EDITORIAL

WILEY

Machine learning to advance the prediction, prevention and treatment of eating disorders

Abstract

Machine learning approaches are just emerging in eating disorders research. Promising early results suggest that such approaches may be a particularly promising and fruitful future direction. However, there are several challenges related to the nature of eating disorders in building robust, reliable and clinically meaningful prediction models. This article aims to provide a brief introduction to machine learning and to discuss several such challenges, including issues of sample size, measurement, imbalanced data and bias; I also provide concrete steps and recommendations for each of these issues. Finally, I outline key outstanding questions and directions for future research in building, testing and implementing machine learning models to advance our prediction, prevention, and treatment of eating disorders.

Highlights

- Machine learning holds significant promise to advance eating disorders research
- Some key considerations for responsible machine learning application in eating disorders research include issues of sample size, measurement, imbalanced data and bias
- Future research should prioritize external validation of machine learning models

1 | INTRODUCTION

Research over the past century has made significant advances in understanding eating disorders (EDs), including identifying genetic (Duncan et al., 2017) and environmental (Culbert et al., 2015) risk factors, mapping their longitudinal course and outcome (Eddy et al., 2017), and developing evidence-based treatments (Wilson & Shafran, 2005). While these *explanatory* approaches have proliferated, less attention has focused on *predicting* EDs. Although many studies use longitudinal data to identify factors associated with EDs at later timepoints, traditional statistical methods are ill-suited to accurately forecast future outcomes. Reflecting this misalignment, conventional approaches have yielded risk factors that often fail to replicate (Vall & Wade, 2015) and response rates for even our best treatments remain low (Kass et al., 2013). Advanced computational methods such as machine learning (ML) are better equipped to advance ED prediction and treatment. These approaches are just emerging in ED research, with promising results (Cyr et al., 2018; Haynos et al., 2020; Sadeh-Sharvit et al., 2020).

As such, this article aims to highlight some key considerations for responsible ML application in EDs. Of note, this paper does not provide a comprehensive overview of all potential challenges, nor is it a primer or tutorial on ML methods. Rather, I hope to provide an accessible introduction to ML for ED researchers and highlight important points to consider before conducting such studies. Although many challenges reviewed are relevant to psychiatry broadly, all examples will be drawn from ED research and suggestions will be tailored to ED

---

investigators, given recent increasing interest in ML methods in this field.

2 | WHAT IS MACHINE LEARNING, AND HOW IS IT DIFFERENT FROM TRADITIONAL STATISTICAL METHODS?

Several excellent tutorials have reviewed ML (Dwyer et al., 2018; Yarkoni & Westfall, 2017) and interested readers can find even greater detail in textbooks (James et al., 2013; Kuhn & Johnson, 2013). Rather than reiterate these details, this section will briefly highlight key differences between traditional statistics and ML to familiarize readers with overarching concepts/terminology. Of note, ML techniques can be broadly grouped into supervised and unsupervised methods. While supervised methods (e.g., elastic net, support vector machines) are applied to *labelled* data to learn functions that best map input to output variables (e.g., predicting ED onset), unsupervised methods (e.g., clustering, dimension reduction) are applied to *unlabelled* data to identify underlying data structures (e.g., identifying subtypes of ED patients). This paper focuses on *supervised* methods which may be most applicable to questions in the field (e.g., predicting ED treatment outcomes).

In brief, there are two critical advantages of ML compared to traditional statistical *methods* and *algorithms*. First, regarding methods, traditional statistical frameworks aim to identify variables significantly (e.g., $p < 0.05$) associated with an outcome by fitting models on entire data sets. However, as any data set is influenced by underlying data-generating mechanisms *and* sampling error, such models will be overfit and likely generalize poorly. Supervised ML methods focus instead on optimizing predictive accuracy for new data, using *internal model validation* to evaluate performance of models (built on subsets of a full data set) in held-out validation data. A common approach is k-fold cross-validation, which involves (1) splitting a data set into k subsets, (2) training/tuning models on $k-1$ subsets, (3) testing on the k held-out subset and (4) repeating k times so each subset serves as the validation set.

