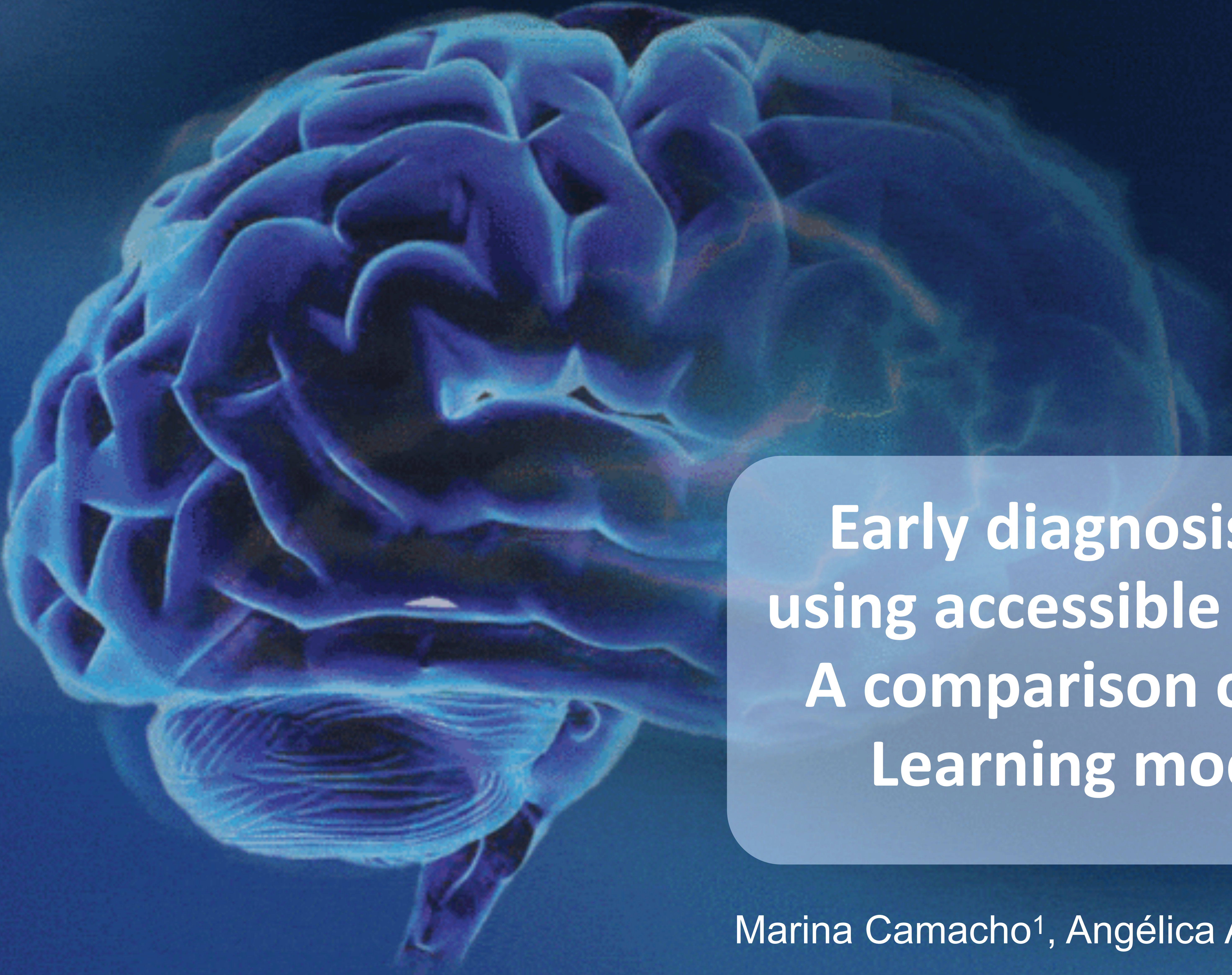


Early diagnosis of Dementia prediction using accessible exposome-based features: A comparison of Statistical and Machine Learning models on the UK Biobank

Marina Camacho[1], Angélica Atehortúa[1], Tim Wilkinson[2] and Karim Lekadir[1]

[1] Universitat de Barcelona

[2] The University of Edinburgh

What are we doing?

- We are using UK-Biobank data (~500,000 individuals).
- To create a predictive test for Dementia affordable and exposome-based.
- We used different algorithms.

Why a Dementia predictor?

Dementia is a syndrome without cure, and with a challenging diagnosis.

That's why currently there is more focus on:

Risk reduction, early intervention and timely diagnosis.

1. Geldmacher, David S., and Peter J. Whitehouse. "Evaluation of dementia." New England Journal of Medicine 335.5 (1996): 330-336.
2. (February 28, 2022) Dementia statistics. Alzheimer's Disease International. https://www.alzint.org/about/ dementia-facts-figures/dementia-statistics/
3. Breitner, John CS. "Dementia—epidemiological considerations, nomenclature, and a tacit consensus definition." Journal of geriatric psychiatry and neurology 19.3 (2006): 129-136.
4. (February 28, 2022) Treatment. NHS. https://www.nhs.uk/conditions/alzheimers-disease/treatment/
5. (February 28, 2022) Treatments for dementia. Alzheimer's Society. https://www.alzheimers.org.uk/about-dementia/treatments

Why an Early predictor?

In clinical practice, Dementia is diagnosed at late stages when symptoms become highly pronounced.

~18 years before a diagnosis

https://www.alzheimers.org.uk/research/care-and-cure-research-magazine/signs-dementia-seen-18-years-diagnosis

Additionally…

FOR ECONOMIC REASONS

Dementia costs in the UK:

