HHS Public Access

Author manuscript

Drug Alcohol Depend. Author manuscript; available in PMC 2021 January 01.

Published in final edited form as:

Drug Alcohol Depend. 2020 January 01; 206: 107605. doi:10.1016/j.drugalcdep.2019.107605.

Analysis of substance use and its outcomes by machine learning I. Childhood Evaluation of Liability to Substance Use Disorder

Yankang Jing[1,*], **Ziheng Hu**[1,*], **Peihao Fan**[1], **Ying Xue**[1], **Lirong Wang**[1], **Ralph E Tarter**[2], **Levent Kirisci**[2], **Junmei Wang**[1,**], **Michael Vanyukov Tarter**[2,**], **Xiang-Qun Xie**[1,**]

[1]Department of Pharmaceutical Sciences, Computational Chemical Genomics Screen Center, School of Pharmacy; NIDA National Center of Excellence for Computational Drug Abuse Research, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213.

[2]Department of Pharmaceutical Sciences, School of Pharmacy, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213.

Abstract

[**]**Corresponding Author:** Xiang-Qun Xie; xix15@pitt.edu, Junmei Wang; junmei.wang@pitt.edu, Michael Vanyukov; mmv@pitt.edu.

Yankang Jing; Department of Pharmaceutical Sciences, School of Pharmacy; NIDA National Center of Excellence for Computational Drug Abuse Research, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213

Ziheng Hu; Department of Pharmaceutical Sciences, School of Pharmacy; NIDA National Center of Excellence for Computational Drug Abuse Research, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213

Peihao Fan; Department of Pharmaceutical Sciences, School of Pharmacy; NIDA National Center of Excellence for Computational Drug Abuse Research, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213

Ying Xue; Department of Pharmaceutical Sciences, School of Pharmacy; NIDA National Center of Excellence for Computational Drug Abuse Research, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213

Lirong Wang; Department of Pharmaceutical Sciences, School of Pharmacy; NIDA National Center of Excellence for Computational Drug Abuse Research, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213

Michael Vanyukov; Department of Pharmaceutical Sciences, School of Pharmacy, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213

Levent Kirisci.; Department of Pharmaceutical Sciences, School of Pharmacy, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213

Ralph E. Tarter.; Department of Pharmaceutical Sciences, School of Pharmacy, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213

Junmei Wang; Department of Pharmaceutical Sciences, School of Pharmacy; NIDA National Center of Excellence for Computational Drug Abuse Research, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213.

Michael Vanyukov; Department of Pharmaceutical Sciences, School of Pharmacy, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213.

Xiang-Qun Xie; Department of Pharmaceutical Sciences, School of Pharmacy; NIDA National Center of Excellence for Computational Drug Abuse Research, University of Pittsburgh, 3501 Terrace St, Pittsburgh, PA, USA, 15213.

Contributors

All co-authors materially participated in the research and/or article preparation. Xiang-Qun Xie, Junmei Wang, Ralph Tarter, Ziheng Hu, and Yankang Jing designed research; Ziheng Hu and Yankang Jing analyzed the data and prepared the first draft of the manuscript. Remaining authors provided feedback to each iteration of the review paper as it progressed until all authors agreed on the final product.

[*] These authors contributed equally to this work

Conflict of Interest

The authors have no conflict declared.

Background: Substance use disorder (SUD) exacts enormous societal costs in the United States, and it is important to detect high-risk youths for prevention. Machine learning (ML) is the method to find patterns and make prediction from data. We hypothesized that ML identifies the health, psychological, psychiatric, and contextual features to predict SUD, and the identified features predict high-risk individuals to develop SUD.

Method: Male (N= 494) and female (N=206) participants and their informant parents were administered a battery of questionnaires across five waves of assessment conducted at 10–12, 12–14, 16, 19, and 22 years of age. Characteristics most strongly associated with SUD were identified using the random forest (RF)algorithm from approximately 1,000 variables measured at each assessment. Next, the complement of features was validated, and the best models were selected for predicting SUD using seven ML algorithms. Lastly, area under the receiver operating characteristic curve (AUROC) evaluated accuracy of detecting individuals who develop SUD+/− up to thirty years of age.

Results: Approximately thirty variables strongly predict SUD. The predictors shift from psychological dysregulation and poor health behavior in late childhood to non-normative socialization in mid to late adolescence. In 10–12-year-old youths, the features predict SUD+/− with 74% accuracy, increasing to 86% at 22 years of age. The RF algorithm optimally detects individuals between 10–22 years of age who develop SUD compared to other ML algorithms.

Conclusion: These findings inform the items required for inclusion in instruments to accurately identify high risk youths and young adults requiring SUD prevention.

Keywords

Substance Use Disorder; Random Forest; Substance Abuse Prevention; Big Data; Screening Addiction Risk

1. Introduction

Hazardous substance use and substance use disorder (SUD) exact enormous societal cost, estimated in the United States to annually exceed seven hundred billion dollars (NIDA, 2017). Considering that consumption of addictive substances usually begins during adolescence, and SUD prevalence declines after thirty years of age (SAMHSA, 2018), it is important to detect high-risk youths and young adults requiring prevention. Toward this goal, the first task requires delineating the characteristics comprising SUD vulnerability.

