A. SPECIFIC AIMS

As of 2020, 42 states and the District of Columbia (D.C.) have partly legalized or decriminalized marijuana for recreational use despite its Schedule I status, increasing the availability, affordability and use of the drug. This is particularly impactful for adolescents, who may begin use at an earlier age. Though scientific research regarding the long-term effects of cannabis use for young people is only at its infancy, early results have already pointed to significant alterations in neurodevelopmental trajectory through toxicity to brain tissue (Joanna Jacobus, 2014), elevated risk for persistent cognitive impairments (Lubman et al., 2015), and increased occurrence of psychotic symptoms (Arseneault et al., 2002). To minimize cannabis' public health risks, it is imperative that the medical community not only gain a better understanding of the precursors to use, but also to utilize such knowledge to identify at-risk adolescent populations for intervention, the latter of which is lacking in the status quo.

The long-term goal of this study is to synthesize existing addiction research to identify at-risk youth populations at earlier ages. Precursors to cannabis use have been identified in many scientific disciplines, and the overall objective of this application is to investigate the combined effect of two such factors: increased prevalence of anxiety disorder diagnoses arising from the psychology literature (Karina Karolina Kedzior, 2014), and increased genetic expression of the short allele of the 5-HTTLPR serotonin transporter gene arising from the neurobiology lexicon (De Pradier et al., 2010). These make up the rationale behind this proposal. **My central hypothesis is that the two factors will be predictive of adolescent initial cannabis use standalone even at the ages of 10-14, through analysis of subject data collected in the Adolescent Brain Cognitive Development (ABCD) study. Considered together, the predictive accuracy will increase.**

Aim 1: To validate the positive relationship between anxiety disorder diagnosis and early consumption of cannabis. During enrollment into the study at age 9-10, and for each year after, subjects were given a diagnostic interview based on DSM-5 criteria, which includes Generalized Anxiety Disorder (GAD), plus substance abuse surveys for marijuana detailing both last and initial use. This study will regress age of first cannabis use against the severity of GAD diagnosis at intake. *I hypothesize that there will be a positive, significant correlation between GAD severity and earlier initial use.*

Aim 2: To validate the positive relationship between genetic expression of the short allele of the 5-HTTLPR serotonin transporter gene and early consumption of cannabis. The ABCD study also collects and sequences enrollees' serum samples yearly. This study will first determine the abundance of the long versus short allele of the 5-HTTLPR gene for each participant, and then regress the age of first cannabis use against these numbers. *I hypothesize that there will be a positive, significant correlation between elevated "s" allele expression and earlier initial use.*

Aim 3: To build a predictive model using regression techniques for adolescent's initial use that takes both their anxiety disorder and 5-HTTLPR genotype as predictors. Splitting the ABCD enrollment data into 70:30 training and validation split, this study will feed a neural net inputs of participant's GAD severity and 5-HTTLPR, and output a prediction for the age of initial use, if applicable. Techniques I plan to attempt include a random forest, simple ordinary least squares model, and a logistic model. *I hypothesize that the two effects will be additive and predictive of initial use to an accuracy above 50%, or random chance.*

This study will create a technological protocol for analysing interdisciplinary risk factors of drug addiction, in addition to verifying the efficacy of using psychiatric diagnoses and genetic composition as flags for potential adolescent drug use. Further, its success will propel my career goals of utilizing big data across scientific disciplines to create tools that inform policy-makers and healthcare workers combating drug use and abuse.

B. SIGNIFICANCE

According to NIH's Monitoring the Future study, past-year vaping of cannabis has more than doubled in the last two years. Cannabis continues to be the most commonly used illicit drug by adolescents. Unfortunately, scientific research regarding the long-term effects of cannabis use for young people is only at its infancy. Early results support the hypothesis that chronic cannabis use leads to significant alterations in neurodevelopment due to toxicity to brain tissue (Joanna Jacobus, 2014), elevated risk for persistent cognitive impairments (Camchong et al., 2017), and increased occurrence of psychotic symptoms (Arsenault et al, 2002). To minimize the public health risk, it is imperative we not only develop a better understanding of the precursors of use, but also identify at-risk adolescent populations for intervention.

