The Journal of Child Psychology and Psychiatry

Journal of Child Psychology and Psychiatry 61:12 (2020), pp 1370–1379 doi:10.1111/jcpp.13226

Machine-Learning prediction of comorbid substance use disorders in ADHD youth using Swedish registry data

Yanli Zhang-James,[1] Qi Chen,[2] Ralf Kuja-Halkola,[2] Paul Lichtenstein,[2] Henrik Larsson,[2,3] and Stephen V. Faraone[1,4]

[1]Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, Syracuse, NY, USA; [2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden; [3]School of Medical Sciences, Örebro University, Örebro, Sweden; [4]Department of Neuroscience and Physiology, SUNY Upstate Medical University, Syracuse, NY, USA

Background: Children with attention-deficit/hyperactivity disorder (ADHD) have a high risk for substance use disorders (SUDs). Early identification of at-risk youth would help allocate scarce resources for prevention programs. **Methods:** Psychiatric and somatic diagnoses, family history of these disorders, measures of socioeconomic distress, and information about birth complications were obtained from the national registers in Sweden for 19,787 children with ADHD born between 1989 and 1993. We trained (a) a cross-sectional random forest (RF) model using data available by age 17 to predict SUD diagnosis between ages 18 and 19; and (b) a longitudinal recurrent neural network (RNN) model with the Long Short-Term Memory (LSTM) architecture to predict new diagnoses at each age. **Results:** The area under the receiver operating characteristic curve (AUC) was 0.73(95%CI 0.70–0.76) for the random forest model (RF). Removing prior diagnosis from the predictors, the RF model was still able to achieve significant AUCs when predicting all SUD diagnoses (0.69, 95%CI 0.66–0.72) or new diagnoses (0.67, 95%CI: 0.64, 0.71) during age 18–19. For the model predicting new diagnoses, model calibration was good with a low Brier score of 0.086. Longitudinal LSTM model was able to predict later SUD risks at as early as 2 years age, 10 years before the earliest diagnosis. The average AUC from longitudinal models predicting new diagnoses 1, 2, 5 and 10 years in the future was 0.63. **Conclusions:** Population registry data can be used to predict at-risk comorbid SUDs in individuals with ADHD. Such predictions can be made many years prior to age of the onset, and their SUD risks can be monitored using longitudinal models over years during child development. Nevertheless, more work is needed to create prediction models based on electronic health records or linked population registers that are sufficiently accurate for use in the clinic. **Keywords:** Machine learning; substance use disorder; attention-deficit hyperactive disorder; comorbidity; risk factor.

Introduction

In recent years, prevalence of substance use disorders (SUDs) has increased significantly (Merikangas & McClair, 2012; Whiteford et al., 2013), magnifying the impact of many adverse consequences (Cain, Bornick, & Whiteman, 2013; Hall, 2015; Hall & Degenhardt, 2014; Karila, Petit, Lowenstein, & Reynaud, 2012; Kuntsche, Kuntsche, Thrul, & Gmel, 2017; Merrin, Davis, Berry, D'Amico, & Dumas, 2016; Moss, 2013). From 2005 to 2015, death due to opioid, cocaine, and amphetamine use disorders increased 30%–68% (Mortality & Causes of Death, 2016). In 2015, over 306,000 deaths were caused by SUD globally, which is 26 times of the total deaths caused by natural disasters and 44% more than all deaths caused by forces of war, violence, and legal interventions (Mortality & Causes of Death, 2016).

Twin studies showed that genes and their interaction with the environment constitute 50%–75% of the liability to develop SUD (Kendler et al., 2012; Tsuang, Bar, Harley, & Lyons, 2001). Many risk-modifying environmental factors have been studied including stress and trauma in early life and family, education, socioeconomic status (SES), and cultural influences (Barr, Silberg, Dick, & Maes, 2018; Crum & Anthony, 2000; Karriker-Jaffe, 2013; Kendler et al., 2012; Mulia & Karriker-Jaffe, 2012; Schnohr et al., 2004; Thompson, Lizardi, Keyes, & Hasin, 2008; Windle & Windle, 2018). Having attention-deficit/hyperactivity disorder (ADHD) is associated with a significantly increased risk for later SUDs (Biederman et al., 2006; Lambert & Hartsough, 1998; Molina & Pelham, 2003). Relatives of individuals diagnosed with ADHD also have a higher risk for SUDs (Skoglund, Chen, Franck, Lichtenstein, & Larsson, 2015). A large study on the UK Biobank data (N = 135,000) found that the polygenic risk for ADHD significantly predicted alcohol and nicotine use (Du Rietz et al., 2018). Furthermore, ADHD symptoms, such as inattention, hyperactivity, and impulsivity, can cause behavioral problems and stress at home and school, which in turn can increase the risk for substance use.

For individuals with ADHD, however, stimulant therapy was found to decrease the rates of smoking and other SUDs later in life (Chang, Lichtenstein,

Conflict of interest statement: See Acknowledgements for full disclosures.

Halldner, et al., 2014; Quinn et al., 2017; Schoenfelder, Faraone, & Kollins, 2014; Wilens, Faraone, Biederman, & Gunawardene, 2003). Behavior therapy was also found to significantly reduce substance use reported by ADHD youth at least to 24-month follow-up (The Multimodal Treatment Study of Children with ADHD (MTA) study, (Molina et al., 2007)). These data suggest that early identification of ADHD youth who are at risk for SUD would allow for more targeted early interventions and possibly prevention of future SUD. Recent studies have demonstrated the feasibility of developing risk prediction models in psychiatry, but the literature is very limited, and most studies are based on small samples (Barak-Corren et al., 2017; Bernardini et al., 2017). Few studies have applied machine-learning algorithms to large-scale data from electronic health records (Simon et al., 2018) or linked population registers, and no previous study using such data has been performed in the context of ADHD. Moreover, many prior machine-learning studies in psychiatry have been limited by small samples sizes and lack of appropriate replication samples (see, e.g., Zhang-James' et al. review of machine-learning applications to neuroimaging data (Zhang-James et al., 2019 (Preprint)).

To address these limitations, the present study sought to develop prediction models using machine-learning algorithms to identify ADHD youth at risk for SUDs. We used information available from the Swedish population registries to construct potential predictors, including the medical history of psychiatric and somatic illnesses for the index children and their immediate family members, as well as their available perinatal records, and socioeconomic, educational, and geographic data. Our goal was to determine if the information from the registries could be used to train a clinically useful machine-learning algorithm to accurately identify youth with ADHD at risk for SUDs.

Methods

Data sources and study population

The study was approved by the Regional Ethical Review Board in Stockholm, Sweden. The sample comprised 19,787 children born between 1989 and 1993 who had a life-time diagnosis of ADHD (National Patient Registry (NPR) ICD-9: 314; ICD-10: F90). We excluded those (a) who died or emigrated prior to age 20, (b) who had no father's information, or (c) who had no socioeconomic records during the ages 0–18. The final dataset contained 19,184 individuals. SUD was defined by either a diagnosis (ICD-9 304, 305, 306; ICD-10: F10-19) or a prescription of medication for SUD treatment (Table S1). We considered the first diagnostic or prescription record as the 'onset' of SUD in this study. Table S2 shows that 10% of the sample had SUD onset prior to age 17, 8.8% had SUD diagnoses recorded during the ages 18 and 19, and 5.9% of the total sample had a SUD onset during age 18–19.