- ~31 billion euros per year

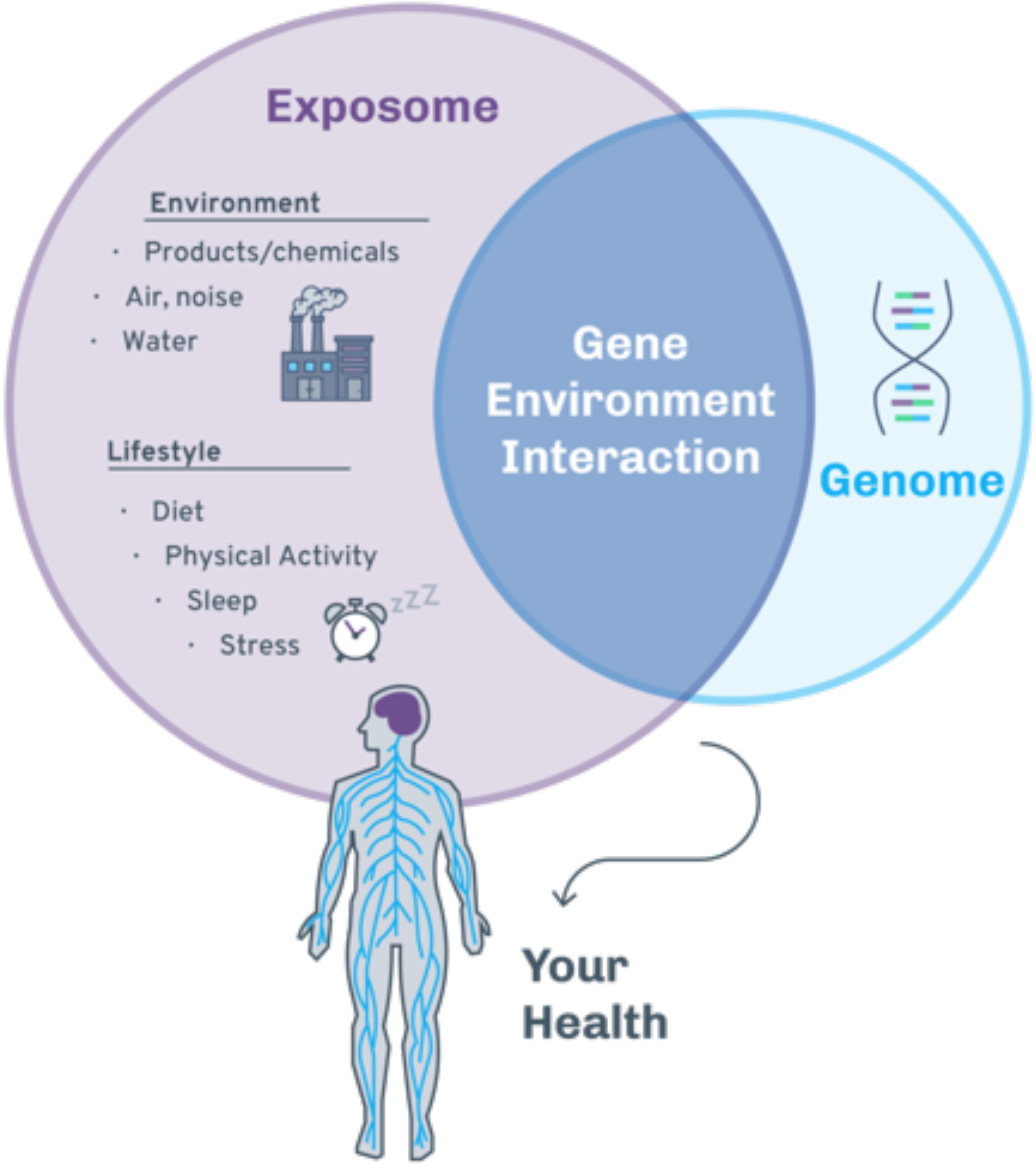
- Diagnosis: Several test and brain images over a period of time

** Additionally, our method could be applied in other cohorts and for other disorders, which will be extremely interesting for low-income countries or those having a private health care system.*

1. GBD 2016 Neurology Collaborators. Global, regional, and national burden of neurological disorders, 1990-2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet Neurol. 2019 May;18(5):459-480. doi: 10.1016/S1474-4422(18)30499-X. Epub 2019 Mar 14. PMID: 30879893; PMCID: PMC6459001.
2. World Health Organization. (2006). Neurological disorders : public health challenges. World Health Organization.
3. Olesen J, Gustavsson A, Svensson M, Wittchen HU, Jönsson B; CDBE2010 study group; European Brain Council. The economic cost of brain disorders in Europe. Eur J Neurol. 2012 Jan;19(1):155-62. doi: 10.1111/j.1468-1331.2011.03590.x. PMID: 22175760.

Why exposome-based?

<< Unlike genetic factors which are stable and unmodifiable, the exposome has large spatiotemporal variability and can be modified at different levels. >>



Milena Gandy, Andreea I. Heriseanu, Joanne Dudeney, Madelyne A. Bisby, Amelia J. Scott, Alana Fisher, Taylor Hathway, Eyal Karin, Nick Titov, Blake Dear, Disability and life satisfaction in neurological disorders: The role of depression and perceived cognitive difficulties,General Hospital Psychiatry, 2021, ISSN 0163-8343,https://doi.org/10.1016/j.genhosppsych.2021.08.01

Previous studies

Papers about: UK Biobank + Exposome + Dementia

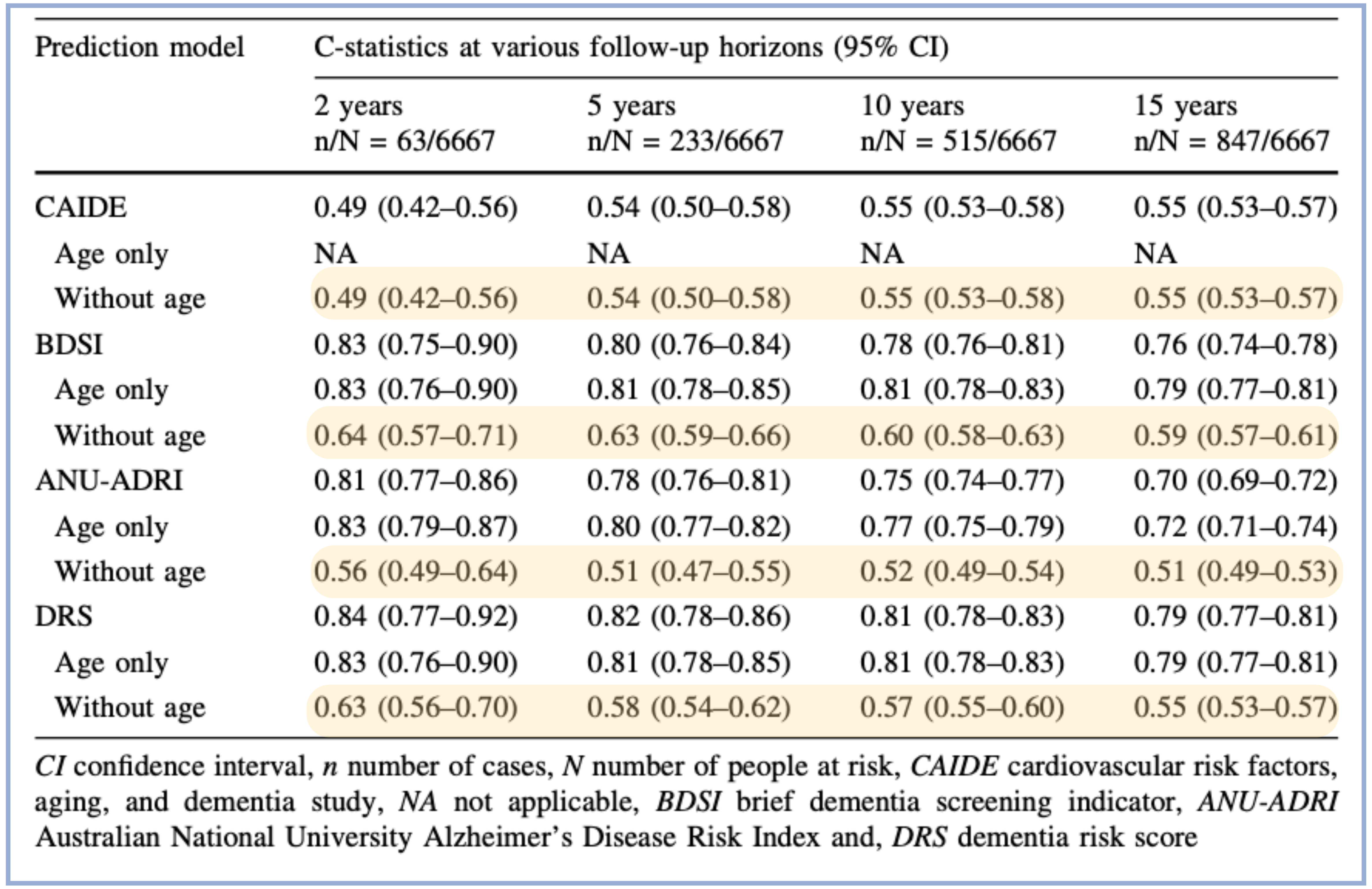
- Sleep, major depressive disorder, and Alzheimer disease
- Meat consumption and risk of incident dementia: cohort study of 493,888 UK Biobank participants
- Is neuroticism differentially associated with risk of Alzheimer's disease, vascular dementia, and frontotemporal
- Diet and Dementia: A Prospective Study
- Identifying dementia outcomes in UK Biobank: a validation study of primary care, hospital admissions and mortality data
- Associations between vascular risk factors and brain MRI indices in UK Biobank
- High coffee consumption, brain volume and risk of dementia and stroke
- Sex differences in the association between major cardiovascular risk factors in midlife and dementia: a cohort study using data from the UK Biobank