Second, ML algorithms are more complex and flexible than traditional statistical models. Although hundreds of algorithms exist (see James et al., 2013), common algorithms in psychiatry can be broadly classified in three groups: regularization (e.g., lasso, elastic net), tree-based methods (e.g., random forests, decision trees) and support vector machines. These algorithms' flexibility is reflected in *tuning parameters,* which control aspects of the algorithm to obtain accurate and generalizable

predictions. For example, lasso can be viewed as an extension of linear regression, where coefficients (βs) are chosen to minimize sum-of-squared errors (SSE) between observed and predicted responses:

$$SSE = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2$$

where y_i = observed outcome and $\hat{y}_i$ = predicted response. This is prone to overfitting by inflating coefficient estimates and cannot handle high collinearity among predictors (Kuhn & Johnson, 2013), quite common in clinical data sets. Lasso addresses these limitations with a penalty term:

$$SSE_{LASSO} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i \right)^2 + \lambda \sum_{j=1}^{P} |\beta|$$

This forces the model to minimize SSE *and* the absolute value of all coefficients. Thus, lasso shrinks coefficients, protects against overfitting and increases likelihood of good prediction in new data. During internal cross-validation, researchers tune λ (controlling shrinkage/regularization) to optimize out-of-sample prediction. Unlike traditional statistical methods, most ML models have one or more tuning parameters.

3 | CHALLENGES AND RECOMMENDATIONS FOR MACHINE LEARNING IN EATING DISORDERS RESEARCH

The promise of ML in psychiatry has generated immense enthusiasm, with a proliferation of papers applying ML to predict depression (Chekroud et al., 2016; Kessler et al., 2016), schizophrenia (Koutsouleris et al., 2016), suicide and self-injury (Fox et al., 2019; Wang et al., 2021), anxiety (Boeke et al., 2020; Bokma et al., 2020) and obsessive–compulsive disorder (Lenhard et al., 2018); ML approaches are also emerging in ED research. For instance, ML has shown promise in predicting longitudinal illness course from self-report data (Haynos et al., 2020), detecting ED symptoms from Internet and/or social media data (Hwang et al., 2020; Sadeh-Sharvit et al., 2020), identifying deteriorating patients in need of medical attention (Ioannidis et al., 2020), and identifying biomarkers from neuroimaging data (Cerasa et al., 2015; Cyr et al., 2018). As these methods become more widely applied in ED research, this section describes challenges and considerations for responsible ML application, including sample size, measurement, imbalanced data

and bias. Although these are general issues in ML for psychiatry, this section focuses specifically on how these issues may arise in ED research, providing steps and recommendations for ED researchers to address each challenge.

3.1 | Sample size

3.1.1 | Challenge

Though small sample sizes are generally a concern in psychiatry (Reardon et al., 2019), they are particularly problematic in ML and can lead to overfitting and vastly inflated estimates of model accuracy. Current recommendations suggest avoiding ML with fewer than several hundred observations (Poldrack et al., 2020). However, it is notoriously difficult to recruit large ED samples, particularly when constrained to specific clinical presentations (e.g., anorexia nervosa [AN] or bulimia nervosa [BN]).

3.1.2 | Recommendation

There are several avenues for increasing sample size. First, online methods are time- and cost-effective for reaching large, representative samples (Smith et al., 2021). Second, collaborative multi-site studies could yield large samples; for instance, Haynos et al. (2020) leveraged data from the McKnight Longitudinal Study of Eating Disorders, a multisite study with 400+ participants. Third, ML models for other clinical concerns are often built using electronic health records with millions of patients (Barak-Corren et al., 2016); such data sources may prove useful for predicting EDs. Fourth, recent large-scale projects have embraced an open science model of publicly releasing data that can be harnessed for ML. For example, the Adolescent Brain Cognitive Development study is following >11,000 children for 10 years and has included ED assessments at each visit, with studies already applying ML (Adeli et al., 2020) and investigating ED (Rozzell et al., 2019) using this data.

When faced with smaller samples, researchers can make methodological choices to protect against overfitting. For instance, prediction accuracy may be evaluated using average cross-validated estimates rather than entirely separate, held-out data (Kuhn & Johnson, 2013). Nested cross-validation methods, which allow for inner and outer cross-validation loops, may also protect against overfitting, particularly in small data sets

(Cearns et al., 2019). Researchers should also avoid complex "black box" algorithms (e.g., deep neural networks) in favour of simpler and more interpretable options (e.g., elastic net, decision trees). Model robustness checks (e.g., comparing fivefold vs. 10-fold cross-validation accuracy) can also help ensure results are not spurious.