Externalizing behaviors and psychiatric disorders, particularly attention deficit hyperactivity disorder and conduct disorder, amplify risk for SUD (Iacono et al., 1999; King et al., 2004; Verdejo-Garcia et al., 2008). In addition, anxiety and depression may also elevate SUD risk (Achenbach, 1995; Grant et al., 2004). In effect, numerous vulnerability characteristics have been described that are consistent with etiological theories of SUD emphasizing disinhibitory behavior and stress relief (i.e., self-medication). Notably, however, externalizing and internalizing propensities are correlated (Winters et al., 2008) and frequently co-occur (Colder et al., 2013), suggesting that SUD is associated with suboptimal psychological self-regulation cardinally featured by behavior under-control which is congruent with deficient modulation of emotions (Tarter et al., 2003).

Research into SUD etiology also focuses on parsing the sources of vulnerability characteristics. For example, informed by genetic research, the nuclear family affords the opportunity to clarify the sources of SUD vulnerability, namely vertical transmission (parent to child), horizontal transmission (sibling to sibling), or extrafamilial influences sources (neighborhood, school, etc.). This line of research has yielded, for example, an interval scale, termed the transmissible liability index (Vanyukov et al., 2009). Whether research into SUD etiology is guided by theory or directed at partitioning the sources of variance (e.g. genetic/non-genetic), the array of vulnerability characteristics remains to be delineated.

A main reason for incomplete understanding of the characteristics comprising SUD vulnerability is the dearth of longitudinal studies containing a) a large set of variables, b) multiple assessment waves, and c) documented SUD outcome. These criteria are satisfied in the dataset consisting of approximately 1,000 variables in each of five assessments spanning childhood to adulthood accrued by the Center for Education and Drug Abuse Research (CEDAR) at the University of Pittsburgh. This resource provides, therefore, a unique opportunity to apply Machine learning (ML) for analyzing the vulnerability characteristics of SUD from a data-driven perspective.

ML is a class of algorithms that learn to perform certain tasks by finding patterns from data. As a data-driven method, ML represents a powerful alternative to hypothesis-driven models for detecting SUD vulnerability (Obermeyer and Emanuel, 2016). It focuses on relating input characteristics (e.g., psychological, health, environment variables) termed features with an outcome variable (e.g., SUD) termed class label (Bishop, 2006). ML methodology can be thus free of investigator biases or assumptions. Whereas ML has been extensively utilized in medical research (Chen and Asch, 2017; Jing et al., 2018; Wernick et al., 2010), its application in SUD has been limited to detecting peripheral biomarkers (Bough and Pollock, 2018) and predictors of treatment outcome (Acion et al., 2017). Two hypotheses are advanced: 1) a small complement of features can be detected from the large pool of variables spanning health, psychological, psychiatric, and contextual/environmental (family, school, schoolwork, neighborhood) characteristics that predict SUD, and, 2) these variables accurately identify youths who develop SUD up to thirty years of age. Confirming these hypotheses provides the empirical foundation for developing age-specific, scalable and efficient screening tools to quantify and temporally monitor SUD risk.

2. Material and Methods

2.1 Participants

Men who were qualified for either lifetime diagnosis of SUD consequent to using an illegal drug, had a non-SUD psychiatric disorder or had no adult-onset psychiatric disorder, and had a 10–12-year-old son (N=494) or daughter (N=260) were identified via advertisement, public service announcements, random digit telephone calls, and posters displayed in public locations. Recruitment was conducted under aegis of the NIDA-funded Center for Education and Drug Abuse Research (CEDAR) (Vanyukov et al., 2009). The children were enrolled in a longitudinal investigation aimed at elucidating the etiology of SUD within a developmental framework. Follow-up evaluations were conducted at 12–14, 16, 19, and 22 years of age. SUD outcome was assessed at each assessment wave and lastly at thirty years of age.

Ethnicity of the sample was 75.6% European-American and 21.2% African-American. The remaining 3.2% self-identified their ethnicity as “other”. Potential participants were excluded from the study if they had a history of neurological disorder, schizophrenia, uncorrectable sensory incapacity, head injury requiring hospitalization, IQ < 70, or chronic physical disability. Informed consent and written assent approved by the University of Pittsburgh IRB were respectively obtained from the parents and their children prior to data collection. At eighteen years of age and thereafter the participants signed informed consent forms.

2.2 Measures and Variables

At each visit, an age-specific battery of questionnaires and interviews (Table 1) containing approximately 1,000 items were administered to the participants and their informant parents to document health, psychological, psychiatric and multiple social environments (family, school, peers, neighborhood, work, etc.) characteristics. The outcome variable, termed class label in ML, was the development of any DSM-III category of SUD (Spitzer et al., 1992). Diagnosis was formulated by a clinical committee based on results of the Structured Clinical Interview for DSM-III in conjunction with information obtained from other aspects of the research protocol and, where available, medical, school and legal records.

2.3 Data Analysis

At the outset, features (i.e., items) were eliminated if (1) the percent of missing responses was 70% or higher; (2) the variable had a variance of <0.1; or (3) the item directly queried substance use.