1. *Development and validation of a predictive algorithm for risk of dementia in the community setting (CANADA)*
2. *Machine learning prediction of incidence of Alzheimer’s disease using large-scale administrative health data (KOREA)*

Previous study

External validation of four dementia prediction models for use in the general community-dwelling population: a comparative analysis from the Rotterdam Study

- All models showed similar discriminative ability when compared to prediction based on age alone. These findings highlight the urgent need for updated or new models to predict dementia risk in the general population.



Prediction model	C-statistics at various follow-up horizons (95% CI)			
	2 years n/N = 63/6667	5 years n/N = 233/6667	10 years n/N = 515/6667	15 years n/N = 847/6667
CAIDE	0.49 (0.42–0.56)	0.54 (0.50–0.58)	0.55 (0.53–0.58)	0.55 (0.53–0.57)
Age only	NA	NA	NA	NA
Without age	0.49 (0.42–0.56)	0.54 (0.50–0.58)	0.55 (0.53–0.58)	0.55 (0.53–0.57)
BDSI	0.83 (0.75–0.90)	0.80 (0.76–0.84)	0.78 (0.76–0.81)	0.76 (0.74–0.78)
Age only	0.83 (0.76–0.90)	0.81 (0.78–0.85)	0.81 (0.78–0.83)	0.79 (0.77–0.81)
Without age	0.64 (0.57–0.71)	0.63 (0.59–0.66)	0.60 (0.58–0.63)	0.59 (0.57–0.61)
ANU-ADRI	0.81 (0.77–0.86)	0.78 (0.76–0.81)	0.75 (0.74–0.77)	0.70 (0.69–0.72)
Age only	0.83 (0.79–0.87)	0.80 (0.77–0.82)	0.77 (0.75–0.79)	0.72 (0.71–0.74)
Without age	0.56 (0.49–0.64)	0.51 (0.47–0.55)	0.52 (0.49–0.54)	0.51 (0.49–0.53)
DRS	0.84 (0.77–0.92)	0.82 (0.78–0.86)	0.81 (0.78–0.83)	0.79 (0.77–0.81)
Age only	0.83 (0.76–0.90)	0.81 (0.78–0.85)	0.81 (0.78–0.83)	0.79 (0.77–0.81)
Without age	0.63 (0.56–0.70)	0.58 (0.54–0.62)	0.57 (0.55–0.60)	0.55 (0.53–0.57)

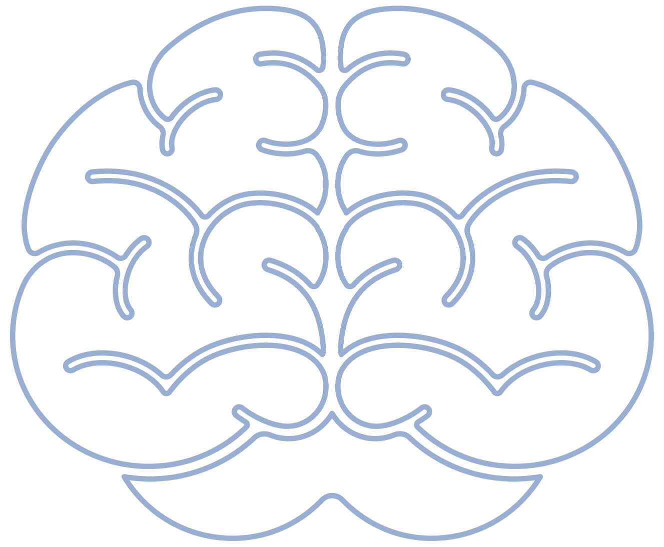
CI confidence interval, *n* number of cases, *N* number of people at risk, *CAIDE* cardiovascular risk factors, aging, and dementia study, *NA* not applicable, *BDSI* brief dementia screening indicator, *ANU-ADRI* Australian National University Alzheimer's Disease Risk Index and, *DRS* dementia risk score

Our study

Early diagnosis of Dementia prediction using accessible exposome-based features: A comparison of Statistical and Machine Learning models on the UK Biobank

Objectives:

1. Demonstrate that accessible exposome variables are powerful enough to predict Dementia.
2. Demonstrate that ML outperforms SL in our classification problem.
3. Demonstrate that our models work well in external validation.
4. Demonstrate that we are not predicting just age.



Dementia ICD-10

1. Alzheimer (AD)
2. Vascular Dementia (VD)
3. Frontotemporal dementia (FTD)
4. Other causes of dementia (OD)

Rejected cases with a Dementia diagnosis before 2011, hence before or during Baseline assessment.

↓

1,523 individuals

Overview

Sample to Build the models and Perform Internal Validation 90%

2.740

Dementia Group after 2011 **1370**

Healthy Control Group **1370**

Assessment Center Locations:

19/22 (All except: Edinburgh, Oxford, Barts)

Mean Birth: 1944.025

Treatment —> **1943.918**

Control —> **1944.132**

For Fairness Purposes:

Treatment —> Female = **623** Male = **747**

Control —> Female = **747** Male = **623**

Sample to perform External Validation 10%

306

Dementia Group after 2011 **153**

Healthy Control Group **153**

Assessment Center Locations:

Edinburgh = **108**

Oxford = **107**

Barts = **91**

Mean Birth: 1943.973

Treatment —> **1943.300**

Control —> **1944.647**

For Fairness Purposes:

Treatment —> Female = **66** Male = **87**

Control —> Female = **87** Male = **66**

Exposome features

1. Physical measurements —> Weight
2. Sociodemographics —> Qualifications
3. Lifestyle —> Sleeplessness
4. Environmental factors —> Average evening noise
5. Early life factors —> Adopted as a child
6. Traumatic events —> Victim of sexual assault
7. Mental health —> Happiness

