Predictive Modeling of Psychiatric Illness using Electronic Health Records and a Novel Machine Learning Approach with Artificial Intelligence

Matthew D. Nemesure, BS*[1,2]; Michael V. Heinz, MD [1,3]; Raphael Huang[1], and Nicholas C. Jacobson; PhD[1,2,4,5]

[1]Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College

[2]Quantitative Biomedical Sciences Program, Dartmouth College

[3]Dartmouth-Hitchcock Medical Center

[4]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College

[5]Department of Psychiatry, Geisel School of Medicine, Dartmouth College

Please cite as: Nemesure, M.D, Heinz, M. V., Huang, R. & Jacobson, N. C. (2021, In Press) Predictive Modeling of Psychiatric Illness using Electronic Medical Records and a Novel Machine Learning Approach with Artificial Intelligence. *Scientific Reports*.

* Corresponding Author:
Matthew D. Nemesure, BS
matthew.d.nemesure.gr@dartmouth.edu
46 Centerra Parkway
Suite 300, Office # 333S
Lebanon, NH 03766
(603) 646-7037

Abstract

Background: Generalized anxiety disorder (GAD) and major depressive disorder (MDD) are highly prevalent and impairing problems, but frequently go undetected, leading to substantial treatment delays. Electronic health records (EHRs) collect a great deal of biometric markers and patient characteristics that could foster the detection of GAD and MDD in primary care settings.

Methods: We approached the problem of predicting MDD and GAD using a novel machine learning pipeline to re-analyze data from an observational study. The pipeline constitutes an ensemble of algorithmically distinct machine learning methods, including deep learning. A sample of 4,184 undergraduate students completed the study, undergoing a general health screening and completing a psychiatric assessment for MDD and GAD. After explicitly excluding all psychiatric information, 59 biomedical and demographic features from the general health survey in addition to a set of engineered features were used for model training.

Results: We assessed the model's performance on a held-out test set and found an AUC of 0.73 (sensitivity: 0.66, specificity: 0.7) and 0.67 (sensitivity: 0.55, specificity: 0.7) for GAD, and MDD, respectively. Additionally, we used advanced techniques (SHAP values) to illuminate which features had the greatest impact on prediction for each disease. The top predictive features for MDD were being satisfied with living conditions and having public health insurance. The top predictive features for GAD were vaccinations being up to date and marijuana use.

Conclusions: Our results indicate moderate predictive performance for the application of machine learning methods in detection of GAD and MDD based on EHR data. By identifying biomarkers of GAD and MDD, these results may be used in future research to aid in the early detection of MDD and GAD.

Predictive Modeling of Psychiatric Illness using Electronic Health Records and a Novel Machine Learning Approach with Artificial Intelligence

Major depressive disorder (MDD) and generalized anxiety disorder (GAD) are prevalent psychiatric disorders that affect 16.2% and 13.3% of U.S. individuals, respectively, over their lifetimes.[1,2] MDD is the leading cause of disability worldwide,[3,4] and anxiety disorders are the sixth leading cause of disability.[5] MDD is characterized by persistent low mood, associated with disturbances with sleep, motivation, energy, appetite, and suicidal thoughts.[6] GAD represents a persistent, uncontrollable pattern of worry occurring in multiple domains of an individual's life.[7] Left untreated, these syndromes often have devastating consequences for affected individuals, their families, and communities.[8,9]

Both MDD and GAD are prevalent in the college population. In a 2015 study, 23% of surveyed college students reported moderate to severe depressive symptoms.[10] Similarly, a 2019 study showed a 20% prevalence of GAD among college students in 2016, representing a 100% increase since 2008.[11] These syndromes negatively impact multiple domains of an individual's functioning, and for college students, this may include interference with class attendance and learning retention.[12] Research among college students found that students with depression are more likely to report drinking-related harms and alcohol abuse.[13]

Two major challenges in adequately addressing MDD and GAD are identifying affected individuals and ensuring appropriate and timely treatment. Because MDD and GAD symptoms are internally experienced, MDD and GAD often go undetected.[14–16] There is an estimated 6 year and 14 year delay between disease onset and intervention for MDD and GAD, respectively, during which time the disease may increase in severity, lowering student quality of life.[17,18]

Early detection and diagnosis is paramount to understanding and addressing mental illness on a populational level. With the rise in electronic health records (EHRs), spurred by initiatives like the Health Information Technology Act (Rights (OCR), 2009), there is increasing potential for addressing previously intractable clinical questions using computational analysis of large data sets. Multiple studies show promise in this area.[19–23]

A 2011 study by Trinh et al.[19] found that an EHR billing diagnosis of "depression" can serve as an effective proxy for identifying clinical depression. Although this study did not exploit advanced statistical models, it demonstrated prediction of psychiatric pathology using structured EHR data, albeit the clinical utility of these predictive models is questionable given that the predictors used were closely related to outcome. Perlis et al.[20] found improvements in prediction of MDD using unstructured clinical narrative features (extracted with NLP) and billing code data, compared with using billing code data alone. A more recent 2019 study by Wang et al.[21] utilized machine learning techniques for prediction of postpartum depression (PPD). The predictors were extracted from the EHR and the model ended up with a good predictive accuracy. Features found to be significant included *depression*, *anxiety*, *use of antidepressant drugs*, and *pain diagnoses*. Geraci et al.[22] used data extracted from psychiatric clinical texts to predict a diagnosis of depression, including both structured or unstructured psychiatric diagnoses. Huang et al.[23] exploit multiple structured features to predict depression, including diagnostic codes and patient prescriptions, which could include psychiatric medications.

Although promising early directions, a common limitation in these studies[19–23] is the use of features highly interdependent with MDD, including psychiatric billing codes or unstructured notes, likely containing explicit diagnostic information. This presents as a major limitation to the potential utility of using these prior studies to close the onset to treatment gap among those with

MDD and GAD. In particular, diagnostic codes could only be obtained from those whose MDD and GAD would have already been detected.