3.2 | Measurement

3.2.1 | Challenge

Questionable measurement practices limit our ability to build accurate ML models. As Jacobucci and Grimm (2020) put it: "throwing the same set of poorly measured variables that have been analysed before into ML algorithms is highly unlikely to produce new insights or findings." For example, take restrictive eating—this construct is often poorly defined, as it is unclear which behaviours it describes (e.g., fasting vs. reducing portion sizes vs. reducing carbohydrates). Because these behaviours may be predicted by different variables, a model considering them as equivalent may find it difficult to determine consistent relationships between predictor and outcome variables. Furthermore, many scales used to measure restrictive eating are misaligned with the theoretical construct (measuring cognitive restraint rather than actual restricting behaviours) (Haynos et al., 2015; Stice et al., 2004, 2007). If these scales emerged as important predictors of treatment response in ML models, interventions may inadvertently target the wrong construct.

3.2.2 | Recommendation

In a seminal paper, Flake and Fried (2020) pose six questions to guide transparent measurement practices. Their first questions involve defining and justifying constructs of interest, which are highly relevant for ML studies in ED research. For instance, should researchers build models predicting any ED diagnosis, a specific diagnosis (e.g., AN vs. BN), severity of overall ED psychopathology or specific symptoms? Typical outcomes include diagnoses and/or ED scale sum scores, but these have critical issues. For instance, EDs are characterized by high diagnostic instability and crossover (Eddy et al., 2008) with boundary conditions between diagnoses often governed solely by BMI; treating these as distinct categories may not be accurately "carving nature at its joints". Moreover, sum scores of global ED

psychopathology make the assumption that all symptoms are equally good and interchangeable indicators of EDs (Fried & Nesse, 2015). As a solution, researchers might consider defining and measuring outcomes at the symptom (rather than diagnostic or sum-score) level to provide more precise predictions about transdiagnostic ED features.

After identifying constructs of interest, it is critical to match the measurement to the construct. How should a construct such as restrictive eating be measured? There are many options ranging from self-report scales to behavioural tasks (e.g., test meal intake) to wearable biosensors, and many choices within each of those categories. While a large body of literature on construct validation (e.g., Clark & Watson, 2019; Strauss & Smith, 2009) cannot be adequately summarized here, researchers should clearly justify and transparently report details of measures used. Despite these measurement details not receiving as much attention in ML research as in more traditional clinical or psychometric studies, they are no less critical; even advanced ML algorithms and big data cannot overcome poor measurement. For example, above, I highlighted the potential of harnessing data from large-scale studies or multi-site collaborations. However, the quality of ED data in such studies relies on researchers' measurement choices. For instance, many popular diagnostic interviews (e.g., MINI, SCID-5) (First et al., 2015; Sheehan et al., 1998) screen for AN based solely on BMI and skip questions about restricting, fear of weight gain and body dissatisfaction if individuals are not underweight. This fails to detect many individuals with ED psychopathology, and may cast doubt on prediction models built using this data. Increased collaboration between ED researchers and investigators from other subfields in developing general psychopathology assessments and designing large-scale studies could improve the validity of ED data available in large data sets for ML.

3.3 | Imbalanced data

3.3.1 | Challenge

In ML, "imbalanced data" refers to a classification problem in which outcomes are not equally likely to occur, leading to overly optimistic performance estimates. For instance, an algorithm predicting suicide risk in the general population could achieve 99% accuracy by simply predicting that no one attempts suicide, but this would not be very useful. ED researchers are likely to encounter many cases of imbalanced data (e.g., predicting ED onset in the general population).

3.3.2 | Recommendation

Several sampling methods can address class imbalance during model training, including up-sampling (simulating or imputing minority class cases), down-sampling (reducing majority class cases) and hybrid methods (Kuhn & Johnson, 2013). It is imperative that subsampling occurs within the training set *only*. Test data sets should never be artificially balanced, but reflect "true" class imbalance anticipated in future data sets. Such subsampling methods were recently applied by Haynos et al. (2020) to predict outcomes including persistence of ED diagnosis and underweight BMI using data that showed significant class imbalance (e.g., 93.4% participants showing persistence of EDs at 1-year follow-up), allowing for more accurate estimates of model performance.

Model performance metrics are also important. As noted above, accuracy (proportion of all correctly classified cases) can be artificially optimistic. Fortunately, numerous other metrics are available (Figure 1). In psychiatry, area under the receiver operating characteristic curve (AUC) has been a standard metric. However, while many models have achieved high AUC (e.g., >0.80 in suicide prediction), their *positive predictive values* (PPV) remain untenably low. Indeed, a review of 64 suicide prediction studies found extremely low (<0.01) PPVs for most models, indicating 99 out of every 100 individuals predicted to attempt suicide will not (Belsher et al., 2019). This illustrates the importance of thoughtfully selecting metrics that capture multiple important characteristics. The internal model validation process requires researchers to select a metric to optimize; many models default to AUC, but other options may be more suitable for imbalanced data (e.g., area under the precision-recall curve).