Predictors and missing information

We used personal identity numbers (PINs) to link different registers. The PIN consists of the date of birth and a four-digit number and was introduced in Sweden in 1947 (the fourth digit was added in 1967) (Ludvigsson, Otterblad-Olausson, Pettersson, & Ekbom, 2009). Therefore, since then, every person residing in Sweden on a permanent basis (i.e., recorded in the Total population register, TPR) is assigned a PIN. The PIN is routinely used by governmental agencies (e.g., tax agency, healthcare providers, prison services, schools). Governmental agencies (such as Statistics Sweden) can merge data from different registers using the PIN. Eight different registers were used in our study to extract linked data using the unique PIN: (a) Medical Birth Register, which was established in 1973 and includes data on prenatal and perinatal measures of all births in Sweden (National Board of Health & Welfare, 2014); (b) National Patient Register, which contains inpatient care since 1964 (psychiatric care since 1973) and outpatient visits to specialty care since 2001 (Ludvigsson et al., 2011); (c) Total Population Register, which provides information on life events such birth, death, migration and family relationships (Ludvigsson et al., 2016); (d) Multi-Generation Register, which constitutes a part of the Total Population Register, but links individuals born in Sweden since 1932 and registered as living in Sweden since 1961 to their biological parents (Ekbom, 2011); (e) Prescribed Drug Register provides information on dispensed drugs to the entire Swedish population since July 2005. Active ingredients of drugs are coded according to the anatomical therapeutic chemical (ATC) classification system (Wettermark et al., 2007); (f) Longitudinal integration database for health insurance and labor market studies (LISA) was established in 1990 and contains annually updated data on highest level of education, civil status, unemployment, social benefits, and income for all Swedish residents aged 16 years or older (Statistics Sweden, 2011); (g) National School Register (NSR) provides individual-level data on final grades from school leaving certificates and eligibility to upper secondary school (The Swedish National Agency for Education, 2017); (h) National Crime Register covers violent and nonviolent crime convictions since 1973 (Chang, Larsson, Lichtenstein, & Fazel, 2015).

Predictors extracted from these eight registers are listed in Table S1, including (a) parental information: age and (maternal) weight at child birth, educational and marital status, criminal records, medical records; (b) family size and the numbers of siblings, full- and half-sibling medical and criminal records; (c) family SES: neighborhood deprivation scores (NDEP, (Sariaslan et al., 2013), family location (metropolitan area or not), family income, and social allowance received; (d) Index child information: perinatal records (child birth complications, APGAR scores, body measurements at birth), medical and criminal records. Medical records included inpatient and outpatient discharge records for 34 disorders and seven categories of prescription drugs. These disorders and prescription records were chosen based on prior knowledge of their relevance to ADHD or SUD. The complete ICD and prescription codes are listed in Table S1. For criminal records, disorder diagnoses, and prescription records, we summed the total numbers of records within each age, or each life periods (prenatal: (prior to birth), perinatal (age 0–2), childhood (2–12), and adolescence/teenage (12–17)) and used the summed totals as predictors.

Missing information is common in register databases. 71.8% subjects had missing prenatal and perinatal information from the Medical Birth Register; 55.2% subjects had one or more years of social economic data missing from the LISA register; and 11.6% children have no records for merit scores or graduation records (Table S1). For some variables, missing status could be informative for prediction. For example, missing father's information and family income could indicate childhood adversity. Instead of removing subjects or variables with missing data, which would lead to a biased and reduced sample size and lost information, we added a new category to

indicate 'missing' status. Some continuous variables were recoded as categorical variables according to the nature of the variables. For example, head circumferences and birth weight were recoded as five categories (normal, above or below 1x and 2x standard deviations) at each gestational age (weeks) according to published standards for the Swedish population (1990–1999; Skjaerven, Gjessing, & Bakketeig, 2000). Family incomes were normalized across the whole sample population by year (between 0 and 1) to eliminate economic differences across years. The normalized family income and NDEP scores were then examined within the index child's three age periods: perinatal (age 0–2), childhood (2–12), and adolescence/teenage (12–17). A small number of subjects ($n = 48$) with complete missing records in any of the three periods were removed. For the remaining subjects, we summarized the mean, minimum and maximum values of each period as predictors. Other continuous variables with a large amount of missing information were coded into seven ordinal categories (between 0%, 5%, 10%, 25%, 75%, 90%, 95%, and 100%) across the whole sample with one additional category to denote missing status. Diagnostic, prescription, and criminal records were summed as the total numbers of records per year for the longitudinal model or summed for the above three periods for the cross-sectional model. Finally, all the categorical variables were one-hot encoded as dummy variables (0 or 1) and continuous variables were scaled between 0 and 1.

Machine-learning models

The sample was randomly divided into training (70%), validation (15%), and test subsets (15%). The training and validation sets were used to optimize the model parameters and hyperparameters. The hyperparameters control structural features of the model such as the number of trees in a random forests model. The test set was reserved only for evaluating the performance of the final models. This is a commonly accepted practice in statistical learning to ensure that different subsets of samples were from the same distribution and that test set were not used in model selection and optimization (Goodfellow, Bengio, & Courville, 2016; Hastie, Tibshirani, & Friedman, 2009; Ng, 2019). We also use the training set to define the minimum and maximum values for each predictor to scale all continuous features between 0 and 1. This scaling function was applied to the validation and test set data.

We designed two types of prediction models: (a) a cross-sectional model to predict SUD diagnoses during age 18–19 and (b) a longitudinal, recurrent neural network (RNN) model to predict new SUD cases each year. For the cross-sectional model, we used the random forest classifier (RF, Pedregosa et al., 2012) based on its stable and superior performance during initial screening of several machine-learning algorithms including support vector machine, multilayer perceptron, gradient boosting, and k-nearest neighbors classifiers (results not shown), as well as its ability to evaluate feature importance, which helps to improve the interpretability of the machine-learning models (Holzinger, Biemann, Pattichis, & Kell, 2017).

We used Scikit-Learn's grid search algorithm to search the hyperparameter spaces for RF models, including numbers of estimators (trees), maximum percentage of features used in each tree, maximum depth, and class weight. Because our case and control classes are highly imbalanced, higher minority class weights encourage more accurate predictions of samples from the under-represented class. Feature selection was incorporated into the RF model using Scikit-Learn's pipeline function in combination with the SelectKBest function, in which '*K*' numbers of best features were selected by their ANOVA *F*-values prior to feeding to the RF models. '*K*' was also tuned as a hyperparameter during the RF grid search.