Based on the limitations of prior studies that utilized psychiatric features to predict GAD and MDD, our study utilized an EHR dataset containing biometric and demographic data from 4,184 undergraduate students. Excluding all psychiatric features, we approach the problem of identification and diagnosis using a novel machine learning pipeline developed for the purpose of this study. The pipeline constitutes an ensemble of multiple algorithmically distinct machine learning methods, including deep learning methods. We trained the model to predict psychiatric illness using varied non-psychiatric input features such as blood pressure, heart rate, housing status, and public insurance. This is to say, unlike all prior studies, we did not use any psychiatric information in predicting diagnosis of GAD or MDD. We hypothesized that using such biomedical data, we could predict MDD and GAD with a level of certainty above chance. Our primary aim was to identify biomarkers for GAD and MDD risk.

Methods

Participants

Four thousand one hundred and eighty four undergraduate students from the University of Nice Sophia-Antipolis underwent a basic medical examination and participated in the current study. All data was publicly available on Dryad and completely de-identified and therefore this research does not meet the federal definition for human subjects research. Additionally, according to the original study, the National Data Protection Authority (NCIL) approved the study.[24] The methods of the study carried out in France were in accordance with the laws of non-interventional clinical research.[24] Due to this being an observational study in compliance with laws that regulate non-interventional clinical research in France (articles L.1121-1 and R.1121-2 of the Public Health Code), informed consent was not required.[24] Additionally, this study received institutional exemption from the Committee for the Protection of Human Subjects at Dartmouth College. These students were 57.4% female and 42.6% male and their ages were split into four categories: less than 18, 18, 19 and 20 or older. The distribution among these categories was as follows: 5%, 36%, 28% and 31%. The outcomes of interest, MDD and GAD, had base rates of 12% and 8% respectively. [24]

Features

A total of 59 features were used including binary, ordinal and continuous variables. Specifically, features included age (4 levels: under 18, 18, 19, over 20), gender, French nationality, field of study, year of university, learning disabilities, difficulty memorizing lessons, professional objective (whether the student indicated an objective), informed about opportunities (whether the student indicated that they felt informed about opportunities at the university), satisfied with living conditions, living with a partner/child, parental home, having only one parent, at least one parent unemployed, siblings (yes/no), long commute, mode of transportation, financial difficulties, grant (yes/no), additional income (yes/no), public health insurance, private health insurance, universal health coverage, irregular rhythm of meals, unbalanced meals, eating junk food, on a diet, irregular rhythm or unbalanced meals, physical activity (3 levels: none, occasional, regular) , physical activity (2 levels: none or occasional, regular), weight (kg), height (cm), overweight and obesity, systolic blood pressure (mmHg), diastolic blood pressure (mmHg), prehypertension or hypertension, heart rate (bpm), abnormal heart rate, distant visual acuity of right eye (score/10), distant visual acuity of left eye (score/10), close visual acuity of right eye (score/10), close visual acuity of left eye (score/10), decreased in distant visual acuity, decreased in close visual acuity, urinalysis (glycosuria), urinalysis (proteinuria), urinalysis (hematuria), urinalysis (leukocyturia), urinalysis (positive nitrite test), abnormal urinalysis, vaccination up to date, control examination needed (whether the student needed a follow-up for any reason), cigarette smoker (5 levels: none, occasional, regular, frequent, heavy), cigarette smoker (3 levels: no, frequent, occasional), drinker (3 levels: no, occasional, regular), drinker (2 levels: no or occasional, regular or heavy), binge drinking, marijuana use, other recreational drugs.

Psychiatric Diagnoses

The outcomes of interest were MDD and GAD. MDD and GAD were each assessed in a multi-stage process. The first stage included a screening questionnaire that assessed four hallmark symptoms of MDD (anhedonia, loss of energy/fatigue, changes in activity and depressed mood) and four hallmark symptoms of GAD (excessive worry, restlessness, fatigue, and irritability). If the assessment indicated possible presence of either disorder (positive answer to two of the four categories), the participants were assessed for full Diagnostic and Statistical Manual of Mental Disorders Fourth Edition (DSM IV) criteria by a medical provider[24].

Data Preprocessing

The preprocessing pipeline included creating dummy variables for ordinal outcomes, normalizing continuous variables, and single imputation for missing values using a Bayesian Ridge approach across features. A total of 20 of the 59 variables included NA values and the percentage missing ranged from <1% to 36%. Total missingness was 5% and median missingness across all variables was 0%

To enhance our model, we used feature engineering, informed by domain specific biomedical knowledge. Feature engineering as used in our study refers to the combination of distinct features into new “engineered” features, which have domain specific meaning and utility. Previous research has shown feature engineering to improve machine learning model performance.[25,26] By combining existing features, we created and used (1) Body Mass Index (BMI),[27] (2) Mean Arterial Pressure (MAP),[28] and (3) Pulse Pressure.[29] BMI is a function of an individual's height and weight. MAP and pulse pressure are clinically meaningful combinations of diastolic and systolic blood pressure.

Data analysis

The first step of analysis was dividing the data into 70% training (*N*=2929) and 30% (*N*=1255) held out testing (see *Figure 1*). The held out test set remained unseen throughout model training and was never used for hyperparameter tuning. The machine learning pipeline included six algorithmically unique machine learning classifiers to inform final predictions. These classifiers were XGBoost, Random Forest, Support Vector Machine, K-nearest-neighbors and a neural network tuned using Bayesian hyperparameter optimization. A 5-fold validation technique was used to train each model. This allowed for each model type (e.g., logistic regression) to make one prediction for each subject in the training set. These predictions were saved to be used as inputs to a “higher level” model that would eventually make final predictions.

The aforementioned “higher level” model was an XGBoost classifier which was trained, again, using 5-fold validation, on the predictions of the original 6 models. Essentially, each “lower level” model made a prediction (i.e. probability of MDD) for each subject and the higher level model decided which model’s predictions were most informative based on the true outcome. Using this information, the higher-level model made a final estimation for the probability of the outcome of interest.

