Psychological Medicine

cambridge.org/psm

Predicting adolescent depression and anxiety from multi-wave longitudinal data using machine learning

Mariah T. Hawes[1], H. Andrew Schwartz[2], Youngseo Son[2] and Daniel N. Klein[1]

[1]Department of Psychology, Stony Brook University, Stony Brook, NY, USA and [2]Department of Computer Science, Stony Brook University, Stony Brook, NY, USA

Original Article

Cite this article: Hawes MT, Schwartz HA, Son Y, Klein DN (2023). Predicting adolescent depression and anxiety from multi-wave longitudinal data using machine learning. *Psychological Medicine* **53**, 6205–6211. https://doi.org/10.1017/S0033291722003452

Received: 25 March 2022
Revised: 3 October 2022
Accepted: 14 October 2022
First published online: 15 November 2022

Key words:
adolescence; anxiety; depression; longitudinal; machine learning; risk assessment

Author for correspondence:
Mariah T. Hawes,
E-mail: hawes2mt@gmail.com

Abstract

Background. This study leveraged machine learning to evaluate the contribution of information from multiple developmental stages to prospective prediction of depression and anxiety in mid-adolescence.

Methods. A community sample ($N = 374$; 53.5% male) of children and their families completed tri-annual assessments across ages 3–15. The feature set included several important risk factors spanning psychopathology, temperament/personality, family environment, life stress, interpersonal relationships, neurocognitive, hormonal, and neural functioning, and parental psychopathology and personality. We used canonical correlation analysis (CCA) to reduce the large feature set to a lower dimensional space while preserving the longitudinal structure of the data. Ablation analysis was conducted to evaluate the relative contributions to prediction of information gathered at different developmental periods and relative to previous disorder status (i.e. age 12 depression or anxiety) and demographics (sex, race, ethnicity).

Results. CCA components from individual waves predicted age 15 disorder status better than chance across ages 3, 6, 9, and 12 for anxiety and 9 and 12 for depression. Only the components from age 12 for depression, and ages 9 and 12 for anxiety, improved prediction over prior disorder status and demographics.

Conclusions. These findings suggest that screening for risk of adolescent depression can be successful as early as age 9, while screening for risk of adolescent anxiety can be successful as early as age 3. Assessing additional risk factors at age 12 for depression, and going back to age 9 for anxiety, can improve screening for risk at age 15 beyond knowing standard demographics and disorder history.

Introduction

Depression and anxiety disorders are among the most common mental disorders and are leading contributors to global disease burden (GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, 2018). Rates of depression and anxiety increase dramatically during adolescence, portending worse outcomes than onsets in adulthood (Beesdo, Knappe, & Pine, 2009; Fleisher & Katz, 2001; Kessler, Chiu, Demler, & Walters, 2005). However, predicting which individuals will experience depression and anxiety in adolescence remains an extremely difficult task. There is increasing recognition of the immense complexity of psychopathology, necessitating shifting away from simple etiological models and toward a complex dynamic systems perspective that recognizes that mental disorders arise from the interplay of numerous interacting components on multiple levels of analysis (Fried & Robinaugh, 2020). This shift in conceptualization, however, is not yet reflected in the dominant statistical paradigms used to study mental illness (Dwyer, Falkai, & Koutsouleris, 2018).

Traditional methods for examining prediction of disorder onsets (e.g. low dimensional linear and logistic regression) are highly limited in the complexity they can accommodate, both in terms of the number of variables and types of relationships (e.g. non-linear, multi-way interactions) that can be modeled simultaneously (Dwyer et al., 2018). Specifically, these methods are liable to overfitting as complexity increases, that is, they produce models that increasingly reflect the idiosyncratic characteristics of a particular sample and thus do not generalize well to other samples (Whelan & Garavan, 2014). Machine learning (ML) methods, on the other hand, are uniquely suited to the task of prediction (Yarkoni & Westfall, 2017). ML is an umbrella term that subsumes a range of flexible mathematical techniques that identify patterns in a (training) dataset with the central goal of producing a model that maximizes prediction of new (test) data. Evaluation of models on their ability to accurately predict out-of-sample data places generalizability and replicability at the core of ML (Coutanche & Hallion, 2019).

ML has the potential to make significant contributions to predicting depressive and anxiety disorders, accounting for the manifold relationships between the biological, cognitive,

© The Author(s), 2022. Published by Cambridge University Press

emotional, interpersonal, and environmental factors that give rise to affective psychopathology. Further, the prioritization of generalizability is critical to translating research findings to real-world applications in clinical settings and to improving replication of findings in a field facing a 'replicability crisis' (Tackett, Brandes, King, & Markon, 2019). This potential is increasingly being realized, as there are already a number of ML applications to depression and (to a lesser extent) anxiety disorders (Shatte, Hutchinson, & Teague, 2019), however, the existing literature remains limited in numerous ways.