For the longitudinal model, we implemented the Long Short-Term Memory (LSTM) model (Hochreiter & Schmidhuber, 1997) to learn the sequential features at each year. The LSTM model uses an RNN architecture with improved gradient-based learning. It was designed to overcome problems caused by lags of unknown duration between important events in a time series (Hochreiter & Schmidhuber, 1997). Such problems exist in register data, for example, missing records/visits. The LSTM model focused on predicting new SUD cases only. At each age, the newly diagnosed SUD were removed from the sample once the prediction was made for that age. This ensures that the model only predicts new onsets at the next age. This process, however, reduces the numbers of positive outcomes at each age. For this reason, we increased the randomly selected and reserved test set to be ~30% of the total sample. Training and validation sets were split with a ratio of 80:20. HyperOpt (Bergstra, Yamins, & Cox, 2013) was used to tune the number of LSTM units. We used TensorFlow library to implement the LSTM network with the adaptive moment estimation (Adam) optimizer and binary cross-entropy as the loss function. Furthermore, stratified balanced sampling was applied to the LSTM training processes, which compensates the class imbalance by oversampling the minority class. Early stopping was implemented to avoid overfitting. Best models were chosen based on the lowest total validation loss and tested on the test set. All machine-learning algorithms were written in Python 3.5.

Diagnostic accuracy statistics

For the test set performance, we computed receiver operating characteristic (ROC) curves and used the area under the ROC curve (AUC) as our primary measure of accuracy. The AUC and its confidence intervals were calculated in Stata 15 using the empirical method and compared with nonparametric approach by DeLong, DeLong, and Clarke-Pearson (1988). We also report the Brier loss for the final model. Hosmer–Lemeshow test was used to estimate the model's goodness-of-fit and visualized with a calibration curve. In addition, the predicted probabilities were used to calculate sensitivity/recall, specificity, positive predictive power (PPP, or precision), negative predictive power (NPP), and F1 scores at various probability cutoff points. Such conditional probability analyses aid in the selection of classification cutoff points that are useful in different clinical situations.

Learning curves

We used learning curve analysis to evaluate the model's bias and variance (Webb et al., 2011). The learning curve plots the training and test set AUCs from different fractions of the training sample up to the total sample size. With increasing sample size, the training and validation scores should gradually converge. Ideally, the training and validation scores converge at a high level of accuracy, indicating that the model learns well from the training set and generalizes well to the test set. If the training and validation scores converge at a lower point, then the model has been underfit, that is, it does not learn sufficiently from the training samples, but it can generalize this low level of fit to the test samples. If the training scores are high, validation scores are low, and the two scores do not converge, then the model is overfitting the data and generalizes poorly. Inspection of the learning curves provides clues to how models might be improved in the future.

Feature importance scores

We computed each feature's importance score for the RF model. The feature importance score is the fraction of the sample's predictions due to the feature. These scores add up to a total of 100% for each model. For individual features, the higher the value, the more important their contribution to the prediction model. We further evaluated the contribution of the

top important features by fitting the models with or without these features using RF classification. We also determined the predictive accuracy of a logistic regression model using only the top features.

Results

Cross-sectional model

The RF cross-sectional model achieved an AUC of 0.73 (95%CI 0.70–0.76) on the test samples when predicting SUD diagnoses at all visits during age 18–19 from prior data (Figure 1). Having had a prior SUD diagnosis before age 17 was the most important predictor, accounting for 25% of predictive accuracy (Table 1). Figure 2, Left shows the feature importance scores across main categories. When prior SUD diagnoses for the index child were not used as a predictive feature, the remaining family members' SUD diagnoses together accounted for 3% of predictive accuracy and other categories such as criminal records and family SES increased their contribution. When prior SUD diagnoses were removed from the prediction model, the AUC decreased to 0.69 (95%CI 0.66–0.72, Figure 1), suggesting that prior diagnosis, although highly informative, is not needed for significant predictions.

For our cross-sectional model, the most useful clinical prediction would be new onsets of SUD during age 18–19. For these individuals, the RF AUC was 0.67 (95%CI: 0.64, 0.71, Figure 3A). Model calibration (Figure 3B) was good, with a nonsignificant Hosmer–Lemeshow test p value ($\chi^2_{(6)} = 9.8$, $p = .13$) and low Brier score 0.086. Figure 3D gives the precision–recall curve, which plots the PPP against sensitivity. Two examples of probability cutoffs were selected, and the corresponding model metrics at these two cutoff points are listed in Figure 3D. The sensitivities/recall was 2.71% and 27.2%, respectively, and their corresponding PPP/precision was 54.6% and 20.4%.

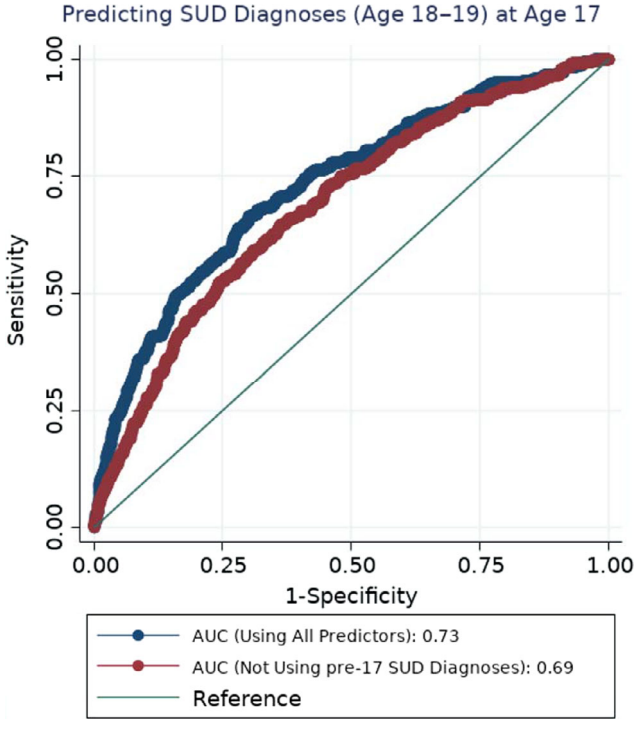


Figure 1 RF cross-sectional model prediction of all SUD diagnoses during age 18–19. Receiver operating characteristic (ROC) curves for the RF model were shown with or without using prior diagnosis of SUD as a predictor [Colour figure can be viewed at wileyonlinelibrary.com]

Top features

In Table 1 lists the top 20 most important features with their relative feature importance scores. Besides a prior SUD diagnosis, the most important features were teenage criminal records (from onset age 15 up to age 17) and a childhood (age 2–12) ADHD diagnosis, followed by stimulant treatment prior to age 18, diagnosis of anxiety disorder and SES indices (such as family income and neighborhood deprivation scores) during teenage years (age 12–17). Father's nonviolent crimes before birth and SES indices, as well as ADHD diagnosis during teenage years, were also among the top 20 list but ranked lower.