Exposome

UNIVERSITAT DE BARCELONA

Methods

BCN-AIM ARTIFICIAL INTELLIGENCE IN MEDICINE

1. Exposome data

502,664 patients from UK Biobank and 128 exposome features

2. Data pre-processing

Selection of Dementia patients, and 1,370 Healthy patients

80% missing value threshold for features leading to 78 exposome features

MissForest for Imputation

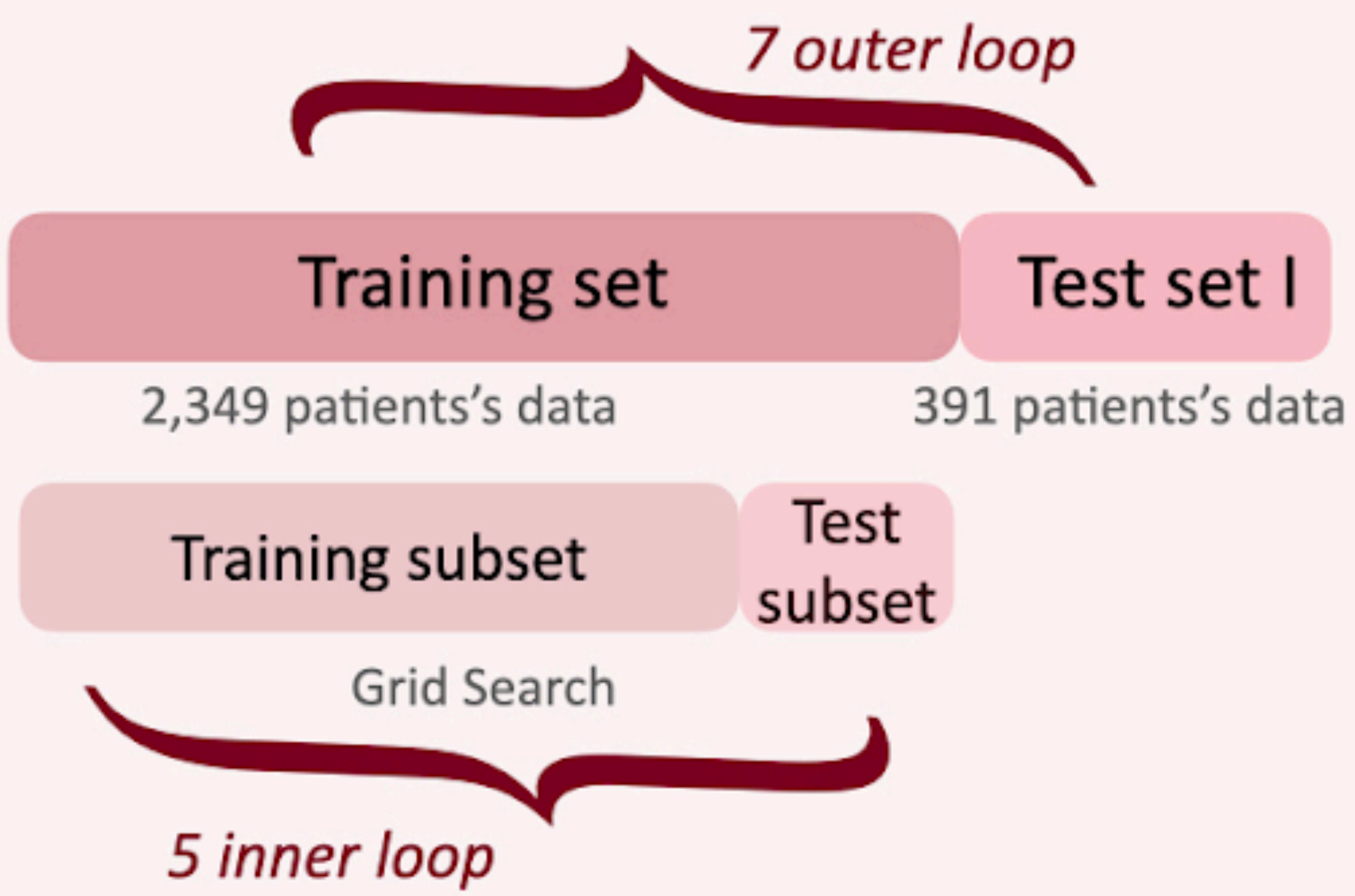
3. Predictive disease modeling

Logistic Regression (SL): linear based model
XGBoost (ML): decision-tree-based ensemble model

3.1. Internal Validation

2,740 patients's data obtained from 19 different assessment centers

Nested cross-validation



7 outer loop

Training set

Test set I

2,349 patients's data

391 patients's data

Training subset

Test subset

Grid Search

5 inner loop

3.2. External Validation

306 patients's data obtained from 3 different assessment centers: Oxford, Edinburgh and Barts.

Test set II

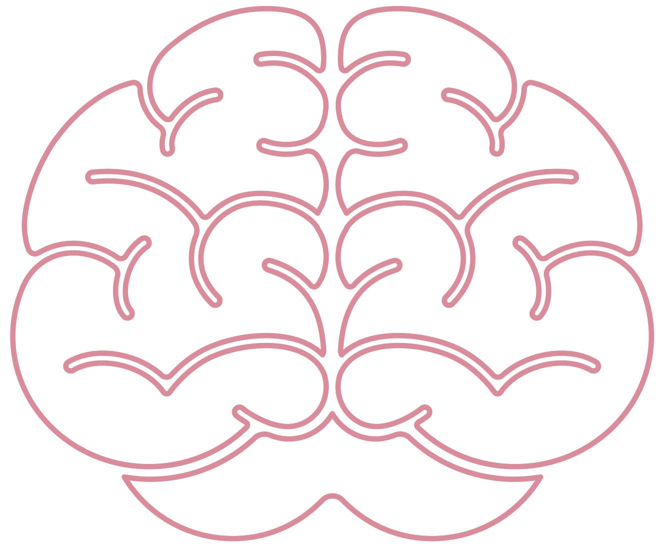
4. Model Interpretability

SL —> Coefficients
ML —> Gini Importance

Flowchart for early prediction of Dementia using exposures data in UK Biobank

Experiments

- **Experiment 1:** Exposome with Age
- **Experiment 2:** Exposome without Age
- **Experiment 3:** Age



- **Experiment 4:** Accessible with Age (30)
- **Experiment 5:** Accessible without Age (30)

Models Performance Analysis

AUC can be interpreted as:

1) No discrimination [AUC = 0.5]
2) Poor discrimination [0.6≥AUC>0.5]
3) Acceptable discrimination [0.7≥AUC>0.6]
4) Excellent discrimination [0.8≥AUC>0.7]
5) Outstanding discrimination [AUC>0.9]

Our Goal:

Obtain an accessible model able to perform an Excellent or Outstanding discrimination.