The proposed study is significant as it aims to translate current knowledge of precursors for cannabis use towards identifying at-risk youth. In the proposed design, mental health and genetic variables of 9-10 year-olds participating, collected from the longitudinal Adolescent Brain Cognitive Development (ABCD) study, will be analyzed to create a predictive algorithm estimating time of initial cannabis use. First and early cannabis use has been linked to long term academic and health problems beyond those of neurodevelopment (Brook et al., 2008), hence the rationale behind predicting first use as an important outcome measure.

This study will go beyond current findings to elucidate whether anxiety can increase susceptibility to cannabis use. We will first examine the relationship between diagnosis and severity of General Anxiety Disorder (GAD) at study intake and age of first cannabis use. Poor mental health, in particular anxiety, depression, and social anxiety, has been linked to sustained long-term cannabis use (Bahorik et al., 2017; Boyle et al., 1992), and research has also shown that successful intervention may be significantly related to reduced long-term substance use problems (Kendall et al., 2004). We will attempt to determine the causal direction of this relationship by regressing the age of first cannabis use against the severity of GAD diagnosis at age 9-10 of ABCD study participants' intake.

We will then examine the relationship between allelic expression of the serotonin-transporter-linked polymorphic region 5-HTTLPR and age of first cannabis use. While increased abundance of short allele for this gene has been linked to cannabis use in adolescents (Otten & Engels, 2013), the predictive quality of this trait has not been established, which this study will examine through regression of first cannabis use against allelic expression of the 5-HTTLPR gene.

Lastly, we will examine the interaction of the 5-HTTLPR genotype and GAD diagnosis in association with initial cannabis use. While hypotheses in the literature exist proposing that 5-HTTLPR genotype mediates anxiety in humans (Munafò et al., 2009), this relationship has not been verified in adolescents, who pose as a more difficult subject group due to the many moving factors associated with undergoing physical and neural development. We will build a predictive model using available machine learning techniques for adolescents' initial cannabis use with both 5-HTTLPR and GAD diagnosis variables as inputs, and I hypothesize that such a model will be more predictive than the sum of univariate regressions for either of these variables, a finding that implies compounding interaction between biological pathways and psychological drivers for substance use.

Thus this study is significant as it will provide a rigorous assessment, through the combination of genomic analysis, statistical regression and artificial neural network techniques, of how genetic information can be used to identify youth at risk of early cannabis use. ***I hypothesize that each of these two factors will predict initiation of cannabis use even at the ages of 10-14, and the two factors combined will be more than additive, through analysis of subject data collected in the Adolescent Brain Cognitive Development (ABCD) study.***

C. INNOVATION

This study is also innovative for two reasons. Firstly, it will be the first to examine the predictive quality of the 5-HTTLPR genotype for initial cannabis use, which the literature has hypothesized but not yet verified. Secondly, it will be the first in the literature to examine the interaction between mental health factors and genetic measurements with respect to cannabis use, and to use both factors to predict early use. If the aims of this study are achieved, the efficacy of using psychiatric diagnoses and genetic composition as flags for potential adolescent drug use will be verified. It will also be a first step towards developing multi-disciplinary, algorithmic measures to pinpoint at-risk youth, which will set the stage for future, larger-scale applications of predictive big data analysis to optimize policy interventions.

D. RESEARCH APPROACH

We propose a research design that consists of two components: 1) statistical analysis to ascertain the relationships between GAD diagnosis, 5-HTTLPR genotype and first cannabis use; 2) machine learning protocol for build a predictive model for adolescents' initial cannabis use that takes in the two above factors as inputs.