To examine the predictive importance of these top features, we compared the AUCs of models predicting new onsets of SUD with or without these features. We also examined logistic regression models using the top 10 or 20 features as predictors. Figure S1 compares these models. For RF model, when we removed the top 10 most important features, prediction accuracy for new SUD cases was

Table 1 Top 20 important features

Rank	Importance (%)	Features
1	25	SUD diagnosis (index child: 12–17)
2	10	NonViolent Crimes (index child: 12–17)
3	6	Violent Crimes (index child: 12–17)
4	3	ADHD Diagnosis (index child: 2–12)
5	2	Psychostimulants treatment (index child: 12–17)
6	2	Family Income (min percentile: 12–17)
7	2	Anxiety diagnosis (index child: 12–17)
8	1	NonViolent Crimes (father: prenatal)
9	1	NDEP (max score:12–17)
10	1	Family Income (max percentile: 12–17)
11	1	Family Income (mean percentile: 12–17)
12	1	Family Income (min percentile: 2–12)
13	1	NDEP (mean score: 0–2)
14	1	Family Income (min percentile: 0–2)
15	1	Family Income (max percentile: 0–2)
16	1	NDEP (max score: 0–2)
17	1	Family Income (mean percentile: 0–2)
18	1	Family Income (mean percentile: 2–12)
19	1	ADHD Diagnosis (index child: 12–17)
20	1	NDEP (max score: 2–12)

Importance score represents percentage of contribution toward the prediction accuracy.
NDEP, Neighborhood deprivation scores.

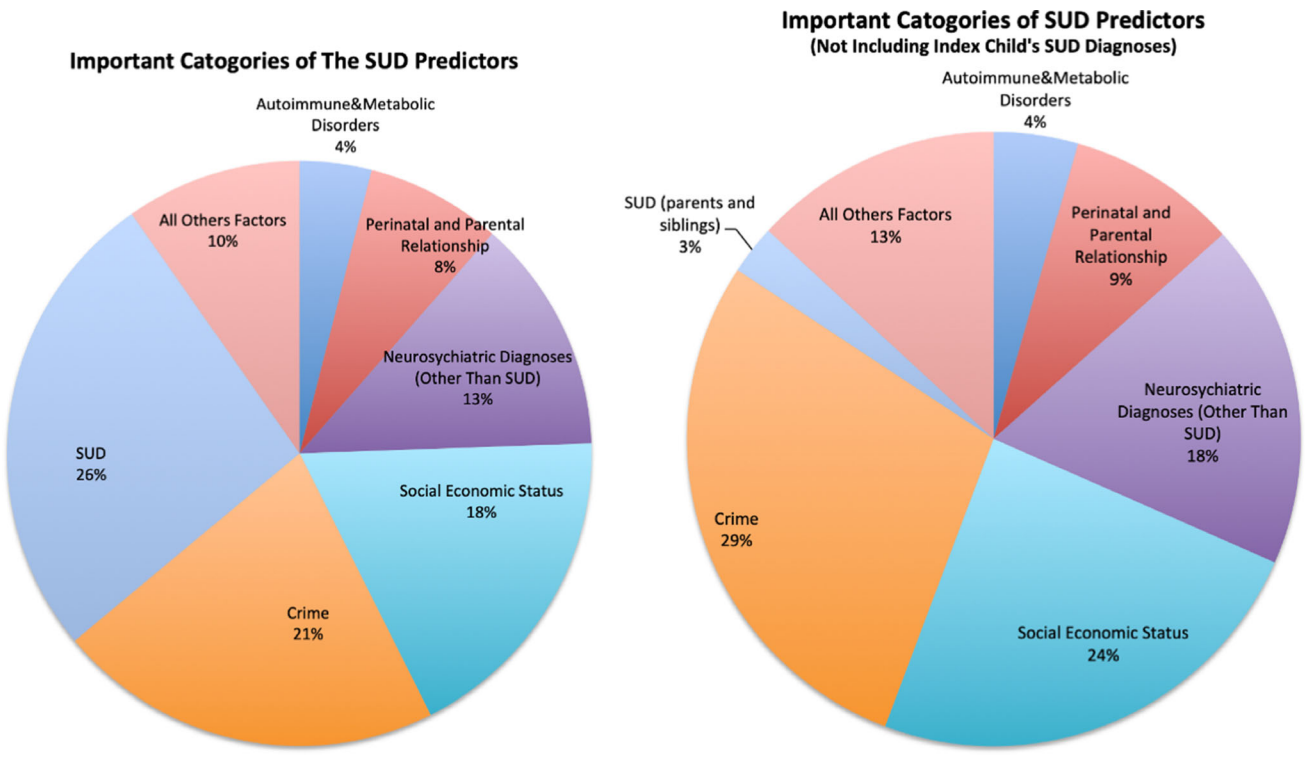
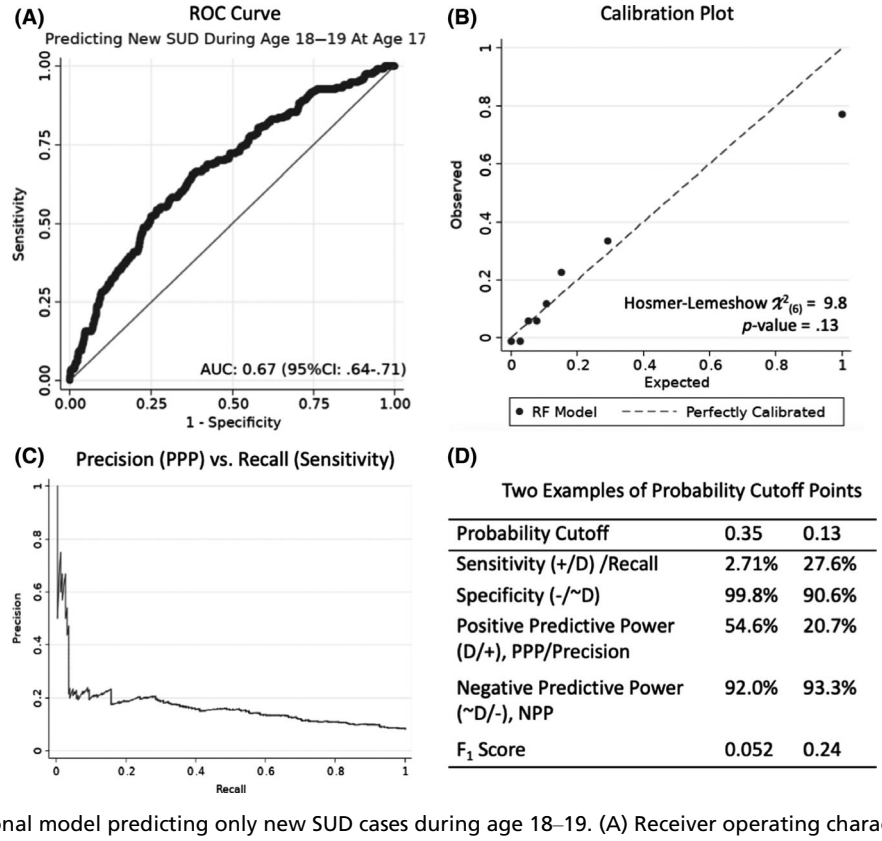


Figure 2 Feature Categories. Features important scores were combined into seven main categories, and their total contribution to the model predictions were plotted for the RF models with (Left) and without (Right) using prior diagnosis as a predictor [Colour figure can be viewed at wileyonlinelibrary.com]