Statistical Learning Results

Logistic Regression: Linear Model

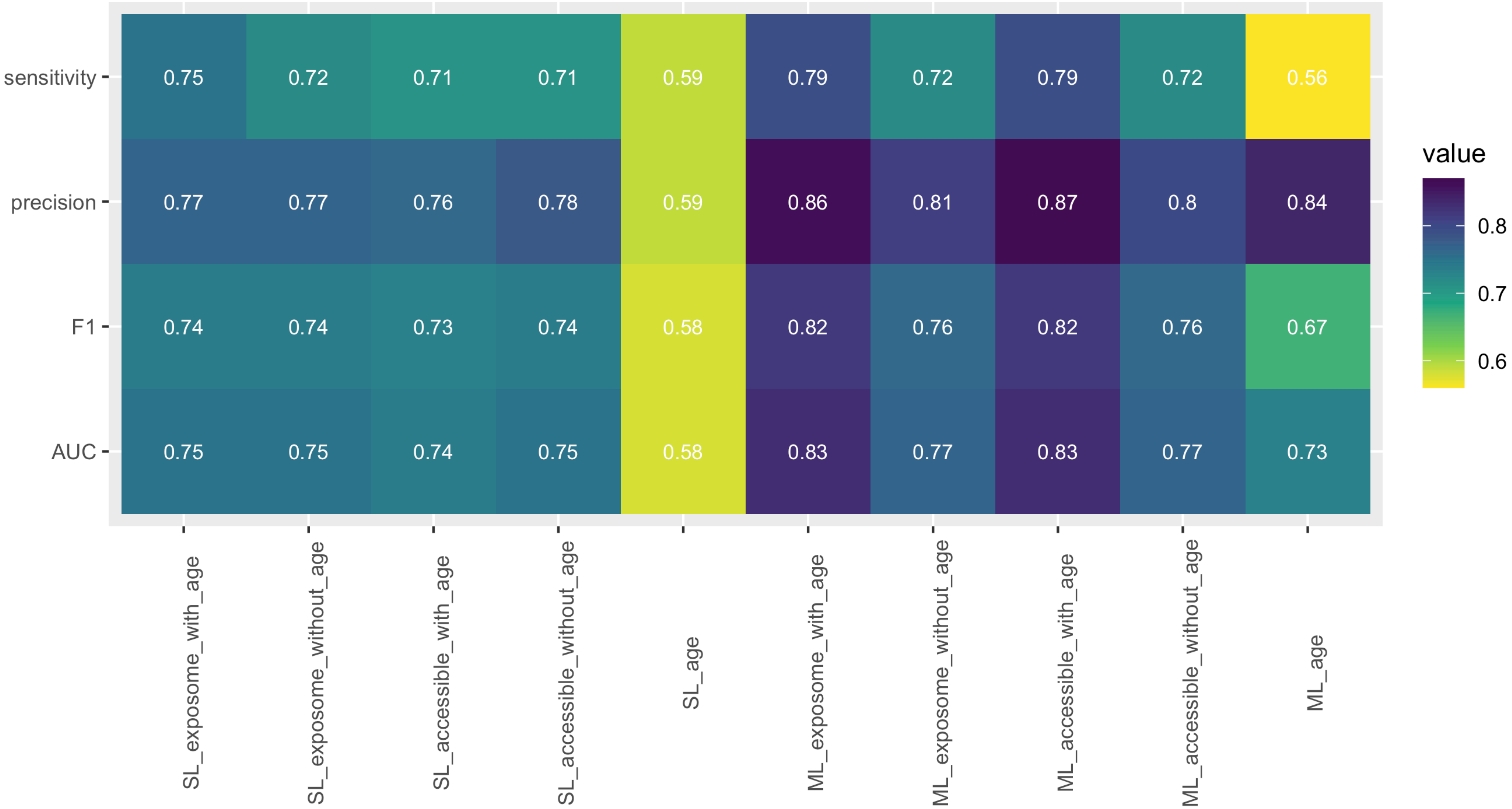
	Exposome with Age	**Exposome without Age**	Age	**Accessible with Age**	**Accessible without Age**
AUC	0.75±0.02|0.77±0.01	0.75±0.01|0.76±0.01	0.58 ± 0.02|0.68±0.03	0.74±0.02|0.75±0.01	0.75±0.02|0.77±0.01
F1	0.74±0.02|0.73±0.01	0.74±0.01|0.73±0.01	0.58 ± 0.03|0.68±0.01	0.73±0.01|0.72±0.01	0.74±0.01|0.75±0.01
precision	0.77±0.02|0.86±0.01	0.77±0.02|0.85±0.01	0.59 ± 0.03|0.68±0.04	0.76±0.03|0.84±0.01	0.78±0.03|0.84±0.01
sensitivity	0.75±0.02|0.64±0.01	0.72±0.02|0.64±0.01	0.59 ± 0.07|0.69±0.02	0.71±0.02|0.63±0.01	0.71±0.02|0.67±0.02

