Psychological Medicine

cambridge.org/psm

Machine learning enhances prediction of illness course: a longitudinal study in eating disorders

Ann F. Haynos[1,*], Shirley B. Wang[2,*], Sarah Lipson[2], Carol B. Peterson[1,3], James E. Mitchell[4], Katherine A. Halmi[5], W. Stewart Agras[6] and Scott J. Crow[1,3]

[1]Department of Psychiatry and Behavioral Sciences, University of Minnesota, Minneapolis, MN, USA; [2]Department of Psychology, Harvard University, Cambridge, MA, USA; [3]The Emily Program, Minneapolis, MN, USA; [4]Department of Psychiatry and Behavioral Science, University of North Dakota School of Medicine and Health Sciences, Fargo, ND, USA; [5]New York Presbyterian Hospital-Westchester Division, Weill Medical College of Cornell University, White Plains, NY, USA and [6]Department of Psychiatry, Stanford University School of Medicine, Stanford, CA, USA

Original Article

*The first two authors contributed equally to this work.

Cite this article: Haynos AF, Wang SB, Lipson S, Peterson CB, Mitchell JE, Halmi KA, Agras WS, Crow SJ (2021). Machine learning enhances prediction of illness course: a longitudinal study in eating disorders. *Psychological Medicine* **51**, 1392–1402. https://doi.org/10.1017/S0033291720000227

Received: 15 October 2019
Revised: 6 January 2020
Accepted: 23 January 2020
First published online: 28 February 2020

Key words:
Anorexia nervosa; binge-eating disorder; bulimia nervosa; computational psychiatry; eating disorder; machine learning

Author for correspondence:
Ann F. Haynos, E-mail: afhaynos@umn.edu

Abstract

Background. Psychiatric disorders, including eating disorders (EDs), have clinical outcomes that range widely in severity and chronicity. The ability to predict such outcomes is extremely limited. Machine-learning (ML) approaches that model complexity may optimize the prediction of multifaceted psychiatric behaviors. However, the investigations of many psychiatric concerns have not capitalized on ML to improve prognosis. This study conducted the first comparison of an ML approach (elastic net regularized logistic regression) to traditional regression to longitudinally predict ED outcomes.

Methods. Females with heterogeneous ED diagnoses completed demographic and psychiatric assessments at baseline ($n = 415$) and Year 1 ($n = 320$) and 2 ($n = 277$) follow-ups. Elastic net and traditional logistic regression models comprising the same baseline variables were compared in ability to longitudinally predict ED diagnosis, binge eating, compensatory behavior, and underweight BMI at Years 1 and 2.

Results. Elastic net models had higher accuracy for all outcomes at Years 1 and 2 [average Area Under the Receiving Operating Characteristics Curve (AUC) = 0.78] compared to logistic regression (average AUC = 0.67). Model performance did not deteriorate when the most important predictor was removed or an alternative ML algorithm (random forests) was applied. Baseline ED (e.g. diagnosis), psychiatric (e.g. hospitalization), and demographic (e.g. ethnicity) characteristics emerged as important predictors in exploratory predictor importance analyses.

Conclusions. ML algorithms can enhance the prediction of ED symptoms for 2 years and may identify important risk markers. The superior accuracy of ML for predicting complex outcomes suggests that these approaches may ultimately aid in advancing precision medicine for serious psychiatric disorders.

The course of psychiatric disorders is complex and heterogeneous (Marquand, Wolfers, Mennes, Buitelaar, & Beckmann, 2016). Clinicians and stakeholders have a pronounced desire to anticipate clinical progression to target intervention selection and intensity appropriately (McMahon, 2014). Therefore, statistical attempts have been made to identify shared predictors of illness outcome, most frequently utilizing regression-based inferential statistics. Yet, the results of these investigations have yielded unreliable or insensitive predictors (Franklin et al., 2017; Suvisaari et al., 2018).

A number of issues associated with conventional statistics limit their utility in predicting complex psychiatric phenomena. Traditional regression-related approaches can only accommodate a restricted number of predictors or else encounter low power, multicollinearity, and poor interpretability (Burke, Ammerman, & Jacobucci, 2019). Additionally, such statistics often rely upon the assumption of linearity, which may not capture the relational patterns between varied predictors and outcomes of interest (Jordan & Mitchell, 2015). These issues may conspire to produce artificially simplistic prediction models inappropriate for describing most psychiatric disorders (Linthicum, Schafer, & Ribeiro, 2019). An additional risk of traditional analytic techniques is over-fitting the data to the measured sample (Jordan & Mitchell, 2015), which could account for the poor between-study correspondence in predictors (Vall & Wade, 2015).

Emerging computational approaches utilizing machine-learning (ML) methods hold promise for optimizing prediction of psychiatric illness course. ML encompasses a collection of data-driven techniques that permit computer algorithms to identify and iteratively refine the optimal parameters to fit complicated patterns between variables (Bzdok & Meyer-Lindenberg, 2018; Linthicum et al., 2019). These approaches are well suited to overcome the historical limitations of predictive analytics. ML approaches accommodate intricate interdependent relations between a large number of variables without formal assumptions regarding the importance or structure of the data, allowing more rigorous predictive models.

CAMBRIDGE UNIVERSITY PRESS

Emerging data demonstrate the advantage of ML over traditional methods in predicting the course of serious mental illnesses, such as major depressive disorder (Kessler et al., 2016), obsessive–compulsive disorder (Askland et al., 2015), and psychotic disorders (Koutsouleris et al., 2016); however, this field is nascent and outcomes from many mental health concerns have not been examined using these novel methods.

Eating disorders (EDs), including anorexia nervosa (AN), bulimia nervosa (BN), and binge-eating disorder (BED), are among the psychiatric disorders most in need of the improved prediction derived through ML techniques. These illnesses are associated with significant medical and psychological morbidity (Marucci et al., 2018; Mitchell & Crow, 2006) and among the highest mortality rates of any psychiatric disorder (Crow et al., 2009). There is considerable unexplained heterogeneity in the lifetime course of EDs. Naturalistic estimates suggest that 50–70% of adults with an ED achieve remission, whereas 30–33% improve, but remain symptomatic, and 20–25% maintain chronic illness or die prematurely (Eddy et al., 2017; Steinhausen, 2002; Steinhausen & Weber, 2009). Conventional regression-based approaches have yielded a wide range of predictors of ED outcomes, including ED diagnosis, symptoms, and behaviors, body mass index (BMI), and comorbid psychiatric symptoms (Carter et al., 2012; Franko et al., 2018; Lock et al., 2013). However, these individual risk factors account for very little variance in clinical outcomes and fail to replicate across samples (Vall & Wade, 2015). Given the potential severity of the consequences associated with EDs, there is a critical need to improve prediction in these populations.