These models were then used to make predictions on the held out test set to ensure there was no overfitting and that the results were meaningful and generalizable. To create the prediction matrix on the held out test set, all 5 saved models for each machine learning method made predictions on each subject. The predictions for each model type were then averaged and filled into the prediction matrix. The high level XGboost model then made final predictions. The area under the receiver operating characteristic curve (AUC) is a measure of how well the model can effectively distinguish between psychiatric diagnosis, reflecting the model performance in optimizing across both sensitivity and specificity. To guide interpretation of the results, please note that an AUC = 0.58 represents a small effect size, AUC = 0.69 represents a medium effect size, and AUC = 0.79 represents a large effect size, based on conversions to Cohen's d values of 0.2, 0.5, and 0.8 respectively.[30] This pipeline was used twice, once with the outcome being GAD and once with the outcome being MDD.

Model Explainability

SHAP (Shapley Additive Explanations) scores were utilized calculate and visualize feature importance this complex model.[31] The SHAP kernel explainer allows for a user to input data and a prediction function and it will return the relative importance for each feature for each subject. The prediction function, in this case, simply took the input data and utilized the trained models from the pipeline to make predictions. These predictions were then averaged across the lower level models and fed into the upper level model. The upper level model returned the final prediction for each subject. With this setup, the kernel explainer would return the SHAP values for each of the features from the original input data based on how it informed the entire pipeline's prediction.

Results

Predictive Performance

The main results of this study are two-fold, the first is the prediction accuracy of the stacked machine learning models and the second is the important features driving those predictions. The validation and test-set AUC for MDD (see Figure 2) and GAD (see *Figure 3*) were (0.70, 0.67) and (0.70, 0.73) respectively. Thus, the ensemble model could predict diagnosis of MDD and GAD well above chance and with a medium effect size. Additionally, when compared to a simple standard logistic regression as run in the original study, the AUCs of the complex machine learning models were increased, on average, by 0.08 (*figure 2B and 3B)*. Given the AUC curve of the model, we can choose thresholds with higher sensitivity at the detriment of specificity. Given the non-invasive nature of secondary screening for each of these illnesses, it seems reasonable to allow a soft threshold for further diagnosis. Specifically, for MDD, the sensitivity and specificity were 55% and 70% respectively. Additionally, the positive predictive value was 20% and the negative predictive value was 92%. For GAD, the sensitivity and specificity were 70% and 66% respectively. The positive predictive value was 16% and the negative predictive value was 96%.

Model Explainability

The second and arguably more important set of results are the important features and how they inform predictions (Figures 4 and 5). The top features (figure 4a and 5a) are the most informative to the model but it is important to note that the impact of features on the outcome was distributed across a large number of features (i.e. the SHAP values for top features were small). This is likely indicative of the complex and heterogenous nature of the disease. To ascertain either MDD or GAD status, it requires a not just a singular biomarker but rather a combination of features and feature interactions to accurately assess the disease state. This exemplifies the necessity for complex models to disentangle the relationships between variables and characterize and assess the disease in any given person.

MDD (See Figure 4): The most important feature driving the prediction of MDD was whether the student was satisfied with their living conditions (4b). High diastolic blood pressure was also indicative of MDD and having public health insurance indicated, for the most part, non-MDD status (4c). In order, living in a parental home, mean arterial pressure and difficulty memorizing lessons made up the remaining important predictors from the top six. Additionally, after further assessing these top features, it was noted that many of them were predictive as part of two-way interactions, such that the relationship between a predictor and an outcome is conditional on another predictor. As seen in *Figure 4d*, typically individuals without public health insurance had lower predictions of MDD, but the extent was conditional on whether they were satisfied with their living conditions. Those who were satisfied with their living conditions seemed to be slightly more informative in telling the model that MDD was not apparent.

GAD (See Figure 5): The most important predictor for GAD was having up to date vaccinations (4b). Another similar and important variable for prediction was the necessity for a control examination. This was essentially a binary indicator for whether or not the student needed to return to the doctor for something unrelated to the psychiatric outcome. The second most important predictor was marijuana use although the effect of this variable on model prediction was clearly impacted by interactions with other subject characteristics (4c). The remaining top six most important predictors were, in order, hypertension or prehypertension, systolic blood pressure and the use of other recreational drugs. These features, overall, were all much closer in importance than in MDD. This further indicated the model's reliance on all features, not just one biomarker. Again, there were very clear two-way interactions between variables when the model was making predictions. Smoking marijuana was clearly more indicative of predicted GAD if the individual was overweight or obese (4d). Other interactions included systolic blood pressure with prehypertension and hypertension and the necessity of a control examination with gender.

Discussion

Our objective was to evaluate the importance and effectiveness of standard clinical data on the prediction of MDD or GAD. We used state-of-the-art novel machine learning methodologies to make predictions. Additionally, SHAP values were generated to explain and clinically validate our findings. We trained our model with >2500 participants and assessed the model's performance on a held-out test set. Although our accuracy metrics are comparable to previous studies predicting psychiatric outcomes, ours is unique in its primary reliance on routine biomedical and demographic features, rather than features with a known correlation to psychiatric outcomes. Previous studies that have looked at EHR to detect MDD have had the significant limitation of including predictive variables that would nullify the clinical utility of the model by relying on features that are directly indicative of known psychiatric illness (e.g. including psychiatric billing codes, which are based upon clinician diagnosis). Thus, this study is the first known study to predict MDD and GAD using EHR data with potential for predictive validity in detecting unknown psychiatric diagnoses.

Studies using magnetic resonance imaging (MRI) have been able to achieve slightly higher predictive performances ranging from 67% to 94%.[32] Nevertheless, perhaps due to the considerable expense of collecting MRI data, a common limitation of these was their small sample sizes. These studies also had considerable range in performance, and the due to their small sample sizes the results are highly inconsistent.[33] Moreover, using MRI to predict MDD is unrealistic when there is no other reason to justify an MRI, especially in an otherwise physically healthy college-age patient.