Machine Learning Results

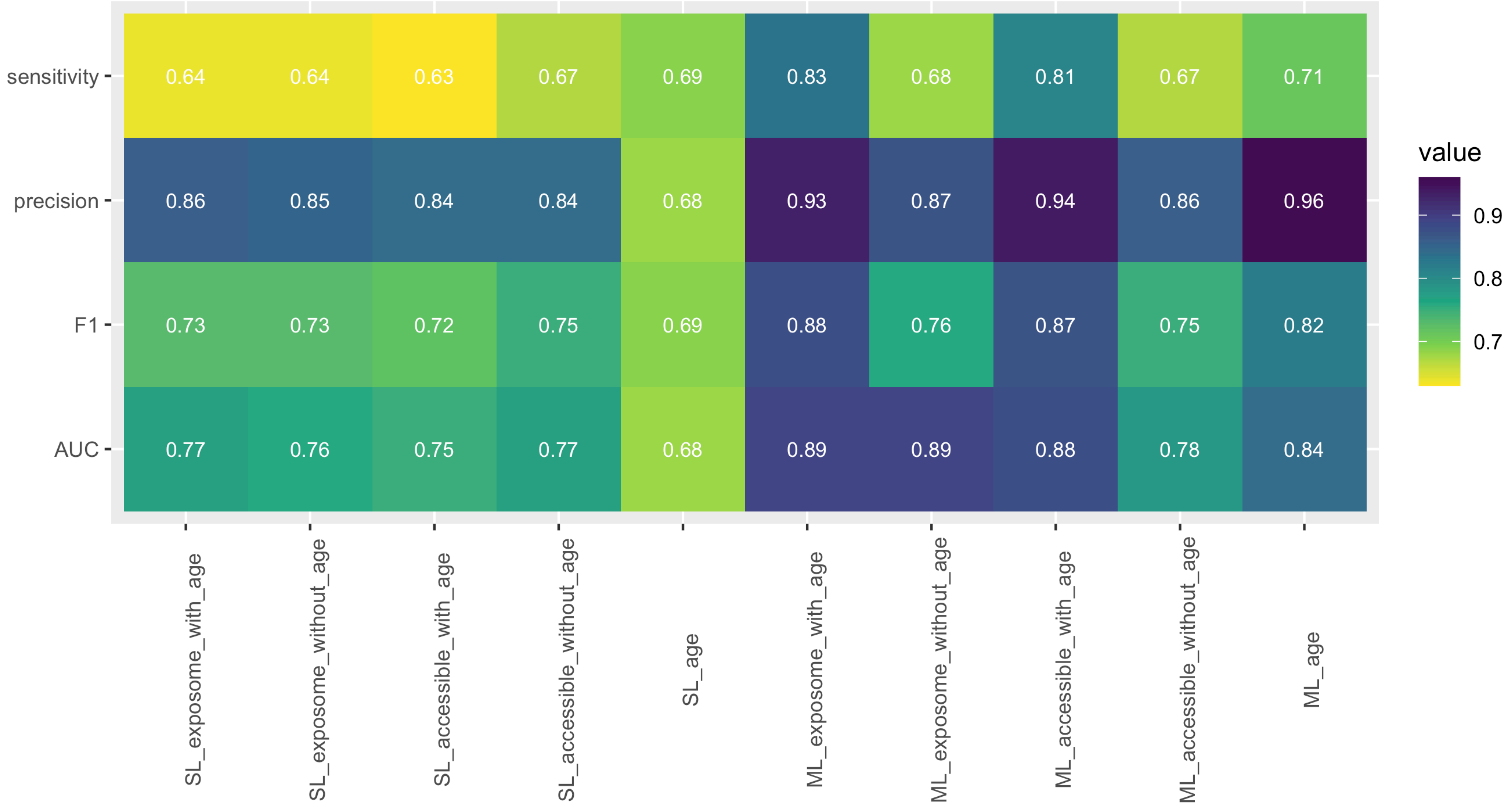
XGBoost: Non-Linear Model

	Exposome with Age	Exposome without Age	Age	Accessible with Age	Accessible without Age
AUC	0.83±0.02|0.89±0.01	0.77±0.01|0.79±0.01	0.73±0.02|0.84±0.00	0.83±0.02|0.88±0.01	0.77±0.01|0.78±0.02
F1	0.82±0.01|0.88±0.01	0.76±0.01|0.76±0.01	0.67±0.02|0.82±0.00	0.82±0.02|0.87±0.01	0.76±0.01|0.75±0.02
precision	0.86±0.02|0.93±0.01	0.81±0.03|0.87±0.02	0.84±0.01|0.96±0.00	0.87±0.04|0.94±0.01	0.80±0.02|0.86±0.03
sensitivity	0.79±0.01|0.83±0.01	0.72±0.02|0.68±0.03	0.56±0.03|0.71±0.00	0.79±0.02|0.81±0.02	0.72±0.02|0.67±0.02

