Supplementary information for Explainable machine learning to predict depression across phenotype definitions integrating genetic and environmental factors

1 Sample description

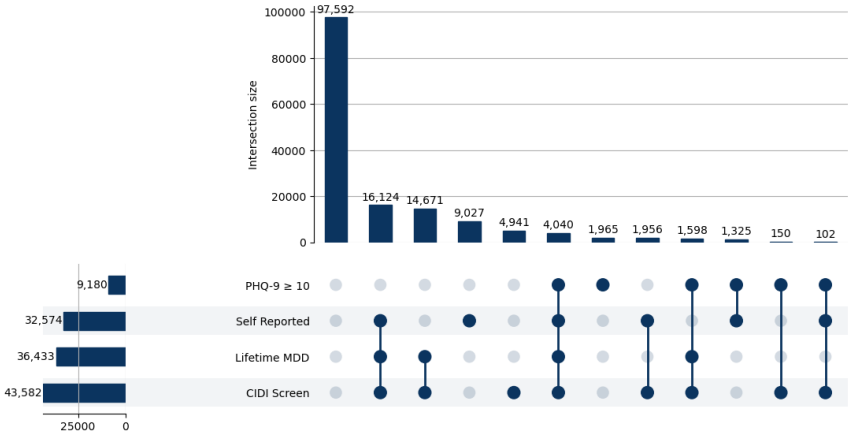


Figure 1: UpSet plot showing participant overlap across the four UK Biobank depression phenotypes. Horizontal bars indicate the total number of cases identified for each phenotype, while vertical bars represent the size of intersections between phenotype definitions. Connected circles denote the phenotype combinations included in each intersection.