3.4 | Bias

3.4.1 | Challenge

Though often heralded as 'objective', algorithms reflect the nature of the data used to train them, and often detect, learn, and perpetuate societal injustices (Birhane & Cummins, 2019). Recidivism algorithms are biased against Black defendants (Angwin et al., 2016), predictive policing software disproportionately target non-White, low-income communities (Lum & Isaac, 2016), and STEM advertisements are less likely to be displayed to women (Lambrecht & Tucker, 2019). There are substantial disparities in the ED field (Sonneville & Lipson, 2018), with White women vastly overrepresented across clinical and research settings. Training ML models using such data
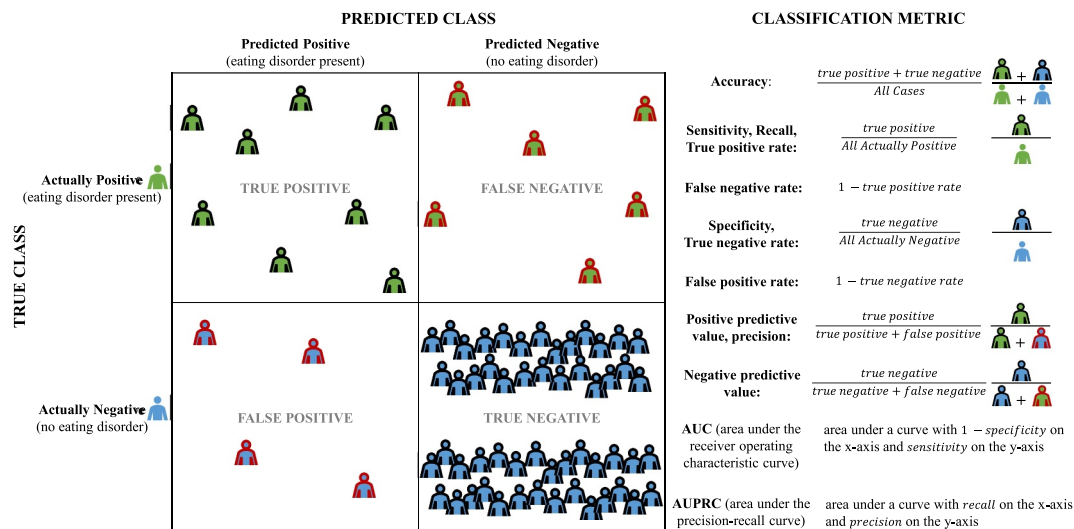
FIGURE 1 Metrics for measuring performance of classifiers [Colour figure can be viewed at wileyonlinelibrary.com]

sets risks perpetuating algorithmic bias with under-detection of EDs in marginalized communities.

3.4.2 | Recommendation

As ML models are built and tested, ED researchers must carefully collect and select data sets diverse in race, ethnicity, gender, socioeconomic status and other key demographics. Researchers should also monitor and test for algorithmic bias at all stages (Gianfrancesco et al., 2018), centre individuals from marginalized communities in the model-building and decision-making process, and consider whether any prediction model would combat or entrench existing injustice (Birhane & Cummins, 2019). Beyond technical solutions, researchers must also recognize the entire construct of EDs—including the criteria used to diagnose them, the institutions that provide treatment, and the clinical interventions offered—result from systems built on foundations of discrimination. For instance, ED screening and assessment tools were primarily developed for and validated among cisgender heterosexual White women and may not accurately capture ED symptoms in racial/ethnic minorities, sexual minorities, transgender, non-binary or gender nonconforming individuals, who have been historically underrepresented in ED research (Burke et al., 2020; Coelho et al., 2019). Addressing these systemic issues will require (but is not limited to) carefully designing and validating culturally sensitive assessments for individuals from historically marginalized backgrounds, intentionally recruiting diverse research samples and reducing barriers to treatment access. As

researchers of colour are more likely than White researchers to publish work highlighting race and recruit more participants of colour into studies (Roberts et al., 2020), increasing workforce diversity is also critical; current estimates demonstrate a striking lack of diversity in the ED field, with only 3% being Black/African-American, 3% Asian/Asian American and 14% Latino/Hispanic (Jennings Mathis et al., 2020).