First, many studies have trained their algorithms on low quality outcome measures (e.g. a self-report questionnaire; Andersson, Bathula, Iliadis, Walter, & Skalkidou, 2021; Su, Zhang, He, & Chen, 2021). Additionally, despite the capacity of ML to accommodate a large number of features (i.e. variables), most studies have considered a limited range of potentially relevant factors. For example, a number of studies have exclusively used neuroimaging data (e.g. Sato et al., 2015), or medical records (e.g. Nemesure, Heinz, Huang, & Jacobson, 2021), with few combining several domains of interest (e.g. clinical records, personality measures, cognitive tests, and biological data). Both of these limitations are, at least in part, a consequence of ML requiring large samples, for which thorough diagnostics (e.g. by a trained interviewer), and extensive assessment of relevant risk factors, is less feasible.

Most existing ML applications in depression and anxiety have used fully cross-sectional data, seeking to improve detection of current depression or anxiety disorders (Guntuku, Yaden, Kern, Ungar, & Eichstaedt, 2017; Kumar, Garg, & Garg, 2020; Liu, Hankey, Cao, & Chokka, 2021). These studies are helpful for aiding in differential diagnosis, flagging individuals already in the health care system or identifying struggling individuals through social media posts; however, they are unlikely to generalize to prediction of future depression and anxiety, as the feature sets (i.e. the collection of predictor variables) include factors that co-occur with or are a consequence of the disorder. Building toward prevention requires training models on data that temporally precede disorder occurrence. A handful of studies have used ML methods to predict future depression or anxiety, although often in specific subgroups (e.g. postpartum; Andersson *et al.* 2021; Zhang, Wang, Hermann, Joly, & Pathak, 2021), over relatively short intervals (e.g. 6 months to 1 year; Bellón *et al.* 2011; Eichstaedt *et al.* 2018; King *et al.* 2008), and, to our knowledge, exclusively in adult samples (Kessler et al., 2016; Rosellini et al., 2020; Wang et al., 2014). Most individuals who will meet criteria for a mental disorder do so by the age of 18, with most first onsets occurring during adolescence (Caspi et al., 2020), so adult samples are similarly limited by either confounding risk factors with consequences of prior occurrence of mental illness or, when focusing exclusively on first onset cases, exclude the majority of individuals who will experience a mental illness in their lifetime.

A final and critical gap in ML applications to depression and anxiety disorders thus far is that the appropriate timing of risk assessment (i.e. at what age, how proximal to disorder occurrence) has been virtually unexplored. Determining the optimal timing of risk screening requires longitudinal assessment across more than two waves of data collection. To our knowledge, no prior study has used features assessed across multiple waves of data collection to predict depression or anxiety at a future wave.

Depression onsets peak in mid-adolescence, and while some anxiety disorders onset in childhood, rates of anxiety also increase drastically during adolescence, with some anxiety disorder onsets occurring during this period-early adulthood (e.g. social anxiety, panic, agoraphobia, generalized anxiety, and obsessive-compulsive disorder; Campbell, Brown, & Grisham, 2003; de Lijster *et al.* 2017; Kessler *et al.* 2005). Adolescent depression and anxiety are associated with a host of negative outcomes, including increased risk of disorder persistence and reoccurrence, increased comorbidity, and worse psychosocial functioning later in life (Essau, Lewinsohn, Olaya, & Seeley, 2014; Fleisher & Katz, 2001; McLeod, Horwood, & Fergusson, 2016; Naicker, Galambos, Zeng, Senthilselvan, & Colman, 2013). Therefore, identifying individuals at risk of experiencing depression and anxiety during this developmental period is particularly critical. In the current study, we used ML to prospectively predict cases of depression and anxiety disorders during adolescence in an unselected community sample ($N = 374$). The feature set included a diverse and large number of potentially important risk factors spanning psychopathology, temperament/personality, family environment, life stress, interpersonal relationships, neurocognitive, hormonal, and neural functioning, and parental psychopathology and personality assessed at 3-year intervals across development from ages 3 to 12. The outcomes – diagnoses of depression and anxiety disorder at the age 15 wave – were assessed through a semi-structured diagnostic interview conducted by trained interviewers.

The primary purpose of this study was to leverage ML methods to evaluate the contribution of information from multiple developmental stages across childhood and early adolescence to prediction of depression and anxiety in mid-adolescence. This allowed us to address questions about the timing of risk assessment, such as how early risk assessment can be fruitful and whether longitudinal assessment provides substantially better prediction of risk than a single assessment at a key developmental stage. We additionally sought to explore the upper bounds of prediction that can be achieved when such a large number of highly relevant features spanning multiple domains are considered and to assess the incremental gains in prediction afforded by including such a volume of information (i.e. features) over a standard minimal risk assessment (specifically, recent disorder history and basic demographics).