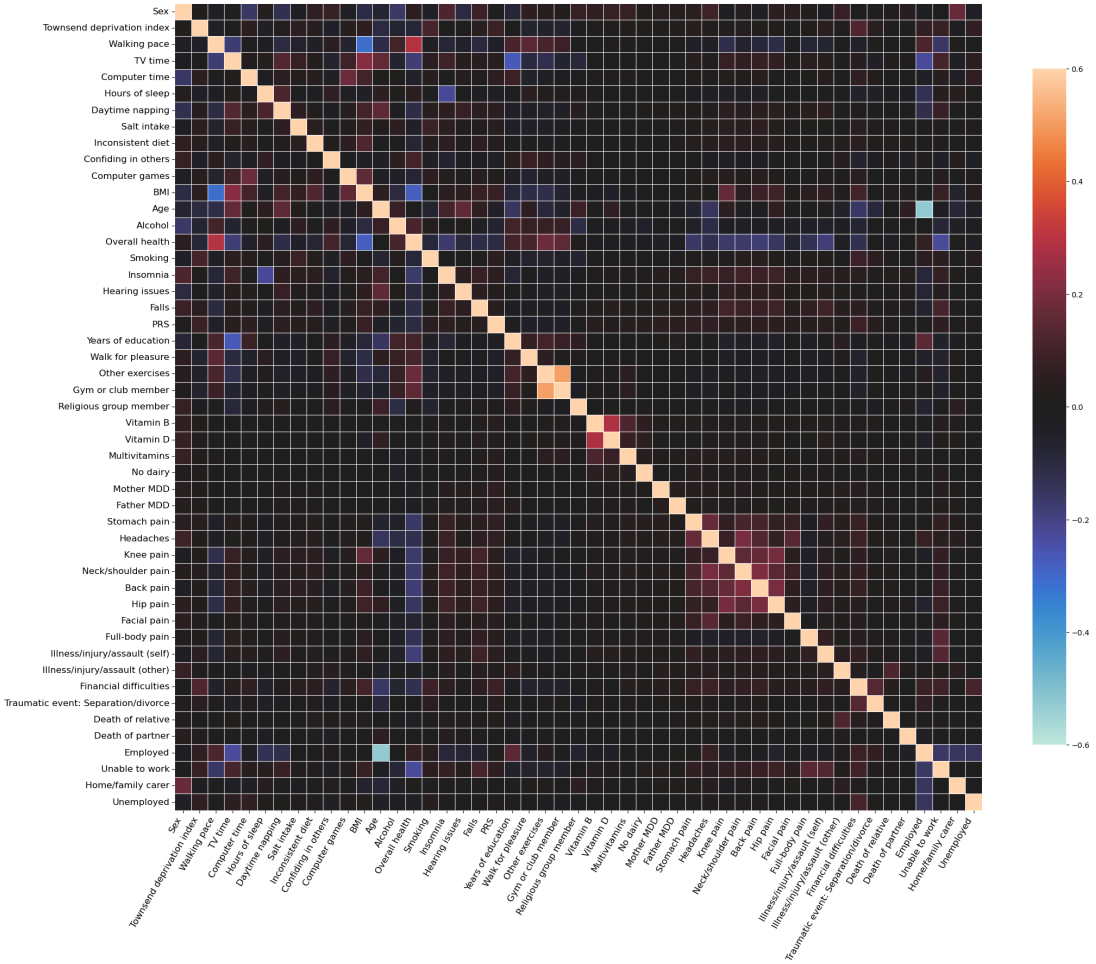


Figure 2: Mixed-type correlation matrix of all variables included in downstream analyses. Pairwise associations were computed using the `dython` package, which automatically selects an appropriate association metric based on variable type. Pearson's correlation coefficient was used for pairs of continuous variables, Cramér's V for pairs of categorical variables, and the correlation ratio (η) for mixed continuous-categorical variable pairs. Warmer colours indicate stronger positive associations, while cooler colours indicate stronger negative associations.

