Principal Investigator

First Name: Martha Ivon
Last Name: Cardenas
Degree: Doctor in Artificial Intelligence
Primary Affiliation: Universitat Politècnica de Catalunya
E-mail: martha.ivon.cardenas@upc.edu
State or Province: Barcelona
Country: Spain

General Information

Key Personnel (other than PI):
First Name: Caroline
Last name: König
Degree: Doctor in Artificial Intelligence
Primary Affiliation: Universitat Politècnica de Catalunya
SCOPUS ID:
Requires Data Access? Yes

Are external grants or funds being used to support this research?: No external grants or funds are being used to support this research.
How did you learn about the YODA Project?: PubMed

Conflict of Interest

https://yoda.yale.edu/wp-content/uploads/2025/02/SV_57KskaKADT3U9Aq-R_8bGfAjU4Coamjwl.pdf
https://yoda.yale.edu/wp-content/uploads/2025/02/COI-FORM-CK.pdf

Certification

Certification: All information is complete; I (PI) am responsible for the research; data will not be used to support litigious/commercial aims.
Data Use Agreement Training: As the Principal Investigator of this study, I certify that I have completed the YODA Project Data Use Agreement Training

1. NCT00574132 - ELN115727-301 - A Phase 3, Multicenter, Randomized, Double-Blind, Placebo-Controlled, Parallel-Group, Efficacy and Safety Trial of Bapineuzumab (AAB-001, ELN115727) In Patients With Mild to Moderate Alzheimer's Disease Who Are Apolipoprotein E4 Non-Carriers

What type of data are you looking for?: Individual Participant-Level Data, which includes Full CSR and all supporting documentation

Research Proposal

Project Title

Machine learning applications for improved interpretability of Alzheimer's clinical trial data in efficient medical decision-making

Narrative Summary:

Alzheimer's disease is a progressive neurodegenerative disorder that affects millions worldwide, yet current clinical trials face significant challenges in identifying the right participants, predicting treatment responses, and interpreting complex medical data. This study aims to improve the efficiency of Alzheimer's clinical trials by applying ML to enhance data interpretability, uncover hidden patient subgroups, and visualize key patterns that influence medical decision-making. Using unsupervised clustering, we will analyze clinical trial data to identify groups of patients with similar characteristics, potentially leading to more personalized treatment strategies.

Scientific Abstract:

Background:Clinical trials in Alzheimer's disease often suffer from inefficiencies in patient selection, response prediction, and data interpretation. Traditional statistical approaches struggle to extract meaningful patterns from large, heterogeneous datasets, necessitating advanced machine learning methods.

Objective:This study aims to improve the interpretability and efficiency of Alzheimer's clinical trials by leveraging unsupervised clustering, interpretable machine learning models, and data visualization techniques.

Study Design:Retrospective analysis of participant-level data from completed Alzheimer's clinical trials.

Participants:Patients diagnosed with Alzheimer's disease from previously conducted trials, meeting inclusion criteria based on demographic, genetic, and biomarker profiles.

Primary and Secondary Outcome Measures:

Primary: Identification of patient subgroups using unsupervised clustering.
Secondary: Improvement in interpretability of treatment responses and biomarkers through machine learning models.

Statistical Analysis:

Unsupervised clustering algorithms (e.g., k-means, hierarchical clustering, t-SNE) for subgroup identification.
Interpretable models (e.g., SHAP values, feature importance analysis) to explain key biomarkers influencing treatment outcomes.
Visualization techniques (e.g., heatmaps, dimensionality reduction plots) to communicate findings effectively.

Brief Project Background and Statement of Project Significance:

Alzheimer's disease clinical trials often struggle with high failure rates due to heterogeneous patient responses, making it challenging to develop effective treatments. Machine learning provides a data-driven approach to stratify patients, uncover patterns, and enhance interpretability, ultimately improving trial efficiency. By incorporating clustering and visualization, this project will make clinical trial data more actionable and transparent for researchers and clinicians.

Specific Aims of the Project:

Specific Aims:

Develop an unsupervised machine learning framework to identify patient subgroups based on biomarker and clinical similarities.
Apply interpretable AI models to explain key factors influencing treatment responses.
Utilize advanced visualization techniques to improve accessibility and comprehension of clinical trial data.

Study Design:

Methodological research

What is the purpose of the analysis being proposed? Please select all that apply.

New research question to examine treatment effectiveness on secondary endpoints and/or within subgroup populations

Participant-level data meta-analysis

Meta-analysis using only data from the YODA Project

Develop or refine statistical methods

Research on clinical prediction or risk prediction

Research Methods

Data Source and Inclusion/Exclusion Criteria to be used to define the patient sample for your study:

Research Methods:

Data Source: Participant-level clinical trial datasets from the YODA Project.
Inclusion Criteria: Alzheimer's patients with available biomarker, cognitive assessment, and treatment response data.
Exclusion Criteria: Patients with incomplete datasets or confounding conditions.