To accomplish these goals, we compared prediction of disorder status at age 15 from disorder status at age 12 along with demographics, both alone and in combination with extensive risk factor data from individual prior waves and combining risk factor data across multiple prior waves. To meet the challenge of working with a large feature set spanning multiple waves, we used canonical correlation analysis (CCA), a multi-view dimensionality reduction technique that preserves the longitudinal structure of the data (Witten, Tibshirani, & Hastie, 2009).

Methods

Procedures

Data were from an ongoing study of the development of psychopathology that has followed children and their families triannually since the participating child was 3 years old (Klein & Finsaas, 2017). Initial recruitment of families with a 3-year-old child living within a 20-mile radius of Stony Brook, New York was conducted via commercial mailing lists. At each wave, families were invited to the lab to complete a battery of assessments. When lab visits were not feasible, questionnaires and interviews were completed remotely. A parent provided written informed

https://doi.org/10.1017/S0033291722003452 Published online by Cambridge University Press

consent at the start of each assessment and the child provided assent starting at the age 9 wave. The Stony Brook University Institutional Review Board approved all study procedures.

Participants

Families were eligible to participate if the primary caretaker spoke English and was the child's biological parent, and if the child did not have a significant medical disorder or developmental disability. Of the total of 559 participants, 374 were included in the current analyses (data exclusion described below). Included participants were predominantly male (53.5%), White (94.7%), and non-Hispanic (90.9%). Excluded participants did not differ significantly from included participants in demographic profile.

Measures

An online Supplementary excel file titled 'List of Features' contains the full list of features included from each wave. Briefly, the features covered a range of important domains, including clinical features (e.g. diagnoses and dimensional symptom scores of all common mental disorders), temperament and personality (e.g. behavioral inhibition, negative and positive emotionality, effortful control, intolerance of uncertainty, rumination), environmental factors (e.g. stressful life events, bullying, parental criticism and support), biological/neurocognitive factors (e.g. pubertal hormones, morning and evening cortisol levels, resting electroencephalography and event-related potentials in a variety of emotion-relevant tasks, executive functions, attentional and memory biases) and a number of parental factors (e.g. parental psychopathology and personality). Each assessment wave included parent and, starting at age 9, child interviews and questionnaires, saliva samples, and laboratory behavioral and neural measures. The features were not identical across waves. This is typical of developmental research, as the relevance of risk factors and appropriateness of modalities of measurement (e.g. self-report *v.* parent report) changes across development, but nevertheless leads to some confounding of age with differences in features.

Prior (age 12) and outcome (age 15) depression and anxiety were diagnosed with a semi-structured diagnostic interview, the Kiddie Schedule for Affective Disorders and Schizophrenia-Present and Lifetime version (K-SADS-PL; Axelson, Birmaher, Zelazny, Kaufman, & Gill, 2009). Diagnoses were based on the interval since the previous assessment (e.g. since the age 12 wave at the age 15 wave) and were used for the baseline and outcome variables. Doctoral students in clinical psychology and master's-level clinicians administered the K-SADS first to the parent (about the child) and then to the child (about themselves). Parent and child report of symptoms were combined into summary ratings, which were used to assign a diagnosis based on either the Diagnostic and Statistical Manual for Mental Disorders 4th (DSM-IV; American Psychiatric Association, 1994) or 5th (DSM-5; American Psychiatric Association, 2013) edition criteria. All cases with a suspected diagnosis were reviewed in a case conference co-led by a child psychiatrist and a clinical psychologist. Diagnosis of depression included major depressive disorder, dysthymic disorder, and depressive disorder-not otherwise specified (NOS; DSM-IV) or other specified depressive disorder (DSM-5); diagnosis of anxiety disorder included specific phobia, social phobia (DSM-IV) or social anxiety (DSM-5), agoraphobia, and panic, generalized anxiety, separation anxiety, obsessive compulsive, post-traumatic stress, and acute stress disorder, and anxiety disorder-NOS (DSM-IV) or other specified anxiety disorder (DSM-5). Interviewers independently rated videotaped interviews to assess inter-rater reliability (kappa = 0.72 and 0.91 for depression and anxiety disorders, respectively).

Data analysis

Preprocessing

The investigators preselected a subset of 429 features from all available data that comprehensively covered the range of constructs assessed in the study while minimizing redundancy (e.g. selecting a scale total score over correlated lower-order subscales). ML methods require complete data, so we excluded cases and features missing ⩾80% of data as well as cases without both outcome variables. Remaining missing values for features were imputed, using the mean for numerical features and the mode for categorical features. Categorical features with more than two levels were transformed into separate dummy-coded variables for each level (final feature set = 517).

Machine learning

This resulted in a feature set that was still very large relative to the number of observations. To mitigate multicollinearity and reduce the number of supervised model parameters (Hastie, Tibshirani, & Jermone, 2009), we used dimensionality reduction to reduce the features to 10 dimensions per wave while maximally preserving information (i.e. variance). To preserve the longitudinal structure (i.e. that groups of variables came from the same wave), we used a multi-view dimensionality reduction, CCA (Witten et al., 2009), to create components (i.e. linear combinations of related features) within a wave that were maximally correlated across waves. This approach allowed us to have a single model yet keep each wave's low dimensional components separate from other waves in time ('views'). We extracted 10 components per wave, as multiples of 10 are conventional and more than 10 components per wave would result in large multi-wave models (>40 features) at risk of overfitting in our relatively small sample.