2 Calibration plots

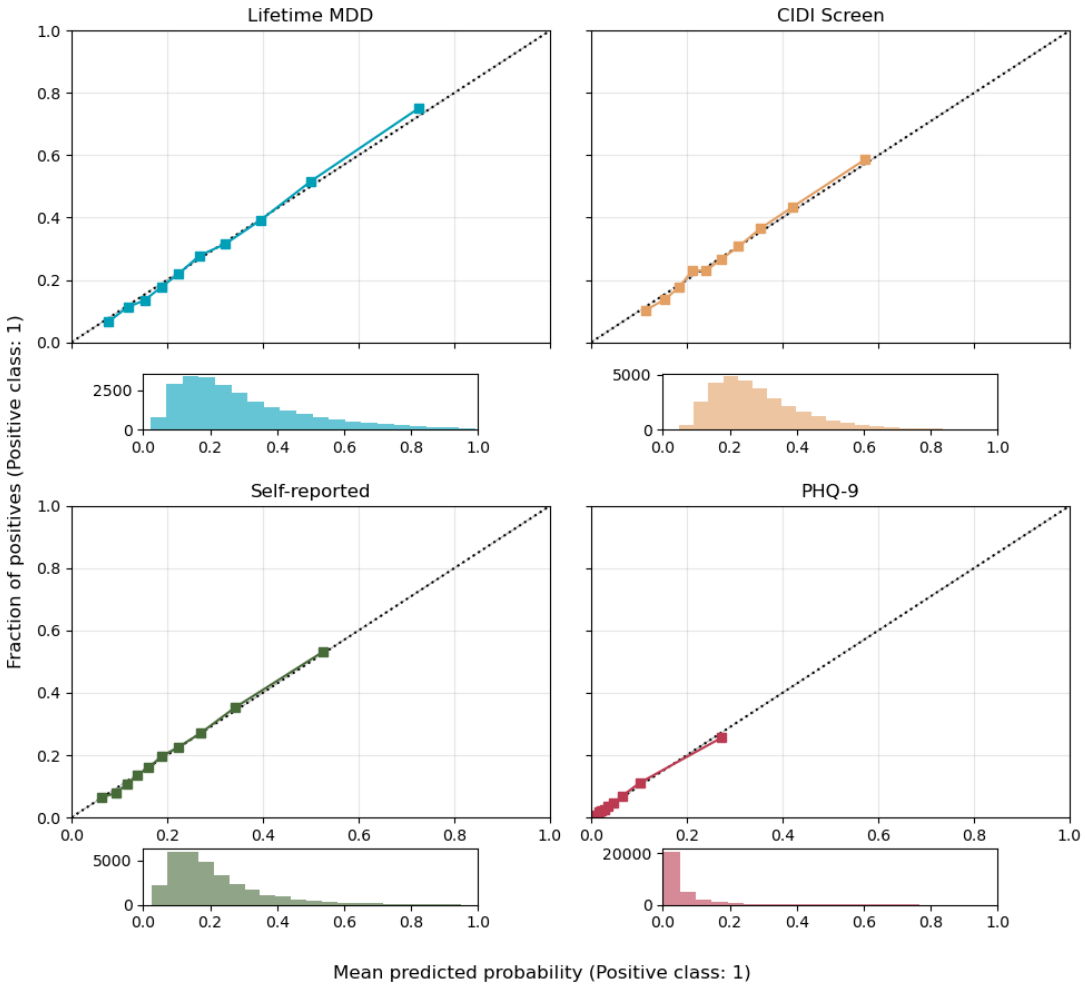


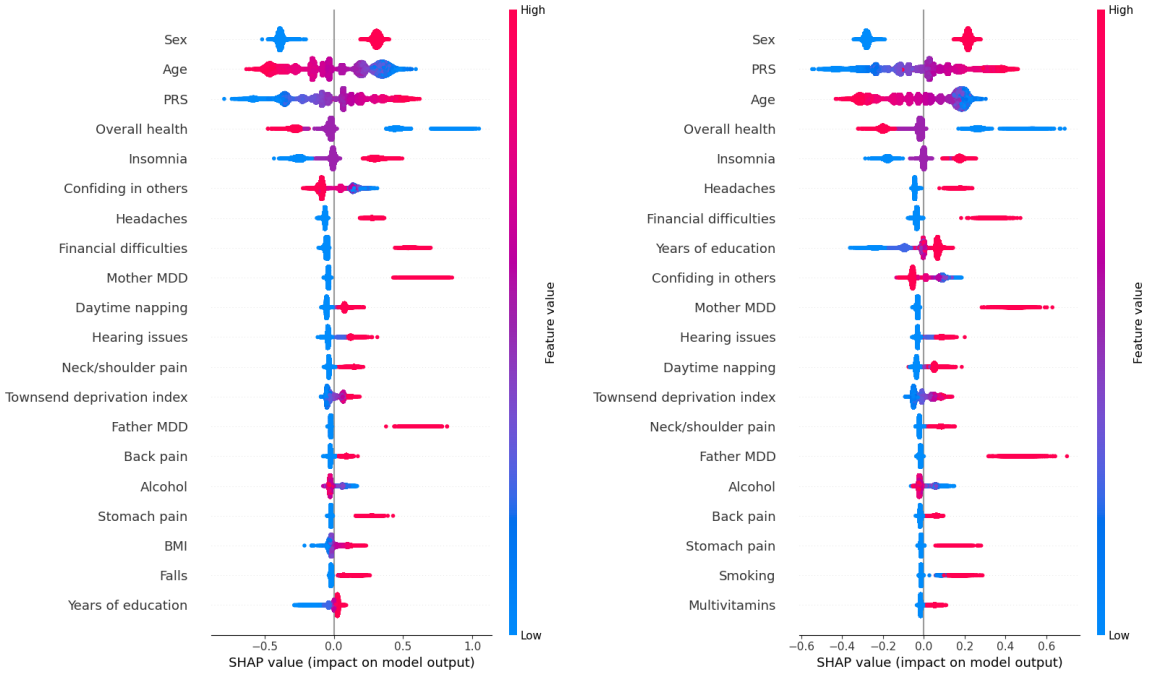
Figure 3: Calibration plots for the CatBoost models across the four depression phenotypes. Calibration curves compare the mean predicted probability with the observed proportion of depression cases using ten quantile-based bins. The dashed line indicates perfect calibration. Histograms below each panel show the distribution of predicted probabilities for the corresponding phenotype.

3 SHAP analysis

Figure 4 contains the beeswarm plots for the twenty most predictive features for each MDD phenotype. In a beeswarm plot, each dot corresponds to a data point; in our case a UKB participant.