Statistical Analysis Plan:

Unsupervised learning algorithms (e.g., k-means, hierarchical clustering, DBSCAN) for subgroup identification.
Model explainability techniques (e.g., SHAP values, LIME) to interpret feature importance.
Data visualization (e.g., PCA, t-SNE, UMAP) for dimensionality reduction and insight generation.

If possible, we are also interested in age, APOE genotype, baseline cognitive scores, biomarker levels, and treatment regimen.

Primary and Secondary Outcome Measure(s) and how they will be categorized/defined for your study:

Primary: Identification of distinct patient clusters.

Secondary: Improved interpretability of biomarkers and treatment response patterns.

Main Predictor/Independent Variable and how it will be categorized/defined for your study:

Main Predictor/Independent Variable:The main independent variable in this study is the patient subgroup classification derived from machine learning clustering techniques. This variable

represents distinct groups of patients identified based on biomarker levels, cognitive assessment scores, and demographic factors.

Categorization & Definition:

Patients will be assigned to subgroups using unsupervised clustering algorithms (e.g., k-means, hierarchical clustering, DBSCAN).
These clusters will be analyzed as categorical variables to assess their impact on primary and secondary outcome measures.
The subgroup classification will be compared across different clustering models to ensure robustness and reliability.

Other Variables of Interest that will be used in your analysis and how they will be categorized/defined for your study:

Demographic Variables:

Age: Continuous variable (in years).
Sex: Categorical variable (Male/Female/Other).
Education Level: Ordinal variable (No formal education, Primary, Secondary, Higher).

Clinical & Cognitive Measures:

Baseline Cognitive Scores (MMSE, ADAS-Cog): Continuous variables (score ranges specific to each test).
Disease Severity: Ordinal variable (Mild, Moderate, Severe based on clinical diagnosis).
Comorbidities: Binary variables (e.g., Diabetes: Yes/No, Hypertension: Yes/No).

Treatment Variables:

Medication Type: Categorical variable (e.g., Donepezil, Rivastigmine, Placebo).
Treatment Duration: Continuous variable (measured in weeks).

Statistical Analysis Plan:

Advanced analyses:

Unsupervised Clustering (e.g., K-means, Hierarchical, DBSCAN): To identify patient subgroups based on biomarker and clinical profiles.
Kaplan-Meier Survival Analysis: To assess time-to-event outcomes, such as disease progression.
Cox Proportional Hazards Model: To evaluate the impact of independent variables on survival outcomes.
Propensity Score Matching: To adjust for potential confounders when comparing treatment groups.
Non-parametric Testing (e.g., Wilcoxon signed-rank test): When data do not meet normality assumptions.

Machine Learning-Based Interpretation:

SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-Agnostic Explanations): To enhance interpretability of machine learning models.
Dimensionality Reduction (PCA, t-SNE, UMAP): For visualization and feature extraction.

Software Used:

Python

Project Timeline:

Project Timeline:The proposed study will follow a structured timeline to ensure timely completion of all milestones within the allocated 12-month period. Below is the estimated timeline:

Month 1-2: Data access approval and preprocessing, including cleaning and structuring participant-level data.

Month 3-4: Exploratory data analysis and descriptive statistics to understand baseline characteristics.

Month 5-6: Implementation of unsupervised clustering methods and evaluation of subgroup classifications.

Month 7-8: Development and validation of interpretable machine learning models to assess treatment responses.

Month 9: Data visualization and interpretation of findings to ensure clarity in communicating results.

Month 10: Drafting of the manuscript, including results, discussion, and conclusions.

Month 11: Submission of the manuscript to a peer-reviewed journal for publication.

Month 12: Final report submission to the YODA Project, including key findings and potential applications.

Dissemination Plan:

JOURNALS
Artificial Intelligence in Medicine
Journal of Biomedical Informatics
Nature Machine Intelligence

Bibliography:

Talpalaru, A., Bhagwat, N., Devenyi, G. A., Lepage, M., & Chakravarty, M. M. (2019). Identifying schizophrenia subgroups using clustering and supervised learning. *Schizophrenia Research, 214*, 51--59. https://doi.org/10.1016/j.schres.2019.05.044

Mellem, I. M., Finnes, T. E., Heldal, K., & Schopf, T. R. (2021). An evaluation of machine learning for predicting hospital admissions based on primary care electronic health records: A systematic review. *BMC Medical Informatics and Decision Making, 21*(1), 162. https://doi.org/10.1186/s12911-021-01510-0

Dor, H., & Hertzberg, L. (2024). Schizophrenia biomarkers: Blood transcriptome suggests two molecular subtypes. *NeuroMolecular Medicine, 26*, 50. https://doi.org/10.1007/s12017-024-08817-x