Internal Validation



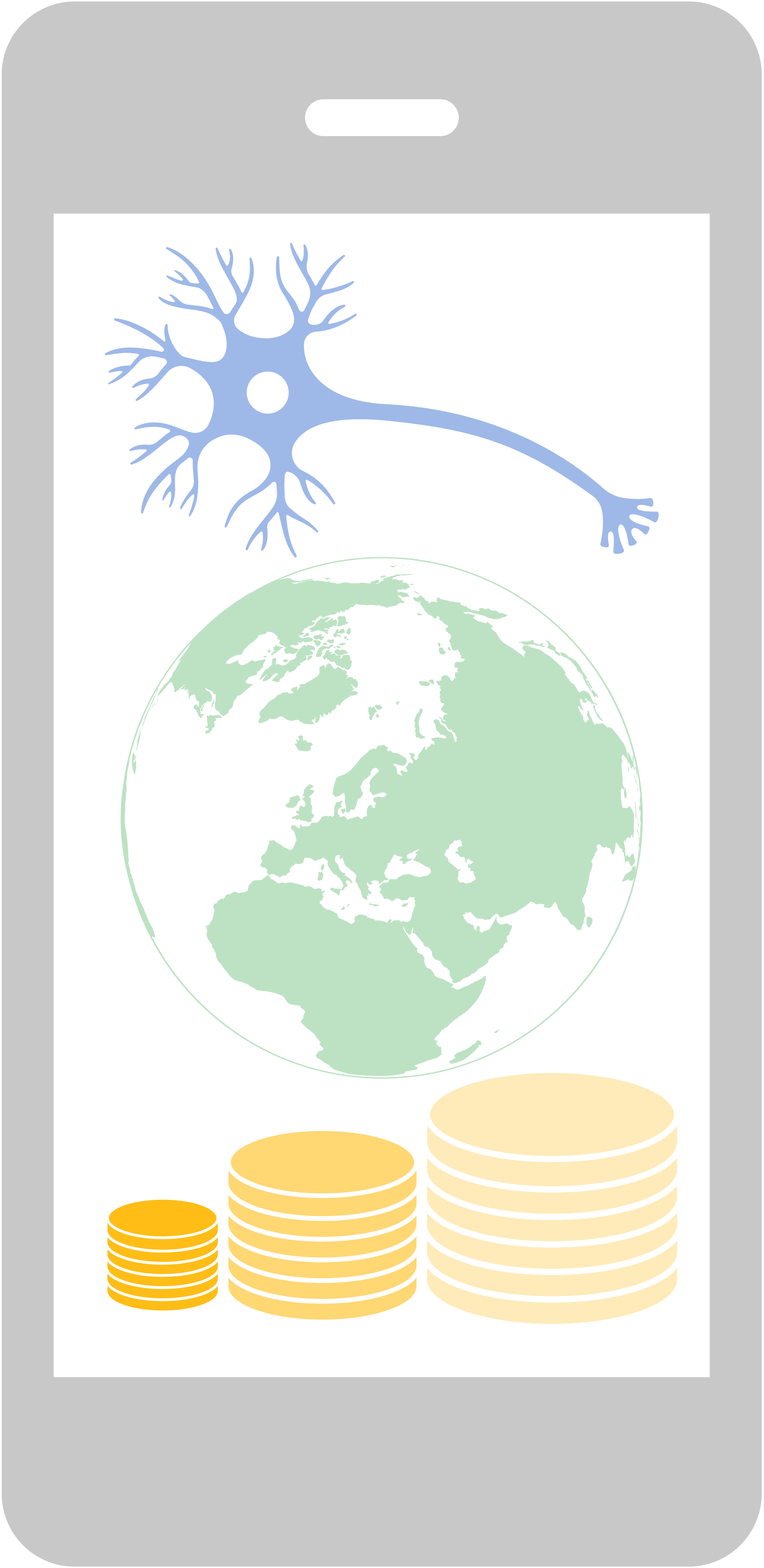
External Validation



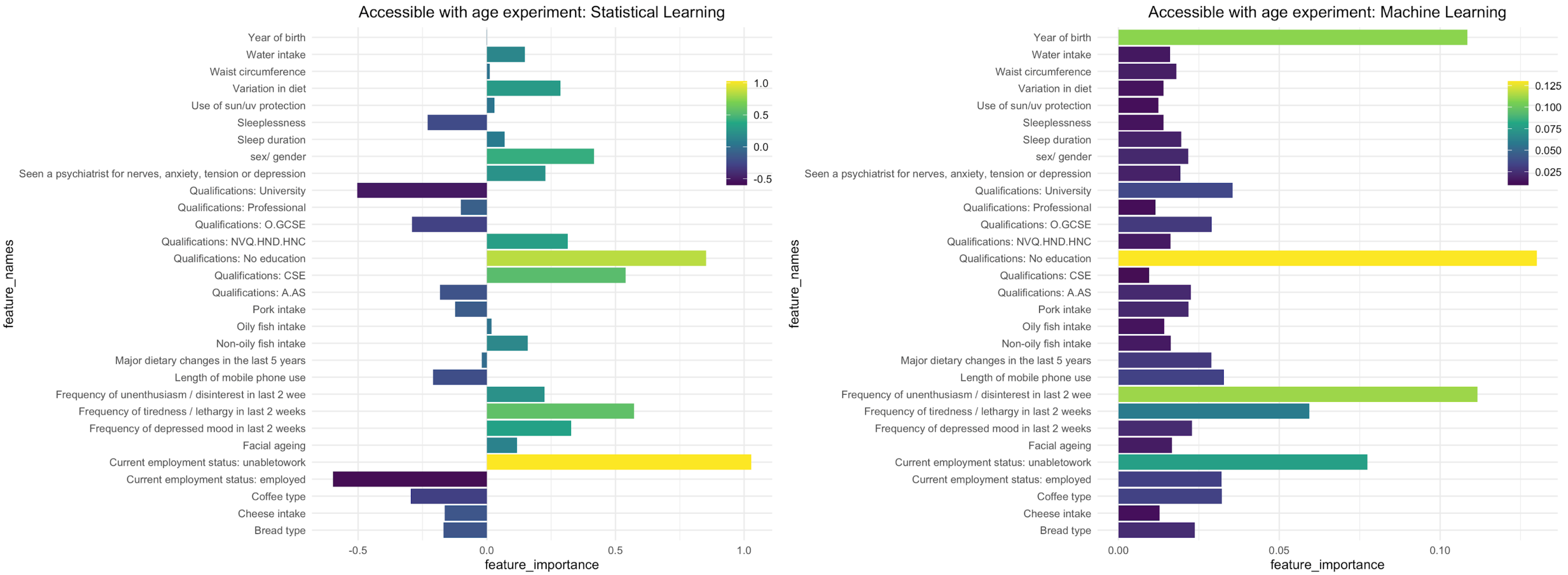
Accesible Models

Create new models with just 30 accessible variables.

noeducation, frequnenthusiasm2weeks, freqtiredness2weeks, unabletowork, university, freqdepressed, coffetype, lengthmobileuse, employed, sex, dietarychange, nonoilyfish, A.AS, NVQ.HND.HNC, professional, oilyfishintake, breadtype, waistcircum, CSE, variationdiet, waterintake, porkintake, partmultiplebirth, sleepduration, sleeplessness, O.GCSE, cheeseintake, facialageing, usesunprotection, standingheight



Accessible with Age



30 accessible variables with age best models and their importance in experiment 4. At right results displayed for Statistical Learning using Logistic Regression, and at left results for machine learning using XGBoost.

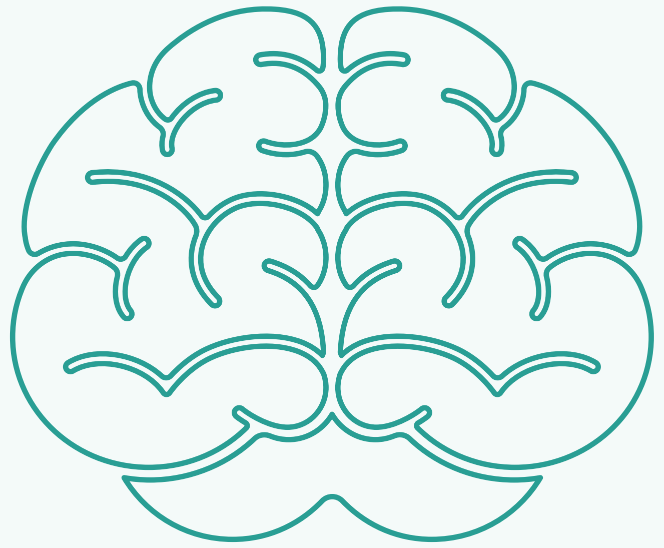
Accessible without Age

Accessible without age experiment: Statistical Learning
Water intake
Waist circumference
Variation in diet
Use of sun/uv protection
Standing height
Sleeplessness
Sleep duration
sex/ gender
Seen a psychiatrist for nerves, anxiety, tension or depression
Qualifications: University
Qualifications: Professional
Qualifications: O.GCSE
Qualifications: NVQ.HND.HNC
Qualifications: No education
Qualifications: CSE
Qualifications: A.AS
Pork intake
Oily fish intake
Non-oily fish intake
Major dietary changes in the last 5 years
Length of mobile phone use
Frequency of unenthusiasm / disinterest in last 2 wee
Frequency of tiredness / lethargy in last 2 weeks
Frequency of depressed mood in last 2 weeks
Facial ageing
Current employment status: unabletowork
Current employment status: employed
Coffee type
Cheese intake
Bread type
feature_names
0
2
4
6
feature_importance
6
4
2
0
Accessible without age experiment: Machine Learning
Water intake
Waist circumference
Variation in diet
Use of sun/uv protection
Standing height
Sleeplessness
Sleep duration
sex/ gender
Seen a psychiatrist for nerves, anxiety, tension or depression
Qualifications: University
Qualifications: Professional
Qualifications: O.GCSE
Qualifications: NVQ.HND.HNC
Qualifications: No education
Qualifications: CSE
Qualifications: A.AS
Pork intake
Oily fish intake
Non-oily fish intake
Major dietary changes in the last 5 years
Length of mobile phone use
Frequency of unenthusiasm / disinterest in last 2 wee
Frequency of tiredness / lethargy in last 2 weeks
Frequency of depressed mood in last 2 weeks
Facial ageing
Current employment status: unabletowork
Current employment status: employed
Coffee type
Cheese intake
Bread type
feature_names
0.00
0.05
0.10
0.15
feature_importance
0.16
0.12
0.08
0.04

30 accessible variables without age best models and their importance in experiment 5. At right results displayed for Statistical Learning using Logistic Regression, and at left results for machine learning using XGBoost.

