Published in final edited form as:

J Psychiatr Res. 2020 February ; 121: 189–196. doi:10.1016/j.jpsychires.2019.12.006.

Developing algorithms to predict adult onset internalizing disorders: An ensemble learning approach

Anthony J. Rosellini, Ph.D., Siyu Liu, B.A., Grace N. Anderson, B.A., Sophia Sbi, B.S., Esther Tung, M.A., Evdokia Knyazhanskaya, B.A.

Center for Anxiety and Related Disorders, Department of Psychological and Brain Sciences, Boston University, Boston, MA, USA

Abstract

A growing literature is utilizing machine learning methods to develop psychopathology risk algorithms that can be used to inform preventive intervention. However, efforts to develop algorithms for internalizing disorder onset have been limited. The goal of this study was to utilize prospective survey data and ensemble machine learning to develop algorithms predicting adult onset internalizing disorders. The data were from Waves 1–2 of the National Epidemiological Survey on Alcohol and Related Conditions ($n = 34,653$). Outcomes were incident occurrence of *DSM-IV* generalized anxiety, panic, social phobia, depression, and mania between Waves 1–2. In total, 213 risk factors (features) were operationalized based on their presence/occurrence at the time of or before Wave 1. For each of the five internalizing disorder outcomes, super learning was used to generate a composite algorithm from several linear and non-linear classifiers (e.g., random forests, k-nearest neighbors). AUCs achieved by the cross-validated super learner ensembles were in the range of 0.76 (depression) to 0.83 (mania), and were higher than AUCs achieved by the individual algorithms. Individuals in the top 10% of super learner predicted risk accounted for 37.97% (depression) to 53.39% (social anxiety) of all incident cases. Thus, the algorithms achieved acceptable-to-excellent prediction accuracy with a high concentration of incident cases observed among individuals predicted to be highest risk. In parallel with the development of effective preventive interventions, further validation, expansion, and dissemination of algorithms predicting internalizing disorder onset/trajectory could be of great value.

Keywords

anxiety; mood; incidence; algorithm; machine learning; risk score

Internalizing disorders (IDs) are the most common form of psychopathology; roughly one-third of the US population experiences an anxiety or mood disorder over their lifetime

Correspondence: Anthony J. Rosellini, Department of Psychological and Brain Sciences, Boston University, 900 Commonwealth Avenue (East) 2nd Floor, Boston, MA, 02215, USA. Phone: 617-353-9610; Fax: 617-353-9609; ajrosellini@gmail.com.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Declarations of interest: None

(Kessler et al., 2005). IDs also are highly chronic (Andrade et al., 2003; Yonkers et al., 2013) and thus consistently identified as a top cause of disease burden (James, 2018). Although IDs frequently onset in childhood and early adolescence, a large proportion of individuals experience first onset after the age of 18 (e.g., >50% of lifetime cases of panic disorder, agoraphobia, generalized anxiety disorder; roughly 75% of all lifetime mood disorder cases; Kessler et al., 2005).

Given their high prevalence and burden, significant efforts have been devoted to identifying ID risk factors. Literature has amassed on the importance of a wide range of factors, many of which can be assessed using self-report (e.g., socio-demographics; pre-existing symptoms of physical and mental disorders, stress/adversity; Grant et al., 2009; Kim-Cohen et al., 2003; McLaughlin et al., 2010; Scott et al., 2007). Although it is common for etiological studies to collect data on numerous ID risk factors, there have been limited efforts to determine how to use/combine all available risk factor data to generate composite algorithms (risk scores) that predict ID onset. The development of algorithms to predict ID onset could be used to inform preventive intervention; that is, to determine who is at highest-risk and possibly in need of (targeted) assessment or preventive intervention.

A recent review identified only four studies that have attempted to develop algorithms predicting the onset of generalized anxiety disorder, panic disorder, or bipolar disorders/mania (Bernardini et al., 2017). More work has been done to develop algorithms predicting the onset of depression (Bellón et al., 2011; King et al., 2013; King et al., 2008). For example, Wang and colleagues (2014) used data from the National Epidemiological Survey on Alcohol and Related Conditions (NESARC) and stepwise regression methods to develop a population-level algorithm for adult onset major depression (Wang et al., 2014). However, a limitation across most ID algorithm development studies (including Wang et al., 2014) has been the reliance on conventional regression methods that (a) are prone to model overfit (i.e., poor performance in hold-out data or an independent sample), particularly in the presence of highly correlated features, and (b) cannot flexibly capture nonlinear associations or interactions among predictors.

An alternative approach for developing algorithms for ID onset is to use (supervised) machine learning methods designed to optimize prediction. A growing literature is utilizing machine learning to develop algorithms for posttraumatic stress disorder/symptoms (e.g., in emergency settings; Galatzer-Levy et al., 2014), but there has been virtually no effort to use machine learning to develop algorithms that predict the onset of other IDs. In particular, *ensemble* methods have been used to develop optimized composite algorithms from several different approaches to prediction and which outperform individual (i.e., non-ensemble) algorithms (Kessler et al., 2014). The goal of the present study was to utilize prospective data from the NESARC and ensemble machine learning methods to develop algorithms predicting adult onset generalized anxiety disorder, panic disorder, social phobia, depression, and mania.