1. Background

Childhood Mental Health and Cannabis Use: Poor mental health, in particular anxiety, depression, and social anxiety, has been linked to sustained long-term cannabis use (Bahorik et al., 2017; Boyle et al., 1992), and even more literature has suggested that prolonged use of cannabis may worsen, or even induce mental health problems such as schizophrenia and psychosis (Bossong & Niesink, 2010). The value of identifying at-risk children and treating their mental ills has its own intrinsic medical value, but research has also shown that successful intervention may be significantly related to reduced long-term substance use problems (Kendall et al., 2004).

5-HTTLPR Short Allele and Cannabis Use: The genetic abundance of 5HTTLPR serotonin transporter short allele may covary with susceptibility for substance use (Kogan et al., 2010). The 5-HTT gene-linked polymorphic region (5-HTTLPR) is the 5' regulatory promoter segment of the serotonin transporter. The short ("s") allele in the 5-HTTLPR is associated with lower transcriptional efficiency of the promoter compared with the long ("l") allele (Lesch et al., 1996). While the biological mechanism linking higher abundance of the "s" allele to substance use is unknown, it is predictive for some adolescent substance use, including alcohol (van der Zwaluw et al., 2010).

5-HTTLPR and Mental Health: Abundance of the short allele for the gene has also been shown to be correlated with depression (Merenäkk et al., 2010; Risch et al., 2009), and mediates human stress response (Hariri et al., 2002). In animal models of mice and rhesus macaques, disrupted HTT and increased expression of the short allele respectively are both associated with decreased serotonergic functions and increased frequency of fear responses (Grant & Bennett, 2003; Murphy et al., 2001). A possible hypothesis is that persons more prone to depression are thus more likely to seek drugs such as cannabis as relief from stress.

2. Methods

This study will build a predictive model for adolescents' initial cannabis use with both 5-HTTLPR and GAD diagnosis variables as inputs. I hypothesize that such a model will be more predictive than the sum of univariate regressions for either of these variables, a finding that implies compounding interaction between biological pathways and psychological drivers for substance use.

Sample Size and Data: Data of adolescents enrolled into the ABCD study will be used. N = 12,000 9-10 year olds were recruited, targeting an equal number of males and females regardless of race and ethnicity. During the intake year and for each year following, participants are interviewed for their mental health corresponding to DSM-5 criteria, and for substance use for both a last use survey and a follow-back survey for lifetime use of a variety of drugs, including cannabis. Biospecimens are also drawn. Of interest to us are the genetically sequenced blood and oral fluid samples.

Relevant variables: Initial use is defined as the year in which the participant first answers affirmative to lifetime use of cannabis. If the participant answers this during the intake survey, we will assume first use occurred at their intake age (9 or 10 years), or the specified age in the survey, whichever is more granular. To measure the degree of GAD diagnosis, the intake diagnostic DSM-5 interview results will be used. To measure the genetic abundance of the 5-HTTLPR short allele, the genetic sequence from their intake blood test will be used.

Genetic Analysis: Upon collection of the blood and oral fluid samples, the DNA is subjected to the Smokescreen Genotyping array (Baurley et al., 2016), which documents over 300,000 SNPs and includes the short/long alleles of 5-HTTLPR. The short and long allelic abundance of this region can therefore be extracted from the dataset directly (Uban et al., 2018).

2.1 Statistical Analysis (Aim 1, 2)

Regression analysis will be conducted using Python 3.7.0 and the statsmodels package, using an ordinary linear regression:

$$Y = \beta_i X_i + \epsilon$$

- Y - age of first use.
- X_1 - severity of GAD diagnosis, where 0 indicates no diagnosis at all
- Epsilon - unexplained variance or the residual

A similar regression will be used to examine Aim 2, with the slight difference of in the X-variable, where X_2 - allelic abundance of the short allele as a proportion of total 5-HTTLPR expression.

