Machine learning to predict depression in the UK Biobank using genetic and environmental factors

Bianca C. Branco[1,2], Sam Bentwood[2], Andrew M. McIntosh[2,3], Peggy Seriès[1], Heather C. Whalley[2], Alex S. F. Kwong[2]

1. School of Informatics, University of Edinburgh; 2. Division of Psychiatry, University of Edinburgh; 3. UKRI Mental Health Platform.

1. Background

Prediction and prevention of depression are difficult due to:

I. **Heterogeneity in risk factors** — complex interaction between environment, lifestyle and genetics[1].
II. **Lack of a unified definition of depression** — some measures capture lifetime experiences, while others capture current symptoms.

Machine learning can help with **I**, but the extent to which prediction accuracies and learned relationships between risk factors are affected by **II** is unknown[2].

2. Aims

- Use environmental and genetic factors to predict four different measures of depression in the UK Biobank[3].
- Investigate which factors consistently emerge as important predictors across the different depression outcomes.

3. Dataset

Predictors

- **Demographic** (e.g. sex, age, employment status, deprivation scores).
- **Lifestyle** (e.g. sleep, exercise, smoking, drinking, screen time).
- **Biological** (e.g. polygenic risk scores, self-rated overall health, pain, parental depression).
- **Traumatic life events** (e.g. financial difficulties, assault/injury, separation/divorce).

biobank uk
Enabling scientific discoveries that improve human health

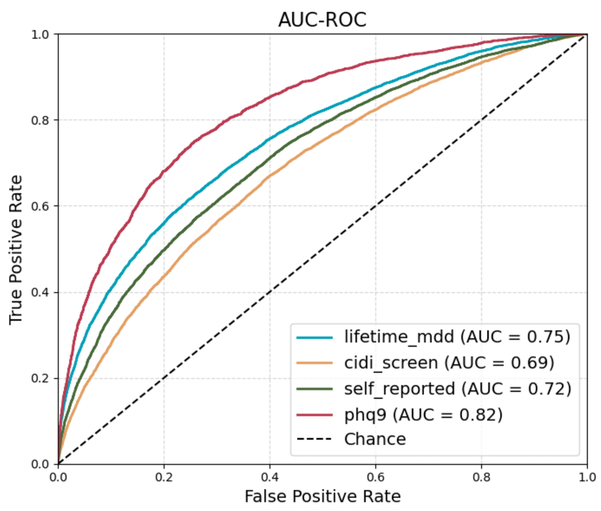
Outcomes

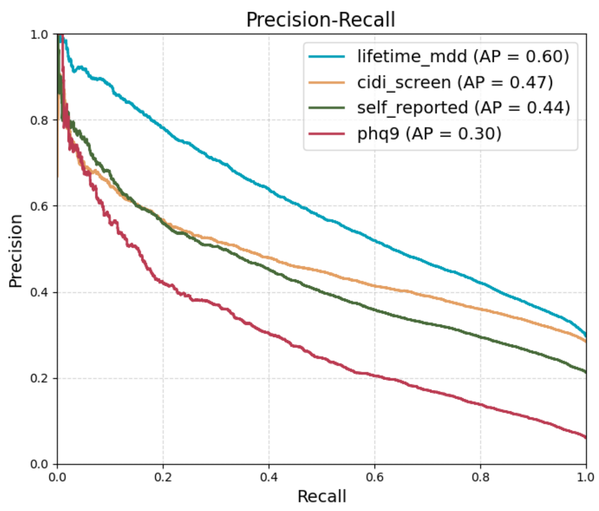
Lifetime MDD
- **N = 126K**
- Captures lifetime experience.
- Combines the CIDI Screen, Self-reported and PHQ-9 measures for more precise definition of positive and negative cases.

CIDI Screen
- **N = 157K**
- Captures lifetime experience.
- Based on CIDI depression module[4].
- Positive case defined by symptoms + their duration and impact on daily life.

Self-reported
- **N = 157K**
- Captures lifetime experience.
- Self-reported measure of whether participant has been diagnosed with depression by a professional.

PHQ-9
- **N = 157K**
- Captures current symptoms.
- Self-reported questionnaire covering the last 2 weeks.
- Positive case defined by a PHQ-9 score of 10 or higher.