Method

Sample

Algorithms were developed using data from Waves 1–2 of the NESARC, a longitudinal representative household survey of US adults (Hasin and Grant, 2015). Although the NESARC focused extensively on substance use, the surveys assessed a wide range of *DSM-IV* mental disorders (including IDs) and associated risk factors. NESARC data previously has been used to develop algorithms for major depression onset (Wang et al., 2014) and is well-suited for supervised machine learning because of its prospective design, large sample, and rich assessment of ID risk factors. The Wave 1 NESARC interview was conducted in 2001–2002 ($n=43,093$). Wave 2 re-interviewed 34,653 of the Wave 1 respondents in 2004–2005. Details of the NESARC design and field procedures are available elsewhere (Grant et al., 2003).

Outcomes

Our goal was to develop algorithms predicting incident onset IDs in adulthood. Mental disorders were diagnosed using the Alcohol Use Disorder and Associated Disability Interview Schedule, *DSM-IV* Version (AUDADIS-IV). The AUDADIS-IV is a reliable fully structured diagnostic interview of *DSM-IV* substance use disorders, IDs, personality disorders, and associated risk factors (e.g., family history of psychopathology) that can be administered by trained lay diagnosticians (Grant et al., 2003). Whereas 12-month and lifetime disorders were assessed at Wave 1, a goal of Wave 2 was to assess disorders first occurring since Wave 1 (i.e., disorder incidence). Our outcomes were defined based on Wave 2 variables indicating incident occurrence (since Wave 1) of five *DSM-IV* disorder outcomes: generalized anxiety disorder (GAD), panic disorder (with *or* without agoraphobia; PDA), social phobia (SOC), depression (major depressive episode/disorder *or* dysthymia, DEP), and mania (manic *or* hypomanic episode). Models were developed for each outcome in a distinct subsample that excluded individuals with disorder incidence prior to Wave 1 (based on Wave 1 lifetime diagnoses): GAD (outcome $n = 1,123$; GAD sample $n = 33,018$), PDA (647; 32,714), SOC (560; 32,902), DEP (1,707; 27,769), mania (970; 32,496).

Features

The NESARC assessed the wide range of established ID risk factors including sociodemographics (Grant et al., 2009), physical health problems (Scott et al., 2007), other (pre-existing) mental disorders/symptoms (Kim-Cohen et al., 2003; Merikangas et al., 2003), personality traits (Krueger, 1999), stressful life events (Hammen, 2005; McLaughlin et al., 2010; Miloyan et al., 2018), childhood adversity/trauma exposure (Lindert et al., 2014; McLaughlin et al., 2010), and a family history of psychopathology (Beardseele et al., 1998). The Wave 1–2 surveys and codebooks were reviewed to identify all plausible risk factors (“features”) within these domains and with minimal missing data (i.e., we did not attempt to operationalize predictors based on questions that were “skipped” for >80% of respondents). Features were operationalized to reflect the presence/severity of the risk factor at the time of or prior to the Wave 1 assessment (prior to incident disorder onset). Features assessed at Wave 2 were determined to be present prior to Wave 1 (a) if they explicitly assessed experiences occurring prior to the age of 18 (Wave 1 respondents were 18), or (b) based on

self-reported age of occurrence (e.g., reported ages of traumatic event occurring prior to age at Wave 1). This approach was used to capture directional associations between features and outcomes (i.e., risk factors preceding outcomes), and is more conservative than the approach used by Wang and colleagues (e.g., Wang et al. defined a predictor for discrimination experiences, which was assessed in Wave 2 and without questions to assess discrimination prior to Wave 1). The 213 features are summarized in Supplemental Tables 1–2.

Analysis method

All analyses were conducted in R (Team, 2013). Super learning (Polley et al., 2018; Van der Laan and Rose, 2011) was used to generate a prediction function for each of the five outcomes. The (a) identifying a user-specified “library” of classifiers (i.e., different approaches to prediction, described below), (b) implementing each classifier in the library using k -fold cross-validation (CV) to generate predicted values (10-fold CV described below), (c) regressing the outcome onto the CV predicted values from each classifier in the library (to determine the best weighted combination of the individual algorithms), and (d) fitting each classifier in the library on the full sample, and combining these fits with the weights generated in the prior step to estimate the super learner predicted values. Super learner analyses were not weighted because the majority of the classifiers included in the super learner library (see below) were not designed for complex survey design.

Classifiers.—Several different classifiers were included in the super learner library, both based on recent recommendations (LeDell et al., 2016) and because it was unknown what type of classifier(s) would result in optimized prediction. Nine classifiers were selected: logistic regression, least absolute shrinkage and selection operator (LASSO) penalized (logistic) regression, random forests, Bayesian main-terms logistic regression, generalized additive modeling, adaptive splines, k -nearest neighbors, linear support vector machines, and linear discriminant analysis. Classifier descriptions are provided in Supplemental Table 3.

In developing the super learner ensemble for each outcome, each classifier was implemented using five different feature subsets that were identified using super learning feature selection methods. Feature selection is a common data processing step in the machine learning literature (Galatzer-Levy et al., 2018). Although reducing the number of features may increase bias (reducing model accuracy), feature selection also mitigates risk of model overfit (reducing variance). Five feature sets were operationalized for each outcome. Four sets of non-collinear features were defined (for each outcome) based on features that had non-zero coefficients in LASSO models tuned to select a maximum of 5, 10, or 20 features, and when permitted to use as many features as possible (four LASSO feature sets were defined for each outcome). The fifth, in comparison, included all features with significant bivariate associations with the outcome (Wald $\chi^2 p < .05$).

