of midlife acute and chronic stressors on Black-White differences in cognitive decline. *J Gerontol B Psychol Sci Soc Sci*. 2023;78(12):2147-2155.
8. NCD Risk Factor Collaboration (NCD-RisC). Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: a pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. *Lancet*. 2017;390(10113):2627-2642.
9. Hales CM, Carroll MD, Fryar CD, Ogden CL. *Prevalence of obesity and severe obesity among adults: United States, 2017–2018. NCHS Data Brief, no 360*. Hyattsville, MD: National Center for Health Statistics. 2020.
10. Després JP, Lemieux I. Abdominal obesity and metabolic syndrome. *Nature*. 2006;444(7121):881-887.
11. Vinciguerra F, Baratta R, Farina MG, et al. Very severely obese patients have a high prevalence of type 2 diabetes mellitus and cardiovascular disease. *Acta Diabetol*. 2013;50(3):443-449.
12. Lloret A, Monllor P, Esteve D, Cervera-Ferri A, Lloret MA. Obesity as a risk factor for Alzheimer's disease: implication of leptin and glutamate. *Front Neurosci*. 2019;13:508.
13. Qizilbash N, Gregson J, Johnson ME, et al. BMI and risk of dementia in two million people over two decades: a retrospective cohort study. *Lancet Diabetes Endocrinol*. 2015;3(6):431-436.
14. Sun Z, Wang ZT, Sun FR, et al. Late-life obesity is a protective factor for prodromal Alzheimer's disease: a longitudinal study. *Aging*. 2020;12(2):2005-2017.
15. Hsu P, Bryant MC, Hayes-Bautista TM, Partlow KR, Hayes-Bautista DE. California and the changing American narrative on diversity, race, and health. *Health Aff*. 2018;37(9):1394-1399.
16. Yang G, Jena PK, Hu Y, et al. The essential roles of FXR in diet and age influenced metabolic changes and liver disease development: a multiomics study. *Biomark Res*. 2023;11(1):20.
17. Yang G, Liu R, Rezaei S, Liu X, Wan YY. Uncovering the gut-liver axis biomarkers for predicting metabolic burden in mice. *Nutrients*. 2023;15(15):3406.
18. Kind AJH, Buckingham WR. Making neighborhood-disadvantage metrics accessible—the neighborhood atlas. *N Engl J Med*. 2018;378(26):2456-2458.
19. Zhou X, Song X. Mediation analysis for mixture Cox proportional hazards cure models. *Stat Methods Med Res*. 2021;30(6):1554-1572.
20. Yu L, Zhou R, Chen R, Lai KK. Missing data preprocessing in credit classification: one-hot encoding or imputation? *Emerg Mark Finan Trade*. 2020;58(2):472-482.
21. Fan Y, Liu M, Sun G. An interpretable machine learning framework for diagnosis and prognosis of COVID-19. *PLoS One*. 2023;18(9):e0291961.
22. Haller N, Kranzinger S, Kranzinger C, et al. Predicting injury and illness with machine learning in elite youth soccer: a comprehensive monitoring approach over 3 months. *J Sports Sci Med*. 2023;22(3):476-487.
23. Liu H, Hou W, Emolyn I, Liu Y. Building a prediction model of college students' sports behavior based on machine learning method: combining the characteristics of sports learning interest and sports autonomy. *Sci Rep*. 2023;13(1):15628.
24. Hui Y, Ma Q, Zhou XR, et al. Immunological characterization and diagnostic models of RNA N6-methyladenosine regulators in Alzheimer's disease. *Sci Rep*. 2023;13(1):14588.
25. Gao Y, Xiong X, Jiao X, et al. PRCTC: a machine learning model for prediction of response to corticosteroid therapy in COVID-19 patients. *Aging*. 2022;14(1):54-72.
26. Kivimäki M, Luukkonen R, Batty GD, et al. Body mass index and risk of dementia: analysis of individual-level data from 1.3 million individuals. *Alzheimers Dement*. 2018;14(5):601-609.
27. Lim AC, Barnes LL, Weissberger GH, et al. Quantification of race/ethnicity representation in Alzheimer's disease neuroimaging research in the USA: a systematic review. *Commun Med*. 2023;3(1):101.
28. Daviet R, Aydogan G, Jagannathan K, et al. Associations between alcohol consumption and gray and white matter volumes in the UK Biobank. *Nat Commun*. 2022;13(1):1175.
29. Livingston G, Huntley J, Sommerlad A, et al. Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet*. 2023;396(10248):413-446.
30. Sandhir R, Gupta S. Molecular and biochemical trajectories from diabetes to Alzheimer's disease: a critical appraisal. *World J Diabetes*. 2015;6(12):1223-1242.
31. Song R, Pan KY, Xu H, et al. Association of cardiovascular risk burden with risk of dementia and brain pathologies: a population-based cohort study. *Alzheimer's Dement*. 2021;17(12):1914-1922.
32. Dalile B, Van Oudenhove L, Vervliet B, Verbeke K. The role of shortchain fatty acids in microbiota-gut-brain communication. *Nat Rev Gastroenterol Hepatol*. 2019;16(8):461-478.

33. Ticinesi A, Tana C, Nouvenne A, Prati B, Lauretani F, Meschi T. Gut microbiota, cognitive frailty and dementia in older individuals: a systematic review. *Clin Interv Aging*. 2018;13:1497-1511.
34. Gilbert MC, Setayesh T, Wan YY. The contributions of bacteria metabolites to the development of hepatic encephalopathy. *Liver Research*. 2023;7:2542-5684.
35. George ES, Sood S, Daly RM, Tan SY. Is there an association between non-alcoholic fatty liver disease and cognitive function? A systematic review. *BMC Geriatr*. 2022;22(1):47.
36. Jena PK, Sheng L, Di Lucente J, Jin LW, Maezawa I, Wan YY. Dysregulated bile acid synthesis and dysbiosis are implicated in Western diet-induced systemic inflammation, microglial activation, and reduced neuroplasticity. *FASEB J*. 2018;32(5):2866-2877.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Liu R, Durbin-Johnson B, Paciotti B, et al. Metabolic dysfunctions predict the development of Alzheimer's disease: Statistical and machine learning analysis of EMR data. *Alzheimer's Dement*. 2024;1-11. https://doi.org/10.1002/alz.14101