Therefore, we tested the ability of ML models to enhance the prediction of clinical outcomes among individuals with EDs. Specifically, we used elastic net regularized regression models to longitudinally predict ED diagnoses and key ED features (e.g. binge eating, compensatory behaviors, and underweight BMI) at 1- and 2-year follow-ups from baseline predictors, including demographic (e.g. age, ethnicity), clinical (e.g. ED behaviors, depressive symptoms), and treatment (e.g. prior hospitalization) variables among a large, transdiagnostic ED sample. We compared predictive accuracy between this ML method and a traditional inferential statistics approach (logistic regression). As further tests of model robustness, we examined whether findings replicated when the most important predictor was removed, or a different ML method (random forests) was applied. We hypothesized that the elastic net models would improve the prediction of long-term ED outcomes compared to logistic regression models, and that these findings would replicate even with the above described tests of model strength. We also conducted an exploratory investigation into the strongest predictor variables in each ML model to inform future investigations. This study constitutes the first to examine an ML approach to improve the prediction of long-term ED outcomes. The results can provide insight into the utility of these innovative computational approaches for advancing research on EDs and other severe psychiatric illnesses in the ultimate direction of data-derived personalized medicine.

Methods

Participants and procedure

Data were used from a longitudinal study conducted across sites in California, New York, and Minnesota (Agras, Crow, Mitchell, Halmi, & Bryson, 2009; Crow, Agras, Halmi, Mitchell, & Kraemer, 2002). Participants were recruited from ED clinics and research studies and local media. Inclusion criteria included being female, between 14 and 50 years old, and meeting full- or sub-threshold diagnoses of AN, BN, or BED according to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV; APA, 2000). Males were excluded due to the elevated ED prevalence in females (Smink, van Hoeken, & Hoek, 2012). Restrictions were not placed on whether participants could be enrolled in treatment. Full study details are described elsewhere (Crow et al., 2002).

Eligible participants completed interviews and self-report questionnaires assessing demographics, ED symptoms, and comorbid psychopathology at baseline and Year 1, 2, 3, and 4 post-baseline follow-ups. At baseline, 415 participants completed assessments. For follow-up, the sample size consisted of 320 participants at Year 1 (77% retention rate), 277 at Year 2 (67% retention rate), 248 at Year 3 (60% retention rate), and 254 at Year 4 (61% retention rate). Because of the higher attrition in Years 3 and 4, we made an *a priori* decision to analyze only Year 1 and 2 data. Institutional Review Boards at all sites approved the study and all participants provided informed consent.

Measures

Predictor variables

Prior investigations have identified links between long-term ED outcomes and baseline demographics (e.g. age, education level), ED symptoms (e.g. binge eating, compensatory behaviors), Axis-I (e.g. depressive disorders) and Axis-II (e.g. borderline personality disorder) diagnoses, treatment history, and psychological symptoms (e.g. depression, self-esteem) (Carter, Blackmore, Sutandar-Pinnock, & Woodside, 2004; Franko et al., 2018; Thompson-Brenner et al., 2013; Vall & Wade, 2015). Thus, the 33 baseline features in our model were selected to reflect these previously identified predictors available within the dataset.

Demographics and psychiatric treatment. Self-reported demographic information included age, living situation (alone, with family, with friends, with significant other), ethnicity (White, Hispanic, Black, American Indian, and Asian),[1†] marital status (single, married once, divorced, divorced/remarried, widowed, widowed/remarried, significant other), number of children, education level (less than grade school, grade 7–12, graduated high school, part college, graduated 2-year college, graduated 4-year college, part graduate school, graduated graduate school), and employment status (employed, unemployed). Participants also reported whether they had received past psychiatric treatment or hospitalization. Although not all demographics included in the model reflected previously identified predictors of ED outcomes, a broad range of demographics were included due to the ease of collection and under-investigation of such variables in prior research (Thompson-Brenner et al., 2013).

ED symptoms and behaviors. Previous research has shown that different measures designed to assess the same ED symptom can predict distinct outcomes (Stice, Fisher, & Lowe, 2004). Therefore, a number of ED symptom measures, including some aiming to assess the same construct, were selected as model features. BMI was calculated using baseline measured height and weight. The Eating Disorder Examination-12 (EDE; Cooper & Fairburn, 1987) assessed baseline ED presentation and diagnosis. The EDE is an investigator-based interview measuring ED

[†]The notes appear after the main text.

https://doi.org/10.1017/S0033291720000227 Published online by Cambridge University Press

psychopathology over the past 28 days, and longer intervals corresponding to diagnostic criteria. The EDE assesses psychological aspects of EDs and the frequency of behavioral symptoms (e.g. binge eating). In this study, we specified baseline EDE subscales (restraint, eating concern, weight concern, shape concern) and behavioral frequency items, including the number of episodes in the past 3 months of objective binge eating and compensatory behavior (vomiting, laxative use, and diuretic use), as ED cognitive and behavioral predictors. The EDE has excellent psychometrics, including inter-rater and test-retest reliability (Berg, Peterson, Frazier, & Crow, 2012a). In this study, inter-rater reliability for the EDE scores and behavioral items ranged from 0.90 to 0.99.

Two questionnaires further measured ED symptoms. The Three-Factor Eating Questionnaire (TFEQ; Stunkard & Messick, 1985) is a 51-item questionnaire assessing restraint (cognitive control over food intake), disinhibition (loss of control over food intake), and hunger (susceptibility to hunger cues), with higher scores indicating greater impairment. In this study, Cronbach's α ranged from 0.86 to 0.89 for these subscales. The Binge Eating Scale (BES; Gormally, Black, Daston, & Rardin, 1982) is a 16-item self-report questionnaire that dimensionally assesses binge eating, with higher scores indicating greater severity. The BES has good reliability, convergent validity, and discriminant validity (Greeno, Marcus, & Wing, 1995). In this study, Cronbach's α was 0.93.

Psychiatric diagnoses and symptoms. The Structured Clinical Interview for DSM-IV Axis-I Disorders (SCID-I; First, Spitzer, Gibbon, & Williams, 1995) and the Structured Clinical Interview for DSM-IV Axis-II Personality Disorders (SCID-II; First, Gibbon, Spitzer, Williams, & Benjamin, 1997) identified baseline comorbid psychopathology. The SCID-I assessed whether participants had a current or past diagnosis of any mood, substance use, obsessive–compulsive, and/or post-traumatic stress disorder or current diagnosis of generalized anxiety disorder. We also used the SCID-II to determine the presence of cluster A (odd/bizarre), cluster B (dramatic/erratic), and cluster C (avoidant/fearful) personality disorders (APA, 2000).

The Beck Depression Inventory (BDI; Beck, Steer, & Brown, 1996) assessed depressive symptomology. The BDI is a 21-item questionnaire evaluating symptoms of major depressive episodes, with higher scores indicating greater depression severity. The BDI has high reliability and internal consistency (Beck, Steer, & Carbin, 1988). In this study, Cronbach's α was 0.92. In addition, the Rosenberg Self-Esteem Scale (RSES; Rosenberg, 1965) assessed global self-esteem. The RSES is a 10-item questionnaire, with higher scores indicating more self-esteem. In this study, Cronbach's α for the RSES was 0.93.