Modelling ED data inherently involves working within a societal and political landscape. ML and other computational methods are not exempt by their data-driven nature and technical sophistication, but also emerge from history ripe with racism, sexism and white supremacy (Birhane & Guest, 2020). Thus, researchers must actively work to change the overwhelming whiteness and masculinity of computational sciences, which excludes minoritized groups. Increasing representation of researchers and clinicians from marginalized backgrounds is critical for advancing assessment, prediction and treatment of EDs in diverse populations.

4 | CONCLUSION AND FUTURE DIRECTIONS

ML is rapidly emerging in the ED field, with promising early results. Where should we go from here? There are several important future directions. First, external model validation is critical before clinical implementation. This is perhaps best illustrated via COVID-19 prediction models: when submitting 22 published models to external validation, *none* beat simple univariate predictors (Gupta

et al., 2020). Despite challenges inherent in collecting large clinical data sets, external model validation is a crucial step.

Although this article focuses on data-driven computational approaches, I also believe theory-driven methods are needed for making meaningful progress. One exciting method involves formalizing theories as mathematical models. Although we have numerous influential ED theories (e.g., cognitive-behavioural theory) (Fairburn, 2008), all theories to date have been instantiated *verbally*, rendering them underspecified due to the inherent imprecision of language. Other areas of science (e.g., ecology, physics) instantiate theories using formal mathematical notation (e.g., differential equations) and computer code, providing greater precision and specificity. Formal theories also allow for simulation of theory-implied data, which can be used to revise theories based on their ability to capture and recapitulate empirical data patterns. These methods have recently been applied to formalize a panic disorder theory (Robinaugh et al., 2019), with emerging efforts to also formalize theories of suicide and EDs, representing an exciting new direction in psychopathology research.

Both theory- and data-driven computational methods are critical for advancing ED research. As ML methods are increasingly applied, I encourage researchers to collaborate with interdisciplinary experts, including computer scientists, statisticians, clinical practitioners, ethicists and those with lived experience. Diverse perspectives are crucial for advancing science and practice, and voices of those who have been historically excluded from the ED field should be included and valued throughout. Collaboration and careful thinking at each step in the data collection, model building, testing and implementation process can allow for meaningful advances in the prediction, prevention and treatment of EDs to reduce their global burden.

KEYWORDS
computational methods, eating disorders, machine learning, prediction

CONFLICTS OF INTEREST
I report no conflicts of interest.

Shirley B. Wang [ID]

Department of Psychology, Harvard University, Cambridge, Massachusetts, USA

Correspondence
Shirley B. Wang, Department of Psychology, Harvard University, 33 Kirkland St, Cambridge, MA 02138, USA.
Email: shirleywang@g.harvard.edu

DATA AVAILABILITY STATEMENT
Data sharing not applicable to this article as no data sets were generated or analysed for this paper.

ORCID
Shirley B. Wang [ID] https://orcid.org/0000-0002-8583-3014

REFERENCES
Adeli, E., Zhao, Q., Zahr, N. M., Goldstone, A., Pfefferbaum, A., Sullivan, E. V., & Pohl, K. M. (2020). Deep learning identifies morphological determinants of sex differences in the pre-adolescent brain. *NeuroImage*, *223*, 117293. https://doi.org/10.1016/j.neuroimage.2020.117293

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias risk assessments in criminal sentencing*. ProPublica. May, 23.

Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Dai, Y., Perlis, R. H., Nock, M. K., Smoller, J. W., & Reis, B. Y. (2016). Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry*, *174*(2), 154–162. https://doi.org/10.1176/appi.ajp.2016.16010077

Belsher, B. E., Smolenski, D. J., Pruitt, L. D., Bush, N. E., Beech, E. H., Workman, D. E., Morgan, R. L., Evatt, D. P., Tucker, J., & Skopp, N. A. (2019). Prediction models for suicide attempts and deaths: A systematic review and simulation. *JAMA Psychiatry*, *76*(6), 642. https://doi.org/10.1001/jamapsychiatry.2019.0174

Birhane, A., & Cummins, F. (2019). Algorithmic injustices: Towards a relational ethics. *ArXiv:1912.07376 [Cs]*. https://arxiv.org/abs/1912.07376

Birhane, A., & Guest, O. (2020). Towards decolonising computational sciences. *ArXiv:2009.14258 [Cs]*. https://arxiv.org/abs/2009.14258