The classifiers were implemented on each of these feature sets rather than all 213 features because (a) some classifiers could perform better using a smaller number of features (e.g., collinearity could negatively impact performance), (b) we wanted to evaluate the extent to which utilizing a smaller number of features would affect model performance, (c) prior efforts to develop ID algorithms in the NESARC have used similar feature selection

methods (e.g., identifying significant bivariate associations, selecting 10–20 with largest odds ratios among collinear sets; Wang et al., 2014), and (d) to mitigate computational intensity (by eliminating irrelevant features). To develop the super learner for each outcome, all nine classifiers were implemented using *each* of the five feature sets (44 individual algorithms were estimated to create the super learner for one outcome¹).

Cross-validation.—Super learning (and thus all classifier-predictor set combinations) was implemented using 10-fold CV. The 10-fold CV approach involved: (a) dividing all observations into 10 mutually exclusive subgroups of roughly equal size in terms of total number and number of outcome cases (for each outcome/sample), (b) fitting the model (i.e., classifier/predictor set combination) in 9 of the 10 blocks within each fold (i.e., the training set), and using the resulting fit to estimate predicted probabilities for observations in the one left-out block (i.e., the validation/test set), and (c) repeating this process until cross-validated (CV) predicted probabilities are estimated for all observations (for any given classifier-predictor set combination) *and* all classifier-predictor set combinations (including the consolidated super learner ensembles). In other words, predicted values were estimated for all observations based on a model developed in a mutually exclusive subsample. In comparison, Wang et al. (2014) split the NESARC into training and validation/test samples based on US consensus region (training: Northeast/Midwest/South regions, test: West region). Ten-fold CV has advantages over split-sample validation including increased sample size (all observations are utilized as both training and test cases) and reduced likelihood of model overfit (because not relying on a single/arbitrary sample split; see James et al., 2013)

Model Performance.—To evaluate and compare the performance of the super learner to the individual classifiers, area under the receiver operating characteristic (ROC) curves (AUC) and their 95% confidence intervals (CI) were calculated using the CV predicted probabilities (expressed to four decimals for increased precision). In addition, two approaches were used to operationalize high risk cut-points. First, cut-point were operationalized for all algorithms based on the threshold at which Youden's Index was maximized (the point on the ROC curve where sensitivity + specificity was highest). Second, super learner risk tiers were created by rank ordering predicted values and dividing respondents into 20 equal groups (ventiles of risk). Operating characteristics were calculated based on Youden's Index and the ventiles of risk. Sensitivity and positive predictive value were prioritized over specificity and negative predictive value given the primary goal of identifying true-positive cases.

Results

Feature selection

Details of which variables were selected in each feature set for each outcome are presented in Supplemental Tables 4–6 (including bivariate odds ratios). The Figure 1 Word Cloud summarizes those selected most frequently across feature sets and outcomes. Five features were selected for all outcomes and in >20 of the 25 total feature sets (in the largest font in

¹Due to concerns of multicollinearity, logistic regression was not implemented using the predictor sets based on all features with significant bivariate associations with the outcomes.

the Word Cloud): having a lifetime PTSD diagnosis prior to Wave 1 (for interpretative purposes, odds ratios [ORs] = 4.9–5.7, $\chi^2_1 = 340.8$ –588.9), number of attention deficit-hyperactivity disorder symptoms in childhood (ORs = 1.1–1.2, $\chi^2_1 = 223.8$ –967.7), SF-12v2 Mental Disability summary score at Wave 1 (higher score indicates better health, ORs = 0.9, $\chi^2_1 = 188.2$ –544.9), number of childhood adversities (ORs = 1.3, $\chi^2_1 = 230.0$ –478.6), and being the victim of sexual assault prior to Wave 1 (ORs = 3.1–4.3, $\chi^2_1 = 222.3$ –297.3). Several other features also were selected for more than one (but not all five) outcomes across the restrictive (5, 10, or 20) LASSO feature sets (see Supplemental Table 6). A limited number of features were selected for a single ID (e.g., subclinical social anxiety fears and SOC; self-perceived health and PDA; unemployment in the past 12-months and DEP; subclinical elevated mood and number of illicit drugs used in past 12-months and mania).

Best performing algorithms

A summary of the best performing individual classifiers using each feature set, and performance of the five super learner ensembles, are presented in Table 1 (see Supplemental Tables 7–11 for super learner weights, AUCs, and operating characteristics for all algorithms based on Youden's Index). Based on AUC, linear discriminant analysis (LDA), generalized additive modeling (GAM), and LASSO were consistently the best performing classifiers. The best performing CV individual algorithm for each outcome was: (a) GAD - LDA (max 20 features, AUC = 0.7988 [0.7863–0.8113]), (b) PDA - LASSO (all bivariate features, AUC = 0.7796 [0.7620–0.7972]), (c) SOC - GAM (max 20 features, AUC = 0.7942 [0.7756–0.8129]), (d) DEP - LASSO (all bivariate features, AUC = 0.7552 [0.7435–0.7668]), and (e) Mania - LASSO (all bivariate features, AUC = 0.8249 [0.8119–0.8379]). AUCs of the best performing algorithms using a maximum of five features was lower and ranged between 0.7281 (0.7155–0.7407; DEP) and 0.7862 (0.7711–0.8013; mania). Several other classifier-feature set combinations resulted in similar but lower AUCs than the best classifier-feature sets combinations (Supplemental Tables 7–11), including the logistic regression models. At least seven (of the nine) total classifiers and 16 (of the 44) total classifier-feature set combinations were represented with non-zero weights in the super learner ensemble for each outcome. The CV super learner ensembles achieved slightly higher AUCs than the individual algorithms, with AUCs ranging between 0.7572 (95% CI = 0.7455–0.7688; DEP) and 0.8261 (0.8131–0.8392; mania). When the super learner coefficients were applied to the full sample (i.e., non-CV), AUCs increased but were not indicative of model overfit (i.e., AUC approaching 1)². Calibration slopes and Brier scores indicated acceptable calibration of the ensembles (slopes = 1.09–1.16, Brier scores = 0.016–0.053; slope near 1.0 and Brier score near 0 indicate better calibration, Van Calster et al. 2016).