Outcome variables

We examined the prediction of the several outcomes at follow-up: (1) meeting diagnostic criteria for any ED using validated DSM-IV EDE diagnostic algorithms (Berg et al., 2012b). According to these algorithms, absence of an ED was determined by no objective or subjective binge-eating episodes or compensatory behaviors in the past 3 months, and scores $\leqslant 2$ out of 6 on items assessing 'Fear of Weight Gain,' 'Importance of Weight,' and 'Importance of Shape'; (2) presence of objective binge eating over the past 3 months on the EDE; (3) presence of compensatory behaviors over the past 3 months on the EDE; and (4) underweight BMI ($<18.5\ kg/m^2$), based on measured height and weight. All outcome variables were assessed at Year 1 and 2 follow-ups.

Data analyses

Sample characteristics

To maximize statistical power, Year 1 outcome analyses included all 320 participants who completed both baseline and Year 1 assessments and Year 2 analyses included 277 participants who completed both baseline and Year 2 assessments. Casewise missing data were not imputed due to the mathematical equivalence of complete-case and imputed data when information is only missing for dependent variables in regression contexts (van Buuren, 2018; Von Hippel, 2017). There was a minimal amount of missing item-level data in the initial EDE interview (1.5% unavailable) due to specific questions not being asked. As the reasons for omitting these questions were not documented by study assessors and these data were not missing at random, we did not impute these data points. To test whether follow-up samples differed from the original sample on baseline characteristics, participants were categorized according to whether they completed follow-up at Year 1 only ($n = 16$), Year 2 only ($n = 56$), both ($n = 261$), or neither ($n = 82$). These groups were compared on baseline predictors using separate linear regression or χ^2 analyses with Bonferroni corrections ($p = 0.002$) for multiple comparisons.

Modeling approach

Analyses were performed in R (R Development Core Team, 2013) via *glm* in base R, and the *caret* (Kuhn, 2008) and *boot* (Canty & Ripley, 2019) packages. To test inferential statistical models, we conducted non-regularized multiple logistic regression models including all baseline variables to predict each outcome. To compare the performance of traditional logistic regressions to ML, we conducted elastic net regularized logistic regressions. Elastic net models (Zou & Hastie, 2005) combine two other regularization/ML methods: ridge (which shrinks coefficients in a model toward zero; Tikhonov, 1963) and lasso (which shrinks some coefficients completely to zero; Tibshirani, 1996). This algorithm was chosen given its well-established accuracy and robustness (Ogutu, Schulz-Streeck, & Piepho, 2012; Zou & Hastie, 2005) and an ability to maintain clinical result interpretability compared to less transparent ML algorithms (e.g. neural networks). However, to ensure that our conclusions about the accuracy of ML models were not idiosyncratic to the elastic net, we also tested our models using a random forests algorithm. Although a complete description of these ML models is beyond the scope of this paper, we refer readers to Kuhn and Johnson (2013) for a full introduction to these models and general ML concepts.

Following recommendations (Kuhn & Johnson, 2013), we used 10-fold cross-validation with three repetitions to select the optimal λ (a shrinkage parameter reflecting the degree of coefficient regularization) and α [a mixing parameter reflecting the balance between ridge ($\alpha = 0$) and lasso ($\alpha = 1$)] parameters for elastic net models, and the optimal mtry parameter (number of random predictors to consider at each split in the model) for random forests. This allowed us to obtain a cross-validation estimate of model performance for non-regularized logistic regression, elastic net, and random forest models. Whereas splitting a dataset into a single training (model building) and testing (model evaluation) set can have limited ability to accurately characterize uncertainty in results for smaller sample sizes, repeating the training and testing process in this manner can provide more reasonable estimates of model accuracy in predicting outcomes in new datasets. We calculated a 95% bias-corrected and accelerated confidence interval of the Area Under the Receiving Operating Characteristics

https://doi.org/10.1017/S0033291720000227 Published online by Cambridge University Press

Curve (AUC) model performance metric, reflecting model fit, using 5000 bootstrapped resamples, and refitted a model using the optimal λ and α parameters to examine variable importance/coefficients for the best-fitting elastic net model. To estimate variable importance for random forest models, we used the varImp() function in the *caret* package. For models with unbalanced outcomes (e.g. more than 70% of cases belonging to one outcome class), we used upsampling in the *caret* package (Kuhn, 2008) to improve the balance across classes. We used this subsampling technique within the resampling process to ensure that the proportions of outcomes in test data sets were not altered (Kuhn, 2008).

Model fit

We calculated the mean cross-validation estimate of AUC as a metric of model performance. AUC is a discrimination metric that ranges from 0.5 = chance to 1.0 = perfect. Guidelines for AUC scores range from extremely poor (0.5–0.59), poor (0.60–0.69), fair (0.70–0.79), good (0.80–0.89), to excellent (>0.90).

Results

Sample characteristics

See Table 1 for baseline characteristics. At Year 1, 299 (93.4%) participants continued to meet the criteria for any ED, 171 (53.4%) reported binge-eating episodes, 118 (36.9%) reported compensatory behaviors, and 46 (14.3%) had an underweight BMI. At Year 2, 242 (87.4%) participants met the criteria for any ED, 119 (43.0%) reported binge-eating episodes, 75 (27.1%) reported compensatory behaviors, and 31 (11.2%) had an underweight BMI. No significant baseline differences were detected between participants who did or did not complete one or both follow-ups (online Supplementary Table S1).

Model performance

Model performance indices for logistic regression and elastic net models are reported in Table 2. At Year 1 and 2 follow-ups, elastic net provided better accuracy in predicting all outcomes (binge eating, compensatory behaviors, ED diagnosis, underweight BMI) than the logistic regression models, with average cross-validated AUC for elastic net models ranging from poor/fair (ED diagnostic status and binge eating) to good (compensatory behaviors and underweight BMI at Year 2) and excellent (underweight BMI at Year 1).

Elastic net model performance did not deteriorate even when the most important predictor was removed (online Supplementary Table S2), demonstrating that the ability of the ML model to improve predictive accuracy was not dependent on any specific input variable. Finally, the accuracy of the random forest models (online Supplementary Table S3) was comparable to elastic net models for all outcomes at both time points, demonstrating that the superior performance of ML compared to standard approaches was not restricted to the selection of elastic net models.

Predictor importance

Coefficients for the variables in each elastic net model are shown in Table 3. Corresponding coefficients for the random forests model are included in online Supplementary Table S4 for comparison.

ED diagnosis

Elastic net models indicated that the most important variables for predicting ED diagnosis at Year 1 were baseline BN diagnosis, no baseline diagnosis of partial BED (this diagnosis was protective), not being of American Indian descent (this ethnicity was protective), and education level (<grade school or >graduate education). At Year 2, the most important variables predicting ED diagnosis included higher baseline EDE restraint, EDE shape concern, and TFEQ disinhibition scores, not being of American Indian descent, lower education level, and living alone relative to living with friends. Some overlap in predictor importance with the random forest model was found; both models identified EDE restraint, EDE shape concern, and TFEQ disinhibition scores to be of high importance to predicting Year 2 ED diagnosis. Evaluation of these metrics should take into account the poor fit of the models predicting ED diagnosis.