Conclusions

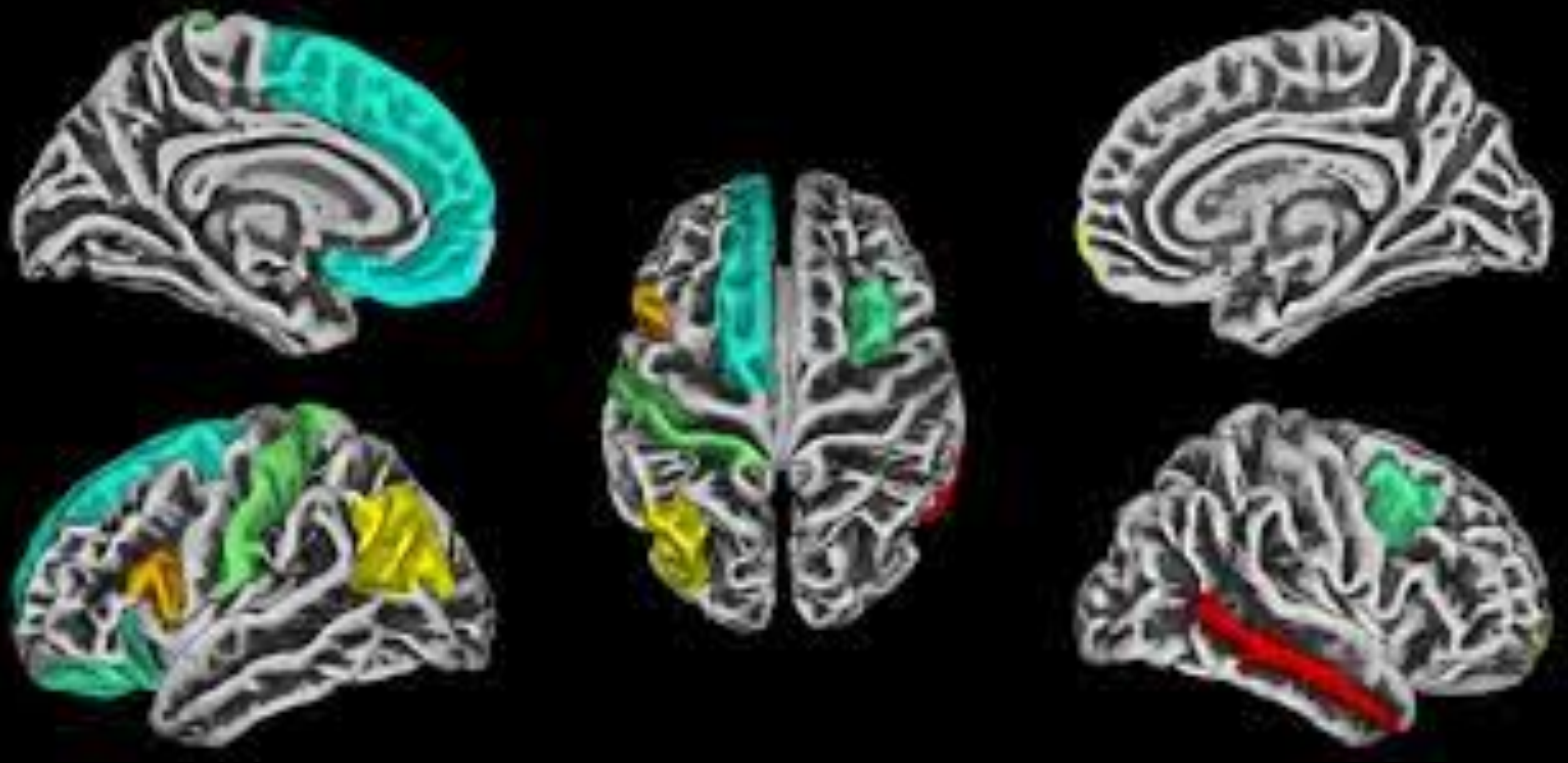
1. If we move from SL to ML for dementia prediction we we gain accuracy.
2. Exposome data is good enough for an accurate prediction of the syndrome.
3. Our models are predicting not just age.
4. We got a high accuracy in external validation.
5. Most relevant exposures found in this study could be used in the future to select risk patients for drug studies.



Possible Future Research

Mapping the exposome effect on the brain

An Artificial Intelligence approach using MRI images from the UK Biobank cohort



↓

● DIET

□ Coffe intake

🟧 Tea intake

…

○ BEHAVIOUR

□ Length of mobile phone use

🟩 Nap during the day

…

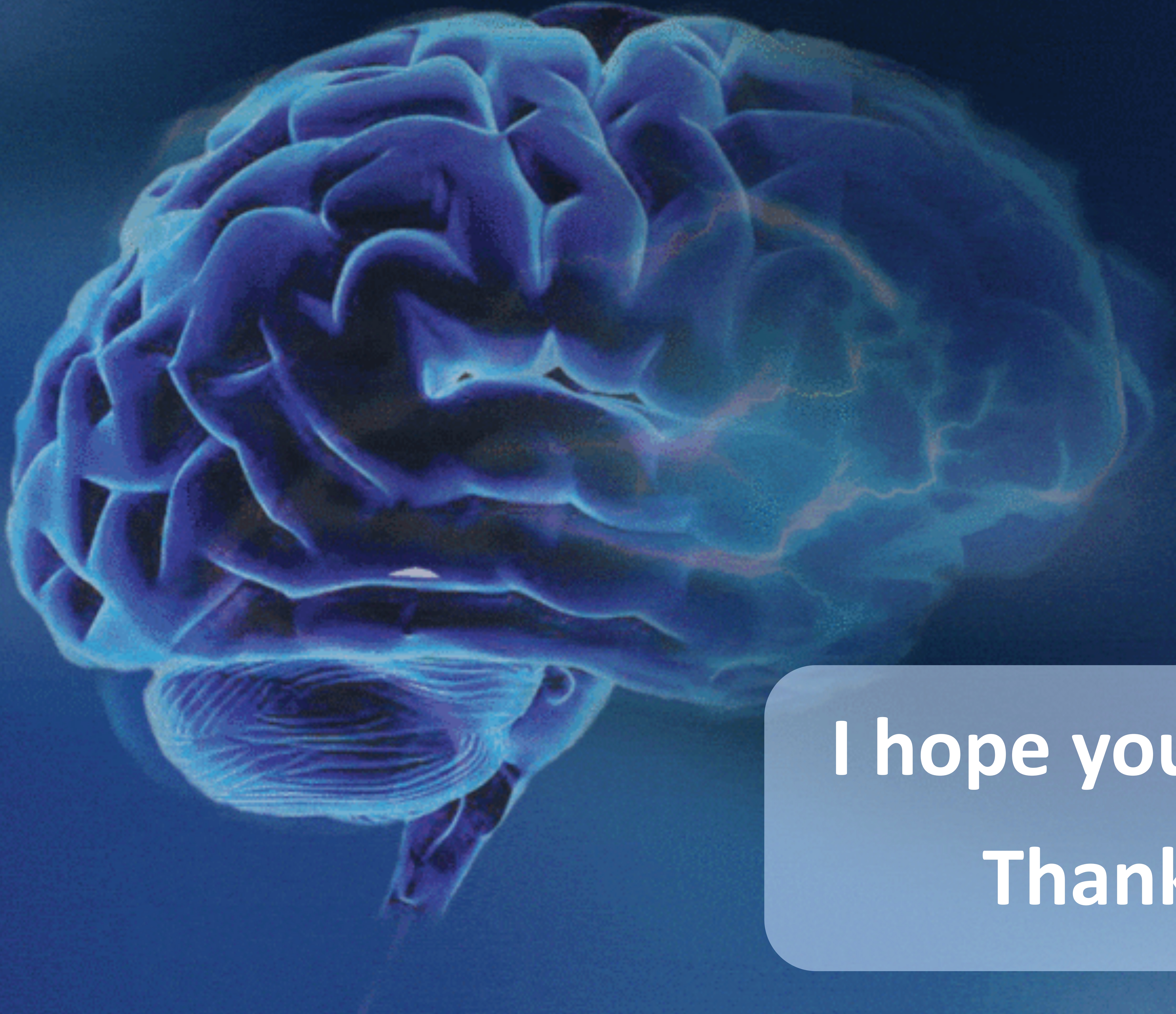
● EARLY LIFE FACTORS

🟨 Maternal smoking around birth

…

○ ENVIROMENTAL FACTORS

…..



I hope you find it interesting.

Thanks for listening!