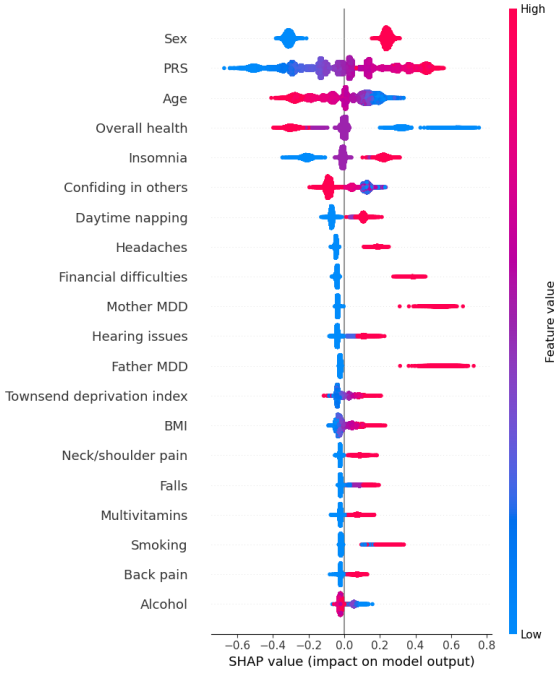
Figure 5 shows the relative feature importance learned by CatBoost models for all predictors and all phenotypes.

Figure 6 contains two examples of SHAP dependence plots between non-PRS predictors in the Lifetime MDD phenotype. Among all predictor relationships, we found that the strongest interaction occurred between age and insomnia, with a SHAP interaction value of 0.012 in the CatBoost Lifetime MDD model. Conversely, the individual SHAP values for age (0.295) and insomnia (0.158) were substantially larger, indicating that their main effects were approximately 24.6 and 13.2 times stronger, respectively, than their interaction effect. Although the interaction effect was modest in magnitude, the age $\times$ insomnia dependence plot (Fig. 6 (a)) reveals a clear age-dependent pattern, with insomnia exerting a stronger association with MDD before age 50. Similarly, we found a sex-stratified pattern for age and MDD (Fig. 6 (b)), with age above 55 presenting stronger positive association with depression in females and negative association in males; consistent with previous studies looking at sex-stratified trajectories in the UK Biobank [1]. These interaction effects highlight that depression risk is not static, and presents high heterogeneity in the interplay of risk factors across the life-course.