Binge eating

The most important variables for predicting binge eating in the elastic net model at Year 1 were baseline BN diagnosis, higher baseline BES and TFEQ disinhibition scores, and lower education level. Similarly, at Year 2, the most important predictors of binge eating were higher baseline BES and TFEQ disinhibition scores, Asian ethnicity, and no history of psychiatric treatment or cluster B personality disorder (these variables were protective). The random forest model also identified BES and TFEQ disinhibition scores as the variables of high importance to predicting binge eating at Years 1 and 2.

Compensatory behavior

The most important variables for predicting compensatory behavior in the elastic net model at Year 1 were baseline diagnosis of sub- or full-threshold BN, no baseline diagnosis of sub- or full-threshold BED (these diagnoses were protective), and living with friends. Similarly, at Year 2, the most important predictors of compensatory behavior were baseline diagnosis of sub- or full-threshold BN, no baseline diagnosis of sub- or full-threshold BED, and baseline diagnosis of generalized anxiety disorder. Similar to the elastic net model, the random forests models identified baseline BN and sub- or full-threshold BED diagnoses to be of high importance to predicting compensatory behavior at Years 1 and 2.

Underweight BMI

The most important variables for predicting underweight BMI in the elastic net model at Year 1 were lower baseline BMI, baseline diagnosis of sub-threshold AN, American Indian ethnicity, lower education level, and history of psychiatric hospitalization. Likewise, at Year 2, the most important predictors were lower baseline BMI, baseline diagnosis of sub-threshold AN, no baseline diagnosis of BED (this diagnosis was protective), history of psychiatric hospitalization, lower baseline TFEQ disinhibition, and current or past diagnosis of obsessive–compulsive disorder. The random forests model also identified baseline BMI to be of high importance to predicting underweight BMI at Years 1 and 2, baseline diagnosis of sub-threshold AN as important to predicting underweight BMI at Year 1, and baseline diagnosis of BED and TFEQ disinhibition score as important to predicting underweight BMI at Year 2.

Discussion

Advances in computational capacity have led the behavioral sciences to increasingly embrace ML techniques to improve the

https://doi.org/10.1017/S0033291720000227 Published online by Cambridge University Press

Table 1. Baseline sample characteristics ($n = 415$)

Variable	*M* (s.d.) *or n* (%)
Age (years)	32.36 (9.18)
Race	
White	380 (91.5%)
Hispanic	17 (4.1%)
Black	9 (2.2%)
American Indian	3 (1.0%)
Asian	4 (1.0%)
Marital status	
Never married	229 (55.2%)
Married once	94 (22.7%)
Divorced	64 (15.4%)
Divorced, remarried	22 (5.3%)
Widowed	1 (0.2%)
Widowed, remarried	4 (1.0%)
Significant other	1 (0.2%)
Number of children	0.67 (1.12)
Living situation	
Alone	93 (22.4%)
Family	211 (50.8%)
Friends	54 (13.0%)
Significant other	54 (13.0%)
Education	
Grade school or less	1 (0.2%)
Grade 7–12	18 (4.3%)
Graduated high school	38 (9.2%)
Part college	145 (34.9%)
Graduated 2-year college	25 (6.0%)
Graduated 4-year college	93 (22.4%)
Part graduate school	40 (9.6%)
Completed graduate school	55 (13.3%)
Employment status	
Employed	294 (70.8%)
Unemployed	121 (29.2%)
Eating disorder diagnosis	
Anorexia nervosa	48 (11.6%)
Bulimia nervosa	96 (23.1%)
Binge-eating disorder	116 (28.0%)
Subthreshold anorexia nervosa	45 (10.8%)
Subthreshold bulimia nervosa	65 (15.7%)
Subthreshold binge-eating disorder	45 (10.8%)
Restraint (EDE)	2.73 (1.62)
Eating concern (EDE)	2.26 (1.45)
Shape concern (EDE)	3.64 (1.46)
Weight concern (EDE)	3.33 (1.47)
Binge-eating episodes (per 3 months)	13.33 (18.30)
Self-induced vomiting episodes (per 3 months)	15.96 (52.09)
Laxative use episodes (per 3 months)	2.34 (9.40)
Diuretic use episodes (per 3 months)	1.05 (5.85)
Body mass index (kg/m^2)	28.04 (10.69)
Binge eating (BES)	25.17 (11.51)
Disinhibited eating (TFEQ)	10.77 (4.31)
Hunger (TFEQ)	8.29 (4.08)
Restraint (TFEQ)	11.89 (5.41)
History of psychiatric treatment	373 (89.9%)
History of psychiatric hospitalization	171 (41.4%)
Current or past mood disorder	298 (71.8%)
Current generalized anxiety disorder	46 (11.1%)
Current or past substance use disorder	166 (40.0%)
Current or past obsessive–compulsive disorder	51 (12.3%)
Current or past post-traumatic stress disorder	58 (14.0%)
Cluster A personality disorder	48 (11.6%)
Cluster B personality disorder	116 (28.0%)
Cluster C personality disorder	135 (32.5%)
Depression (BDI)	16.63 (10.69)
Self-esteem (RSES)	24.79 (6.45)

BDI, Beck Depression Inventory (Beck et al., 1996); BES, Binge Eating Scale (Gormally et al., 1982); EDE, Eating Disorder Examination (Cooper & Fairburn, 1987); RSES, Rosenberg Self-Esteem Scale (Rosenberg, 1965); TFEQ, Three-Factor Eating Questionnaire (Stunkard & Messick, 1985).

prediction of complex psychiatric phenomena (Askland et al., 2015; Fox et al., 2019; Kessler et al., 2016; Koutsouleris et al., 2016; Ribeiro, Huang, Fox, Walsh, & Linthicum, 2019). However, gaps in the adoption of these approaches remain. Despite the urgent need to improve prognosis related to EDs given their high morbidity and mortality rates (Crow et al., 2009; Mitchell & Crow, 2006), heterogeneous outcomes (Eddy et al., 2017; Steinhausen & Weber, 2009), and modest treatment response (Berkman et al., 2006), modeling techniques have not been used to maximize prediction for these populations. In this study, we conducted an initial investigation to determine whether ML models (e.g. elastic net) could improve upon traditional analytical techniques (logistic regression) in predicting key ED outcomes, such as continued ED diagnosis, binge eating, compensatory behaviors, and underweight BMI. Results demonstrated that, for each outcome at both Year 1 and 2 follow-ups, the elastic net model outperformed the logistic regression model, improving classification by up to 19%. In fact, in all but one comparison, the elastic net analyses moved the AUC estimate into a qualitatively higher accuracy category (e.g. fair to good). These results highlight the promise of ML for enhancing the ability to accurately predict complex outcomes for individuals with serious psychiatric disorders using only a set of mostly simple self-report measures.

https://doi.org/10.1017/S0033291720000227 Published online by Cambridge University Press

Table 2. Model performance of non-regularized and elastic net regularized logistic regressions