Probability Cutoff	0.35	0.13
Sensitivity (+/D) /Recall	2.71%	27.6%
Specificity (-/~D)	99.8%	90.6%
Positive Predictive Power (D/+), PPP/Precision	54.6%	20.7%
Negative Predictive Power (~D/-), NPP	92.0%	93.3%
F_1 Score	0.052	0.24

Figure 3 RF cross-sectional model predicting only new SUD cases during age 18–19. (A) Receiver operating characteristic (ROC) curve. (B) Calibration curve. (C) precision–recall curve. (D) Sensitivities, specificities, PPPs, NPPs, and F_1 Scores at two example probability cutoffs

significantly reduced (AUC 0.59, 95%CI: 0.56–0.63, $\chi^2_{(1)} = 16.2$, $p = .0001$), although the prediction was still significantly above chance (at AUC 0.5). Excluding 10 more features (Top 20) did not significantly reduce further the AUC (0.58, 95%CI 0.54–0.62). The importance of the top 10 ranked features

was further confirmed by the significance and magnitude of the AUCs when using only these 10 features (0.66, 95%CI 0.62–0.70) and similar results when using the top 20 features in the RF model (0.67, 95%CI: 0.64–0.71). Logistic regression models using the top 10 or top 20 features derived from the RF analyses produced slightly but significantly lower AUCs when compared with the corresponding RF models (0.63, 95%CI 0.60–0.67, $\chi^2_{(1)} = 3.96$, $p = .047$ and 0.64, 95%CI 0.60–0.68, $\chi^2_{(1)} = 4.75$, $p = .03$).

Longitudinal model

Figure 4A illustrates the longitudinal model design depicting the input predictors at each age. The prediction AUCs at each age are shown in Figure 4B, for predicting new SUD diagnoses one, two, five, or ten years in the future. An inserted table lists the numbers of diagnosed SUD cases at each age, starting from age 12, and the total numbers of new cases and their prevalence (%) at each age in the test set. The average AUCs were similar (0.65 ~ 0.66) for all the intervals and majority of the AUCs were significant with their 95% CI intervals above 0.5. We compared the two-year outlook prediction of new SUD cases at age 17 for the cross-sectional and longitudinal models. Figure S2 shows that both models have significant AUCs that were above 0.5. However, the cross-sectional model had a significantly higher AUC than the longitudinal model ($\chi^2_{(1)} = 6.60$, $p = .01$).

Learning curve analysis

The learning curve was plotted for the RF cross-sectional model (Figure S3). Training and validation AUCs gradually converged with the increased samples used. However, their final scores did not fully converge. In addition, the validation AUCs plateaued rather quickly. The learning curve characteristics suggest that more training samples and additional informative features are both needed to improve the prediction accuracy and reduce overfitting.

Discussion

Using Swedish population register data, we applied different ML models to predict SUDs in ADHD youth. The cross-sectional model significantly predicted the probability of having SUD during ages 18–19. The longitudinal model predicted short- and long-term risks for future new diagnoses at each age. Both models yielded significant predictions. Notably, the longitudinal model was able to predict future SUD diagnoses at young ages, many years prior to their ages at first SUD diagnosis.

This study is the first to apply machine-learning algorithms to predict a serious and public health relevant outcome in the context of ADHD. We evaluated the potential clinical utility of the prediction models by computing sensitivity/recall, specificity, PPP/precision, NPP and F_1 scores at various cutoff points (Figure 3B–D). Ideally, a prediction model would identify most patients who would go on to future substance use (high sensitivity/recall) and few who would not (high PPP/precision). Although our reported metrics showed that our model performance was not ideal, they do, however, indicate that large-scale data combined with machine learning may eventually arrive at clinically useful prediction models. Analyses of the full electronic health records ascertained from the actual healthcare system or even more detailed predictor information from linked population register are two strategies to improve the predictive power.

Future research also needs to carefully consider clinically useful cutoff points for the obtained risk scores. For example, Figure 3B shows that using a cutoff point with a sensitivity of 2.7% defines a sample in which 54.6% (the PPP at that cutoff point) will have a subsequent SUD diagnosis. If we instead use a cutoff point with a sensitivity of 27.2%, the PPP decreases to 20.4%, which means that only 1 in 5 patients defined by the model as being at risk for SUD are truly at risk. Although this cutoff point gives a low PPP, it could be useful because the burden of data collection is low (the data are already available in the medical record) and the results can be used for economical and noninvasive interventions such as parent and patient education or more frequent monitoring of high-risk patients.

Explainable machine-learning models are extremely important in translational medicine as they facilitate transparent and trustworthy implementations (Holzinger et al., 2017). In contrast to the difficult to interpret 'Black Box' underlying most machine-learning methods, random forest's feature importance scores aid the interpretability of the prediction models and serve as an effective feature selection method. Our examination of all features extracted from eight different registers showed that only 10 features are needed for obtaining significant predictions. These features mainly include criminal records, prior diagnoses of SUD, ADHD and anxiety, ADHD stimulant treatment and family social economic status (Table 1). In addition, having a handful of useful features, rather than hundreds or thousands of registry records, significantly eases the difficulty of any clinical implementation in the future. In fact, we showed that a logistic regression model with these identified top features produced almost equally significant prediction accuracies.

However, it is also important to fully understand the limitations of RF feature importance scores. Top features in the RF models are not adjusted for confounders and should be interpreted as useful predictors but not, based on our analyses, etiologic risk factors. Furthermore, although the feature