In addition to the complex machine learning approach and our carefully curated feature set, we are providing insights to the complex clinical appearance of MDD. Our pipeline, using SHAP values to visualize feature importance, provides not only the outcome prediction but the possible characteristics that a physician can identify when making a decision. These characteristics including mean arterial pressure, blood pressure, markers for low SES and general health markers have been shown to be previously associated with depression and anxiety.[34,35]

In further investigation of the predictors for generalized anxiety disorder, vaccination status may be reflective of overall poorer health outcomes in individuals with GAD [36]. Regarding the “marijuana use”, prior research demonstrates high comorbidity between anxiety disorders and substance use disorders. [37] With regard to the most important features driving major depressive disorder, there is research supporting overall poorer life satisfaction in individuals with MDD, [38]which may certainly include dissatisfaction with living conditions. Low interest and energy, DSM criteria for MDD, may contribute to difficulties maintaining satisfactory living conditions. Robust research to date indicates that individuals of lower socioeconomic status are more likely to have MDD. [39] “Difficulty memorizing lessons” may be related to concentration difficulties, also identified by the DSM as a clinical feature of MDD. An additional top predictive feature for both MDD and GAD is hypertension. Research to date corroborates this finding by demonstrating that individuals with either MDD or GAD are more likely to have hypertension. [40,41]

This information has the potential to allow health care providers to make informed recommendations for further screening regardless of whether the patient discusses or even recognizes his or her symptoms. This is important because as previously mentioned, it can take on average 6 or 14 years from onset of illness until diagnosis for MDD and GAD respectively. [17] Our study is one of the first of its kind to tackle this issue by not relying on previous psychiatric diagnoses or expensive imagine techniques to capture the disease in an early stage.

This study has several important limitations which deserve mention. One is that the original screening for the outcomes of MDD and GAD may not have captured all cases within the population. This, in addition to the study population, limits the generalizability of the results. Our dataset comes from French college aged students, who likely have baseline differences from other populations with psychiatric illness. Despite this limitation, our study still serves to show the predictive ability of mainly non-psychiatric variables for psychiatric illness. Such variables, further analyzed individually for their connection to psychiatric pathology, may prove the basis of further research. Another limitation of our study, which is fairly ubiquitous in mental health research is the low prevalence of anxiety and depression in our study population, as well as our sample size. Although this is a limitation in many studies of psychiatric nature, we were able to enhance our predictive power using a stacked ensemble model pipeline. Additionally, the lack of qualitative information (i.e., severity, subtype, etc.) regarding mental health diagnoses was not available to allow for a severity prediction analysis. Thus, future research should examine the potential for these biomarkers to predict severity and subtype of MDD and GAD.

This research is an important step in the direction towards identifying potentially difficult to diagnose illnesses with readily available and easy to obtain information. Our tool, using an optimal sensitivity/specificity split would be able to capture two out of every three subjects with GAD and one out of two MDD cases while only incurring a 30% false positive rate. Because there are detrimental outcomes to both the patient and provider in a false positive, looking at the efficacy of case identification while requiring 70% specificity gives a reasonable idea of how many cases would be captured if this model were to be deployed in a clinical setting. These findings have shown promise on multiple fronts: Ability to use easy to obtain information to inform possible detection of MDD and GAD, further understanding of the demographic and biological characteristics associated with illness, and both the success and necessity for computational tools to inform psychological medicine. We believe, given a larger and more heterogeneous sample, this modeling technique could be used to elucidate the drivers of psychological illness and provide a tool that indicates the necessity of treatment with high precision and accuracy.

References

1. Bystritsky A, Khalsa SS, Cameron ME, Schiffman J. Current Diagnosis and Treatment of Anxiety Disorders. *Pharm Ther*. 2013;38(1):30-57.

2. Kessler RC, Berglund P, Demler O, et al. The Epidemiology of Major Depressive Disorder: Results From the National Comorbidity Survey Replication (NCS-R). *JAMA*. 2003;289(23):3095-3105. doi:10.1001/jama.289.23.3095

3. Mathers C, Fat DM, Boerma JT, World Health Organization, eds. *The Global Burden of Disease: 2004 Update*. World Health Organization; 2008.

4. Reddy MS. Depression: The Disorder and the Burden. *Indian J Psychol Med*. 2010;32(1):1-2. doi:10.4103/0253-7176.70510

5. Friedrich MJ. Depression Is the Leading Cause of Disability Around the World. *JAMA*. 2017;317(15):1517-1517. doi:10.1001/jama.2017.3826

6. Fava M, Kendler KS. Major Depressive Disorder. *Neuron*. 2000;28(2):335-341. doi:10.1016/S0896-6273(00)00112-4

7. Lader M. Generalized Anxiety Disorder. In: Stolerman IP, Price LH, eds. *Encyclopedia of Psychopharmacology*. Springer; 2015:699-702. doi:10.1007/978-3-642-36172-2_317

8. Bonari L, Pinto N, Ahn E, Einarson A, Steiner M, Koren G. Perinatal Risks of Untreated Depression during Pregnancy: *Can J Psychiatry*. Published online November 1, 2004. doi:10.1177/070674370404901103

9. Ghio L, Gotelli S, Cervetti A, et al. Duration of untreated depression influences clinical outcomes and disability. *J Affect Disord*. 2015;175:224-228. doi:10.1016/j.jad.2015.01.014

10. Beiter R, Nash R, McCrady M, et al. The prevalence and correlates of depression, anxiety, and stress in a sample of college students. *J Affect Disord*. 2015;173:90-96. doi:10.1016/j.jad.2014.10.054

11. Scheffler R. Impact of Anxiety and Depression on Student Academic Progress. IBCCES. Published May 1, 2019. Accessed October 19, 2019. https://ibcces.org/blog/2019/05/01/impact-anxiety-depression-student-progress/

12. Alonso J, Liu Z, Evans-Lacko S, et al. Treatment gap for anxiety disorders is global: Results of the World Mental Health Surveys in 21 countries. *Depress Anxiety*. 2018;35(3):195-208. doi:10.1002/da.22711

13. Weitzman ER. Poor Mental Health, Depression, and Associations With Alcohol Consumption, Harm, and Abuse in a National Sample of Young Adults in College. *J Nerv Ment Dis*. 2004;192(4):269–277. doi:10.1097/01.nmd.0000120885.17362.94

14. Kessler D, Bennewith O, Lewis G, Sharp D. Detection of depression and anxiety in primary care: follow up study. *BMJ*. 2002;325(7371):1016-1017.