The goal of this analysis is to come up with the two betas associated with the X-variables, which will represent the relationship between GAD diagnostic severity and 5-HTTLPR allelic abundance with age of first cannabis use respectively. Further, to understand the interaction between these two factors, the OLS will be tweaked to the following:

$$Y = \beta_1{}'X_1 + \beta_2{}'X_2 + \epsilon$$

Importantly, the beta-tilde values may not be the same as the beta values. The difference between the beta-tilde and beta values will indicate the relationship between the two X-variables.

2.2 Machine Learning Protocol for Predictive Model Development (Aim 3)

All supervised machine learning analysis will be conducted using TensorFlow, a Python package for machine learning. The model will be trained on 70% of the data, randomly chosen. It will be given examples (inputs) as tuples of each participant's GAD diagnosis and 5-HTTLPR short allele abundance, and labels (outputs) as the participant's age of initial cannabis use. We will start with a 2-layer model, where the first layer has a linear activation, and the second layer has a Regularized Linear Unit (ReLU) activation. This is to ensure that the model outputs are greater than zero. Other activation functions can be explored as the need arises, such as the sigmoid.

All models will be evaluated and trained using the metric of mean squared error, and optimized using Adam (Kingma & Ba, 2014). This optimizer is preferred as it computes adaptive learning rates for individual parameters, which is beneficial in this case as the predictive power of the two X-variables may be different.

3. Analyses and Predictions

3.1 Statistical Analysis Evaluation

A sample of the statistical output from running an OLS in Python is shown in Figure 1. The "coef" column is the beta for the X-variable. As we expect that the severity of GAD diagnosis and allelic abundance of the short allele for 5-HTTLPR are both correlated with an earlier age for initial cannabis use, all betas should be negative.

Aside from the beta value, a few key metrics will be analyzed. For statistical significance, the P-value for the X-variable must be smaller than 0.05 threshold, as is the standard in the literature.

```
In [88]: print(str(model.summary()))
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.338
Model:                            OLS   Adj. R-squared:                  0.272
Method:                 Least Squares   F-statistic:                     5.113
Date:                Tue, 30 Jan 2018   Prob (F-statistic):             0.0473
Time:                        14:21:22   Log-Likelihood:                -41.442
No. Observations:                  12   AIC:                             86.88
Df Residuals:                      10   BIC:                             87.85
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
const        176.6364     20.546      8.597      0.000       130.858   222.415
x             -0.3572      0.158     -2.261      0.047        -0.709    -0.005
==============================================================================
Omnibus:                        1.934   Durbin-Watson:                   1.182
Prob(Omnibus):                  0.380   Jarque-Bera (JB):                1.010
Skew:                          -0.331   Prob(JB):                        0.603
Kurtosis:                       1.742   Cond. No.                     1.10e+03
==============================================================================
```

Figure 1 | Sample OLS Results

The R-squared value will also be important, and can be interpreted as the amount of variation in initial age of cannabis use that can be explained by the X-variable. Since we expect these two values to be additive, the third OLS should yield the highest R-squared and adjusted R-squared value (i.e. lower epsilons), while still retaining significant p-values for both coefficients, meaning that the two coefficients combined explain more of the variation in age of initial cannabis use than one of them alone. If the two X-variables interact and covariate, then the beta-tilde values should be smaller than the corresponding beta values.

3.2 Machine Learning Evaluation

To evaluate model quality, training error and testing error will be plotted against epochs. Both errors should decrease over epochs (Figure 2). If after plateauing, the test error increases while training error decreases or remains constant, it indicates that overfitting has occurred.

The model's performance will also be validated with the 30% holdout set. The model's output will be rounded to the nearest integer, and compared to the actual initial age of cannabis use for each participant. Accuracy is defined as a match (i.e. +/- 0.5 of the true age). A successful model will have an accuracy rate greater than 50%, indicating that it performs better than chance.