Boeke, E. A., Holmes, A. J., & Phelps, E. A. (2020). Toward robust anxiety biomarkers: A machine learning approach in a large-scale sample. *Biological Psychiatry: Cognitive Neuroscience*

and Neuroimaging, 5(8), 799–807. https://doi.org/10.1016/j.bpsc.2019.05.018

Bokma, W. A., Zhutovsky, P., Giltay, E. J., Schoevers, R. A., Penninx, B. W. J. H., van Balkom, A. L. J. M., Batelaan, N. M., & van Wingen, G. A. (2020). Predicting the naturalistic course in anxiety disorders using clinical and biological markers: A machine learning approach. *Psychological Medicine*, 1–11. https://doi.org/10.1017/S0033291720001658

Burke, N. L., Schaefer, L. M., Hazzard, V. M., & Rodgers, R. F. (2020). Where identities converge: The importance of intersectionality in eating disorders research. *International Journal of Eating Disorders*, 53(10), 1605–1609. https://doi.org/10.1002/eat.23371

Cearns, M., Hahn, T., & Baune, B. T. (2019). Recommendations and future directions for supervised machine learning in psychiatry. *Translational Psychiatry*, 9(1), 271. https://doi.org/10.1038/s41398-019-0607-2

Cerasa, A., Castiglioni, I., Salvatore, C., Funaro, A., Martino, I., Alfano, S., Donzuso, G., Perrotta, P., Gioia, M. C., Gilardi, M. C., & Quattrone, A. (2015). Biomarkers of eating disorders using support vector machine analysis of structural neuroimaging data: Preliminary results. *Behavioural Neurology*, 10, e924814. https://doi.org/10.1155/2015/924814

Chekroud, A. M., Zotti, R. J., Shehzad, Z., Gueorguieva, R., Johnson, M. K., Trivedi, M. H., Cannon, T. D., Krystal, J. H., & Corlett, P. R. (2016). Cross-trial prediction of treatment outcome in depression: A machine learning approach. *The Lancet Psychiatry*, 3(3), 243–250. https://doi.org/10.1016/S2215-0366(15)00471-X

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological Assessment*, 31(12), 1412–1427. https://doi.org/10.1037/pas0000626

Coelho, J. S., Suen, J., Clark, B. A., Marshall, S. K., Geller, J., & Lam, P.-Y. (2019). Eating disorder diagnoses and symptom presentation in transgender youth: A scoping review. *Current Psychiatry Reports*, 21(11), 107. https://doi.org/10.1007/s11920-019-1097-x

Culbert, K. M., Racine, S. E., & Klump, K. L. (2015). Research Review: What we have learned about the causes of eating disorders – a synthesis of sociocultural, psychological, and biological research. *Journal of Child Psychology and Psychiatry*, 56(11), 1141–1164. https://doi.org/10.1111/jcpp.12441

Cyr, M., Yang, X., Horga, G., & Marsh, R. (2018). Abnormal frontostriatal activation as a marker of threshold and subthreshold Bulimia Nervosa. *Human Brain Mapping*, 39(4), 1796–1804. https://doi.org/10.1002/hbm.23955

Duncan, L., Yilmaz, Z., Gaspar, H., Walters, R., Goldstein, J., Anttila, V., Bulik-Sullivan, B., Ripke, S., Thornton, L., Hinney, A., Daly, M., Sullivan, P. F., Zeggini, E., Breen, G., Bulik, C. M., Duncan, L., Yilmaz, Z., Gaspar, H., Walters, R., ... Bulik, C. M. (2017). Significant locus and metabolic genetic correlations revealed in genome-wide association study of anorexia nervosa. *American Journal of Psychiatry*, 174(9), 850–858. https://doi.org/10.1176/appi.ajp.2017.16121402

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry.

Annual Review of Clinical Psychology, 14(1), 91–118. https://doi.org/10.1146/annurev-clinpsy-032816-045037

Eddy, K. T., Dorer, D. J., Franko, D. L., Tahilani, K., Thompson-Brenner, H., & Herzog, D. B. (2008). Diagnostic crossover in anorexia nervosa and bulimia nervosa: Implications for DSM-V. *American Journal of Psychiatry*, 165(2), 245–250. https://doi.org/10.1176/appi.ajp.2007.07060951

Eddy, K. T., Tabri, N., Thomas, J. J., Murray, H. B., Keshaviah, A., Hastings, E., Edkins, K., Krishna, M., Herzog, D. B., Keel, P. K., & Franko, D. L. (2017). Recovery from anorexia nervosa and bulimia nervosa at 22-year follow-up. *The Journal of Clinical Psychiatry*, 78(02), 184–189. https://doi.org/10.4088/JCP.15m10393

Fairburn, C. G. (2008). *Cognitive behavior therapy and eating disorders.* https://www.semanticscholar.org/paper/Cognitive-Behavior-Therapy-and-Eating-Disorders-Fairburn/c25dfa46ef03eced6c48e35e0acf50bcb7e1538f

First, M., Williams, J., Karg, R., Spitzer, R., & others (2015). *Structured clinical interview for DSM-5—Research version (SCID-5 for DSM-5, research version; SCID-5-RV)* (pp. 1–94). American Psychiatric Association.