15. Kessler D, Lloyd K, Lewis G, Gray DP. Cross sectional study of symptom attribution and recognition of depression and anxiety in primary care. *BMJ*. 1999;318(7181):436-440.

16. Löwe B, Gräfe K, Zipfel S, et al. Detecting panic disorder in medical and psychosomatic outpatients: comparative validation of the Hospital Anxiety and Depression Scale, the Patient Health Questionnaire, a screening question, and physicians' diagnosis. *J Psychosom Res*. 2003;55(6):515-519. doi:10.1016/s0022-3999(03)00072-2

17. Kessler RC, Olfson M, Berglund PA. Patterns and Predictors of Treatment Contact After First Onset of Psychiatric Disorders. *Am J Psychiatry*. 1998;155(1):62-69. doi:10.1176/ajp.155.1.62

18. Thompson A, Issakidis C, Hunt C. Delay to Seek Treatment for Anxiety and Mood Disorders in an Australian Clinical Sample. *Behav Change*. 2008;25(2):71-84. doi:10.1375/bech.25.2.71

19. Trinh N-HT, Youn SJ, Sousa J, et al. Using electronic medical records to determine the diagnosis of clinical depression. *Int J Med Inf*. 2011;80(7):533-540. doi:10.1016/j.ijmedinf.2011.03.014

20. Perlis RH, Iosifescu DV, Castro VM, et al. Using electronic medical records to enable large-scale studies in psychiatry: treatment resistant depression as a model. *Psychol Med*. 2012;42(1):41-50. doi:10.1017/S0033291711000997

21. Wang S, Pathak J, Zhang Y. Using Electronic Health Records and Machine Learning to Predict Postpartum Depression. Published 2019. Accessed December 4, 2019. http://ebooks.iospress.nl/publication/52116

22. Geraci J, Wilansky P, de Luca V, Roy A, Kennedy JL, Strauss J. Applying deep neural networks to unstructured text notes in electronic medical records for phenotyping youth depression. *Evid Based Ment Health*. 2017;20(3):83-87. doi:10.1136/eb-2017-102688

23. Huang SH, LePendu P, Iyer SV, Tai-Seale M, Carrell D, Shah NH. Toward personalizing treatment for depression: predicting diagnosis and severity. *J Am Med Inform Assoc*. 2014;21(6):1069-1075. doi:10.1136/amiajnl-2014-002733

24. Tran A, Tran L, Geghre N, et al. Health assessment of French university students and risk factors associated with mental health disorders. *PLOS ONE*. 2017;12(11):e0188187. doi:10.1371/journal.pone.0188187

25. Garla VN, Brandt C. Ontology-guided feature engineering for clinical text classification. *J Biomed Inform*. 2012;45(5):992-998. doi:10.1016/j.jbi.2012.04.010

26. Xu Y, Hong K, Tsujii J, Chang EI-C. Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries. *J Am Med Inform Assoc*. 2012;19(5):824-832. doi:10.1136/amiajnl-2011-000776

27. Stensland SH, Margolis S. Simplifying the calculation of body mass index for quick reference. *J Am Diet Assoc*. 1990;90(6):856.

28. MEANEY E, ALVA F, MOGUEL R, MEANEY A, ALVA J, WEBEL R. Formula and nomogram for the sphygmomanometric calculation of the mean arterial pressure. *Heart*. 2000;84(1):64. doi:10.1136/heart.84.1.64

29. Franklin Stanley S., Khan Shehzad A., Wong Nathan D., Larson Martin G., Levy Daniel. Is Pulse Pressure Useful in Predicting Risk for Coronary Heart Disease? *Circulation*. 1999;100(4):354-360. doi:10.1161/01.CIR.100.4.354

30. Salgado JF. Transforming the Area under the Normal Curve (AUC) into Cohen's d, Pearson's r pb , Odds-Ratio, and Natural Log Odds-Ratio: Two Conversion Tables. *Eur J Psychol Appl Leg Context*. 2018;10(1):35-47. doi:10.5093/ejpalc2018a5

31. Lundberg SM, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell*. 2020;2(1):56-67. doi:10.1038/s42256-019-0138-9

32. Patel MJ, Khalaf A, Aizenstein HJ. Studying depression using imaging and machine learning methods. *NeuroImage Clin*. 2016;10:115-123. doi:10.1016/j.nicl.2015.11.003

33. Toenders YJ, van Velzen LS, Heideman IZ, Harrison BJ, Davey CG, Schmaal L. Neuroimaging predictors of onset and course of depression in childhood and adolescence: A systematic review of longitudinal studies. *Dev Cogn Neurosci*. 2019;39:100700. doi:10.1016/j.dcn.2019.100700

34. Licht CMM, de Geus EJC, Seldenrijk A, et al. Depression Is Associated With Decreased Blood Pressure, but Antidepressant Use Increases the Risk for Hypertension. *Hypertension*. 2009;53(4):631-638. doi:10.1161/HYPERTENSIONAHA.108.126698

35. Defoe IN, Farrington DP, Loeber R. Disentangling the relationship between delinquency and hyperactivity, low achievement, depression, and low socioeconomic status: Analysis of repeated longitudinal data. *J Crim Justice*. 2013;41(2):100-107. doi:10.1016/j.jcrimjus.2012.12.002

36. Louise P, Siobhan O, Louise M, Jean G. The burden of generalized anxiety disorder in Canada. *Health Promot Chronic Dis Prev Can Res Policy Pract*. 2017;37(2):54-62.