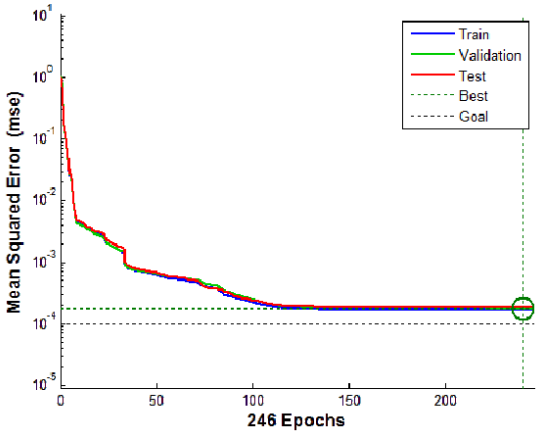

Figure 2 | Sample Train and Test Error

REFERENCES

Arseneault, L., Cannon, M., Poulton, R., Murray, R., Caspi, A., & Moffitt, T. E. (2002). Cannabis use in adolescence and risk for adult psychosis: longitudinal prospective study. *BMJ : British Medical Journal*, *325*(7374), 1212.

Bahorik, A. L., Leibowitz, A., Sterling, S. A., Travis, A., Weisner, C., & Satre, D. D. (2017). Patterns of marijuana use among psychiatry patients with depression and its impact on recovery. *Journal of Affective Disorders*, *213*, 168–171.

Baurley, J. W., Edlund, C. K., Pardamean, C. I., Conti, D. V., & Bergen, A. W. (2016). Smokescreen: a targeted genotyping array for addiction research. *BMC Genomics*, *17*, 145.

Bossong, M. G., & Niesink, R. J. M. (2010). Adolescent brain maturation, the endogenous cannabinoid system and the neurobiology of cannabis-induced schizophrenia. *Progress in Neurobiology*, *92*(3), 370–385.

Boyle, M. H., Offord, D. R., Racine, Y. A., Szatmari, P., Fleming, J. E., & Links, P. S. (1992). Predicting substance use in late adolescence: results from the Ontario Child Health Study follow-up. *The American Journal of Psychiatry*, *149*(6), 761–767.

Brook, J. S., Stimmel, M. A., Zhang, C., & Brook, D. W. (2008). The Association Between Early Marijuana Use and Subsequent Academic Achievement and Health Problems: A Longitudinal Study. *The American Journal on Addictions / American Academy of Psychiatrists in Alcoholism and Addictions*, *17*(2), 155.

Camchong, J., Lim, K. O., & Kumra, S. (2017). Adverse Effects of Cannabis on Adolescent Brain Development: A Longitudinal Study. *Cerebral Cortex* , *27*(3), 1922–1930.

De Pradier, M., Gorwood, P., Beaufils, B., Adès, J., & Dubertret, C. (2010). Influence of the serotonin transporter gene polymorphism, cannabis and childhood sexual abuse on phenotype of bipolar disorder: a preliminary study. *European Psychiatry: The Journal of the Association of European Psychiatrists*, *25*(6), 323–327.

Grant, K. A., & Bennett, A. J. (2003). Advances in nonhuman primate alcohol abuse and alcoholism research. *Pharmacology & Therapeutics*, *100*(3), 235–255.

Hariri, A. R., Mattay, V. S., Tessitore, A., Kolachana, B., Fera, F., Goldman, D., Egan, M. F., & Weinberger, D. R. (2002). Serotonin Transporter Genetic Variation and the Response of the Human Amygdala. *Science*, *297*(5580), 400–403.

Joanna Jacobus, S. F. T. (2014). Effects of Cannabis on the Adolescent Brain. *Current Pharmaceutical Design*, *20*(13), 2186.

Karina Karolina Kedzior, L. T. L. (2014). A positive association between anxiety disorders and cannabis use or cannabis use disorders in the general population- a meta-analysis of 31 studies. *BMC Psychiatry*, *14*, 136.