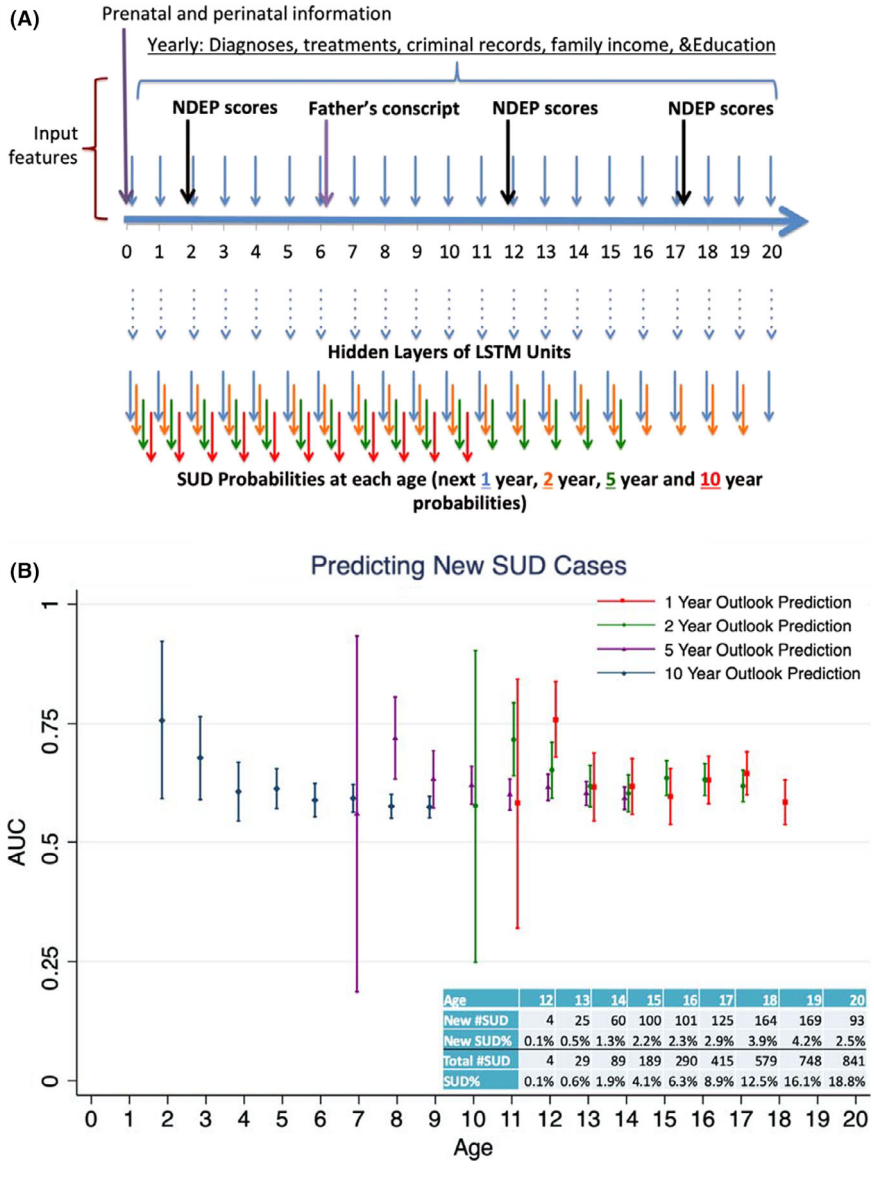


Figure 4 Longitudinal model predicting new SUD diagnoses at each age. (A) Model architecture. (B) AUC at each age for 1-, 2-, 5-, and 10-year outlook predictions [Colour figure can be viewed at wileyonlinelibrary.com]

importance scores from the RF algorithm will extract important features, we cannot conclude that features not extracted are not relevant to SUD. For example, if two features are highly correlated, the one which is most predictive will be deemed important. The other one will have a low importance score because it is not useful after the first one has been selected into the model (Breiman, 1984). Therefore, although the importance scores are useful for explaining the performance of our models, they should not be used to make relative comparisons between features with regards the degree of risk they impart for SUD. In other words, the top features may mask other smaller but important predictors of SUDs. Because of such correlations, when we dropped the prior diagnoses of SUD from our models, their AUCs did not drop dramatically, indicating a substantial amount of redundant information in the remaining feature set.

Consistent with prior work (Havnes, Clausen, Brux, & Middelthon, 2014; Lichtenstein et al., 2012; Stoddard et al., 2015), we found that prior committed crimes (both nonviolent and violent crimes) during teenage years contributed 16% to the predictive accuracy. Indeed, in our sample, those who committed crimes during teenage years had three times higher SUD prevalence (29.4%) during age 18–19 than those who did not have criminal history (prevalence 9.6%). Many previous studies have reported that low family socioeconomic status is associated with SUDs (e.g., Barr et al., 2018; Butterworth, Becker, Degenhardt, Hall, & Patton, 2018). Indeed, numerous SES features in our study, including family income and neighborhood deprivation scores, had high feature importance scores. Altogether, features from the SES category accounted for 18%–24% of predictive accuracy (Figure 2B). Having had an ADHD diagnosis in childhood was ranked fourth in feature importance.

Interestingly, children who had been diagnosed with ADHD during childhood had lower risks for SUD (5.2% would have SUD during 18–19) than those who were diagnosed with ADHD during adolescence (9.4%). Those who were diagnosed with ADHD during 18–19 had the highest SUD prevalence 14.3%. This could reflect changes in diagnostic and treatment practices with calendar time but it is possible that delayed diagnoses of ADHD increase the risk for SUD due to delayed treatment. This idea is supported by many studies showing that the treatment of ADHD in youth leads to lower risks for outcomes such as criminality (Lichtenstein et al., 2012), traffic accidents (Chang, Lichtenstein, D'Onofrio, Sjolander, & Larsson, 2014), smoking (Schoenfelder et al., 2014), and SUDs (Chang, Lichtenstein, Halldner, et al., 2014; Quinn et al., 2017).

One limitation of the study concerns the reliability and accuracy of diagnoses by ICD codes in registry data. Although formal validity studies are lacking for clinically diagnosed SUD in the Swedish national patient registry, most validated diagnoses in the register have a positive predictive value of about 85–95% when compared with research diagnoses (Ludvigsson et al., 2011). The national patient registry has been used substantially in prior research about SUD to generate research findings that fit with findings from studies using other types of measures for SUD (Fazel, Langstrom, Hjern, Grann, & Lichtenstein, 2009). For ADHD, recent validation checks indicate low numbers of false-positive diagnoses of ADHD in the Swedish patient registry (Larsson et al., 2013). The ascertainment of individuals with ADHD was predominantly based on ICD-10 diagnoses. The ICD-10 definition of ADHD is somewhat stricter compared with that in DSM-5, meaning that generalizations to cases of less severe ADHD symptoms should be made with caution. We also do not know if our model predictions would generalize to the population, because our sample only contains patients with an ICD-10 diagnosis of ADHD. Future studies to predict the SUD in the entire population would be needed.

There are several other limitations in our study. First, our learning curve analyses suggest that improvements in predictive accuracy will require additional features and a larger sample size. It is also possible that major improvements to prediction will require a different source of data, such as biomarker or behavioral assays, which are not available in the registries. Second, we have only information available up to age 20 for all patients. Therefore, we are only predicting the SUD onset risks up to age 20 and cannot draw conclusions about predictive accuracy for older ages. Finally, our feature importance analysis, albeit informative for prediction mechanisms, does not provide direct evidence for etiological risk factors for SUD. However, future research based on new ideas derived from our feature importance analysis could provide evidence clarifying if there are any causal mechanisms and perhaps discover novel links.

Despite these limitations, our results suggest that population registry data are useful for machine-learning algorithms to predict the future onset of SUDs when the actions taken based on the predictions are neither costly nor invasive. Future work should focus on improving the sensitivity and positive predictive value by including more detailed information from predictors.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article:

Figure S1. Effects of top features on prediction of new SUD Cases.
Figure S2. Predicting new SUD cases at age 17: cross-sectional model versus longitudinal model.
Figure S3. Learning curves plotted for the RF cross-sectional model.
Table S1. Complete list of registers and predictors extracted from each register, as well as the percentage of the missing data.
Table S2. Retained samples from the 1989 to 1993 Birth Cohort and their SUD diagnostic status pre- and post-age 17.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme grant agreement No 667302. This publication reflects only the authors' views and the European Commission is not responsible for any use that may be made of the information it contains. H.L. acknowledges financial support from the Swedish Research Council (2018-02599) and the Swedish Brain Foundation (FO2018-0273). Additional support was provided by Shire Development, Inc. The authors would like to thank Machine2Learn for programming support. In the past year, S.V.F. received income, potential income, travel expenses continuing education support and/or research support from Vallon, Tris, Otsuka, Arbor, Ironshore, Shire, Akili Interactive Labs, VAYA, Ironshore, Sunovion, Supernus, and Genomind. With his institution, S.V.F. has US patent US20130217707 A1 for the use of sodium-hydrogen exchange inhibitors in the treatment of ADHD. H.L. has served as a speaker for Evolan Pharma and Shire and has received research grants from Shire; all outside the submitted work. The remaining authors have declared that they have no competing or potential conflicts of interest.