Classification was performed using L2-penalized logistic regression, a regularization method that shrinks the regression coefficients by imposing a penalty on the maximum likelihood parameter estimate based on the squared magnitude of the coefficients as is standard in ML to guard against overfitting (James, Witten, Hastie, & Tibshirani, 2013). The L2 penalty, also known as the 'ridge', adjusts for the collinearity between variables which has been found beneficial especially in longitudinal studies where multiple waves of similar variables covary (Eliot, Ferguson, Reilly, & Foulkes, 2011; Miché et al., 2020). Algorithms were trained on depression and anxiety outcomes using *k*-fold cross-validation (CV) with 10 folds. Briefly, 10-fold CV partitions all observations into 10 roughly equally sized, mutually exclusive, and randomized subgroups (folds). The algorithm is trained on 9/10 of the folds and the resulting model is used to predict the fold that was left out (i.e. the test set). Thus, the data used to train the algorithm are never contaminated with information about the data when their accuracy is evaluated. This process is repeated until predictions have been made for all 10 folds (Koul, Becchio, & Cavallo, 2018). Performance was indexed using the area under the receiver operating characteristics curve (AUC). AUC values were computed for each fold and then averaged across folds to produce a more stable estimate of out-of-sample performance.

https://doi.org/10.1017/S0033291722003452 Published online by Cambridge University Press

Ablation analysis

We conducted an ablation analysis to evaluate the relative contributions to prediction of information gathered at different developmental periods and relative to prior disorder status (i.e. age 12 diagnosis of depression or anxiety disorder, depending on the outcome). Ablation analysis is a process of training algorithms on different configurations of features and then comparing performance metrics across configurations to assess which features are contributing to prediction (Fawcett & Hoos, 2016). We compared models containing CCA components from each individual wave (i.e. age 12, 9, 6, and 3) and in cumulative combinations (i.e. ages 12–3, 12–6, and 12–9) alongside prior disorder status and demographics (A12 Dx + Demos) to a model containing just A12 + Demos. Demographic features included sex, race, and ethnicity. To determine whether these comparisons had statistically significant differences in AUCs, we used a permutation test.

Sensitivity analyses

To ensure that our conclusions were robust to choice of classifier, we tested two additional classification algorithms: random forests and neural networks. Additionally, to demonstrate the advantage L2 penalization affords to prediction, we also fit models using traditional logistic regression as a benchmark of conventional statistical approaches in psychology and psychiatry. Details and results of these analyses are presented in the online Supplementary section S2.

All data analyses were conducted using Python 3.7 with libraries DLATK v1.1 (Schwartz et al., 2017) and scikit-learn v22.2 (Pedregosa et al., 2011).

Results

Depressive disorders

Table 1 displays prediction accuracy results for models predicting age 15 depressive disorders. The top section of Table 1 displays the AUCs for the models excluding A12 Dx + Demos (i.e. CCA components only), and the *p* values comparing these models to chance. Across models with CCA components from individual waves and combinations of successive waves, all but the models with only age 3 components (AUC = 0.556) and only age 6 components (AUC = 0.608) performed significantly better than chance (AUCs = 0.669–0.751).

The bottom section of Table 1 displays the AUCs for the models combining age 12 depression and demographics with the CCA components (i.e. CCA components + A12 Dx + Demos), and the *p* values comparing these models to chance and to the comparison model without CCA components (i.e. A12 Dx + Demos alone). All of these models performed better than chance, except for the individual wave model with age 3 components (AUC = 0.599). All models combining components across successive waves performed significantly better than the comparison model (AUCs = 0.739–0.748). The only model including components from an individual wave that performed significantly better than the comparison model was the model with age 12 components (AUC = 0.744).

The comparison model including A12 Dx + Demos produced an AUC of 0.633, which was significantly better than chance (0.500). Without the demographics, age 12 depression status did not predict age 15 depression better than chance (AUC = 0.522).

Anxiety disorders

Table 2 displays prediction accuracy results for all models predicting age 15 anxiety disorders. The top section of Table 2 displays the AUCs for the models excluding A12 Dx + Demos (i.e. CCA components only), and the *p* values comparing these models to chance. Across models with CCA components from individual waves and combinations of successive waves, all models performed significantly better than chance (AUCs = 0.621–0.788).

The bottom section of Table 2 displays the AUCs for the models combining age 12 anxiety and demographics with the CCA components (i.e. CCA components + A12 Dx + Demos), and the *p* values comparing these models to chance and to the comparison model (i.e. A12 Dx + Demos alone). All models performed better than chance and models combining components across successive waves performed significantly better than the comparison model (AUCs = 0.807–0.812). The only models including components from an individual wave that performed significantly better than the comparison model were the models with age 12 (AUC = 0.810) and age 9 (AUC = 0.805) components.