2.3.1 Missing Data Imputation.—

Imputation of missing data was performed using the k-nearest-neighbors algorithm (kNN) (Beretta and Santaniello, 2016). The rationale underlying the kNN algorithm is that the missing value of a characteristic for one participant can be substituted with values of “closest” cases (neighbors) within the entire sample. In this study, the “closest” three neighbors (k = 3) for each participant were used. The proximity between any two participants was calculated using the equation:

$$S_{ij} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left[w_k\left(v_{ik} - v_{jk}\right)\right]^2},$$

where n is the number of features without missing data for subjects i and j, w_k is the weight of feature k, $v_{ik,}$ and v_{jk} are the normalized values of feature k. The following two criteria needed to be satisfied during the difference score calculations: (1) n must be no smaller than 40% of total features, and, (2) a feature is disqualified if the missing data are greater than 30%. If the k-th feature of subject i, v_{ik}, is missing, three subjects whose profiles are most similar to subject i are first identified, that is, their difference score S_{ij} is the smallest. Lastly, the mean of the three v_{jk} values is assigned to v_{ik}.

2.3.2 Features (Items) Selection.—

Selection of features in ML enables deriving the most parsimonious model by removing from prediction the items that are either irrelevant or redundant (Guyon and Elisseeff, 2003; Liu and Zhao, 2012). We adopted the random forest

(RF) method for features selection because it is widely used to analyze diverse types of high-dimensional data (Genuer et al., 2010). In RF-based feature selection, each feature can be denoted an importance score that is calculated based on the concept of information entropy (Shannon, 1948), and this score represents the feature's contribution to prediction accuracy. Next, all features are ranked according to their importance scores, followed by their sequential entry into the model until reaching the maximum accuracy for predicting SUD+/−. Pearson's χ^2 test was also performed to assess the relationship between each feature and outcome class (presence/absence of SUD).

2.3.3 Model Construction Using ML Algorithms.—ML models were developed for predicting the risk of developing SUD by age 30 for young at each assessment wave (10–12, 12–14, 16, 19, and 22 years of age) based on the selected features. Seven algorithms were used: 1) logistic regression (Kleinbaum et al., 2002), 2) RF (Breiman, 2001), 3) adaptive boosting (AdaBoost) (Ma et al., 2011; Solomatine and Shrestha, 2004), 4) naïve Bayes (Murphy, 2006), 5) support vector machine (SVM) (Steinwart and Christmann, 2008), 6) kNN (Altman, 1992) and 7) deep neural network (DNN) (Myint et al., 2012; Schmidhuber, 2015). The Scikit-learn Python package (Pedregosa et al., 2011) was utilized to develop the models. Lastly, for comparison, models developed from the entire set of features before feature selection (N approx. 1,000) were compared for all seven ML algorithms.

2.3.4 Model Evaluation Using Cross-Validation.—We performed 10-fold cross-validation by resampling the dataset to evaluate forecasting accuracy of the seven models (Kohavi, 1995). This validation procedure is less biased than other methods (e.g., simple train/test split). The entire dataset was randomly divided into ten subsets having approximately equal size. A single fold was used as the validation set, and the remaining nine folds were combined and used as the training set. This procedure was repeated ten times until every single fold serves as the test set. This repeating ensures that every observation from the original dataset has the chance of appearing in training and test set, and the overall accuracy of the model is the mean of accuracies derived from the 10 rounds.

2.3.5 Receiver Operating Characteristics (ROC) Analysis.—ROC analysis was applied to evaluate model performance (Hanley and McNeil, 1982). The ROC is a curve which is plotted with false negative rate (1-specificity) against true positive rate (sensitivity) of a classification model at all classification thresholds. The two-dimensional area under the curve (AUC) under the entire ROC curve represents the degree or measure of separability. This AUC (also called AUROC) ranging from 0.5 to 1 specifies accuracy of the model for classifying the individual according to presence/absence of current or future SUD.

3. Results

3.1 Selected Features for Predicting SUD Individuals

As illustrated in Figure 1, accuracy at all five visits reached a plateau when the number of the features (items) used for building models was approximately thirty. These features were selected, therefore, to generate models for predicting SUD in each assessment. Table 2 lists the top 30 features selected for model at 10–12 years of age. Almost half (N=14) were

ratings provided by the parent. This finding concurs with the observation that young children are not the best informants about themselves. At subsequent visits (Supplementary Table S1–S4*), all of the best features were responses provided by the participants. Overall, the best features at 10–12 years of age are indicators of psychological self-regulation spanning behavior control, emotion modulation, daily routine, and mental concentration. In subsequent follow-ups social maladjustment and interpersonal interaction problems had more prominence. These results, considered from the ontogenetic perspective, indicate that the most prognostic features of SUD shift from psychological dysregulation during childhood and early adolescence to non-normative socialization in late adolescence and thereafter.

