APPLYING NEW MACHINE LEARNING ANALYSES TO PREDICT RISK FACTORS FOR ANOREXIA AND BULIMIA NERVOSA: FINDINGS FROM A MULTI-CENTRE EUROPEAN PROJECT

Krug, I., , Linardon, J., Greenwood, C.,Youssef, G., Treasure, J., Fernandez-Aranda, F., Karwautz, A., Wagner, G., Collier, D., Anderluh, M., Tchanturia, K., Ricca, V., Sorbi, S., Nacmias, B., Bellodi, L., Fuller-Tyszkiewicz, M.

THE UNIVERSITY OF MELBOURNE

INTRODUCTION

*A wide range of risk & protective factors have been implicated in development of eating disorders (EDs)
*However, much of this research has proceeded in a piecemeal fashion.
*The Cross Cultural Risk Factor Questionnaire (CCQ), an extensive measure of risk/protective factors, provides a more comprehensive test of the key drivers of ED onset.
*There are also challenges in modelling as the number of predictors necessitates larger sample sizes to achieve suitable power & also carries risk of multicollinearity.
*These characteristics (large number of variables, possible instability of results) are suitable grounds for machine learning (ML) approaches to ascertain the key risk factors for ED onset.

OBJECTIVES

*To compare 3 different statistical approaches to gain greater insight into the key risk/protective predictors for ED onset:
1.) A *standard logistic regression* with **all factors entered simultaneously** and retained in the model;
2.) A *Least Absolute Shrinkage and Selection Operator* (LASSO) regression approach, which enters **all factors simultaneously, but shrinks small risk factor contributions to zero** – this ML approach seeks to balance parsimony with overall model performance & is equipped to handle concerns about multicollinearity;
3.) A *decision tree approach*, which evaluates interactions among predictors in an automated fashion, thus being sensitive to a range of interactions.

METHODS

Participants
* 626 ED patients (333 AN, 255 BN, 38 other ED)
* 776 controls.

Measures
* CCQ, which assesses a range of risk factors
* All risk factors were asked retrospectively < age 12 yrs

RESULTS

Selected indicators non-ED vs ED (excluding diagnosis predictor)

Figure 1: LASSO (non-ED v ED)

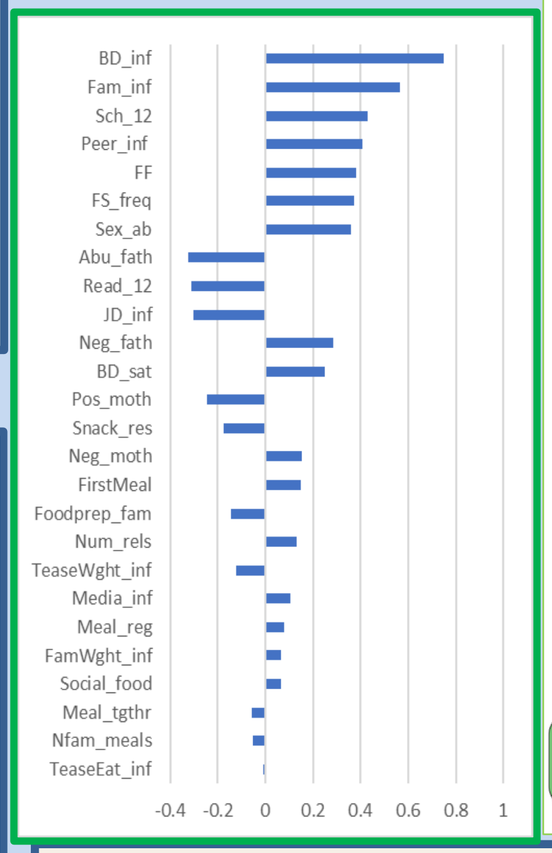
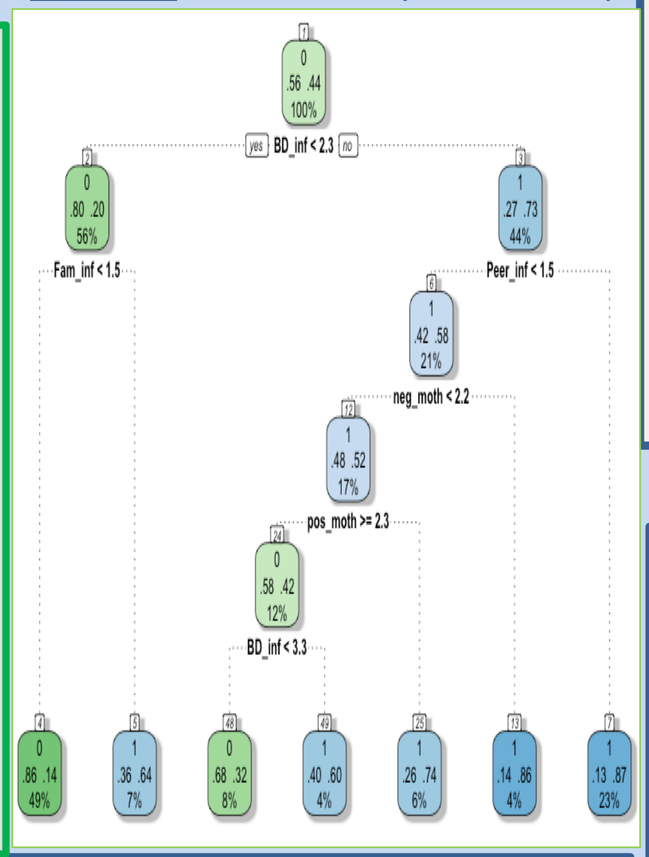


Figure 2: Decision Tree (non-ED v ED)



Accuracy of prediction – excluding ED diagnosis predictor
1.) Logistic regression – 82% ED status & 68% AN vs BN cases.
2.) LASSO - 83% ED status & 67% AN vs BN cases.
3.) Decision Tree: 89 % ED status, and 68% AN vs BN cases.

Best predictors - excluding ED diagnosis:

2.) Lasso

ED versus non-ED: 26 selected predictors.
1.) Body dissatis (BD) influenced eating (BD_inf)
2.) Family relation influenced eating (Fam_inf)
3.) School-work at school (Sch_12).

AN vs BN model: 26 selected predictors
1.) Played computer games before 12 (CompG_12)
2.) Physical appearance influenced eating (BD_inf
3.) Eating first meal before lessons (FirstMeal)

3.) Decision Tree

ED versus non-ED: 5 selected predictors
1.) BD during childhood (BD_inf)
2.) Relationship with family (Fam_inf)
3.) Relationship with friends (Peer_inf)

AN vs BN model: 8 predictors
1.) BD during childhood (BD_inf)
2.) Teasing about weight (TeaseWght_inf)
3.) Eating first meal of day before lessons (FirstMeal)

DISCUSSION

*We found very high accuracy for the ED versus control models for all 3 statistical approaches (with Decision Tree most accurate and most parsimonious).
*Our findings provide important insights into aetiological models of EDs using novel statistical approaches with the aim of improving prevention and intervention for EDs
*Further confirmatory studies are needed to test these exploratory hypotheses with rigorous prospective designs

Correspondence

Isabel.krug@unimelb.edu.au