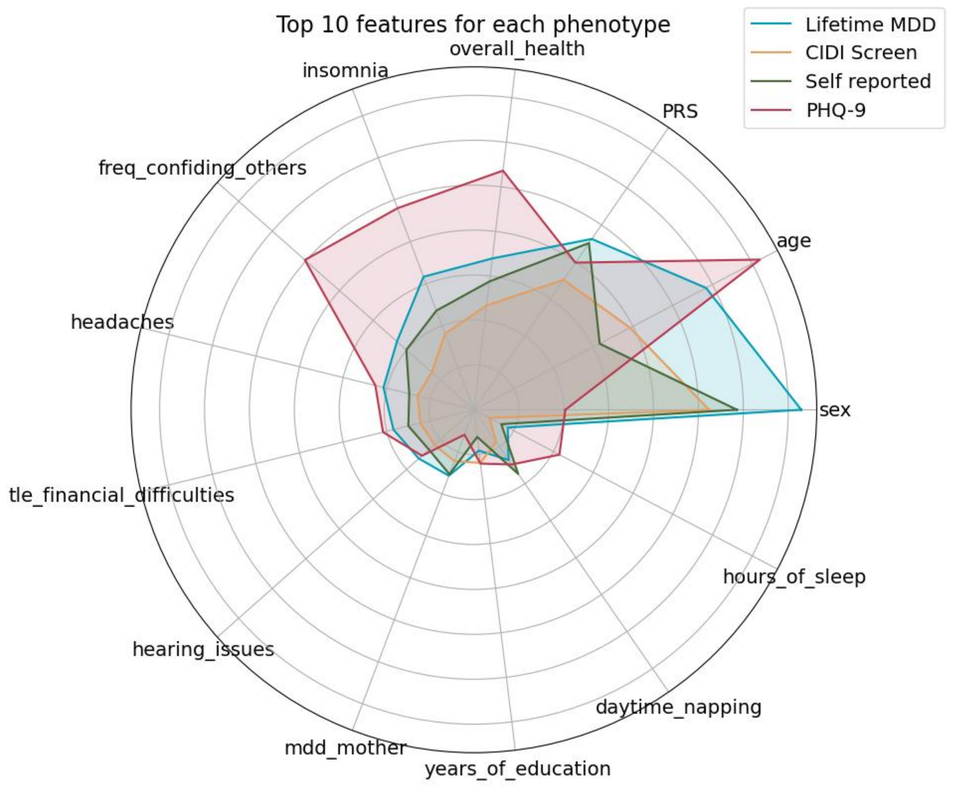


4. Methods

- We compared **six machine learning classifiers** (logistic regression, random forest, XGBoost, GBM, LightGBM, and CatBoost) in their performance at predicting case vs control as defined by each phenotype.
- We then selected the best-performing models and used **SHAP analysis** to look for the **most predictive factors** for each depression outcome.

5. Conclusions

- The CatBoost models performed best across the four phenotypes. We can predict depression with **70-80% accuracy**, but achieving high values for precision and recall simultaneously is challenging — especially in the PHQ-9 outcome.
- **Predictive machine learning models generalise well across different measures of lifetime depression, but less so between lifetime diagnoses and recent symptom questionnaires.**
  - Lifetime depression outcomes showed consistency among its top predictors (**sex, age, polygenic risk score, insomnia, and overall health**).
  - Recent feelings of depression were more strongly associated with lifestyle factors (**frequency confiding in others, financial difficulties and hours of sleep**) in addition to age, polygenic risk and overall health.



References

1. Lynch et al. (2020). Causes and consequences of diagnostic heterogeneity in depression: paths to discovering novel biological depression subtypes. *Biological psychiatry*.
2. Chekroud et al. (2021). The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*.
3. Sudlow et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*.
4. *Kessler et al. (1998)*. The World Health Organization composite international diagnostic interview short-form (CIDI-SF). *International journal of methods in psychiatric research*.

Questions? Get in touch!
b.branco@sms.ed.ac.uk
bianca-c-branco

UK Research and Innovation
AI4Bi
UKRI AI Centre for Doctoral Training in Biomedical Innovation
THE UNIVERSITY of EDINBURGH