Super learner operating characteristics

Between 21.09% and 35.57% of each outcome-specific sample had a predicted value above the optimized Youden's Index risk threshold (Table 2). Individuals with predicted values above the optimized risk thresholds accounted for 69.01% (DEP) to 81.75% (GAD) of all

²To explore sex-differences in model performance, AUCs of the final super learner ensembles were calculated separately among men and women (for each super learner/outcome). AUCs were generally similar for men and women for the PDA (men AUC = 0.8017, women AUC = 0.8069), SOC (0.8345, 0.8362), and mania (0.8515, 0.8571) ensembles. There were somewhat larger discrepancies in AUC for GAD (0.8260, 0.7982) and DEP (0.8043, 0.7448).

incidence cases of the outcomes (sensitivity). The rate of the outcomes among individuals above the risk threshold (PPV) ranged from 5.16% (PDA) to 14.19% (DEP). Sensitivity and PPV were 2.30 (GAD) to 3.41 (SOC) times greater than their expected values. In comparison, specificity (range = 66.05% [GAD] to 79.79% [SOC]) and NPV (range = 97.28% [DEP] to 99.40% [SOC]) were 1.01 (PDA; SOC) to 1.04 (DEP) times greater than their expected values.

Individuals in the top 5% of predicted risk accounted for 24.55% (DEP) to 37.52% (SOC) of all incidence cases of the outcomes (Figure 2), whereas those in the top 10% of predicted risk accounted for 36.97% (DEP) to 53.39% (SOC) of all incident cases (sensitivity). The outcome rates (PPV) in the top 5% of risk ranged from 12.70% (SOC) to 30.17% (DEP) (Figure 3) and in the top 10% of risk from 9.05% (PDA) to 22.72% (DEP). Sensitivity/PPV was 4.91–7.50 times the expected value (5%) among individuals in the top 5% of predicted risk, 3.70–5.34 times the expected value (10%) among those in the top 10% (the “top decile”), and remained above the expected value across the top five ventiles for all five IDs.

Discussion

This is the first study to use ensemble machine learning to develop algorithms predicting the first onset of several IDs. The CV and full-sample super learner ensembles achieved AUCs indicating *acceptable* (AUC 0.70) to *excellent* (AUC 0.80) prediction (Hosmer et al., 2013). As expected, the CV super learners achieved higher AUCs than the best performing CV individual algorithms. However, the magnitude of these differences were relatively small for several classifier-feature set combinations (e.g., AUC < .01), including when the classifiers were restricted to a maximum of five features. Along these lines, the comparable but slightly lower AUCs of the logistic regression models based on non-collinear/optimized LASSO feature sets underscores the utility of modeling “traditional” linear associations between survey features and ID outcomes. Further, a wide range of algorithms contributed to the ensembles with non-zero super learner weights. Collectively, these findings suggest that (a) a number of different classifiers can be used to predict ID onset from self-report/survey data with an acceptable level of accuracy, even with a selective number of features, and (b) the incremental predictive validity of combining individual algorithms via ensemble learning may be small. Although reasonable to question whether these improvements in AUC warrants the use of all available features and computational intense ensemble procedure such as super learning, *any* improvement in prediction accuracy is important when the goal is forecasting debilitating outcomes such as IDs. In addition, the differential/incremental performance of individual classifier-feature set combinations and super learning may be more pronounced in more complex data structures (e.g., including neuroimaging, psychophysiological, or audio features).

Several features appeared to have transdiagnostic relevance by being selected in feature sets for IDs, particularly indicators of prior trauma/stress (e.g., number of childhood adversities, sexual assault victimization) and mental disorders/symptoms (e.g., PTSD diagnosis, SF-12v2 Mental Disability). In comparison, subclinical features of some outcomes only were selected in feature sets for that outcome (e.g., subclinical social anxiety fears and social phobia; subclinical elevated mood and mania). These findings are in line with research

implicating both distal (childhood trauma/adversity, Carre et al., 2013; Hoppen and Chalder, 2018) and proximal (recent stress, Hammen, 2005; Miloyan et al., 2018) environmental factors as important transdiagnostic predictors of IDs, as well as evidence that experiencing some mental disorders/symptoms (including subthreshold symptoms) increases risk of developing additional disorders/symptoms (Karsten et al., 2011; Kim-Cohen et al., 2003; Merikangas et al., 2003). Similar predictors have been included in previously developed algorithms for ID onset (King et al., 2008; Moreno-Peral et al., 2014; Wang et al., 2014).