3.2 Model Performance, Selection and Validation

Performance of the different ML algorithms for predicting SUD at different ages is illustrated in Figure 2. As can be seen, the models using the top 30 features (black bar) are generally superior to the models using the entire dataset (striped bar). However, regardless of the method, used to construct the predictive model, forecasting accuracy unsurprisingly increases with chronological age. Although only a modest difference between the seven ML models is observed, the Naïve Bayes, SVM and RF models perform similarly and somewhat better than the other four ML methods. From among these latter three equally performing ML algorithms, the RF model is preferred (Pranckevičius and Marcinkevičius, 2017) because it is tree-based and therefore can be utilized to assist “if-then” decision making (Chen et al., 2018). As shown in Figure 3, accuracy of the RF model for predicting SUD across chronological age rises from 0.74 at age 10–12 to 0.86 at age 22.

4. Discussion

The results of this prospective study demonstrate that the RF algorithm detects important psychological, health, and environment features in childhood and early adolescence, and subsequently non-normative socialization features in late adolescence onward, that predict SUD up to thirty years of age. At 10–12 years of age, the features detect youths who develop SUD with 74% accuracy. This level of accuracy compares favorably with 65% for neurobehavior disinhibition (Kirisci et al., 2006) and 68% for transmissible risk (Vanyukov et al., 2009). The results also reveal that the strongest indicator of SUD risk is swearing, followed by poor play behavior and irritability. This finding underscores the salience of affective dysregulation and social interaction problems during late childhood on risk for SUD. Moreover, daily health behavior routines are suboptimal (e.g., eating and sleeping), raising the prospect that irregular circadian rhythms also constitute an important dimension of SUD vulnerability (Logan et al., 2014).

The findings in this study additionally highlight two important issues pertinent to SUD etiology research. First, both individual phenotypic characteristics and environmental factors rank among the best features predicting SUD. Whereas most researches into SUD etiology distinguish and separate variables according to either characterizing the individual or the

*Supplementary material can be found by accessing the online version of this paper at http://dx.doi.org and by entering doi: https://doi.org/10.1016/j.drugalcdep.2019.107605

environment, ML methodology joins both etiological dimensions and quantifies their salience for SUD prediction. This latter attribute of ML directly forms prioritization of prevention tactics. And second, the selected best features (Tables S2–S5) in later visits (ages 16, 19, and 22) include facets of non-normative socialization. In effect, as SUD liability unfolds during adolescent development the strongest predictors of SUD shift in emphasis from psychological dysregulation and health problems to social maladjustment. These findings demonstrate the heuristic utility of ML for comprehensively characterizing the ontogenetic patterning of SUD liability.

Several limitations in this study are noted. Because the high-risk paradigm was used (i.e., oversampling children having affected parents) the results may not generalize to the broader population. Accordingly, testing model performance in a random sample is warranted. It is also noteworthy that the standard deviations of the accuracy (AUROC) across the 10-fold cross-validation are large, indicating that while the models are adequate their prediction accuracy can be potentially improved by using a larger dataset with a more balanced distribution. Finally, the ML prediction of the SUD outcome based on the vulnerability traits cannot be interpreted as causal effects, and it offers little insight into the longitudinal development of SUD during adolescence through the prodrome phase. This topic will be addressed in the companion paper (Hu et al.).

Notwithstanding these limitations, the findings point to the feasibility of using ML algorithms to comprehensively delineate the psychological, health and environmental characteristics associated with the vulnerability for SUD. Once the optimum complement of robust features is delineated it is feasible to derive and psychometrically validate accurate age-specific assessments to quantify and monitor SUD risk.

5. Conclusions

The RF algorithm identified thirty psychological, health, environmental and social behavior features that predict SUD in each of five assessments conducted at 10–12, 12–14, 16, 19, and 22 years of age. The complement of features accurately detects youth and young adults who are at high risk for SUD. It is thus concluded that ML methodology is heuristic for deriving scalable unobtrusive screening tools tailored to the respondent's age to quantify risk for SUD.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Role of Funding Source

This work was supported by the National Institutes of Health [P30 DA-035778-01A1 (XQX), DA-P50-05605 (XQX); R01GM79383 (JW); R21GM097617-01 (JW)]; the Department of Defense [W81XWH-1N6-1-0490:412288 (XQX)]. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funding organizations.

References

Achenbach TM, 1995 Empirically based assessment and taxonomy: Applications to clinical research. Psychological Assessment 7(3), 261–274.

Acion L, Kelmansky D, van der Laan M, Sahker E, Jones D, Arndt S, 2017 Use of a machine learning framework to predict substance use disorder treatment success. PLoS One 12(4), e0175383. [PubMed: 28394905]

Altman NS, 1992 An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician 46(3), 175–185.

Beretta L, Santaniello A, 2016 Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak 16 Suppl 3(3), 74. [PubMed: 27454392]

Bishop CM, 2006 Pattern recognition and machine learning. springer.

Bough KJ, Pollock JD, 2018 Defining Substance Use Disorders: The Need for Peripheral Biomarkers. Trends Mol Med 24(2), 109–120. [PubMed: 29396146]

Breiman L, 2001 Random forests. 45(1), 5–32.

Chen JH, Asch S.M.J.T.N.E.j.o.m., 2017 Machine learning and prediction in medicine—beyond the peak of inflated expectations. 376(26), 2507.