The comparison model including A12 Dx + Demos produced an AUC of 0.774, which was significantly better than chance. Without the demographics, age 12 anxiety status still predicted age 15 anxiety disorder better than chance (AUC = 0.720).

Sensitivity analyses

Results for the sensitivity analyses using different classification algorithms (i.e. neural networks, random forests, and logistic regression without regularization) are displayed in online Supplementary Tables S1 and S2 for depression and anxiety, respectively. Results of the main analyses were also included in these tables for easy comparison. The pattern of results was

Table 1. Depression prediction accuracy results

	Combined waves			Individual waves			
Models	A12 9 6 3	A12 9 6	A12 9	A12	A9	A6	A3
CCA components alone	0.743	0.746	0.751	0.745	0.669	0.608	0.556
p (*v.* chance)	<0.001	<0.001	<0.001	<0.001	0.003	0.114	0.413
CCA components + A12 Dx + Demos	0.739	0.740	0.748	0.744	0.679	0.639	0.599
p (*v.* chance)	<0.001	<0.001	<0.001	<0.001	0.002	0.013	0.082
p (*v.* A12 Dx + Demos alone)	0.008	0.009	0.007	0.009	0.123	0.432	0.897

A, age; A12 Dx, age 12 depression diagnostic status; Demos, demographics (sex, race, and ethnicity).
Note: Cells contain area under the receiver operating characteristics curve (AUC) values.

https://doi.org/10.1017/S0033291722003452 Published online by Cambridge University Press

Table 2. Anxiety prediction accuracy results

	Combined waves			Individual waves			
Models	A12 9 6 3	A12 9 6	A12 9	A12	A9	A6	A3
CCA components alone	0.777	0.784	0.787	0.788	0.749	0.711	0.621
p (*v*. chance)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
CCA components + A12 Dx + Demos	0.807	0.811	0.812	0.810	0.805	0.799	0.762
p (*v*. chance)	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
p (*v*. A12 Dx + Demos alone)	0.046	0.039	0.025	0.027	0.022	0.069	0.840

A, age; A12 Dx, age 12 anxiety diagnostic status; Demos, demographics (sex, race, and ethnicity).
Note: Cells contain area under the receiver operating characteristics curve (AUC) values.

generally consistent with the main findings, with a few exceptions. For example, not all combinations of CCA components from consecutive waves improved prediction of age 15 depression over A12 Dx + Demos using the additional classifiers, while all combinations did so for the L2-penalized logistic regression models. Additionally, none of the models using the additional classifiers improved prediction of age 15 anxiety over A12 Dx + Demos, while for L2-penalized logistic regression models, all combinations of components from successive waves and from the individual age 9 and 12 waves improved prediction. Notably, L2-penalized regression generally produced higher accuracy rates than equivalent models in the sensitivity analyses, particularly the logistic regression models without regularization; however, among the additional ML classifiers tested in the sensitivity analyses, the highest performing classifier differed by model, though differences were mostly within the margin of error.

Discussion

The current study used ML to predict depression and anxiety disorders in mid-adolescence using information from multiple waves of assessment across childhood and early adolescence in an unselected community sample. Our primary aim was to determine the relative contributions to prediction of information from different and multiple developmental stages across childhood and early adolescence. We also sought to explore the upper bounds of prediction that can be achieved when such a large number of highly relevant features spanning multiple domains are considered and assess the incremental gains in prediction of including such a volume of information (i.e. features) afforded over knowing prior disorder status and basic demographics.

In regards to timing of risk assessment, our results comparing model performance to chance suggest that screening for risk of adolescent anxiety can be successful as early as age 3, whereas for depression, screening may only be successful starting at age 9. Accuracy estimates were higher at the more proximal waves, which is unsurprising as greater delay between risk assessment and disorder occurrence increases the odds that new risk and resilience factors come into play, decreasing the influence of vulnerabilities assessed earlier. In addition, youth self-report is not feasible until later ages, expanding the range of constructs that can be assessed.

It is notable that combining information across waves either only marginally improved prediction or worsened prediction relative to the individual wave models with components from the most proximal wave (age 12). This demonstrates the limitations of smaller sample sizes highly typical in psychology and psychiatry. Specifically, our multi-wave models contain magnitudes more parameters than the individual wave models, which translates to increased noise and a higher benchmark for detecting generalizable signals, in effect underfitting the data after shrinkage as compared to a model with less parameters (Hastie et al., 2009). In other words, it is more difficult to separate the reliable signals from the non-reliable noise in data with many features. However, larger samples could offset this and enable the detection of smaller effects that may be present in the multi-wave data.