	Logistic regression		Elastic net			
	Mean AUC (95% CI)	Model fit	Mean AUC (95% CI)	Model fit	λ	α
Year 1						
Eating disorder diagnosis	0.48 (0.41–0.56)	Extremely poor	0.62 (0.53–0.71)	Poor	0.070	0.2
Binge eating	0.69 (0.65–0.73)	Poor	0.77 (0.73–0.81)	Fair	0.082	0.8
Compensatory behaviors	0.83 (0.80–0.86)	Good	0.88 (0.87–0.90)	Good	0.309	0.1
Underweight BMI	0.80 (0.74–0.85)	Good	0.93 (0.91–0.95)	Excellent	0.025	1
Year 2						
Eating disorder diagnosis	0.47 (0.40–0.55)	Extremely poor	0.61 (0.56–0.67)	Poor	0.110	0.3
Binge eating	0.64 (0.61–0.68)	Poor	0.71 (0.68–0.74)	Fair	0.150	0.3
Compensatory behaviors	0.72 (0.66–0.77)	Fair	0.85 (0.81–0.89)	Good	0.100	0.5
Underweight BMI	0.70 (0.65–0.77)	Fair	0.89 (0.86–0.91)	Good	0.178	0.3

AUC, Area Under the Receiving Operating Characteristics Curve; BMI, body mass index; CI, confidence interval.
The λ (penalization) and α (mixing) parameters were identified through 10-fold cross-validation repeated three times for the elastic net models.

Although the logistic regression analyses occasionally produced good models (e.g. correctly classifying 83% of Year 1 compensatory behaviors), overwhelmingly these models yielded fair to extremely poor AUCs. In fact, the ability to classify whether an individual would maintain an ED diagnosis at Year 1 and 2 follow-ups was no better than chance in these models. These findings are somewhat unsurprising, given that logistic regression is not designed to accommodate multiple, highly correlated predictors (Burke et al., 2019). However, the results highlight the insufficiency of the traditional analytic techniques that comprise the majority of predictive research in addressing a critical question facing psychiatric care: how likely are individuals with different clinical presentations to improve? In contrast, an elastic net produced fair to excellent models in all but two cases (ED diagnosis at Years 1 and 2). In the strongest models, predicting compensatory behaviors and underweight, 85–93% of cases were correctly classified. This is especially important given established links between low weight and purging and premature mortality in EDs (Crow et al., 2009). The AUCs reflected in these findings parallel the accuracy of ML models for other psychiatric phenomena (Fox et al., 2019; Ribeiro et al., 2019). Further, the fit of each model was not altered drastically with the removal of the most impactful variable, and was replicated using an alternative ML algorithm, highlighting the flexibility, reliability, and power of these approaches.

Improved prediction of psychiatric outcomes could have important long-term clinical implications. More individuals are affected with psychiatric disorders, including EDs, than the current care model can accommodate (Kazdin, Fitzsimmons-Craft, & Wilfley, 2017), thus healthcare access for these disorders remains limited (Guarda, Wonderlich, Kaye, & Attia, 2018). However, a sizable portion of individuals can improve with lower intensity interventions, such as guided self-help (Traviss-Turner, West, & Hill, 2017). Enhanced ability to detect the likelihood of clinical improvement for a particular individual could lead to optimized personalized decision-making regarding the appropriate intervention intensity. As such, ML models hold significant promise as scalable approaches to providing highly accurate, less burdensome, guidance in clinical settings in the future. However, it is important to acknowledge that the specific models developed in this study are not expected to yield direct practice utility and that additional research is needed before ML approaches can inform individualized clinical judgment (Challen et al., 2019). Rather, this study provides proof of concept for a class of analytic approaches with the potential to improve the models of psychiatric outcomes in subsequent research and, perhaps, clinical care.

However, even our ML models could be improved. Few produced excellent prediction, and prediction of ED diagnosis at Years 1 and 2 remained quite poor. This limitation could result from several factors. First, despite the potency of data-driven analytics, these approaches are constrained by their input. Although the predictors in this study reflect many of the most common psychiatric and ED assessments, they are not diverse and may have excluded important factors. For instance, clinical prognosis decisions often incorporate a number of variables not examined in this study (e.g. motivation, medical stability, support). Additionally, many of the indices in this study reflected symptoms or diagnoses, despite the increasing gravitation toward dimensional, mechanistic variables for predicting mental health outcomes (Bilder, Howe, & Sabb, 2013). Further, all the measures relied upon participant self-report. Acknowledging self-report limitations, there is a movement to include multi-modal assessment, incorporating behavioral (e.g. neuropsychological tasks) and biological (e.g. neuroimaging) features into ML models (Bilder et al., 2013). Because this study collected naturalistic data of long-term outcomes, all predictors were derived at baseline and dependent variables were collected infrequently. Future ML models may be improved by incorporating frequent and continuous data collection, including information gathered throughout clinical care. Finally, although behavioral indices often define ED outcomes, converging evidence suggests that cognitive symptoms (e.g. weight and shape overvaluation) may be of greater centrality to ED pathology (Bardone-Cone, Hunt, & Watson, 2018; DuBois, Rodgers, Franko, Eddy, & Thomas, 2017). More research is needed to determine if different or more comprehensive measures could further enhance the prediction of a variety of ED outcomes.

We also conducted exploratory analyses into which variables contributed the greatest variance to the ML models. There is a

https://doi.org/10.1017/S0033291720000227 Published online by Cambridge University Press

Table 3. Coefficients for variables in elastic net models

	Year 1				Year 2			
	ED diagnosis	Binge eating	Compensatory behaviors	Underweight BMI	ED diagnosis	Binge eating	Compensatory behaviors	Underweight BMI
Age	–	–	−0.01	–	–	–	–	–
Ethnicity (Hispanic)	–	–	–	–	–	–	–	–
Ethnicity (Black)	–	–	–	–	–	–	–	–
Ethnicity (American Indian)	−2.96	–	–	0.35	−0.96	–	–	–
Ethnicity (Asian)	–	–	–	–	–	0.01	–	–
Marital status (married once)	0.04	–	–	–	–	–	–	–
Marital status (divorced)	–	–	–	–	–	–	–	–
Marital status (divorced, remarried)	–	–	–	–	–	–	–	–
Marital status (widowed)	–	–	–	–	–	–	–	–
Marital status (widowed, remarried)	–	–	–	–	–	–	–	–
Marital status (significant other)	–	–	–	–	–	–	–	–
Number of children	–	–	–	–	–	–	–	–
Living situation (family)	–	–	–	–	–	–	–	–
Living situation (friends)	–	–	0.18	–	−0.03	–	–	–
Living situation (significant other)	–	–	–	–	–	–	–	–
Education level (grade 7–12)	−0.29	–	–	−1.36	–	–	–	–
Education level (graduated high school)	–	–	–	–	−0.11	–	–	–
Education level (part college)	–	–	0.06	–	–	–	–	–
Education level (graduated 2-year college)	−1.04	−0.15	–	–	−0.08	–	–	–
Education level (graduated 4-year college)	–	–	–	–	–	–	–	–
Education level (part graduate school)	–	–	–	–	–	–	–	–
Education level (graduated graduate school)	0.12	–	–	–	–	–	–	–
Employment status (unemployed)	–	–	–	–	–	–	–	–
Baseline ED diagnosis (BN)	0.25	0.56	0.77	–	–	–	1.32	–
Baseline ED diagnosis (BED)	–	–	−0.46	–	–	–	−0.58	−0.05
Baseline ED diagnosis (subthreshold AN)	–	–	–	0.39	–	–	–	0.38
Baseline ED diagnosis (subthreshold BN)	–	–	0.40	–	–	–	0.21	–
Baseline ED diagnosis (subthreshold BED)	−0.14	–	−0.37	–	–	–	−0.17	–
Restraint (EDE)	–	–	0.04	–	0.03	–	–	–
Eating concern (EDE)	–	–	–	–	–	–	–	–