Chen W, Zhang S, Li R, Shahabi H, 2018 Performance evaluation of the GIS-based data mining techniques of best-first decision tree, random forest, and naive Bayes tree for landslide susceptibility modeling. Sci Total Environ 644, 1006–1018. [PubMed: 30743814]

Colder CR, Scalco M, Trucco EM, Read JP, Lengua LJ, Wieczorek WF, Hawk LW Jr., 2013 Prospective associations of internalizing and externalizing problems and their co-occurrence with early adolescent substance use. J Abnorm Child Psychol 41(4), 667–677. [PubMed: 23242624]

Genuer R, Poggi J-M, Tuleau-Malot C, 2010 Variable selection using random forests. Pattern Recognition Letters 31(14), 2225–2236.

Grant BF, Stinson FS, Dawson DA, Chou SP, Dufour MC, Compton W, Pickering RP, Kaplan K, 2004 Prevalence and co-occurrence of substance use disorders and independent mood and anxiety disorders: results from the National Epidemiologic Survey on Alcohol and Related Conditions. Arch Gen Psychiatry 61(8), 807–816. [PubMed: 15289279]

Guyon I, Elisseeff A, 2003 An introduction to variable and feature selection, in: Kaelbling LP (Ed.). MIT Press, US, pp. 1157–1182.

Hanley JA, McNeil BJ, 1982 The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143(1), 29–36. [PubMed: 7063747]

Iacono WG, Carlson SR, Taylor J, Elkins IJ, McGue M, 1999 Behavioral disinhibition and the development of substance-use disorders: Findings from the Minnesota Twin Family Study. Dev Psychopathol 11(4), 869–900. [PubMed: 10624730]

Jing Y, Bian Y, Hu Z, Wang L, Xie X-QS, 2018 Deep Learning for Drug Design: an Artificial Intelligence Paradigm for Drug Discovery in the Big Data Era. The AAPS Journal 20(3), 58. [PubMed: 29603063]

King SM, Iacono WG, McGue M, 2004 Childhood externalizing and internalizing psychopathology in the prediction of early substance use. Addiction 99(12), 1548–1559. [PubMed: 15585046]

Kirisci L, Tarter RE, Reynolds M, Vanyukov M, 2006 Individual differences in childhood neurobehavior disinhibition predict decision to desist substance use during adolescence and substance use disorder in young adulthood: A prospective study. Addictive Behaviors 31(4), 686–696. [PubMed: 15964148]

Kleinbaum DG, Dietz K, Gail M, Klein M, 2002 Logistic regression. Springer.

Kohavi R, 1995 A study of cross-validation and bootstrap for accuracy estimation and model selection, Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2 Morgan Kaufmann Publishers Inc, Montreal, Quebec, Canada, pp. 1137–1143.

Liu H, Zhao Z, 2012 Manipulating Data and Dimension Reduction Methods: Feature Selection, in: Meyers RA (Ed.) Computational Complexity. Springer New York, New York, NY, pp. 1790–1800.

Logan RW, Williams WP 3rd, McClung CA, 2014 Circadian rhythms and addiction: mechanistic insights and future directions. Behav Neurosci 128(3), 387–412. [PubMed: 24731209]

Ma C, Wang L, Xie XQ, 2011 Ligand Classifier of Adaptively Boosting Ensemble Decision Stumps (LiCABEDS) and its application on modeling ligand functionality for 5HT-subtype GPCR families. J Chem Inf Model 51(3), 521–531. [PubMed: 21381738]

Murphy KP, 2006 Naive bayes classifiers. 18, 60.

Myint KZ, Wang L, Tong Q, Xie XQ, 2012 Molecular fingerprint-based artificial neural networks QSAR for ligand biological activity predictions. Mol Pharm 9(10), 2912–2923. [PubMed: 22937990]

National Institute on Drug Abuse, 2017 Trends & Statistics National Institute on Drug Abuse, Bethesda, MD.

Obermeyer Z, Emanuel EJ, 2016 Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. N Engl J Med 375(13), 1216–1219. [PubMed: 27682033]

Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, 2011 Scikit-learn: Machine learning in Python. Journal of machine learning research 12(10), 2825–2830.

Pranckevičius T, Marcinkevičius V, 2017 Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. Baltic Journal of Modern Computing 5(2), 221–232.

SAMHSA, 2018 Key substance use and mental health indicators in the United States: Results from the 2017 National Survey on Drug Use and Health Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration, Rockville, MD.

Schmidhuber J, 2015 Deep learning in neural networks: An overview. Neural Networks 61, 85–117. [PubMed: 25462637]

Shannon CE, 1948 A Mathematical Theory of Communication. At&T Tech J 27(4), 623–656.

Solomatine DP, Shrestha DL, 2004 AdaBoost.RT: a boosting algorithm for regression problems, 2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No.04CH37541). pp. 1163–1168 vol.1162.

Spitzer RL, Williams JB, Gibbon M, First MB, 1992 The Structured Clinical Interview for DSM-III-R (SCID). I: History, rationale, and description. Arch Gen Psychiatry 49(8), 624–629. [PubMed: 1637252]

Steinwart I, Christmann A, 2008 Support vector machines. Springer Science & Business Media.