37. Alegría AA, Hasin DS, Nunes EV, et al. Comorbidity of Generalized Anxiety Disorder and Substance Use Disorders: Results From the National Epidemiologic Survey on Alcohol and Related Conditions. *J Clin Psychiatry*. 2010;71(09):1187-1195. doi:10.4088/JCP.09m05328gry

38. Rissanen T, Viinamäki H, Lehto SM, et al. The role of mental health, personality disorders and childhood adversities in relation to life satisfaction in a sample of general population. *Nord J Psychiatry*. 2013;67(2):109-115. doi:10.3109/08039488.2012.687766

39. Everson SA, Maty SC, Lynch JW, Kaplan GA. Epidemiologic evidence for the relation between socioeconomic status and depression, obesity, and diabetes. *J Psychosom Res*. 2002;53(4):891-895. doi:10.1016/S0022-3999(02)00303-3

40. Härter MC, Conway KP, Merikangas KR. Associations between anxietydisorders and physical illness. *Eur Arch Psychiatry Clin Neurosci*. 2003;253(6):313-320. doi:10.1007/s00406-003-0449-y

41. Wu E-L, Chien I-C, Lin C-H, Chou Y-J, Chou P. Increased risk of hypertension in patients with major depressive disorder: A population-based study. *J Psychosom Res*. 2012;73(3):169-174. doi:10.1016/j.jpsychores.2012.07.002

Acknowledgements

This work was funded by National Institute of Drug Abuse (NIDA P30DA02992610).

Author Contributions

M.D.N and N.C.J designed the study. M.D.N. and M.V.H. prepared and analyzed the data and wrote sections of the Manuscript. R.H. Wrote part of the manuscript. N.C.J. was a mentor throughout the project and assisted in writing the manuscript.

Financial Disclosures

Dr. Jacobson is the owner of a free application entitled “Mood Triggers”. He does not receive any direct or indirect revenue from his ownership of the application (i.e. the application is free, there are no advertisements, and the data is only being used for research purposes).

Matthew Nemesure has no Financial disclosures or conflicts of interest. Michael V. Heinz has no Financial disclosures or conflicts of interest. Raphael Huang has no financial disclosures or conflicts of interest.

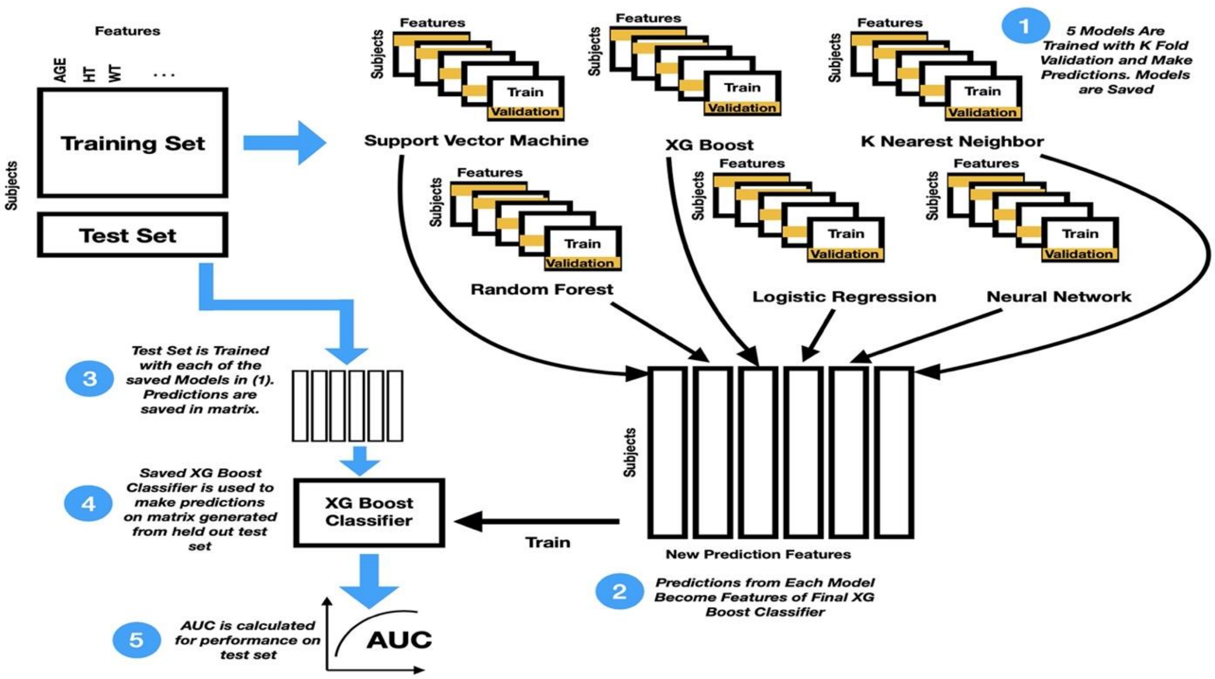
Figure 1. This is the pipeline used to train the machine learning models and generate predictions. The training set is sent through 5-fold training for each model type to generate a prediction for each training sample. These predictions are then used to train a higher level model to predict a final outcome given the predictions from the 5-fold training. Each of the 6 models from each fold then predicts on the held out test set and the average prediction for the probability of depression is stored. The higher level model then makes final predictions on the held out test set.