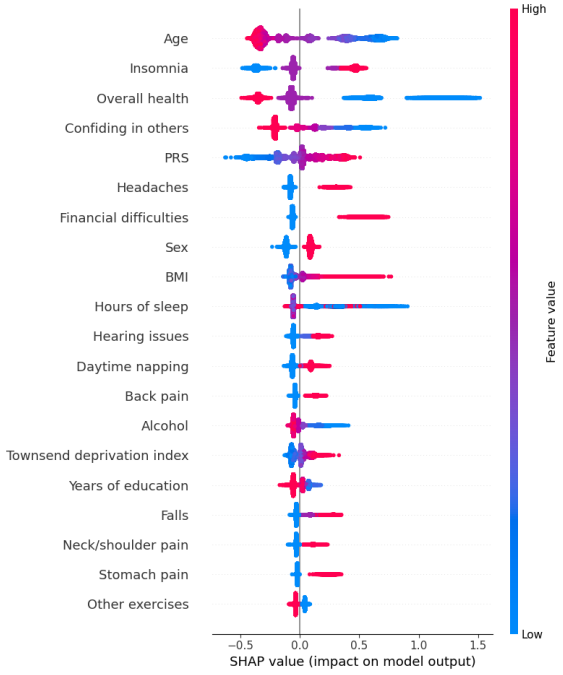


(a) Lifetime MDD

(b) CIDI screen



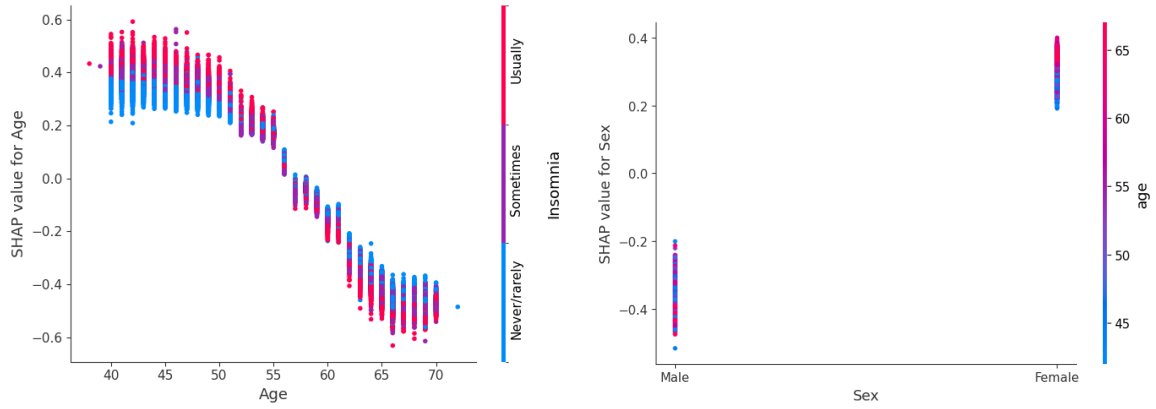
(c) Self-reported



(d) PHQ-9

Figure 4: SHAP summary plots for each depression phenotype. Features are ordered by mean absolute SHAP value, representing their overall contribution to model predictions across all individuals. Each point corresponds to a single participant, with position along the x-axis indicating the direction and magnitude of the feature's contribution to the predicted outcome. Feature values are colour-coded from low (blue) to high (red). Positive SHAP values indicate increased predicted depression risk, while negative values indicate decreased predicted risk.

Figure 5: Heatmap showing relative feature importance across CatBoost models trained on each depression phenotype. Values represent the percentage contribution of each feature to the overall model importance, normalised within phenotype to sum to 100%. Across phenotypes, sex, age, PRS, overall health, and insomnia consistently ranked among the most influential predictors, although notable differences in feature importance profiles were observed between depression definitions.



(a) Age and insomnia

(b) Sex and age

Figure 6: SHAP dependence plots illustrating interaction effects identified within the CatBoost models. (a) Relationship between age and SHAP values for age, coloured by insomnia frequency. Increasing age was generally associated with lower predicted depression risk, while higher insomnia frequency amplified positive SHAP contributions particularly among participants younger than 50 years. (b) Relationship between sex and SHAP values for sex, coloured by age. Male individuals showed a noisier age-related pattern, while for female individuals increasing age was associated with higher predicted depression risk.