Tarter RE, Kirisci L, Mezzich A, Cornelius JR, Pajer K, Vanyukov M, Gardner W, Blackson T, Clark D, 2003 Neurobehavioral Disinhibition in Childhood Predicts Early Age at Onset of Substance Use Disorder. American Journal of Psychiatry 160(6), 1078–1085. [PubMed: 12777265]

Vanyukov MM, Kirisci L, Moss L, Tarter RE, Reynolds MD, Maher BS, Kirillova GP, Ridenour T, Clark DB, 2009 Measurement of the risk for substance use disorders: phenotypic and genetic analysis of an index of common liability. Behav Genet 39(3), 233–244. [PubMed: 19377872]

Verdejo-Garcia A, Lawrence AJ, Clark L, 2008 Impulsivity as a vulnerability marker for substance-use disorders: review of findings from high-risk research, problem gamblers and genetic association studies. Neurosci Biobehav Rev 32(4), 777–810. [PubMed: 18295884]

Wernick MN, Yang Y, Brankov JG, Yourganov G, Strother SC, 2010 Machine Learning in Medical Imaging. IEEE signal processing magazine 27(4), 25–38. [PubMed: 25382956]

Winters KC, Stinchfield RD, Latimer WW, Stone A, 2008 Internalizing and externalizing behaviors and their association with the treatment of adolescents with substance use disorder. J Subst Abuse Treat 35(3), 269–278. [PubMed: 18328664]

Highlights

- We identified behavioral and health characteristics at five ages spanning childhood to adulthood that are prognostic of substance use disorder using machine learning methodology.
- We derived a model that accurately detects youths who develop substance use disorder.
- We found that the salience of SUD risk characteristics shifts from psychological dysregulation in childhood to non-normative socialization during adolescence and thereafter.

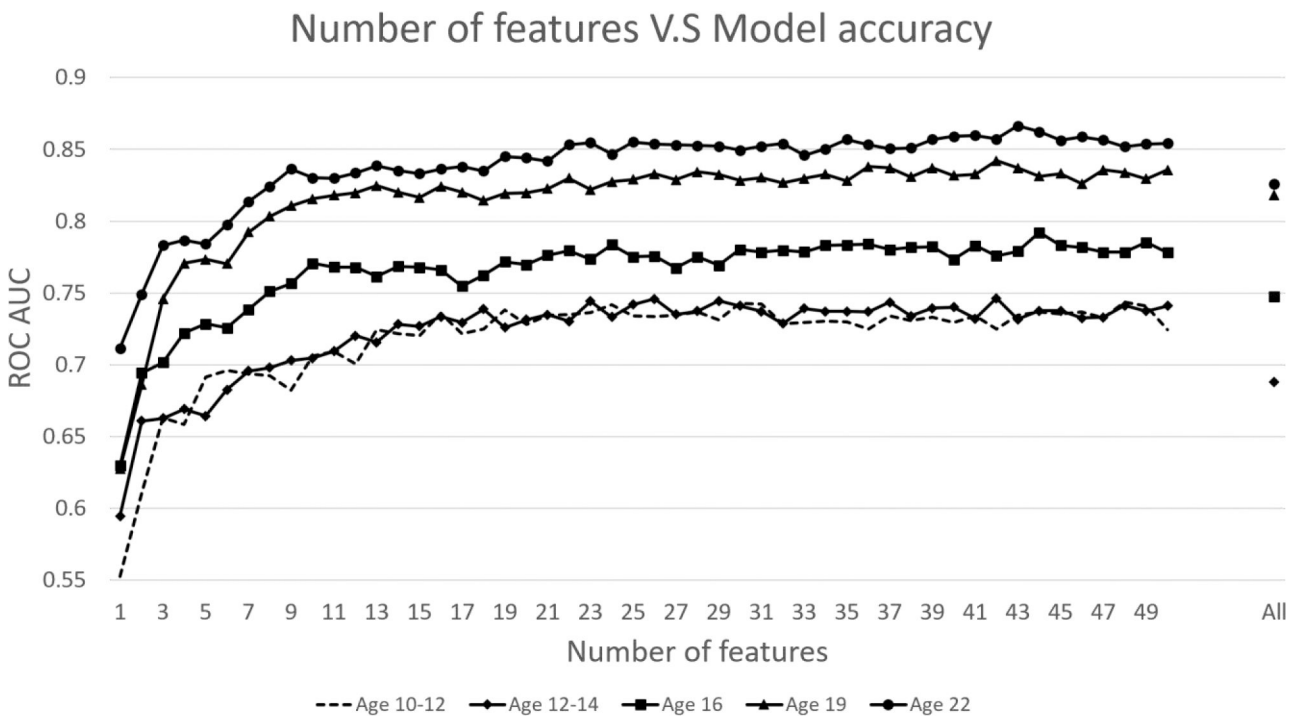

Figure 1.

Relationship between number of features and predictive power of the model using the RF algorithm in all visits. Predictive power was scaled using AUROC in the 10-fold cross validation.