Correspondence

Stephen V. Faraone, Department of Psychiatry and Behavioral Sciences, SUNY Upstate Medical University, 750 E Adam St, Syracuse, NY 13210, USA; Email: sfaraone@childpsychresearch.org

Key points

- Population registry data and linked electronic health records can be used to predict at-risk comorbid SUDs in individuals with ADHD.
- Risk predictions can be made many years prior to the first diagnosis.
- Risk monitoring over years during child development can be achieved using longitudinal deep learning models.
- Although promising, more work is needed to improve the prediction accuracy to be sufficient for use in the clinic.

References

Barak-Corren, Y., Castro, V. M., Javitt, S., Hoffnagle, A. G., Dai, Y., Perlis, R. H., … & Reis, B. Y. (2017). Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry*, *174*, 154–162.

Barr, P. B., Silberg, J., Dick, D. M., & Maes, H. H. (2018). Childhood socioeconomic status and longitudinal patterns of alcohol problems: Variation across etiological pathways in genetic risk. *Social Science and Medicine*, *209*, 51–58.

Bergstra, J., Yamins, D., & Cox, D. D. (2013). Making a Science of Model Search: Hyperparameter Optimizationin Hundreds of Dimensions for Vision Architectures. Paper presented at the 30th International Conference on Machine Learning (ICML 2013), Atlanta, Gerorgia.

Bernardini, F., Attademo, L., Cleary, S. D., Luther, C., Shim, R. S., Quartesan, R., & Compton, M. T. (2017). Risk prediction models in psychiatry: Toward a new frontier for the prevention of mental illnesses. *Journal of Clinical Psychiatry*, *78*, 572–583.

Biederman, J., Monuteaux, M. C., Mick, E., Spencer, T., Wilens, T. E., Silva, J. M., … & Faraone, S. V. (2006). Young adult outcome of attention deficit hyperactivity disorder: A controlled 10-year follow-up study. *Psychological Medicine*, *36*, 167–179.

Breiman, L. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Butterworth, P., Becker, D., Degenhardt, L., Hall, W. D., & Patton, G. C. (2018). Amphetamine use in the fourth decade of life: Social profiles from a population-based Australian cohort. *Drug Alcohol Rev*, *37*, 743–751.

Cain, M. A., Bornick, P., & Whiteman, V. (2013). The maternal, fetal, and neonatal effects of cocaine exposure in pregnancy. *Clinical Obstetrics and Gynecology*, *56*, 124–132.

Chang, Z., Larsson, H., Lichtenstein, P., & Fazel, S. (2015). Psychiatric disorders and violent reoffending: A national cohort study of convicted prisoners in Sweden. *Lancet Psychiatry*, *2*, 891–900.

Chang, Z., Lichtenstein, P., D'Onofrio, B. M., Sjolander, A., & Larsson, H. (2014). Serious transport accidents in adults with attention-deficit/hyperactivity disorder and the effect of medication: A population-based study. *JAMA Psychiatry*, *71*, 319–325.

Chang, Z., Lichtenstein, P., Halldner, L., D'Onofrio, B., Serlachius, E., Fazel, S., … & Larsson, H. (2014). Stimulant ADHD medication and risk for substance abuse. *Journal of Child Psychology and Psychiatry*, *55*, 878–885.

Crum, R. M., & Anthony, J. C. (2000). Educational level and risk for alcohol abuse and dependence: Differences by race-ethnicity. *Ethnicity and Disease*, *10*, 39–52.

DeLong, E. R., DeLong, D. M., & Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, *44*, 837–845.

Du Rietz, E., Coleman, J., Glanville, K., Choi, S. W., O'Reilly, P. F., & Kuntsi, J. (2018). Association of polygenic risk for attention-deficit/hyperactivity disorder with co-occurring traits and disorders. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*, 635–643.

Ekbom, A. (2011). The Swedish multi-generation register. *Methods in Molecular Biology*, *675*, 215–220.

Fazel, S., Langstrom, N., Hjern, A., Grann, M., & Lichtenstein, P. (2009). Schizophrenia, substance abuse, and violent crime. *JAMA*, *301*, 2016–2023.

GBD 2015 Mortality and Causes of Death Collaborators (2016). Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: A systematic analysis for the Global Burden of Disease Study 2015. *Lancet*, *388*, 1459–1544.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. Cambridge, MA: The MIT Press.

Hall, W. (2015). What has research over the past two decades revealed about the adverse health effects of recreational cannabis use? *Addiction*, *110*, 19–35.

Hall, W., & Degenhardt, L. (2014). The adverse health effects of chronic cannabis use. *Drug Test Anal*, *6*, 39–45.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd edn). New York: Springer.

Havnes, I. A., Clausen, T., Brux, C., & Middelthon, A. L. (2014). The role of substance use and morality in violent crime: A qualitative study among imprisoned individuals in opioid maintenance treatment. *Harm Reduction Journal*, *11*, 24.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*, 1735–1780.

Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? arXiv e-prints.

Karila, L., Petit, A., Lowenstein, W., & Reynaud, M. (2012). Diagnosis and consequences of cocaine addiction. *Current Medicinal Chemistry*, *19*, 5612–5618.

Karriker-Jaffe, K. J. (2013). Neighborhood socioeconomic status and substance use by U.S. adults. *Drug and Alcohol Dependence*, *133*, 212–221.

Kendler, K. S., Sundquist, K., Ohlsson, H., Palmer, K., Maes, H., Winkleby, M. A., & Sundquist, J. (2012). Genetic and familial environmental influences on the risk for drug abuse: A national Swedish adoption study. *Archives of General Psychiatry*, *69*, 690–697.

Kuntsche, E., Kuntsche, S., Thrul, J., & Gmel, G. (2017). Binge drinking: Health impact, prevalence, correlates and interventions. *Psychology and Health*, *32*, 976–1017.

Lambert, N. M., & Hartsough, C. S. (1998). Prospective study of tobacco smoking and substance dependencies among samples of ADHD and non-ADHD participants. *Journal of Learning Disabilities*, *31*, 533–544.

Larsson, H., Ryden, E., Boman, M., Langstrom, N., Lichtenstein, P., & Landen, M. (2013). Risk of bipolar disorder and

schizophrenia in relatives of people with attention-deficit hyperactivity disorder. *British Journal of Psychiatry*, *203*, 103–106.

Lichtenstein, P., Halldner, L., Zetterqvist, J., Sjolander, A., Serlachius, E., Fazel, S., … & Larsson, H. (2012). Medication for attention deficit-hyperactivity disorder and criminality. *The New England Journal of Medicine*, *367*, 2006–2014.