Shape concern (EDE)	–	–	–	–	0.13	–	–	–
Weight concern (EDE)	–	–	–	–	0.02	–	–	–
Binge-eating episodes	–	–	–	–	–	–	–	–
Self-induced vomiting episodes	–	–	–	–	–	–	–	–
Laxative episodes	–	–	0.01	–	–	–	–	–
Diuretic episodes	–	–	–	–	−0.01	–	–	–
Body mass index	–	–	−0.01	−0.21	–	–	−0.02	−0.02
Binge eating (BES)	0.02	0.02	–	–	–	0.02	–	–
Disinhibited eating (TFEQ)	0.01	0.10	–	−0.04	0.03	0.05	–	−0.03
Hunger (TFEQ)	–	–	–	–	–	–	–	–
Restraint (TFEQ)	–	–	0.01	–	–	–	–	–
History of psychiatric treatment	0.05	–	–	–	–	−0.12	–	–
History of psychiatric hospitalization	–	–	0.14	0.14	–	–	–	0.16
Current or past mood disorder	–	–	–	–	–	–	–	–
Current generalized anxiety disorder	–	–	0.07	–	–	–	0.04	–
Current or past substance use disorder	–	–	–	–	–	–	–	–
Current or past obsessive–compulsive disorder	–	–	0.13	–	–	–	–	0.02
Current or past post-traumatic stress disorder	–	–	0.12	–	–	–	–	–
Cluster A personality disorder	–	–	–	–	–	–	–	–
Cluster B personality disorder	–	–	0.14	–	–	−0.06	–	–
Cluster C personality disorder	–	–	–	–	–	–	–	–
Depression (BDI)	–	–	–	–	–	–	–	–
Self-esteem (RSES)	0.02	–	–	–	–	–	–	–

AN, anorexia nervosa; BDI, Beck Depression Inventory (Beck et al., 1996); BED, binge-eating disorder; BES, Binge Eating Scale (Gormally et al., 1982); BN, bulimia nervosa; ED, eating disorder; EDE, Eating Disorder Examination (Cooper & Fairburn, 1987); RSES, Rosenberg Self-Esteem Scale (Rosenberg, 1965); TFEQ, Three-Factor Eating Questionnaire (Stunkard & Messick, 1985).
Elastic net coefficients represent the final model, identified through 10-fold cross-validation repeated three times to identify the optimal λ (penalization) and α (mixing) parameters. Dashes represent values of zero. Reference groups for categorical variables are: anorexia nervosa (baseline eating disorder diagnosis), white (ethnicity), married (marital status), alone (living situation), less than grade school (education level), and employed (employment status).

significant debate about the degree to which individual ML predictors should be interpreted given the complexity of the interactions modeled in ML (Ribeiro et al., 2019). However, data-driven approaches reduce reliance on subjective interpretations that sometimes bias hypothesis-driven investigations (Burke et al., 2019) and may permit the identification of variables with unanticipated influences on outcomes that warrant examination in future predictive research. In this study, many of the important model features, especially those replicating between ML methods, had strong face validity (e.g. disinhibition predicting binge eating; sub-threshold AN predicting underweight BMI). However, other unexpected patterns (e.g. level of educational attainment contributing to most elastic net models) may warrant investigation in future research. Although these findings highlight the ability of ML to construct models that overcome clinical biases in the service of accurately calculating risk, these approaches hold the greatest promise for precise prediction, as opposed to mechanism identification. Additionally, particular caution is encouraged in interpreting feature importance for predicting ED diagnosis at Years 1 and 2, given the poor fit of these models.

Our study strengths include multi-site recruitment of a large transdiagnostic sample, extended length of follow-up, and collection of a sizeable number of assessments that could be readily administered in a real-world setting. However, there were also limitations. The sample size was adequate for an elastic net model, but was somewhat small for ML, which often requires very large sample sizes for the most rigorous prediction (Bzdok & Meyer-Lindenberg, 2018). Additionally, there are numerous ML approaches besides elastic net and random forests, each with a unique set of strengths and limitations (Cho, Yim, Choi, Ko, & Lee, 2019); it is possible that other models could yield better accuracy. It is also possible that alternative modeling choices may have yielded different outcomes. For instance, although cross-validation is an accepted ML practice for investigations with smaller sample sizes and we conducted upsampling within the training (but not test) datasets (Kuhn & Johnson, 2013), it has the potential to yield overly optimistic models (Smith, Seaman, Wood, Royston, & White, 2014). Similarly, AUC is a frequently-used metric for testing model performance; however, some recommend the use of precision-recall curves when outcomes are imbalanced (He & Garcia, 2009). Further, as with any longitudinal investigation, this study experienced attrition. No baseline differences were found between individuals who did and did not complete follow-ups, but it is possible that these participants varied in unmeasured ways. The sample also lacked diversity in certain demographics (e.g. gender, race/ethnicity) and may not be generalizable to the broader ED population. Additionally, some model features were derived from relatively time-consuming assessments requiring extensive training (e.g. diagnostic interviews), which may limit the ease in clinical dissemination. Finally, the method of determining ED diagnosis, though based on established algorithms (Berg et al., 2012b), may have been outdated (based on DSM-IV) and biased toward inflating clinical cases due to the high standard for cognitive symptom improvement. Further research is needed to test multiple ML models with larger, more diverse ED samples and objective markers of clinical outcome.

Conclusion

This investigation confirmed that ML can enhance the long-term prediction of persistent ED symptoms and may also identify novel markers of risk. These results encourage future testing of novel ML algorithms to predict illness course, as well as to determine treatment response and to develop adaptively tailored psychiatric interventions. The complexity, flexibility, and accuracy of ML make these approaches well suited to ultimately advance precision medicine for serious mental health concerns, including EDs, and thereby to contribute toward mitigating the severe and life-threatening consequences of these illnesses.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291720000227.

Acknowledgements. The investigators express gratitude to the McKnight Longitudinal Study research staff at the University of Minnesota, Cornell University, and Stanford University for their contributions to this investigation. The investigators also thank Dr Patrick Mair, who provided statistical guidance and consulting on machine-learning models.

Financial support. This work was supported in part by the McKnight Foundation; National Institute of Mental Health of the National Institutes of Health (Award numbers: K23MH112867, T32MH082761); National Science Foundation (Award number DGE-1745303); Klarman Family Foundation; Hilda and Preston Davis Foundation; and Minnesota Obesity Center. These funding agencies did not influence the design of the study, collection, analysis, and interpretation of data, or writing of the manuscript

Conflict of interest. None.