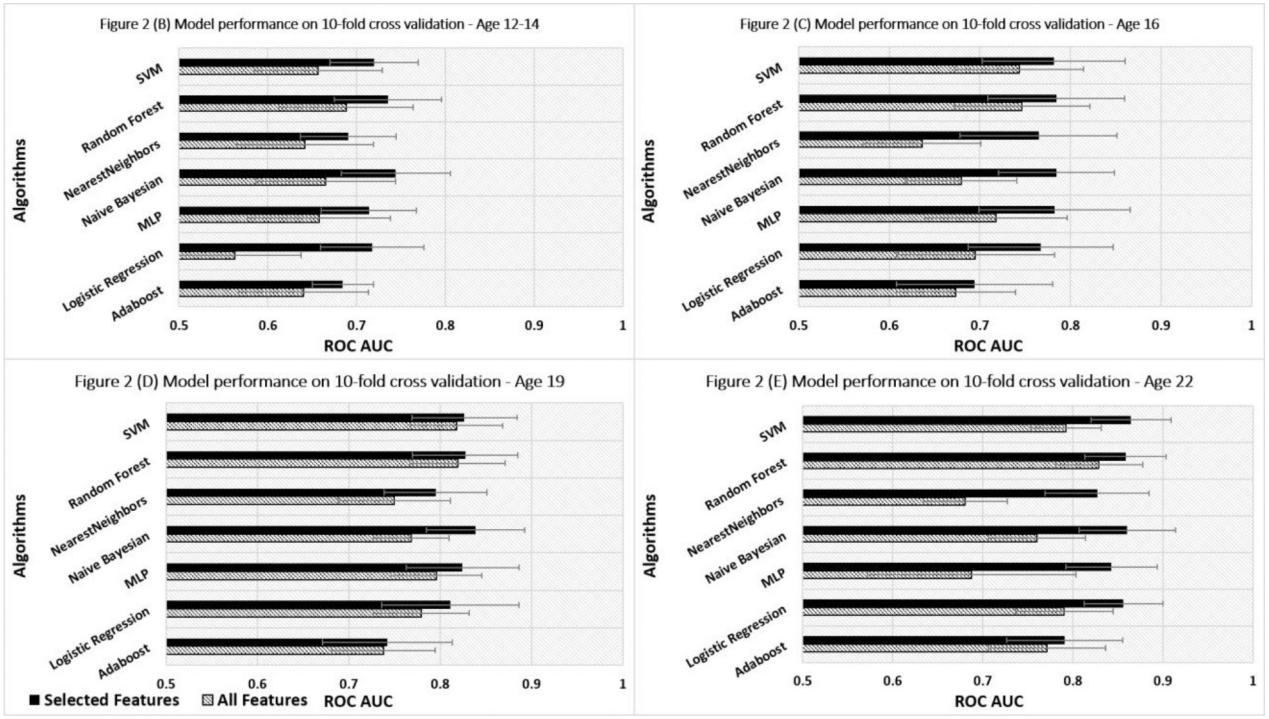

Figure 2.

Comparing performance of seven SUD prediction algorithms at the four follow-up visits according to thirty selected features and entire dataset. Abbreviations: RF, random forest; SVM, support vector machine; Bayesian, naïve Bayes; AdaBoost, adaptive boost; MLP, multilayer perceptron; AUROC, area under the receiver operating characteristic curve; HS, high severity.

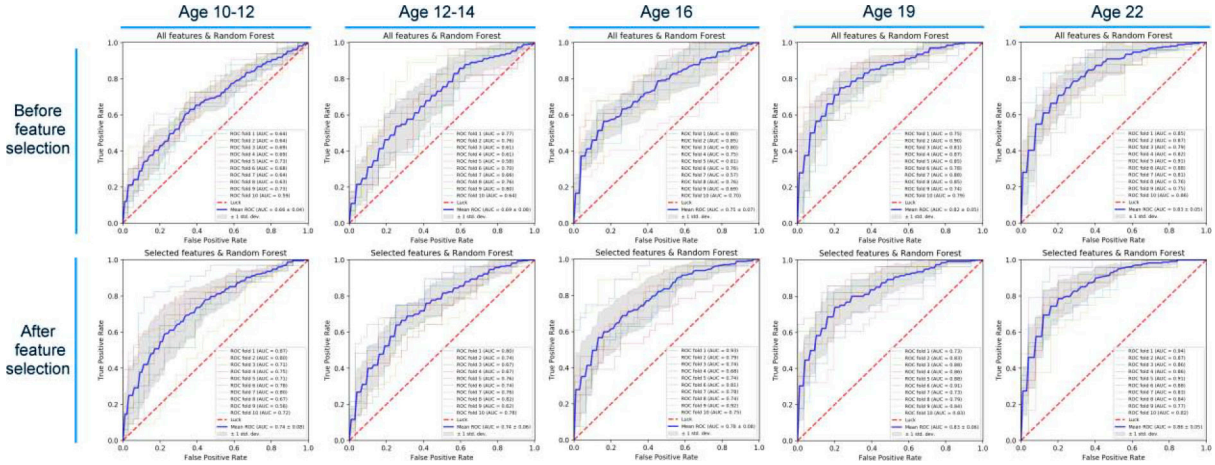

Figure 3.

Random forest (RF) prediction before and after features selection. Top figures depict models generated using all the features in the dataset. Bottom figures depict the performances of models using the selected (n=30) features. In each chart, the blue line shows the average ROC curve in the 10-fold cross validation and the gray areas shows the standard deviation. ROC curves. The other colors show the detailed performances of the models in the cross validation.