Our sensitivity analyses provide another example of the trade-off between complexity and generalizability in smaller samples. The models using neural networks and random forests performed equivalently or slightly worse than models using L2-penalized regression. The former two ML methods capture more complex relationships than the latter, practically translating into more model parameters and thus facing the same limitation (i.e. difficulty separating signal from noise). Much larger samples are needed to leverage the more complex features of ML (Hastie et al., 2009). Notably, however, the L2-penalized logistic regression models produced higher accuracy across the board than their counterparts using logistic regression without regularization, the more traditional statistical approach in psychology and psychiatry. This highlights the benefits of using ML methods, even in the relatively small sample sizes common in psychopathology research.

Often, the goal of ML studies is to develop a prediction algorithm that can be translated to applied settings (e.g. risk screening in a hospital). The current study does not share this goal, as the volume and variety of features could not be practically and economically assessed in any applied setting. Rather, this uniquely comprehensive feature set allows us to estimate the upper bounds of prediction that can be achieved in this idealized risk assessment context. Our highest performing model for predicting depression at age 15, which included all information (i.e. components) from age 9 and 12 waves, achieved an AUC of 0.751. For anxiety, our highest performing model achieved an AUC of 0.812 and was the model combining information from the age 9 and 12 waves alongside prior disorder status and basic demographic (i.e. sex, race, and ethnicity). For reference, these approximately correspond to Cohen's *d* values of 0.96 and 1.26, respectively, which are considered large effects (Rice & Harris, 2005).

These findings are highly consistent with another prospective study using a similarly diverse collection of risk factors to predict depression and anxiety disorders in a mixed-age adult sample over an approximately 3-year follow-up period. Using data from the National Epidemiological Survey on Alcohol and Related Conditions (NESARC), Rosellini et al. (2020) obtained AUCs of

https://doi.org/10.1017/S0033291722003452 Published online by Cambridge University Press

0.775 for depression and 0.780–0.799 for individual anxiety disorders. Achieving substantially better model performance may require more sophisticated ML techniques and use of less traditional types of data (e.g. social media; Guntuku et al., 2017), for which much larger samples can more realistically be obtained.

Using our comparison model ('A12 Dx + Demos alone') as a reference point, we were able to evaluate whether gathering additional information beyond a basic risk assessment (demographics and history of disorder) is helpful. Our findings from models combing information across waves suggest that assessing additional risk factors longitudinally across development can improve upon a basic risk assessment for both depression and anxiety in adolescence; however, risk screening at any one timepoint earlier than age 12 for depression, and age 9 for anxiety, may not improve prediction beyond knowing recent disorder history and basic demographics.

It may not be surprising that information from more distal waves did not improve prediction over prior disorder status. Both depression and anxiety demonstrate a moderate degree of homotypic continuity across development (Beesdo et al., 2009; Kessler et al., 2005), so it is likely that the vulnerabilities captured with the additional features are conferring risk for both the earlier and later instances of the disorder. In a statistical sense, age 12 disorder status is capturing nearly the same variance that is important for predicting age 15 disorder status, so it is difficult for prediction to improve. Further, as noted previously, most onsets of mental disorders occur during adolescence (Caspi et al., 2020), and rates of depression and many types of anxiety increase substantially during this time (Campbell et al., 2003; de Lijster et al., 2017; Kessler et al., 2005). Although the specific mechanisms producing this developmental pattern are unclear and likely manifold, it can reasonably be assumed that factors specific to the early-mid adolescence period (e.g. divergence in brain development, hormonal changes, increased relational and academic stressors), that would only have been captured in the older assessments, play an important role in influencing risk. An additional consideration is that we determined age 12 disorder status through a semi-structured interview administered by a trained interviewer. Such a thorough diagnostic assessment is often not practical in applied settings, which may only have the time and resources to administer screening questionnaires. In light of this, it is fairly remarkable that we were able to improve upon prediction by prior disorder status and demographics.

A final noteworthy result to address is the finding that baseline disorder status alone (i.e. excluding demographics) did not predict age 15 depressive disorders better than chance. We observed a fairly low rate of depressive disorder diagnoses at age 12 ($N = 22$), which is entirely consistent with its later age of onset (Kessler et al., 2005). A number of individuals in our sample did not develop depression until age 15, and many more will experience first onset in the following years. Thus, we tested models excluding prior disorder status and compared performance to chance (i.e. results in the top section of Tables 1 and 2) for this reason. These models represent assessment contexts in which age 12 disorder history is not known and/or cannot be known (i.e. assessment prior to age 12).