Our evaluation of model performance based on AUC is consistent with prior efforts to develop algorithms predicting the onset of depressive (Bellón et al., 2011; King et al., 2013; King et al., 2008; Wang et al., 2014), bipolar (Gan et al., 2011), and anxiety disorders (King et al., 2011; Moreno-Peral et al., 2014). Our CV super learner for DEP achieved a higher AUC than was obtained by the Wang and colleagues algorithm for incident onset of major depression developed using NESARC data and parametric/stepwise regression methods (hold-out sample AUC = 0.7259). Several ID onset prediction models also have been developed through two multi-site cohort studies of primary care patients in Europe and South America; “PredictD” (Bellón et al., 2011; King et al., 2013; King et al., 2008) and “PredictA” (King et al., 2011; Moreno-Peral et al., 2014). CV performance of the super learner ensembles was generally higher than test sample performance of the “PredictD” model for major depression (AUC = 0.728, King et al., 2013) and “PredictA” model for generalized anxiety/panic (AUCs = 0.707 in one sample but 0.811 in another, King et al., 2011).

AUC is limited in that it is a metric of overall prediction accuracy that does not capture performance using a specific high-risk threshold that might be used to make decisions/recommendations for preventive intervention (Wald and Bestwick, 2014). Accordingly, we also evaluated the operating characteristics of various high-risk thresholds. Using Youden’s Index to operationalize high-risk thresholds resulted in the correct classification of over two-thirds of all cases for each outcome. However, this performance (sensitivity) was only 2–3 times better than would be expected by chance because >20% of each sample had predicted probabilities surpassing the Youden’s Index thresholds. In comparison, the top ventile of risk accounted for roughly one-quarter to one-third (up to 7.5 times the expected values) and top decile of risk between one-third and one-half (up to 5.3 times the expected values) of all incident cases of IDs. Whereas the top decile (10%) of super learner risk accounted for 36.97% of all incident cases of DEP (3.69 times the expected value), the top decile of Wang and colleagues (Wang et al., 2014) NESARC algorithm accounted for 33.9% of all cases (3.39 times the expected value). Other prior ID algorithm development studies did not report sensitivity and/or the size of high-risk groups (Bellón et al., 2011; King et al., 2011, 2013; Moreno-Peral et al., 2014), making comparisons difficult (e.g., if 75% of all individuals are classified as high risk, sensitivity of 75% would be poor).

Although additional validation of the current models is needed, there could be several ways to disseminate ID risk algorithms. As has been done for cardiovascular disease (<http://static.heart.org/riskcalc/app/index.html#!/baseline-risk>), publicly available online risk calculators and preventive intervention recommendations could be made available and accessed independently or in collaboration with healthcare professionals (e.g., during

appointments). Utilization of ID algorithms also could be useful during ubiquitously-experienced life stages of stress/risk during which it might be feasible to assess ID risk and provide recommendations for preventive intervention (e.g., during college orientation). Numerous ID preventive interventions could be recommended for individuals predicted to be at highest risk (e.g., exercise, Harvey et al., 2017; self-guided internet prevention Deady et al., 2017; ongoing assessment, psychotherapy/pharmacotherapy referrals). However, the costs and anticipated benefits (e.g., number needed to treat) of the recommended prevention approach would need to be carefully considered. One path toward integrating ID risk algorithms in a cost-effective way could be by using algorithms to identify and recruit individuals to participate in prevention research (cf. recruitment based on family history), and expanding the models to predict who among the highest risk would benefit from a specific prevention approach (precision preventive medicine).

This study has several limitations. The data were exclusively self-report and primarily categorical/dichotomous measures. It is possible that super learner performance, including incremental performance over the individual algorithms, could be improved with information on additional risk factors assessed dimensionally (e.g., severity of subthreshold symptoms) and using other units of analysis (e.g., neurophysiological reactivity, Meyer et al., 2015). Relying on categorical *DSM-IV* diagnoses as the primary outcome measures also is limited. As has been done in the PTSD algorithm literature (Galatzer-Levy et al., 2014), future research developing ID algorithms should assess symptom severity repeatedly over time with the goal of predicting symptom trajectories (e.g., increasing, chronic, and remitting ID symptoms). Super learning also has limitations; for example, it requires significant computational time, particularly when implemented with CV. Further, although our use of 10-fold CV is a study strength (cf. split sample validation), replication of model performance in independent samples is needed.

Within the context of these limitations, the current study is the first to use ensemble learning to develop algorithms for adult onset IDs. The algorithms achieved acceptable-to-excellent overall prediction accuracy and a high concentration of outcome cases were observed among individuals predicted to be highest risk. As the number of empirically-supported ID preventive interventions continues to grow, further validation, expansion, and dissemination of such algorithms could be of great value for the broader public health goal of reducing the incidence of ID psychopathology.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgement:

The authors thank Eujene Chung for her assistance with this study. This study was supported by Grant MH106710 from the National Institute of Mental Health to the first author. The contents are solely the responsibility of the authors and do not necessarily represent the views of the National Institutes of Health.