4 PRS×E analysis

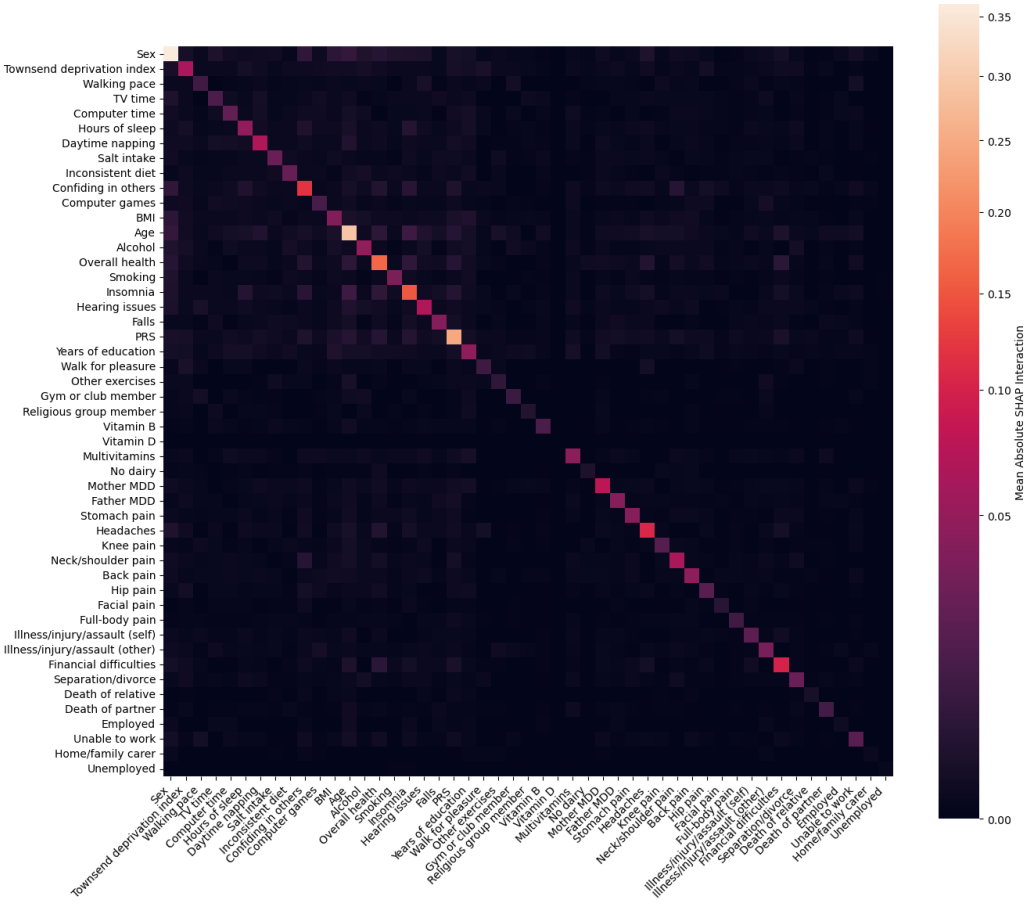


Figure 7: SHAP interaction matrix showing the mean absolute pairwise interaction strength between all model features for the lifetime MDD CatBoost model. SHAP interaction values are computed by decomposing each feature's SHAP contribution into main effects and pairwise interaction effects with all other features. Values represent the average absolute interaction contribution across all participants, with larger values indicating stronger joint influence on model predictions. Diagonal elements correspond to feature main effects, while off-diagonal elements represent pairwise interaction effects.

Figure 8: Heatmap of PRS × feature interaction coefficients across all depression phenotypes. Values represent the interaction term coefficients obtained from logistic regression models including PRS, the feature of interest, their interaction term, and covariates. Positive coefficients indicate that higher feature values amplified the association between PRS and depression risk, whereas negative coefficients indicate attenuation of PRS effects. Statistical significance is denoted as * $p < 0.05$, ** $p < 0.01$, and *** $p < 0.001$, with Benjamini-Hochberg false discovery rate (FDR) correction applied within each phenotype.

5 Sensitivity analysis

To assess the impact of missing data handling, all analyses were repeated using the sensitivity analysis dataset after excluding participants with missing values. Performance metrics are reported in Supplementary Table 5, with corresponding feature importance and PRS analyses shown in Fig. 9 and Fig. 10.

Overall, results were highly consistent between the primary imputed dataset and the sensitivity analysis dataset. CatBoost remained the highest-performing classifier across depression phenotypes, with only minor changes in predictive performance. AUC-ROC values differed by less than 0.01 across all outcomes, while AUC-PR values showed similarly small changes. AUC-ROC values were 0.76 for Lifetime MDD, 0.68 for CIDI Screen, 0.71 for Self-reported and 0.81 for PHQ-9, differing by less than 0.01 from the primary analysis across all outcomes. AUC-PR values were 0.60 for Lifetime MDD, 0.47 for CIDI Screen, 0.42 for Self-reported and 0.28 for PHQ-9, and showed similarly small differences. These findings indicate that model performance was robust to the handling of missing data.

The relative importance of the most predictive features also remained largely unchanged (Fig. 9). Sex, age, PRS, overall health, and insomnia continued to be the strongest predictors across the lifetime depression phenotypes. As observed in the primary analysis, the PHQ-9 phenotype displayed a distinct predictive profile characterised by a greater contribution of lifestyle factors, including insomnia and social connectedness. Similarly, the inclusion of PRS information resulted in consistent improvements in predictive performance across phenotypes (Fig. 10). Relative improvements ranged from 0.63-1.84% for AUC-ROC and 2.27-4.08% for AUC-PR, with the largest gains observed for the Self-reported phenotype. Together, these results support the robustness of the main findings and suggest that the observed predictor rankings and genetic contributions were not driven by the imputation procedure.