Ethical standards. The authors assert that all procedures contributing to this work comply with the ethical standards of the relevant national and institutional committees on human experimentation and with the Helsinki Declaration of 1975, as revised in 2008.

Note

1 Data were collected between 1995 and 1998; therefore, the terminology used to describe race/ethnicity may not conform to current standards.

References

Agras, W. S., Crow, S., Mitchell, J. E., Halmi, K. A., & Bryson, S. (2009). A 4-year prospective study of eating disorder NOS compared with full eating disorder syndromes. *The International Journal of Eating Disorders*, *42*, 565–570. doi: 10.1002/eat.20708

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders: DSM-IV-TR* (4th ed.). text revision. Washington, DC: American Psychiatric Association.

Askland, K. D., Garnaat, S., Sibrava, N. J., Boisseau, C. L., Strong, D., Mancebo, M., … Eisen, J. (2015). Prediction of remission in obsessive compulsive disorder using a novel machine learning strategy. *International Journal of Methods in Psychiatric Research*, *24*, 156–169. doi: 10.1002/mpr.1463

Bardone-Cone, A. M., Hunt, R. A., & Watson, H. J. (2018). An overview of conceptualizations of eating disorder recovery, recent findings, and future directions. *Current Psychiatry Reports*, *20*, 79. doi: 10.1007/s11920-018-0932-9

Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. San Antonio, TX: Psychological Corporation.

Beck, A. T., Steer, R. A., & Carbin, M. G. (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. *Clinical Psychology Review*, *8*, 77–100. doi: 10.1016/0272-7358(88)90050-5

Berg, K. C., Peterson, C. B., Frazier, P., & Crow, S. J. (2012a). Psychometric evaluation of the eating disorder examination and eating disorder examination-questionnaire: A systematic review of the literature. *The International Journal of Eating Disorders*, *4*, 428–438. doi: 10.1002/eat.20931

Berg, K. C., Stiles-Shields, E. C., Swanson, S. A., Peterson, C. B., Lebow, J., & Le Grange, D. (2012b). Diagnostic concordance of the interview and

questionnaire versions of the eating disorder examination. *The International Journal of Eating Disorders*, *45*, 850–855. doi: 10.1002/eat.20948

Berkman, N. D., Bulik, C. M., Brownley, K. A., Lohr, K. N., Sedway, J. A., Rooks, A., & Gartlehner, G. (2006). Management of eating disorders. *Evidence Report/Technology Assessment*, *135*, 1–166.

Bilder, R. M., Howe, A. G., & Sabb, F. W. (2013). Multilevel models from biology to psychology: Mission impossible? *Journal of Abnormal Psychology*, *122*, 917–927. doi: 10.1037/a0032263

Burke, T. A., Ammerman, B. A., & Jacobucci, R. (2019). The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *Journal of Affective Disorders*, *245*, 869–884. doi: 10.1016/j.jad.2018.11.073

Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *3*, 223–230. doi: 10.1016/j.bpsc.2017.11.007

Canty, A., & Ripley, B. (2019). *boot: Bootstrap R (S-Plus) Functions. Version 1.3-22*.

Carter, J. C., Blackmore, E., Sutandar-Pinnock, K., & Woodside, D. B. (2004). Relapse in anorexia nervosa: A survival analysis. *Psychological Medicine*, *34*, 671–679.

Carter, J. C., Mercer-Lynn, K. B., Norwood, S. J., Bewell-Weiss, C. V., Crosby, R. D., Woodside, D. B., & Olmsted, M. P. (2012). A prospective study of predictors of relapse in anorexia nervosa: Implications for relapse prevention. *Psychiatry Research*, *200*, 518–523. doi: 10.1016/j.psychres.2012.04.037

Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, *28*, 231–237. doi: 10.1136/bmjqs-2018-008370

Cho, G., Yim, J., Choi, Y., Ko, J., & Lee, S.-H. (2019). Review of machine learning algorithms for diagnosing mental illness. *Psychiatry Investigation*, *16*, 262–269. doi: 10.30773/pi.2018.12.21.2

Cooper, Z., & Fairburn, C. (1987). The eating disorder examination: A semi-structured interview for the assessment of the specific psychopathology of eating disorders. *The International Journal of Eating Disorders*, *6*, 1–8. doi: 10.1002/1098-108X(198701)6:1<1::AID-EAT2260060102>3.0.CO;2-9

Crow, S. J., Agras, W. S., Halmi, K., Mitchell, J. E., & Kraemer, H. C. (2002). Full syndromal versus subthreshold anorexia nervosa, bulimia nervosa, and binge eating disorder: A multicenter study. *The International Journal of Eating Disorders*, *32*, 309–318. doi: 10.1002/eat.10088

Crow, S. J., Peterson, C. B., Swanson, S. A., Raymond, N. C., Specker, S., Eckert, E. D., & Mitchell, J. E. (2009). Increased mortality in bulimia nervosa and other eating disorders. *The American Journal of Psychiatry*, *166*, 1342–1346. doi: 10.1176/appi.ajp.2009.09020247

DuBois, R. H., Rodgers, R. F., Franko, D. L., Eddy, K. T., & Thomas, J. J. (2017). A network analysis investigation of the cognitive-behavioral theory of eating disorders. *Behavior Research and Therapy*, *97*, 213–221. doi: 10.1016/j.brat.2017.08.004

Eddy, K. T., Tabri, N., Thomas, J. J., Murray, H. B., Keshaviah, A., Hastings, E., … Franko, D. L. (2017). Recovery from anorexia nervosa and bulimia nervosa at 22-year follow-up. *The Journal of Clinical Psychiatry*, *78*, 184–189. doi: 10.4088/JCP.15m10393

First, M. B., Gibbon, M., Spitzer, R. L., Williams, J. B. W., & Benjamin, L. S. (1997). *Structured Clinical Interview for DSM-IV Axis II Personality Disorders* (*SCID-II*). Washington, DC: American Psychiatric Press, Inc.

First, M. B., Spitzer, R. L., Gibbon, M., & Williams, J. B. W. (1995). Structured Clinical Interview for DSM-IV Axis I Disorders: Patient Edition (SCID-I/P). Version 2.0. Biometrics Research Department, New York State Psychiatric Institute: New York.