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456–465. https://doi.org/10.1177/2515245920952393

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. https://doi.org/10.1037/ccp0000421

Fried, E. I., & Nesse, R. M. (2015). Depression sum-scores don't add up: Why analyzing specific depression symptoms is essential. *BMC Medicine*, 13(1), 72. https://doi.org/10.1186/s12916-015-0325-4

Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine*, 178(11), 1544. https://doi.org/10.1001/jamainternmed.2018.3763

Gupta, R. K., Marks, M., Samuels, T. H. A., Luintel, A., Rampling, T., Chowdhury, H., Quartagno, M., Nair, A., Lipman, M., Abubakar, I., van Smeden, M., Wong, W. K., Williams, B., & Noursadeghi, M. (2020). Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: An observational cohort study. *European Respiratory Journal*, 56, 2003498. https://doi.org/10.1183/13993003.03498-2020

Haynos, A. F., Field, A. E., Wilfley, D. E., & Tanofsky-Kraff, M. (2015). A novel classification paradigm for understanding the positive and negative outcomes associated with dieting: PSYCHO-BEHAVIORAL DIETING PARADIGM. *International Journal of Eating Disorders*, 48(4), 362–366. https://doi.org/10.1002/eat.22355

Haynos, A. F., Wang, S. B., Lipson, S., Peterson, C. B., Mitchell, J. E., Halmi, K. A., Agras, W. S., & Crow, S. J. (2020). Machine learning enhances prediction of illness course: A longitudinal

study in eating disorders. *Psychological Medicine*, *51*, 1–1402. https://doi.org/10.1017/S0033291720000227

Hwang, Y., Kim, H. J., Choi, H. J., & Lee, J. (2020). Exploring abnormal behavior patterns of online users with emotional eating behavior: Topic modeling study. *Journal of Medical Internet Research*, *22*(3), e15700. https://doi.org/10.2196/15700

Ioannidis, K., Serfontein, J., Deakin, J., Bruneau, M., Ciobanca, A., Holt, L., Snelson, S., & Stochl, J. (2020). Early warning systems in inpatient anorexia nervosa: A validation of the MARSIPAN-based modified early warning system. *European Eating Disorders Review*, *28*(5), 551–558. https://doi.org/10.1002/erv.2753

Jacobucci, R., & Grimm, K. J. (2020). Machine learning and psychological research: The unexplored effect of measurement. *Perspectives on Psychological Science*, *15*(3), 809–816.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 103). Springer. https://doi.org/10.1007/978-1-4614-7138-7

Jennings Mathis, K., Anaya, C., Rambur, B., Bodell, L. P., Graham, A. K., Forney, K. J., Anam, S., & Wildes, J. E. (2020). Workforce diversity in eating disorders: A multi-methods study. *Western Journal of Nursing Research*, *42*(12), 1068–1077. https://doi.org/10.1177/0193945920912396

Kass, A. E., Kolko, R. P., & Wilfley, D. E. (2013). Psychological treatments for eating disorders. *Current Opinion in Psychiatry*, *26*(6), 549–555. https://doi.org/10.1097/YCO.0b013e328365a30e

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., Ebert, D. D., Hwang, I., Li, J., De Jonge, P., Nierenberg, A. A., Petukhova, M. V., Rosellini, A. J., Sampson, N. A., Schoevers, R. A., Wilcox, M. A., & 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*(10), 1366–1371. https://doi.org/10.1038/mp.2015.198

Koutsouleris, N., Kahn, R. S., Chekroud, A. M., Leucht, S., Falkai, P., Wobrock, T., Derks, E. M., Fleischhacker, W. W., & Hasan, A. (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*(10), 935–946. https://doi.org/10.1016/S2215-0366(16)30171-7

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer-Verlag. https://doi.org/10.1007/978-1-4614-6849-3

Lambrecht, A., & Tucker, C. (2019). Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads. *SSRN Electronic Journal*, *65*(7), 2966–2981. https://doi.org/10.2139/ssrn.2852260