References

- American Heart Association/American College of Cardiology, 2018 2018 Prevention Guidelines Tool CV Risk Calculator. <http://static.heart.org/riskcalc/app/index.html#!/baseline-risk> (accessed 6 August 2019).
- Andrade L, Caraveo-anduaga JJ, Berglund P, Bijl RV, Graaf RD, Vollebergh W, ... Wittchen H-U, 2003 The epidemiology of major depressive episodes: Results from the International Consortium of Psychiatric Epidemiology (ICPE) surveys. *Int. J. Methods. Psychiatr. Res* 12(1), 3–21. 10.1002/mpr.138 [PubMed: 12830306]
- Beardselee WR, Versage EM, Giadstone TRG, 1998 Children of affectively ill parents: a review of the past 10 years. *J. Am. Acad. Child. Adolesc. Psychiatry* 37(11), 1134–1141. 10.1097/00004583-199811000-00012 [PubMed: 9808924]
- Bellón JÁ, de Dios Luna J, King M, Moreno-Küstner B, Nazareth I, Montón-Franco C, ... Torres-González F, 2011 Predicting the onset of major depression in primary care: International validation of a risk prediction algorithm from Spain. *Psychol. Med* 41(10), 2075–2088. 10.1017/s0033291711000468 [PubMed: 21466749]
- Bernardini F, Attademo L, Cleary SD, Luther C, Shim RS, Quartesan R, Compton MT, 2017 Risk prediction models in psychiatry: toward a new frontier for the prevention of mental illnesses. *J. Clin. Psychiatry* 78(5), 572–583. 10.4088/JCP.15r10003 [PubMed: 27337225]
- Carr CP, Martins CMS, Stingel AM, Lemgruber VB, Jurueña MF, 2013 The role of early life stress in adult psychiatric disorders: a systematic review according to childhood trauma subtypes. *J. Nerv. Ment. Dis* 201(12), 1007–1020. [PubMed: 24284634]
- Deady M, Choi I, Calvo RA, Glozier N, Christensen H, Harvey SB, 2017 eHealth interventions for the prevention of depression and anxiety in the general population: a systematic review and meta-analysis. *BMC. Psychiatry* 17(1), 310. [PubMed: 28851342]
- Galatzer-Levy IR, Karstoft K-I, Statnikov A, Shalev AY, 2014 Quantitative forecasting of PTSD from early trauma responses: A machine learning application. *J. Psychiatr. Res* 59, 68–76. 10.1016/j.jpsychires.2014.08.017 [PubMed: 25260752]
- Gan Z, Diao F, Wei Q, Wu X, Cheng M, Guan N, ... Zhang J, 2011 A predictive model for diagnosing bipolar disorder based on the clinical characteristics of major depressive episodes in Chinese population. *J. Affect. Disord* 134(1–3), 119–125. 10.1016/j.jad.2011.05.054 [PubMed: 21684010]
- Grant BF, Dawson DA, Stinson FS, Chou PS, Kay W, Pickering R, 2003 The Alcohol Use Disorder and Associated Disabilities Interview Schedule-IV (AUDADISIV): Reliability of alcohol consumption, tobacco use, family history of depression and psychiatric diagnostic modules in a general population sample. *Drug. Alcohol. Depend* 71(1), 7–16. 10.1016/S0376-8716(03)00070-X [PubMed: 12821201]
- Grant BF, Goldstein RB, Chou SP, Huang B, Stinson FS, Dawson DA, ... Compton WM, 2009 Sociodemographic and psychopathologic predictors of first incidence of DSM-IV substance use, mood, and anxiety disorders: results from the Wave 2 National Epidemiologic Survey on Alcohol and Related Conditions. *Mol. Psychiatry* 14(11), 1051–1066. 10.1038/mp.2008.41 [PubMed: 18427559]
- Grant BF, Moore TC, Shepard J, Kaplan K, 2003 Source and accuracy statement: Wave 1 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC). Bethesda. MD. Natl. Inst. Alcohol. Abuse. Alcohol. 52.
- Hammen C, 2005 Stress and depression. *Annu. Rev. Clin. Psychol* 1, 293–319. [PubMed: 17716090]
- Harvey SB, Øverland S, Hatch SL, Wessely S, Mykletun A, Hotopf M, 2017 Exercise and the prevention of depression: results of the HUNT Cohort Study. *Am. J. Psychiatry* 175(1), 28–36. [PubMed: 28969440]
- Hasin DS, Grant BF, 2015 The National Epidemiologic Survey on Alcohol and Related Conditions (NESARC) Waves 1 and 2: Review and summary of findings. *Soc. Psychiatry. Psychiatr. Epidemiol* 50(11), 1609–1640. 10.1007/s00127-015-1088-0 [PubMed: 26210739]
- Hoppen TH, Chalder T, 2018 Childhood adversity as a transdiagnostic risk factor for affective disorders in adulthood: a systematic review focusing on biopsychosocial moderating and mediating variables. *Clin. Psychol. Rev* 65, 81–151. [PubMed: 30189342]

