Amsterdam UMC
University Medical Centers

Department of Psychiatry Amsterdam

Using Proteomics and Lipidomics Data to Improve Individual Prediction of Chronicity in Depression

Philippe C. Habets[1,2,3]*, Rajat M Thomas[1], Guido A van Wingen[1], Brenda WJH Penninx[2], Onno C Meijer[3], Christiaan H. Vinkers[1,2]

1. Department of Anatomy & Neurosciences, Amsterdam UMC, Amsterdam, The Netherlands 2. Department of Psychiatry, Amsterdam UMC/GGZ inGeest, Amsterdam, The Netherlands. 3. Department of Internal Medicine, section Endocrinology, Leiden University Medical Center, Leiden, The Netherlands

Introduction

Depressive disorder is a heterogenous condition that differs widely in both therapy response and prognosis. The ability to individually predict disease course early on is essential for optimal treatment planning. Here, we use a data-driven machine learning approach to test the potential of combining two sets of -omics data in addition to easy-to-acquire clinical baseline variables for the prediction of two-year chronicity in major depressive disorder.

Objective

The aim was to test whether predictions for two-year chronicity in major depressive disorder could be improved from a previous predictive model (1) by adding proteomics and/or lipidomics data to clinical baseline data, using non-linear predictive models.

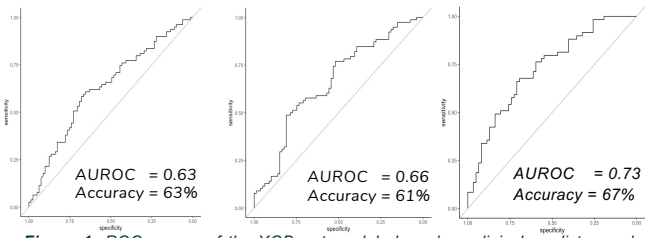

Figure 1: *ROC curves of the XGBoost models based on clinical predictors only (left), clinical + lipidomic predictors (middle), and clinical + proteomic predictors (right).*

Methods

Proteomics (243 analytes) and lipidomics (231 variables) assays were performed on whole blood serum at baseline in patients with baseline depression (n=611 for proteomics, n=790 for metabolomics, n=608 for combined data) from the NESDA cohort (2). Two classes of outcome (two-year chronicity or remission) were balanced (49% vs 51%) and showed no significant differences in possible confounding factors (age, gender, bmi, years of education, antidepressant use at baseline, months of antidepressant use between baseline and follow-up). XGboost implementation in R was used with an 80-20 train-test split. An inner 10 times repeated 10-fold cross validation loop was used to optimize hyperparameters, mitigate overfitting issues and select for the best generalizable model (1000 hyperparameter-grid combinations were used) in the train set. Final models using a) clinical, psychological, and clinical/demographical data only (10 variables), b) added proteomics data and c) added lipidomics data were tested on an outheld test set and evaluated by their respective AUROC. Variable importance analysis was performed using SHAP (3).

Model	Data	AUROC (accuracy)
elastic net	IDS/bigFive/demographics	0.65 (0.64)
elastic net	**Proteomics added**	**0.68 (0.60)**
SVM	IDS/bigFive/demographics	0.63 (0.62)
SVM	**Proteomics added**	**0.69 (0.65)**
Random Forest	IDS/bigFive/demographics	0.60 (0.59)
Random Forest	**Proteomics added**	**0.70 (0.66)**
XGBoost	IDS/bigFive/demographics	0.63 (0.63)
XGBoost	**Proteomics added**	**0.73 (0.67)**

Table 1: *using different algorithms to predict 2-year chronicity consistently shows improved performance when including proteomic data.*

Results

Using non-biological data only, our model was able to predict two-year chronicity (AUROC = 0.63, accuracy = 63%). Addition of proteomics yielded substantially increased predictive performance (AUROC = 0.73, accuracy = 67%). Using lipidomics in combination with the non-biological data did not substantially improve predictive value (AUROC = 0.66, accuracy = 61%). Adding proteomics data to the lipidomics and non-biological data slightly increased model performance, but still showed moderate performance (AUROC = 0.60, accuracy = 58%). In terms of absolute AUROC and balanced accuracy, the proteomics-only trained model showed optimal performance metrics (AUROC= 0.73, balanced accuracy = 70%). The lipidomics-only trained model showed poor performance metrics (AUROC = 0.57, accuracy = 53%), indicating limited lipidomic predictive power. Variable importance analysis indicated that for the model trained on both proteomic and non-biological data, symptom severity was the most important predictor, followed by proteomic analytes involved in coagulation and the immune system. Repeated analysis on proteomics data with a penalized logistic regression model (elastic net) showed lower performance and limited overlap of most-predictive variables, indicating the importance of detection of non-linear patterns in the included biological data.

Conclusion

Proteomics but not lipidomics data were able to augment the performance of clinical variables regarding depression chronicity 2 years later. Even though proteomics-informed predictions showed good performance (AUROC=0.73), improvement in accuracy and feasibility of data-acquirement are needed to warrant clinical implementation.

Amsterdam Neuroscience

References: 1. Dinga R, Marquand AF, Veltman DJ, et al. Predicting the naturalistic course of depression from a wide range of clinical, psychological, and biological data: a machine learning approach. *Transl Psychiatry*. 2018;8(1):241. Published 2018 Nov 5. doi:10.1038/s41398-018-0289-1 **2.** Penninx BWJH, Beekman ATF, Smit JH, et al. The Netherlands Study of Depression and Anxiety (NESDA): rationale, objectives and methods. *Int J Methods Psychiatr Res*. 2008;17(3):121-140. doi:10.1002/mpr.256 **3.** Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67. doi:10.1038/s42256-019-0138-9