Table 1.

Questionnaires summary for different visits.

Questionnaires Name	Age 10–12	Age 12–14	Age 16	Age 19	Age 22
Antisocial Personality Disorder Interview	No	No	No	Yes	Yes
Andrew's Scale of Severity and History of Offenses	No	No	No	Yes	Yes
Dysregulation Inventory	Yes	Yes	Yes	No	No
Conner's Behavioral Rating Scale	Yes	Yes	Yes	No	No
Irritability Scale	No	Yes	No	No	No
TC Child Behavior Checklist	Yes	Yes	Yes	No	No
Constructive Thinking Inventory	No	No	Yes	No	No
Disruptive Behavior Disorder Scale	Yes	No	No	No	No
Diagnostic Instrument (K-SADS-E)	Yes	Yes	Yes	Yes	Yes
Dimensions of Temperament Survey	Yes	No	Yes	No	No
Drug Use Screening Questionnaire	No	Yes	Yes	Yes	Yes
Emotional Susceptibility Scale	No	Yes	No	No	No
Hostility Guilt Inventory	No	Yes	No	No	No
Health Problem Checklist	No	No	No	Yes	No
Multidimensional Personality Questionnaire	No	No	Yes	Yes	Yes
Sensation Seeking Scale	No	No	No	No	Yes
Tarter Childhood Questionnaire	Yes	No	No	No	No
Child Health and Illness Profile (Chip-AE)	No	No	No	No	Yes
Young Adult Self Report	No	No	No	No	Yes
Youth Self-Report	No	No	Yes	No	Yes
Number of Overall Questionnaires	7	8	9	6	9

Table 2.

Top 30 Items for predicting high substance use risk at age 10–12.

Questions	Feature Importance	Importance rank	chi2	p-value
Do you often swear or use bad language?	0.0069	1	42.3804	0.0000
Do you have difficulty playing quietly?	0.0055	2	39.0885	0.0000
My child eats about the same amount at breakfast from day to day (Parents)	0.0047	3	10.9055	0.0010
Are you touchy or easily annoyed by others?	0.0043	4	26.8952	0.0000
About how many times a week does your child do things with any friends outside of regular school hours? (Parents)	0.0042	5	4.7421	0.0294
Do you have difficulty staying in line in the supermarket or waiting for your turn while you were playing with other children?	0.0042	6	38.1248	0.0000
Do you deliberately refuse adults, or do you refuse to do your chores at home or disobey rules a lot?	0.0041	7	34.7887	0.0000
Do you often argue with adults?	0.0038	8	30.2260	0.0000
How many jobs, chores do your child has? (Parents)	0.0036	9	3.6153	0.0573
Is your child hard to be distracted? (Parents)	0.0033	10	7.8940	0.0050
Does your child get very restless If he/she has to stay in one place for a long time? (Parents)	0.0032	11	11.0743	0.0009
Does your child get hungry about the same time each day? (Parents)	0.0031	12	5.4537	0.0195
Do you get very fidgety after a few minutes if you’re supposed to sit still?	0.0029	13	14.0798	0.0002
Does your child get very fidgety after a few minutes Een when he/she is supposed to be still? (Parents)	0.0028	14	9.3343	0.0022
How many organizations, clubs, teams or groups does your child belongs to? (Parents)	0.0028	15	9.0944	0.0026
Within the past 6 months, does your child, hangs around with other who get in troubles? (Parents)	0.0027	16	19.4536	0.0000
Compared to others of his/her age, how well does your child play and work alone? (Parents)	0.0027	17	3.2823	0.0700
No matter when your child goes to sleep, does he/she wake up at the same time the next morning? (Parents)	0.0027	18	8.0267	0.0046
Does your child have difficulty following through on instructions from others (not due to oppositional behavior or failure of comprehension), e.g., fails to finish chores? (Parents)	0.0027	19	42.7693	0.0000
Does failure at a task or in school make your work harder?	0.0026	20	3.7005	0.0544
Can you read a book for half an hour before you get restless?	0.0026	21	6.6266	0.0100
Do you get into trouble because you would do things without thinking about them first, for example running into the street without looking?	0.0025	22	29.7495	0.0000
Do you get very restless when you have to stay in one place for a long time?	0.0025	23	8.9215	0.0028
Does your child wake up the same time each day when he/she is away from home? (Parents)	0.0024	24	8.0571	0.0045
Do your heart beats fast for a long time when you get stirred up?	0.0023	25	4.4068	0.0358
Do you have so much energy that you just can’t stop moving?	0.0023	26	8.2014	0.0042
Do you get so excited that I remain very excited for a long time after watching an action show?	0.0023	27	6.5546	0.0105
Are you easily distracted?	0.0023	28	6.9223	0.0085

Questions	Feature Importance	Importance rank	chi2	p-value
Compared to others of the same age, about how much time does your child spend in hobbies, activities and games other than sports? (Parents)	0.0023	29	0.6293	0.4276
Do you develop a plan for all your important goals?	0.0022	30	3.3211	0.0684