EDITORIAL

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

Abbreviations: AN, anorexia nervosa; AUC, area under the receiver operating characteristic curve; BN, bulimia nervosa; ED, eating disorder; ML, machine learning; PPV, positive predictive value; SSE, sum-of-squared errors.
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 labeled 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 $$\lambda$$ (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 (Sonveille & Lipson, 2018), with White women vastly overrepresented across clinical and research settings. Training ML models using such data
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

**ACKNOWLEDGEMENTS**
I thank Kathryn Fox, Ruben Van Genugten, Rowan Hunt, Patrick Mair and Matthew Nock for their helpful discussion and feedback on earlier drafts of this manuscript. Shirley B. Wang, is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1745303. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Science Foundation.

**FUNDING**
Shirley B. Wang, is supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1745303. The content is solely the responsibility of the author and does not necessarily represent the official views of the National Science Foundation.

**CONFLICTS OF INTEREST**
I report no conflicts of interest.

Shirley B. Wang

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 https://orcid.org/0000-0002-8583-3014

**REFERENCES**


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