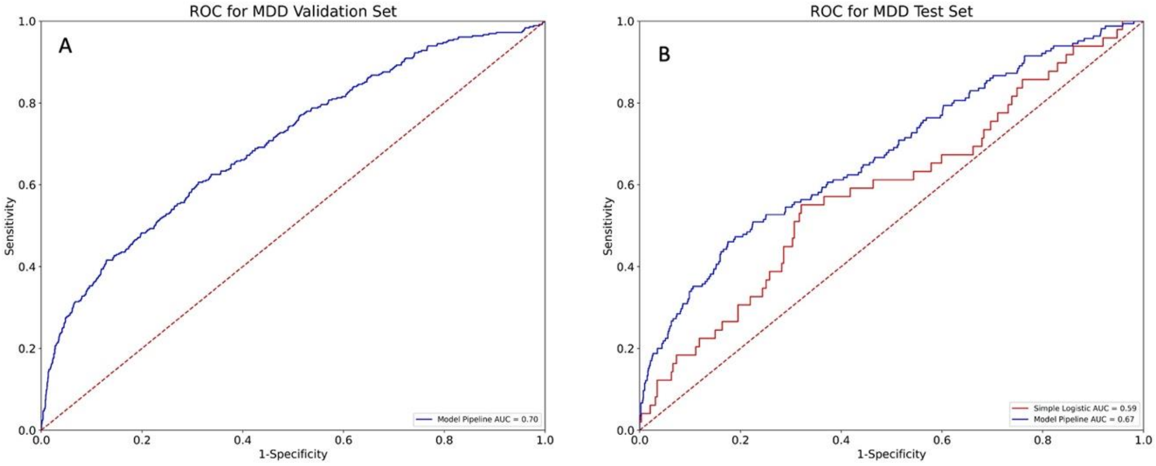
Figure 2. A: AUC for prediction of MDD in the training set. B: AUC for the prediction of depression in the held-out test set using both a simple logistic regression and our novel pipeline. These curves show the sensitivity and specificity at different thresholds for prediction.

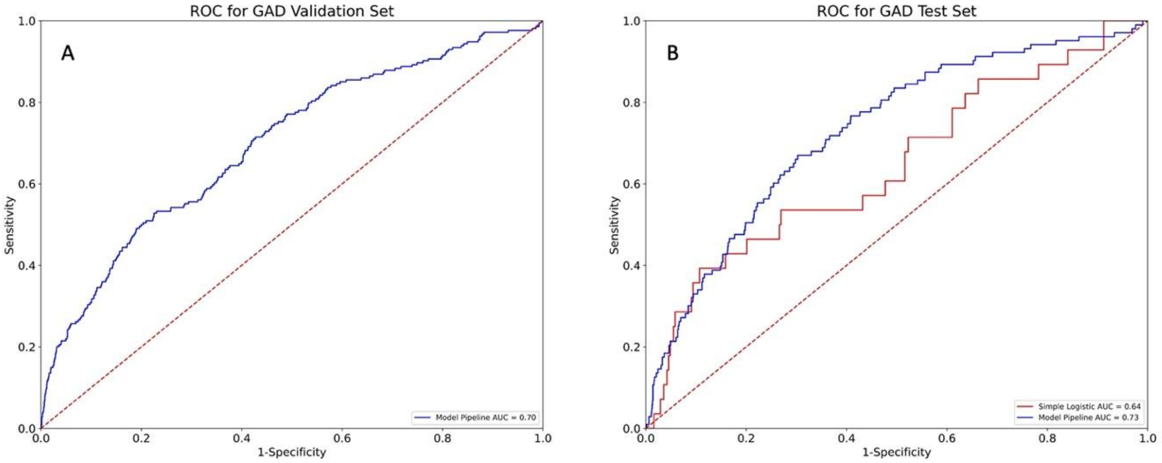
Figure 3. A: AUC for prediction of GAD in the training set. B: AUC for the prediction of anxiety in the held-out test set using both a simple logistic regression and our novel pipeline for prediction. These curves show the sensitivity and specificity at different thresholds for prediction.

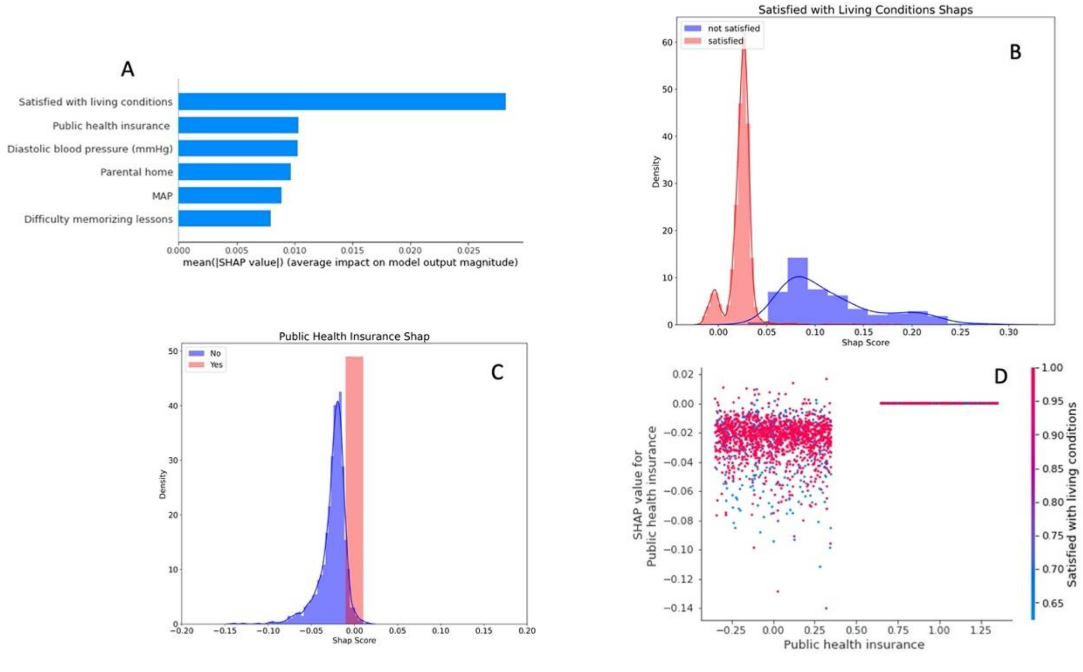
Figure 4. A: This plot shows the top six most important features for predicting MDD. This is displayed as the mean of the absolute value of SHAP scores across all subjects for that given feature. A higher SHAP value indicates that the feature was important in informing the models prediction. B: This plot displays the density distribution of SHAP values for the top performing feature in predicting depression. C: This plot also displays the density distribution of SHAP values for the second most important feature in predicting MDD. Positive SHAP score indicates that the feature was indicative of the subject having MDD. D: This is an interaction plot showing the effect of two features working together to inform the model. Here it is apparent that when a student does not have public health insurance, living conditions can partially inform prediction.