This study possessed several strengths, including an unprecedented number and variety of important risk factors, a multi-wave longitudinal design allowing us to compare risk assessment across critical developmental periods, use of a multi-view dimensionality reduction technique allowing us to preserve the longitudinal structure of the data, and a rigorous outcome measure. A few important limitations should also be acknowledged. First the sample is relatively small by ML standards, limiting our investigation to less sophisticated ML techniques (i.e. L2-penalized logistic regression). While we used *k*-fold CV to increase generalizability to new data while maximizing test data, we did not test our models on truly independent data and thus accuracy estimates are likely slightly inflated. Features were not identical across waves, as is typical for developmental studies because few measures and risk factors are appropriate across development (e.g. young children cannot provide reliable self-report and peer relationships become more important in later childhood/adolescence). Nevertheless, the impact of differences in feature sets across waves cannot be fully teased apart from developmental differences in the relevance of vulnerabilities to risk. Additionally, through CCA we were able to impose an a priori structure based on time (wave of assessment) but, as can be seen in the online Supplementary file displaying the top features of each component, the components are not easily interpretable, combining features from multiple conceptual domains. We did not separate first onsets from recurrent and persisting cases because rates of disorders were relatively low, although typical of a community sample. The timing of optimal risk assessment may differ for first-onset cases. Finally, the sample is relatively geographically and racially homogeneous, limiting generalizability to more diverse populations.

Conclusion

In this study, we leveraged ML to prospectively predict adolescent depression and anxiety risk assessment across development. Progress in translating research to reduced burden of mental health has been stifled by overreliance on statistical approaches that cannot meet the challenge of capturing such complex phenomena. This study demonstrates the potential of ML, which can accommodate a large number and variety of relationships while prioritizing generalizability, to contribute to efforts to reduce suffering from mental health problems.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S0033291722003452

Acknowledgements. This work was supported by the National Institute of Mental Health Grant R01 MH069942.

Conflict of interest. We have no known conflict of interest to disclose.

References

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders: DSM-IV*. Washington, DC: American Psychiatric Association.

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders* (*DSM-5*®). Washington, DC: American Psychiatric Publishing.

Andersson, S., Bathula, D. R., Iliadis, S. I., Walter, M., & Skalkidou, A. (2021). Predicting women with depressive symptoms postpartum with machine learning methods. *Scientific Reports*, *11*(1), 1–15.

Axelson, D., Birmaher, B., Zelazny, J., Kaufman, J., & Gill, M. (2009). The schedule for affective disorders and schizophrenia-present and lifetime version (K-SADS-PL) 2009 working draft. Advanced Centre for Intervention and Services Research, Western Psychiatric Institute and Clinics.

Beesdo, K., Knappe, S., & Pine, D. S. (2009). Anxiety and anxiety disorders in children and adolescents: Developmental issues and implications for DSM-V. *Psychiatric Clinics*, *32*(3), 483–524.

Bellón, J., de Dios Luna, J., King, M., Moreno-Küstner, B., Nazareth, I., Montón-Franco, C., … Vicens, C. (2011). Predicting the onset of major depression in primary care: International validation of a risk prediction algorithm from Spain. *Psychological Medicine*, *41*(10), 2075–2088.

Campbell, L. A., Brown, T. A., & Grisham, J. R. (2003). The relevance of age of onset to the psychopathology of generalized anxiety disorder. *Behavior Therapy*, *34*(1), 31–48.

Caspi, A., Houts, R. M., Ambler, A., Danese, A., Elliott, M. L., Hariri, A., … Ramrakha, S. (2020). Longitudinal assessment of mental health disorders and comorbidities across 4 decades among participants in the Dunedin birth cohort study. *JAMA Network Open*, *3*(4), e203221–e203221.

Coutanche, M., & Hallion, L. (2019). Machine learning for clinical psychology and clinical neuroscience. In A. Wright & M. Hallquist (Eds.), *The Cambridge handbook of research methods in clinical psychology* (pp. 467–482). Cambridge: Cambridge University Press.

de Lijster, J. M., Dierckx, B., Utens, E. M., Verhulst, F. C., Zieldorff, C., Dieleman, G. C., & Legerstee, J. S. (2017). The age of onset of anxiety disorders: A meta-analysis. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *62*(4), 237.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*, 91–118.

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., … Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, *115*(44), 11203–11208.

Eliot, M., Ferguson, J., Reilly, M. P., & Foulkes, A. S. (2011). Ridge regression for longitudinal biomarker data. *The International Journal of Biostatistics*, *7*(1), 1–11.

Essau, C. A., Lewinsohn, P. M., Olaya, B., & Seeley, J. R. (2014). Anxiety disorders in adolescents and psychosocial outcomes at age 30. *Journal of Affective Disorders*, *163*, 125–132.

Fawcett, C., & Hoos, H. H. (2016). Analysing differences between algorithm configurations through ablation. *Journal of Heuristics*, *22*(4), 431–458.

Fleisher, W. P., & Katz, L. Y. (2001). Early onset major depressive disorder. *Paediatrics & Child Health*, *6*(7), 444–448.

Fried, E. I., & Robinaugh, D. J. (2020). Systems all the way down: Embracing complexity in mental health research. *BMC Medicine*, *18*(205), 1–4.

GBD 2017 Disease and Injury Incidence and Prevalence Collaborators (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: A systematic analysis for the Global Burden of Disease Study 2017. *The Lancet*, *392*(10159), 1789–1858.

Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, *18*, 43–49.

Hastie, T., Tibshirani, R., & Jermone, F. (2009). *The elements of statistical learning: Data mining, inference and prediction* (2nd ed.). New York, NY: Springer-Verlag.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. *112*). New York, NY: Springer.

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the national comorbidity survey replication. *Archives of General Psychiatry*, *62*(6), 617–627.

Kessler, R. C., van Loo, H. M., Wardenaar, K. J., Bossarte, R. M., Brenner, L. A., Cai, T., … de Jonge, P. (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.

King, M., Walker, C., Levy, G., Bottomley, C., Royston, P., Weich, S., … Rotar, D. (2008). Development and validation of an international risk prediction algorithm for episodes of major depression in general practice attendees: The PredictD study. *Archives of General Psychiatry*, *65*(12), 1368–1376.

Klein, D. N., & Finsaas, M. C. (2017). The stony brook temperament study: Early antecedents and pathways to emotional disorders. *Child Development Perspectives*, *11*(4), 257–263.

Koul, A., Becchio, C., & Cavallo, A. (2018). Cross-validation approaches for replicability in psychology. *Frontiers in Psychology*, *9*, 1117.

Kumar, P., Garg, S., & Garg, A. (2020). Assessment of anxiety, depression and stress using machine learning models. *Procedia Computer Science*, *171*, 1989–1998.

Liu, Y., Hankey, J., Cao, B., & Chokka, P. (2021). Screening for major depressive disorder in a tertiary mental health centre using EarlyDetect: A machine learning-based pilot study. *Journal of Affective Disorders Reports*, *3*, 100062.

McLeod, G. F., Horwood, L. J., & Fergusson, D. M. (2016). Adolescent depression, adult mental health and psychosocial outcomes at 30 and 35 years. *Psychological Medicine*, *46*(7), 1401–1412.

Miché, M., Studerus, E., Meyer, A. H., Gloster, A. T., Beesdo-Baum, K., Wittchen, H.-U., & Lieb, R. (2020). Prospective prediction of suicide attempts in community adolescents and young adults, using regression methods and machine learning. *Journal of Affective Disorders*, *265*, 570–578.

Naicker, K., Galambos, N. L., Zeng, Y., Senthilselvan, A., & Colman, I. (2013). Social, demographic, and health outcomes in the 10 years following adolescent depression. *Journal of Adolescent Health*, *52*(5), 533–538.

Nemesure, M. D., Heinz, M. V., Huang, R., & Jacobson, N. C. (2021). Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific Reports*, *11*(1), 1–9.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning research*, *12*, 2825–2830.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r*. *Law and Human Behavior*, *29*(5), 615–620.

Rosellini, A. J., Liu, S., Anderson, G. N., Sbi, S., Tung, E. S., & Knyazhanskaya, E. (2020). Developing algorithms to predict adult onset internalizing disorders: An ensemble learning approach. *Journal of Psychiatric Research*, *121*, 189–196.

Sato, J. R., Moll, J., Green, S., Deakin, J. F., Thomaz, C. E., & Zahn, R. (2015). Machine learning algorithm accurately detects fMRI signature of vulnerability to major depression. *Psychiatry Research: Neuroimaging*, *233*(2), 289–291.

Schwartz, H. A., Giorgi, S., Sap, M., Crutchley, P., Ungar, L., & Eichstaedt, J. (2017). *Dlatk: Differential language analysis toolkit*. Paper presented at the Proceedings of the 2017 conference on empirical methods in natural language processing: System demonstrations.

Shatte, A. B., Hutchinson, D. M., & Teague, S. J. (2019). Machine learning in mental health: A scoping review of methods and applications. *Psychological Medicine*, *49*(9), 1426–1448.

Su, D., Zhang, X., He, K., & Chen, Y. (2021). Use of machine learning approach to predict depression in the elderly in China: A longitudinal study. *Journal of Affective Disorders*, *282*, 289–298.

Tackett, J. L., Brandes, C. M., King, K. M., & Markon, K. E. (2019). Psychology's replication crisis and clinical psychological science. *Annual Review of Clinical Psychology*, *15*, 579–604.

Wang, J., Sareen, J., Patten, S., Bolton, J., Schmitz, N., & Birney, A. (2014). A prediction algorithm for first onset of major depression in the general population: Development and validation. *Journal of Epidemiology & Community Health*, *68*(5), 418–424.

Whelan, R., & Garavan, H. (2014). When optimism hurts: Inflated predictions in psychiatric neuroimaging. *Biological Psychiatry*, *75*(9), 746–748.

Witten, D. M., Tibshirani, R., & Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics (Oxford, England)*, *10*(3), 515–534.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122.

Zhang, Y., Wang, S., Hermann, A., Joly, R., & Pathak, J. (2021). Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. *Journal of Affective Disorders*, *279*, 1–8.

https://doi.org/10.1017/S0033291722003452 Published online by Cambridge University Press