Consistent patterns were also observed in the sensitivity analysis of candidate PRS×E interactions (Fig. 11). Of the ten highest-ranked candidate interactions identified through SHAP interaction values, following FDR correction, evidence remained for PRS interactions with age (OR = 0.97, 95% CI [0.96, 0.99], $p < 0.01$), insomnia (OR = 1.03, 95% CI [1.01, 1.05], $p < 0.05$), overall health (OR = 0.97, 95% CI [0.95, 0.99], $p < 0.05$). The direction and magnitude of these effects were consistent with the primary analysis, with higher PRS effects observed among younger participants, those reporting insomnia, and individuals reporting poorer overall health. In contrast, interactions involving social support (OR = 0.99, 95% CI [0.98, 1.00]) and neck/shoulder pain (OR = 1.04, 95% CI [1.00, 1.08]) did not replicate in the sensitivity analysis. This suggests that the age, overall health, and insomnia interactions represent the most robust signals identified in the study.

Figure 9: Relative importance of the ten most predictive features across depression phenotypes in the sensitivity analysis datasets. Feature importance values are normalised to sum to 100% within each phenotype to facilitate comparison across outcomes.

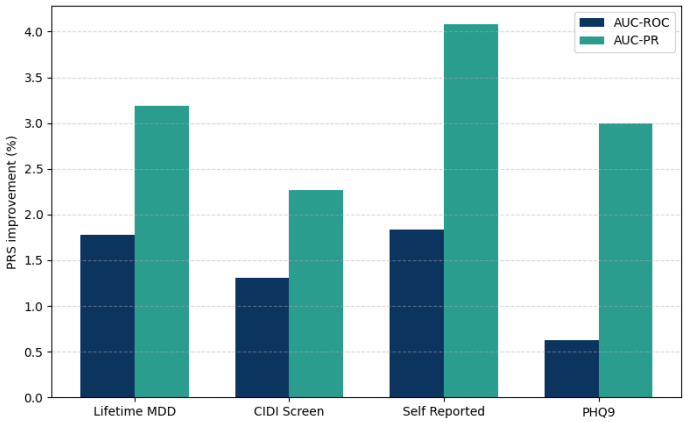


Figure 10: Relative improvement in predictive performance following inclusion of depression PRS across the four depression phenotypes in the sensitivity analysis datasets. Bars represent the percentage increase in AUC-ROC and AUC-PR relative to otherwise identical CatBoost models excluding PRS.

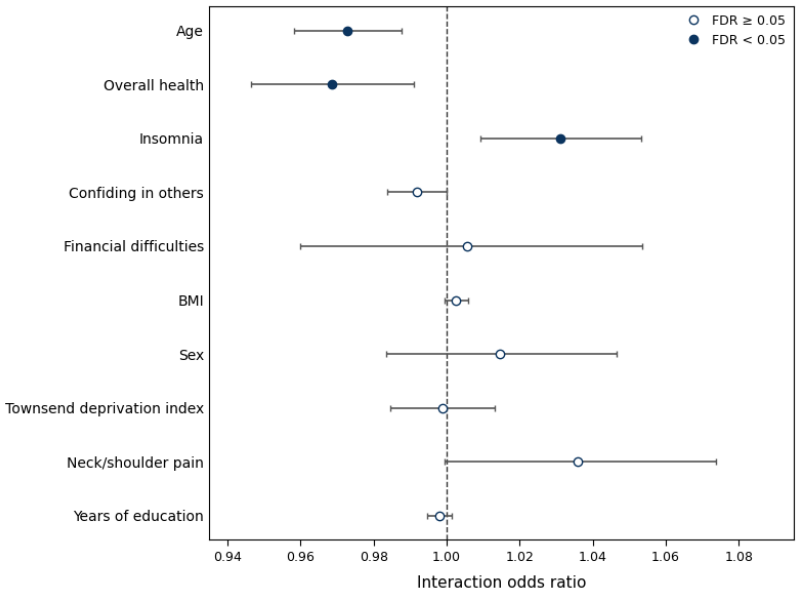


Figure 11: Sensitivity analysis of candidate PRS×E interactions identified through SHAP interaction values. Points represent interaction odds ratios from logistic regression models and error bars indicate 95% confidence intervals. Filled points denote interactions remaining significant after FDR correction, while open points indicate non-significant associations.

References

[1] Arias de la Torre J, Vilagut G, Ronaldson A, Dregan A, Ricci-Cabello I, Hatch SL, et al. Prevalence and age patterns of depression in the United Kingdom. A population-based study. Journal of Affective Disorders. 2021 Jan;279:164-72. Available from: `https://www.sciencedirect.com/science/article/pii/S0165032720328263`.