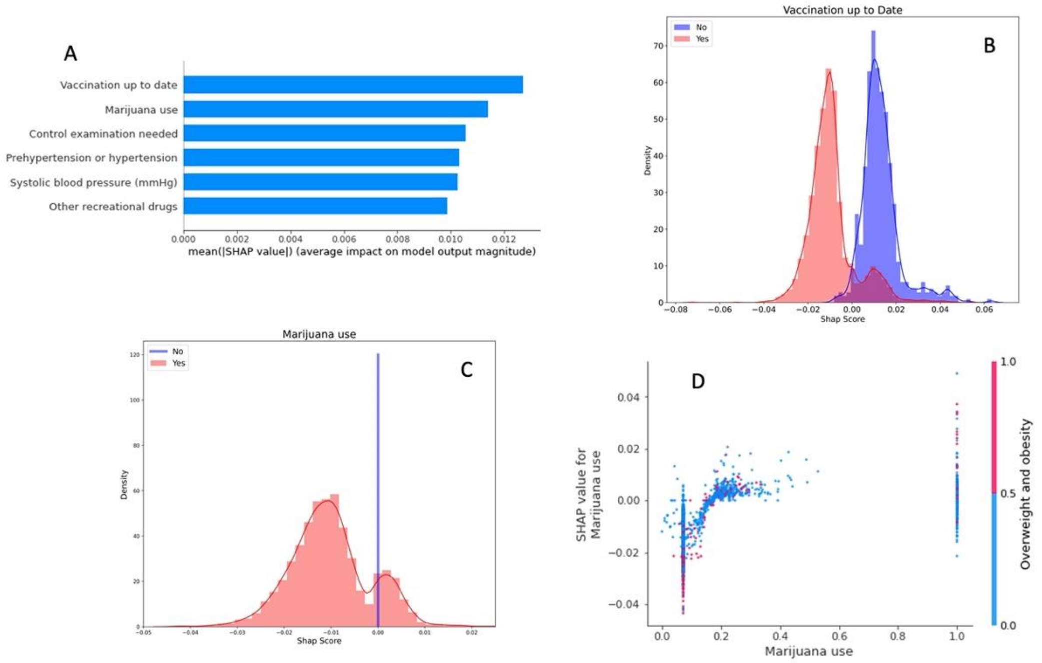
Figure 5. A: This plot shows the top six most important features for predicting GAD. This is displayed as the mean of the absolute value of SHAP scores across all subjects for that given feature. A higher SHAP value indicates that the feature was important in informing the models prediction. B: This plot displays the density distribution of SHAP values for the top performing feature in predicting GAD. C: This plot also displays the density distribution of SHAP values for the second most important feature in predicting GAD. Positive SHAP score indicates that the feature was indicative of the subject having GAD. D: This is an interaction plot showing the effect of two features working together to inform the model. Here it is apparent that marijuana use is more predictive of GAD in overweight students.


Features
AGE
HT
WT
. . .
Subjects
Training Set
Test Set
Features
Subjects
Train
Validation
Support Vector Machine
XG Boost
K Nearest Neighbor
Random Forest
Logistic Regression
Neural Network
1
5 Models Are Trained with K Fold Validation and Make Predictions. Models are Saved
2
Predictions from Each Model Become Features of Final XG Boost Classifier
New Prediction Features
Train
3
Test Set is Trained with each of the saved Models in (1). Predictions are saved in matrix.
4
Saved XG Boost Classifier is used to make predictions on matrix generated from held out test set
XG Boost Classifier
5
AUC is calculated for performance on test set
AUC


ROC for MDD Validation Set
A
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Model Pipeline AUC = 0.70
ROC for MDD Test Set
B
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Simple Logistic AUC = 0.59
Model Pipeline AUC = 0.67


ROC for GAD Validation Set
A
Sensitivity
1-Specificity
0.0
0.2
0.4
0.6
0.8
1.0
Model Pipeline AUC = 0.70
ROC for GAD Test Set
B
Sensitivity
1-Specificity
Simple Logistic AUC = 0.64
Model Pipeline AUC = 0.73


A
Satisfied with living conditions
Public health insurance
Diastolic blood pressure (mmHg)
Parental home
MAP
Difficulty memorizing lessons
0.000
0.005
0.010
0.015
0.020
0.025
mean(|SHAP value|) (average impact on model output magnitude)
Satisfied with Living Conditions Shaps
B
not satisfied
satisfied
Density
Shap Score
Public Health Insurance Shap
C
No
Yes
Density
Shap Score
D
SHAP value for
Public health insurance
Public health insurance
Satisfied with living conditions


A
Vaccination up to date
Marijuana use
Control examination needed
Prehypertension or hypertension
Systolic blood pressure (mmHg)
Other recreational drugs
0.000
0.002
0.004
0.006
0.008
0.010
0.012
mean(|SHAP value|) (average impact on model output magnitude)
Vaccination up to Date
No
Yes
B
Density
0
10
20
30
40
50
60
70
−0.08
−0.06
−0.04
−0.02
0.00
0.02
0.04
0.06
Shap Score
Marijuana use
No
Yes
C
Density
0
20
40
60
80
100
120
−0.05
−0.04
−0.03
−0.02
−0.01
0.00
0.01
0.02
Shap Score
D
SHAP value for Marijuana use
0.04
0.02
0.00
−0.02
−0.04
0.0
0.2
0.4
0.6
0.8
1.0
Marijuana use
Overweight and obesity
1.0
0.5
0.0