Fox, K. R., Huang, X., Linthicum, K. P., Wang, S. B., Franklin, J. C., & Ribeiro, J. D. (2019). Model complexity improves the prediction of nonsuicidal self-injury. *Journal of Consulting and Clinical Psychology*, *87*, 684–692. doi: 10.1037/ccp0000421

Franklin, J. C., Ribeiro, J. D., Fox, K. R., Bentley, K. H., Kleiman, E. M., Huang, X., … Nock, M. K. (2017). Risk factors for suicidal thoughts and behaviors: A meta-analysis of 50 years of research. *Psychological Bulletin*, *143*, 187–232. doi: 10.1037/bul0000084

Franko, D. L., Tabri, N., Keshaviah, A., Murray, H. B., Herzog, D. B., & Thomas, J. J. (2018). Predictors of long-term recovery in anorexia nervosa and bulimia nervosa: Data from a 22-year longitudinal study. *Journal of Psychiatry Research*, *96*, 183–188. doi: 10.1016/j.jpsychires.2017.10.008

Gormally, J., Black, S., Daston, S., & Rardin, D. (1982). The assessment of binge eating severity among obese persons. *Addictive Behaviors*, *7*, 47–55. doi: 10.1016/0306-4603(82)90024-7

Greeno, C. G., Marcus, M. D., & Wing, R. R. (1995). Diagnosis of binge eating disorder: Discrepancies between a questionnaire and clinical interview. *The International Journal of Eating Disorders*, *17*, 153–160. doi: 10.1002/1098-108x(199503)17:2<153::aid-eat2260170208>3.0.co;2-v

Guarda, A. S., Wonderlich, S., Kaye, W., & Attia, E. (2018). A path to defining excellence in intensive treatment for eating disorders. *The International Journal of Eating Disorders*, *51*, 1051–1055. doi: 10.1002/eat.22899

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*, 1263–1284. doi: 10.1109/TKDE.2008.239

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*, 255–260. doi: 10.1126/science.aaa8415

Kazdin, A. E., Fitzsimmons-Craft, E. E., & Wilfley, D. E. (2017). Addressing critical gaps in the treatment of eating disorders. *The International Journal of Eating Disorders*, *50*, 170–189. doi: 10.1002/eat.22670

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., … Zaslavsky, A. M. (2016). Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. *Molecular Psychiatry*, *21*, 1366–1371. doi: 10.1038/mp.2015.198

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., & Wobrock, T. (2016). Multisite prediction of 4-week and 52-week treatment outcomes in patients with first-episode psychosis: A machine learning approach. *The Lancet. Psychiatry*, *3*, 935–946. doi: 10.1016/S2215-0366(16)30171-7

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*, 1–26. doi: 10.18637/jss.v028.i05

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer.

Linthicum, K. P., Schafer, K. M., & Ribeiro, J. D. (2019). Machine learning in suicide science: Applications and ethics. *Behavioral Science and the Law*, *37*, 214–222. doi: 10.1002/bsl.2392

Lock, J., Agras, W. S., Le Grange, D., Couturier, J., Safer, D., & Bryson, S. W. (2013). Do end of treatment assessments predict outcome at follow-up in eating disorders? *The International Journal of Eating Disorders*, *46*, 771–778. doi: 10.1002/eat.22175

Marquand, A. F., Wolfers, T., Mennes, M., Buitelaar, J., & Beckmann, C. F. (2016). Beyond lumping and splitting: A review of computational approaches for stratifying psychiatric disorders. *Biological Psychiatry. Cognitive Neuroscience and Neuroimaging*, *1*, 433–447. doi: 10.1016/j.bpsc.2016.04.002.

Marucci, S., Ragione, L. D., De Iaco, G., Mococci, T., Vicini, M., Guastamacchia, E., & Triggiani, V. (2018). Anorexia nervosa and comorbid psychopathology. *Endocrine, Metabolic, and Immune Disorders Drug Targets*, *18*, 316–324. doi: 10.2174/1871530318666180213111637

McMahon, F. J. (2014). Prediction of treatment outcomes in psychiatry – where do we stand? *Dialogues in Clinical Neuroscience*, *16*, 455–464.

Mitchell, E., & Crow, E. (2006). Medical complications of anorexia nervosa and bulimia nervosa. *Current Opinions in Psychiatry*, *19*, 438–443. doi: 10.1097/01.yco.0000228768.79097.3e

Ogutu, J. O., Schulz-Streeck, T., & Piepho, H.-P. (2012). Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. *BMC Proceedings*, *6*(Suppl 2), S10. doi: 10.1186/1753-6561-6-S2-S10

R Development Core Team (2013). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Ribeiro, J. D., Huang, X., Fox, K. R., Walsh, C. G., & Linthicum, K. P. (2019). Predicting imminent suicidal thoughts and nonfatal attempts: The role of complexity. *Journal of Consulting and Clinical Psychology*, *87*, 684–692. doi: 10.1177/2167702619838464

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University Press.

Smink, F. R., van Hoeken, D., & Hoek, H. W. (2012). Epidemiology of eating disorders: Incidence, prevalence and mortality rates. *Current Psychiatry Reports*, *14*, 406–414. doi: 10.1007/s11920-012-0282-y

Smith, G. C. S., Seaman, S. R., Wood, A. M., Royston, P., & White, I. R. (2014). Correcting for optimistic prediction in small data sets. *American Journal of Epidemiology*, *180*, 318–324. doi: 10.1093/aje/kwu140

Steinhausen, H.-C. (2002). The outcome of anorexia nervosa in the 20th century. *The American Journal of Psychiatry*, *159*, 1284–1293. doi: 10.1176/appi.ajp.159.8.1284

Steinhausen, H.-C., & Weber, S. (2009). The outcome of bulimia nervosa: Findings from one-quarter century of research. *The American Journal of Psychiatry*, *166*, 1331–1341. doi: 10.1176/appi.ajp.2009.09040582

Stice, E., Fisher, M., & Lowe, M. R. (2004). Are dietary restraint scales valid measures of acute dietary restriction? Unobtrusive observational data suggest not. *Psychological Assessment*, *16*, 51–59. doi: 10.1037/1040-3590.16.1.51

Stunkard, A. J., & Messick, S. (1985). The three-factor eating questionnaire to measure dietary restraint, disinhibition and hunger. *Journal of Psychosomatic Research*, *29*, 71–83. doi: 10.1016/0022-3999(85)90010-8

Suvisaari, J., Mantere, O., Keinänen, J., Mäntylä, T., Rikandi, E., Lindgren, M., … Raij, T. T. (2018). Is it possible to predict the future in first-episode psychosis? *Frontiers in Psychiatry*, *9*, 580. doi: 10.3389/fpsyt.2018.00580

Thompson-Brenner, H., Franko, D. L., Thompson, D. R., Grilo, C. M., Boisseau, C. L., Roehrig, J. P., … Wilson, G. T. (2013). Race/ethnicity, education, and treatment parameters as moderators and predictors of outcome in binge eating disorder. *Journal of Consulting and Clinical Psychology*, *81*, 710–721. doi: 10.1037/a0032946

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* (*Methodological*), *58*, 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics*, *4*, 1035–1038.

Traviss-Turner, G. D., West, R. M., & Hill, A. J. (2017). Guided self-help for eating disorders: A systematic review and metaregression. *European Eating Disorder Review*, *25*, 148–164. doi: 10.1002/erv.2507

Vall, E., & Wade, T. D. (2015). Predictors of treatment outcome in individuals with eating disorders: A systematic review and meta-analysis. *The International Journal of Eating Disorders*, *48*, 946–971. doi: 10.1002/eat.22411

van Buuren, S. (2018). *Flexible imputation of missing data* (2nd ed.). Boca Raton, FL: CRC/Chapman & Hall.

Von Hippel, P. T. (2017). Regression with missing ys: An improved strategy for analyzing multiply imputed data. *Sociological Methodology*, *37*, 83–117. doi: 10.1111/j.1467-9531.2007.00180.x

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B*, *67*, 301–320.