Ludvigsson, J. F., Almqvist, C., Bonamy, A. K., Ljung, R., Michaelsson, K., Neovius, M., … & Ye, W. (2016). Registers of the Swedish total population and their use in medical research. *European Journal of Epidemiology*, *31*, 125–136.

Ludvigsson, J. F., Andersson, E., Ekbom, A., Feychting, M., Kim, J. L., Reuterwall, C., … & Olausson, P. O. (2011). External review and validation of the Swedish national inpatient register. *BMC Public Health*, *11*, 450.

Ludvigsson, J. F., Otterblad-Olausson, P., Pettersson, B. U., & Ekbom, A. (2009). The Swedish personal identity number: Possibilities and pitfalls in healthcare and medical research. *European Journal of Epidemiology*, *24*, 659–667.

Merikangas, K. R., & McClair, V. L. (2012). Epidemiology of substance use disorders. *Human Genetics*, *131*, 779–789.

Merrin, G. J., Davis, J. P., Berry, D., D'Amico, E. J., & Dumas, T. M. (2016). The longitudinal associations between substance use, crime, and social risk among emerging adults: A longitudinal within and between-person latent variables analysis. *Drug and Alcohol Dependence*, *165*, 71–78.

Molina, B. S., Flory, K., Hinshaw, S. P., Greiner, A. R., Arnold, L. E., Swanson, J. M., … & Wigal, T. (2007). Delinquent behavior and emerging substance use in the MTA at 36 months: Prevalence, course, and treatment effects. *Journal of the American Academy of Child and Adolescent Psychiatry*, *46*, 1028–1040.

Molina, B., & Pelham, W. (2003). Childhood predictors of adolescent substance use in a longitudinal study of children with ADHD. *Journal of Abnormal Psychology*, *112*, 497–507.

Moss, H. B. (2013). The impact of alcohol on society: A brief overview. *Social Work in Public Health*, *28*, 175–177.

Mulia, N., & Karriker-Jaffe, K. J. (2012). Interactive influences of neighborhood and individual socioeconomic status on alcohol consumption and problems. *Alcohol and Alcoholism*, *47*, 178–186.

National Board of Health and Welfare (2014). *Yearbook of the Swedish medical birth registry*. Stockholm, Sweden.

Ng, A. (2019). Machine Learning YearningTechnical Strategy for AI Engneers, In the Era of Deep Learning.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Louppe, G. (2012). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Quinn, P. D., Chang, Z., Hur, K., Gibbons, R. D., Lahey, B. B., Rickert, M. E., … & D'Onofrio, B. M. (2017). ADHD medication and substance-related problems. *American Journal of Psychiatry*, *174*, 877–885.

Sariaslan, A., Langstrom, N., D'Onofrio, B., Hallqvist, J., Franck, J., & Lichtenstein, P. (2013). The impact of neighbourhood deprivation on adolescent violent criminality and substance misuse: A longitudinal, quasi-experimental study of the total Swedish population. *International Journal of Epidemiology*, *42*, 1057–1066.

Schnohr, C., Hojbjerre, L., Riegels, M., Ledet, L., Larsen, T., Schultz-Larsen, K., … & Gronbaek, M. (2004). Does educational level influence the effects of smoking, alcohol, physical activity, and obesity on mortality? A prospective population study. *Scandinavian Journal of Public Health*, *32*, 250–256.

Schoenfelder, E. N., Faraone, S. V., & Kollins, S. H. (2014). Stimulant treatment of ADHD and cigarette smoking: A meta-analysis. *Pediatrics*, *133*, 1070–1080.

Simon, G. E., Johnson, E., Lawrence, J. M., Rossom, R. C., Ahmedani, B., Lynch, F. L., … & Shortreed, S. M. (2018). Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *American Journal of Psychiatry*, *175*, 951–960.

Skjaerven, R., Gjessing, H. K., & Bakketeig, L. S. (2000). Birthweight by gestational age in Norway. *Acta Obstetricia Et Gynecologica Scandinavica*, *79*, 440–449.

Skoglund, C., Chen, Q., Franck, J., Lichtenstein, P., & Larsson, H. (2015). Attention-deficit/hyperactivity disorder and risk for substance use disorders in relatives. *Biological Psychiatry*, *77*, 880–886.

Statistics Sweden (2011). *Longitudinell integrationsdatabas för Sjukförsäkrings- och Arbetsmarknadsstudier (LISA) 1990–2009*. Sweden: Statistics Sweden.

Stoddard, S. A., Epstein-Ngo, Q., Walton, M. A., Zimmerman, M. A., Chermack, S. T., Blow, F. C., … & Cunningham, R. M. (2015). Substance use and violence among youth: A daily calendar analysis. *Substance Use and Misuse*, *50*, 328–339.

The Swedish National Agency for Education (2017). An overview of the Swedish education system. Available from: https://www.skolverket.se/om-skolverket/andra-sprak/in-english/the-swedish-education-system/an-overview-of-the-swedish-education-system-1.72184 [last accessed 5 November 2018].

Thompson Jr, R. G., Lizardi, D., Keyes, K. M., & Hasin, D. S. (2008). Childhood or adolescent parental divorce/separation, parental history of alcohol problems, and offspring lifetime alcohol dependence. *Drug and Alcohol Dependence*, *98*, 264–269.

Tsuang, M. T., Bar, J. L., Harley, R. M., & Lyons, M. J. (2001). The Harvard twin study of substance abuse: What we have learned. *Harvard Review of Psychiatry*, *9*, 267–279.

Webb, G. I., Sammut, C., Perlich, C., Horváth, T., Wrobel, S., Korb, K. B., … & Raedt, L. D. (2011). Learning curves in machine learning. 577–580.

Wettermark, B., Hammar, N., Fored, C. M., Leimanis, A., Otterblad Olausson, P., Bergman, U., … & Rosen, M. (2007). The new Swedish Prescribed Drug Register–opportunities for pharmacoepidemiological research and experience from the first six months. *Pharmacoepidemiology and Drug Safety*, *16*, 726–735.

Whiteford, H. A., Degenhardt, L., Rehm, J., Baxter, A. J., Ferrari, A. J., Erskine, H. E., … & Vos, T. (2013). Global burden of disease attributable to mental and substance use disorders: Findings from the Global Burden of Disease Study 2010. *Lancet*, *382*, 1575–1586.

Wilens, T. E., Faraone, S. V., Biederman, J., & Gunawardene, S. (2003). Does stimulant therapy of attention-deficit/hyperactivity disorder beget later substance abuse? A meta-analytic review of the literature. *Pediatrics*, *111*, 179–185.

Windle, M., & Windle, R. C. (2018). Parental divorce and family history of alcohol disorder: Associations with young adults' alcohol problems, marijuana use, and interpersonal relations. *Alcoholism, Clinical and Experimental Research*, *42*, 1084–1095.

Zhang-James, Y., Helminen, E., Liu, J., Franke, B., Hoogman, M., Faraone, S. V., & ENIGMA-ADHD Working Group (2019, Preprint). Evidence for similar structural brain anomalies in youth and adult attention-deficit/hyperactivity disorder: A machine learning analysis. *bioRxiv*. https://doi.org/10.1101/546671.

Accepted for publication: 28 January 2020