- Hosmer DW, Lemeshow S, Sturdivant RX, 2013 Applied logistic regression (Vol. 398). John Wiley & Sons.
- James G, Witten D, Hastie T, Tibshirani R, 2013 An introduction to statistical learning with Applications in R. Springer.
- James L, 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. *Lancet. Lond. Engl* 392(10159), 1789–1858. 10.1016/S0140-6736(18)32279-7
- Karsten J, Hartman CA, Smit JH, Zitman FG, Beekman AT, Cuijpers P, ... Penninx BW, 2011 Psychiatric history and subthreshold symptoms as predictors of the occurrence of depressive or anxiety disorder within 2 years. *Br. J. Psychiatry* 198(3), 206–212. [PubMed: 21357879]
- Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE, 2005 Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* 62(6), 593–602. 10.1001/archpsyc.62.6.593 [PubMed: 15939837]
- Kessler RC, Rose S, Koenen KC, Karam EG, Stang PE, Stein DJ, ... Carmen Viana M, 2014 How well can post-traumatic stress disorder be predicted from pre-trauma risk factors? An exploratory study in the WHO World Mental Health Surveys. *World Psychiatry*, 13(3), 265–274. 10.1002/wps.20150 [PubMed: 25273300]
- Kim-Cohen J, Caspi A, Moffitt TE, Harrington H, Milne BJ, Poulton R, 2003 Prior juvenile diagnoses in adults with mental disorder: developmental follow-back of a prospective-longitudinal cohort. *Arch. Gen. Psychiatry* 60(7), 709–717. 10.1001/archpsyc.60.7.709 [PubMed: 12860775]
- King M, Bottomley C, Bellón-Saameño JA, Torres-Gonzalez F, Švab I, Rifel J, ... Nazareth I, 2011 An international risk prediction algorithm for the onset of generalized anxiety and panic syndromes in general practice attendees: PredictA. *Psychol. Med* 41(8), 1625–1639. 10.1017/S0033291710002400 [PubMed: 21208520]
- King M, Bottomley C, Bellón-Saameño J, Torres-Gonzalez F, Švab I, Rotar D, ... Nazareth I, 2013 Predicting onset of major depression in general practice attendees in Europe: Extending the application of the PredictD risk algorithm from 12 to 24 months. *Psychol. Med* 43(09), 1929–1939. 10.1017/s0033291712002693 [PubMed: 23286278]
- King M, Walker C, Levy G, Bottomley C, Royston P, Weich S, ... Nazareth I, 2008 Development and validation of an international risk prediction algorithm for episodes of major depression in general practice attendees: the PredictD Study. *Arch. Gen. Psychiatry* 65(12), 1368–1376. 10.1001/archpsyc.65.12.1368 [PubMed: 19047523]
- Krueger RF, 1999 Personality traits in late adolescence predict mental disorders in early adulthood: a perspective-epidemiological study. *J. Pers* 67(1), 39–65. 10.1111/1467-6494.00047 [PubMed: 10030020]
- LeDell E, Van der Laan MJ, Petersen M, 2016 AUC-maximizing ensembles through metalearning. *Int. J. Biostat* 12(1), 203–218. 10.1515/ijb-2015-0035 [PubMed: 27227721]
- Lindert J, von Ehrenstein OS, Grashow R, Gal G, Braehler E, Weisskopf MG, 2014 Sexual and physical abuse in childhood is associated with depression and anxiety over the life course: Systematic review and meta-analysis. *Int. J. Public. Health* 59(2), 359–372. 10.1007/s00038-013-0519-5 [PubMed: 24122075]
- McLaughlin KA, Conron KJ, Koenen KC, Gilman SE, 2010 Childhood adversity, adult stressful life events, and risk of past-year psychiatric disorder: A test of the stress sensitization hypothesis in a population-based sample of adults. *Psychol. Med* 40(10), 1647–1658. 10.1017/S0033291709992121 [PubMed: 20018126]
- Merikangas KR, Zhang H, Avenevoli S, Acharyya S, Neuenschwander M, Angst J, 2003 Longitudinal trajectories of depression and anxiety in a prospective community study: the Zurich Cohort Study. *Arch. Gen. Psychiatry* 60(10), 993–1000. [PubMed: 14557144]
- Meyer A, Hajcak G, Torpey-Newman DC, Kujawa A, Klein DN, 2015 Enhanced error-related brain activity in children predicts the onset of anxiety disorders between the ages of 6 and 9. *J. Abnorm. Psychol* 124(2), 266–274. 10.1037/abn0000044 [PubMed: 25643204]
- Miloyan B, Bienvenu OJ, Brilot B, Eaton WW, 2018 Adverse life events and the onset of anxiety disorders. *Psychiatry. Res* 259, 488–492. [PubMed: 29154170]

- Moreno-Peral P, Luna J. de D., Marston L, King M, Nazareth I, Motrico E, ... Bellón JÁ, 2014 Predicting the onset of anxiety syndromes at 12 months in primary care attendees. The PredictA-Spain study. *PloS One*, 9(9), e106370 10.1371/journal.pone.0106370 [PubMed: 25184313]
- Polley E, LeDell E, Kennedy C, Lendle S, Laan M.van der, 2018 SuperLearner: Super learner prediction (Version 2.0–24). Retrieved from <https://CRAN.R-project.org/package=SuperLearner>
- Scott KM, Bruffaerts R, Tsang A, Ormel J, Alonso J, Angermeyer MC, ... Von Korff M, 2007 Depression–anxiety relationships with chronic physical conditions: Results from the World Mental Health surveys. *J. Affect. Disord* 103(1–3), 113–120. 10.1016/j.jad.2007.01.015 [PubMed: 17292480]
- Team R. C., 2013 R: A language and environment for statistical computing. R. Found. Stat. Comput. Vienna
- Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW, 2016 A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol* 74, 167–176. [PubMed: 26772608]
- Van der Laan MJ, Rose S, 2011 Targeted learning: causal inference for observational and experimental data. Springer Science & Business Media.
- Wald NJ, Bestwick JP, 2014 Is the area under an ROC curve a valid measure of the performance of a screening or diagnostic test? *J. Med. Screen* 21(1), 51–56. 10.1177/0969141313517497 [PubMed: 24407586]
- 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. *J. Epidemiol. Community. Health* 68(5), 418–424. 10.1136/jech-2013-202845 [PubMed: 24391206]
- Yonkers KA, Bruce SE, Dyck IR, Keller MB, 2003 Chronicity, relapse, and illness-course of panic disorder, social phobia, and generalized anxiety disorder: findings in men and women from 8 years of follow-up. *Depress. Anxiety* 17(3), 173–179. 10.1002/da.10106 [PubMed: 12768651]

Figure 1.