Lenhard, F., Sauer, S., Andersson, E., Månsson, K. N., Mataix-Cols, D., Rück, C., & Serlachius, E. (2018). Prediction of outcome in internet-delivered cognitive behaviour therapy for paediatric obsessive-compulsive disorder: A machine learning approach. *International Journal of Methods in Psychiatric Research*, *27*(1), e1576. https://doi.org/10.1002/mpr.1576

Lum, K., & Isaac, W. (2016). To predict and serve? *Significance*, *13*(5), 14–19. https://doi.org/10.1111/j.1740-9713.2016.00960.x

Poldrack, R. A., Huckins, G., & Varoquaux, G. (2020). Establishment of best practices for evidence for prediction: A review. *JAMA Psychiatry*, *77*(5), 534–540. https://doi.org/10.1001/jamapsychiatry.2019.3671

Reardon, K. W., Smack, A. J., Herzhoff, K., & Tackett, J. L. (2019). An N-pact factor for clinical psychological research. *Journal of Abnormal Psychology*, *128*(6), 493–499. https://doi.org/10.1037/abn0000435

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*(6), 1295–1309. https://doi.org/10.1177/1745691620927709

Robinaugh, D., Haslbeck, J. M. B., Waldorp, L., Kossakowski, J. J., Fried, E. I., Millner, A., McNally, R. J., van Nes, E. H., Scheffer, M., Kendler, K. S., & Borsboom, D. (2019). Advancing the network theory of mental disorders: A computational model of panic disorder [preprint]. *PsyArXiv*. https://doi.org/10.31234/osf.io/km37w

Rozzell, K., Moon, D. Y., Klimek, P., Brown, T., & Blashill, A. J. (2019). Prevalence of eating disorders among US children aged 9 to 10 Years: Data from the adolescent brain cognitive development (ABCD) study. *JAMA Pediatrics*, *173*(1), 100–101. https://doi.org/10.1001/jamapediatrics.2018.3678

Sadeh-Sharvit, S., Fitzsimmons-Craft, E. E., Taylor, C. B., & Yom-Tov, E. (2020). Predicting eating disorders from Internet activity. *International Journal of Eating Disorders*, *53*(9), 1526–1533. https://doi.org/10.1002/eat.23338

Sheehan, D. V., Lecrubier, Y., Sheehan, K. H., Amorim, P., Janavs, J., Weiller, E., Hergueta, T., Baker, R., & Dunbar, G. C. (1998). The Mini-International Neuropsychiatric Interview (M.I.N.I): The development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. *The Journal of Clinical Psychiatry*, *59*(Suppl 20), 22–33.

Smith, D. M. Y., Lipson, S. M., Wang, S. B., & Fox, K. R. (2021). Online methods in adolescent self-injury research: Challenges and recommendations. *Journal of Clinical Child & Adolescent Psychology*, 1–12. https://doi.org/10.1080/15374416.2021.1875325

Sonneville, K. R., & Lipson, S. K. (2018). Disparities in eating disorder diagnosis and treatment according to weight status, race/ethnicity, socioeconomic background, and sex among college students. *International Journal of Eating Disorders*, *51*(6), 518–526. https://doi.org/10.1002/eat.22846

Stice, E., Cooper, J., Schoeller, D., Tappe, K., & Lowe, M. (2007). Are dietary restraint scales valid measures of moderate- to long-term dietary restriction? Objective biological and behavioral data suggest not. *Psychological Assessment*, *19*, 449–458. https://doi.org/10.1037/1040-3590.19.4.449

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*.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology*, *5*, 1–25. https://doi.org/10.1146/annurev.clinpsy.032408.153639

Vall, E., & Wade, T. D. (2015). Predictors of treatment outcome in individuals with eating disorders: A systematic review and meta-analysis. *International Journal of Eating Disorders*, *48*(7), 946–971. https://doi.org/10.1002/eat.22411

Wang, S. B., Coppersmith, D. D. L., Kleiman, E. M., Bentley, K. H., Millner, A. J., Fortgang, R., Mair, P., Dempsey, W., Huffman, J. C., & Nock, M. K. (2021). A pilot study using frequent inpatient Assessments of suicidal thinking to predict short-term postdischarge suicidal behavior. *JAMA Network Open*, *4*(3), e210591. https://doi.org/10.1001/jamanetworkopen.2021.0591

Wilson, G. T., & Shafran, R. (2005). Eating disorders guidelines from NICE. *The Lancet*, *365*(9453), 79–81. https://doi.org/10.1016/S0140-6736(04)17669-1

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393