Kendall, P. C., Safford, S., Flannery-Schroeder, E., & Webb, A. (2004). Child anxiety treatment: outcomes in adolescence and impact on substance use and depression at 7.4-year follow-up. *Journal of Consulting and Clinical Psychology*, *72*(2), 276–287.

Kingma, D. P., & Ba, J. (2014). *Adam: A Method for Stochastic Optimization*. http://arxiv.org/abs/1412.6980

Kogan, S. M., Beach, S. R. H., Philibert, R. A., Brody, G. H., Chen, Y.-F., & Lei, M.-K. (2010). 5-HTTLPR status moderates the effect of early adolescent substance use on risky sexual behavior. *Health Psychology: Official Journal of the Division of Health Psychology, American Psychological Association*, *29*(5), 471–476.

Lesch, K.-P., Bengel, D., Heils, A., Sabol, S. Z., Greenberg, B. D., Petri, S., Benjamin, J., Müller, C. R., Hamer, D. H., & Murphy, D. L. (1996). Association of Anxiety-Related Traits with a Polymorphism in the Serotonin Transporter Gene Regulatory Region. *Science*, *274*(5292), 1527–1531.

Lubman, D. I., Cheetham, A., & Murat, Y. (2015). Cannabis and adolescent brain development. *Pharmacology & Therapeutics*, *148*, 1–16.

Merenäkk, L., Mäestu, J., Nordquist, N., Parik, J., Oreland, L., Loit, H.-M., & Harro, J. (2010). Effects of the serotonin transporter (5-HTTLPR) and α 2A -adrenoceptor (C-1291G) genotypes on substance use in children and adolescents: a longitudinal study. *Psychopharmacology*, *215*(1), 13–22.

Munafò, M. R., Freimer, N. B., Ng, W., Ophoff, R., Veijola, J., Miettunen, J., Järvelin, M.-R., Taanila, A., & Flint, J. (2009). 5-HTTLPR genotype and anxiety-related personality traits: a meta-analysis and new data. *American Journal of Medical Genetics. Part B, Neuropsychiatric Genetics: The Official Publication of the International Society of Psychiatric Genetics*, *150B*(2), 271–281.

Murphy, D. L., Li, Q., Engel, S., Wichems, C., Andrews, A., Lesch, K. P., & Uhl, G. (2001). Genetic perspectives on the serotonin transporter. *Brain Research Bulletin*, *56*(5), 487–494.

Otten, R., & Engels, R. C. M. E. (2013). Testing bidirectional effects between cannabis use and depressive symptoms: moderation by the serotonin transporter gene. *Addiction Biology*, *18*(5), 826–835.

Risch, N., Herrell, R., Lehner, T., Liang, K.-Y., Eaves, L., Hoh, J., Griem, A., Kovacs, M., Ott, J., & Merikangas, K. R. (2009). Interaction Between the Serotonin Transporter Gene (5-HTTLPR), Stressful Life Events, and Risk of Depression: A Meta-analysis. *JAMA: The Journal of the American Medical Association*, *301*(23), 2462–2471.

Uban, K. A., Horton, M. K., Jacobus, J., Heyser, C., Thompson, W. K., Tapert, S. F., Madden, P. A. F., Sowell, E. R., & Adolescent Brain Cognitive Development Study. (2018). Biospecimens and the ABCD study: Rationale, methods of collection, measurement and early data. *Developmental Cognitive Neuroscience*, *32*, 97–106.

van der Zwaluw, C. S., Engels, R. C. M. E., Vermulst, A. A., Rose, R. J., Verkes, R. J., Buitelaar, J., Franke, B., & Scholte, R. H. J. (2010). A serotonin transporter polymorphism (5-HTTLPR) predicts the development of adolescent alcohol use. *Drug and Alcohol Dependence*, *112*(1-2), 134–139.