Word cloud of the features selected most frequently across predictor sets and outcomes (larger font indicates higher frequency of selection).

Figure 2.
Sensitivity (proportion of outcome cases) across ventiles of super learner predicted risk.

Figure 3.
Positive predictive value (outcome rate) across ventiles of super learner predicted risk.

Table 1.

Best performing 10-fold cross-validated models across feature sets and outcomes

	Generalized anxiety disorder	Panic disorder	Social phobia	Depression	Mania
Base Rate/Incidence, %	3.40	1.98	1.70	6.15	2.99
Best Individual Algorithm/Performance					
Max 5 Features [/]	LDA	LDA	GAM	GAM	LDA
AUC (95% CI)	0.7669 (0.7529–0.7810)	0.7752 (0.7572–0.7933)	0.7350 (0.7118–0.7581)	0.7281 (0.7155–0.7407)	0.7862 (0.7711–0.8013)
SL Weight	n/a	n/a	n/a	n/a	n/a
Max 10 Features [/]	LDA	LASSO	GAM	GAM	GAM
AUC (95% CI)	0.7821 (0.7687–0.7956)	0.7634 (0.7447–0.7822)	0.7739 (0.7536–0.7942)	0.7498 (0.7380–0.7616)	0.8169 (0.8036–0.8302)
SL Weight	n/a	0.0214	<.001	0.0400	<.001
Max 20 Features [/]	LDA	LDA	GAM	GAM	GAM
AUC (95% CI)	0.7988 (0.7863–0.8113)	0.7752 (0.7572–0.7933)	0.7942 (0.7756–0.8129)	0.7504 (0.7387–0.7620)	0.8231 (0.8103–0.8360)
SL Weight	0.1283	n/a	0.1604	0.0428	0.0548
All LASSO Features (n=101–104) [/]	LDA	LASSO	LASSO	Splines	LDA
AUC (95% CI)	0.7869 (0.7734–0.8004)	0.7678 (0.7492–0.7865)	0.7867 (0.7670–0.8063)	0.7472 (0.7354–0.7590)	0.8128 (0.7988–0.8269)
SL Weight	0.1022	0.0361	0.1039	0.0673	0.1119
All Significant Features (n=151–180) [/]	LASSO	LASSO	LASSO	LASSO	LASSO
AUC (95% CI)	0.7975 (0.7849–0.8101)	0.7796 (0.7620–0.7972)	0.7751 (0.7548–0.7954)	0.7552 (0.7435–0.7668)	0.8249 (0.8119–0.8379)
SL Weight	0.0586	0.1142	n/a	0.0829	0.1142
Ensemble Performance					
Number of classifiers used in the ensemble (max 9)	9	9	7	9	9
Number of algorithms used in the ensemble (max 44)	25	27	16	26	28
Superlearner CV AUC					
AUC (95% CI)	0.7991 (0.7865–0.8117)	0.7813 (0.7638–0.7988)	0.7990 (0.7803–0.8177)	0.7572 (0.7455–0.7688)	0.8261 (0.8131–0.8392)
Superlearner Full Sample AUC					
AUC (95% CI)	0.8174 (0.8057–0.8292)	0.8163 (0.8008–0.8319)	0.8369 (0.8206–0.8531)	0.7779 (0.7668–0.7890)	0.8553 (0.8444–0.8662)

Abbreviations. AUC, area under the receiver operating characteristic curve; 95% CI, 95% confidence interval; SL weight, super learner weight (n/a indicates weight of 0); LDA, linear discriminant analysis; GAM, generalized additive modeling; LASSO, least absolute shrinkage and selection operator; Splines, multivariate adaptive polynomial splines.

See Supplemental Tables 4–6 for the variables selected for each feature set and outcome J

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2.

Super learner performance based on optimized risk cut-point (Youden's Index)

	Optimal risk threshold ¹	% sample above threshold	Sensitivity	PPV	(O/E) ²	Specificity	NPV	(O/E) ³
Generalized Anxiety Disorder	.021	35.57	81.75	7.82	2.30	66.05	99.04	1.03
Panic Disorder	.015	29.18	76.20	5.16	2.61	71.77	99.34	1.01
Social Phobia	.018	21.09	71.96	5.81	3.41	79.79	99.40	1.01
Depression	.052	29.89	69.01	14.19	2.31	72.68	97.28	1.04
Mania	.021	26.79	79.69	8.88	2.97	74.84	99.17	1.02

Abbreviations. PPV, positive predictive value; NPV, negative predictive value.¹Optimal risk threshold is the optimal predicted value (0–1 range) multiplied by 100 for increased precision (0–100 range)²Observed (O) / Expected (E) sensitivity is calculated by dividing the observed sensitivity (above the optimized threshold) by the expected sensitivity (i.e., the % of the sample above the optimized risk threshold). O/E PPV is calculated by dividing the observed PPV (above the optimized risk threshold) by the overall rate/incidence (true positive rate) of the outcome and is equivalent to O/E sensitivity.³O/E specificity is calculated by dividing the observed specificity (below the optimized threshold) by the expected sensitivity (i.e., % of sample below the optimized risk threshold). O/E NPV is calculated by dividing the observed NPV (above the optimized risk threshold) by the overall true negative rate of the outcome and is equivalent to O/E specificity.