Preventive Medicine Reports 25 (2022) 101674

Contents lists available at ScienceDirect

Preventive Medicine Reports

journal homepage: www.elsevier.com/locate/pmedr

Joint risk prediction for hazardous use of alcohol, cannabis, and tobacco among adolescents: A preliminary study using statistical and machine learning

Thanthirige Lakshika Maduwanthi Ruberu [a], Emily A. Kenyon [b], Karen A. Hudson [b], Francesca Filbey [c], Sarah W. Feldstein Ewing [b], Swati Biswas [a,*], Pankaj K. Choudhary [a,*]

[a] Department of Mathematical Sciences, University of Texas at Dallas, USA
[b] Department of Psychology, University of Rhode Island, USA
[c] School of Behavioral and Brain Sciences, University of Texas at Dallas, USA

ARTICLE INFO

Keywords:
Adolescents
Machine learning
MCGLM
Multiple outcomes
Multivariate lasso
Risk prediction
Statistical learning
Substance use

ABSTRACT

For some, substance use during adolescence may be a stepping stone on the way to substance use disorders in adulthood. Risk prediction models may help identify adolescent users at elevated risk for hazardous substance use. This preliminary analysis used cross-sectional data (n = 270, ages 13–18) from the baseline dataset of a randomized controlled trial intervening with adolescent alcohol and/or cannabis use. Models were developed for jointly predicting quantitative scores on three measures of hazardous substance use (Rutgers Alcohol Problems Index, Adolescent Cannabis Problem Questionnaire, and Hooked on Nicotine Checklist) based on personal risk factors using two statistical and machine learning methods: multivariate covariance generalized linear models (MCGLM) and penalized multivariate regression with a lasso penalty. The predictive accuracy of a model was evaluated using root mean squared error computed via leave-one-out cross-validation. The final proposed model was an MCGLM model. It has eleven risk factors: age, early life stress, age of first tobacco use, age of first cannabis use, lifetime use of other substances, age of first use of other substances, maternal education, parental attachment, family cigarette use, family history of hazardous alcohol use, and family history of hazardous cannabis use. Different subsets of these risk factors feature in the three outcome-specific components of this joint model. The quantitative risk estimate provided by the proposed model may help identify adolescent substance users of cannabis, alcohol, and tobacco who may be at an elevated risk of developing hazardous substance use.

1. Introduction

Substance use disorders (SUDs) are a major public health issue in the United States (US); in 2019, more than 2.5 million Americans died due to drug- or alcohol-related causes (CDC Wonder, 2020)(CDC WONDER, 2020). Many adults who develop SUDs report initiating substance use during adolescence (NIDA, 2020). Three substances — alcohol, tobacco, and cannabis — have particularly widespread use among adolescents (Johnston et al., 2019). Polysubstance use, or the consumption of more than one substance simultaneously, is also common during the developmental period of adolescence. For example, 34% of adolescents reported using two or more substances from among alcohol, cigarettes, and cannabis prior to age 16 (Moss et al., 2014). In another study, 41.9% of adolescents (average age = 17) reported alcohol and marijuana co-use (Choi et al., 2018).

Several factors have been identified to be associated with increased risk of substance use and its severity among adolescents such as lower levels of parental monitoring, higher levels of parental substance use, family history of substance use, and lower levels of parental education (Lee et al., 2018; Rusby et al., 2018; Yule et al., 2018). Initiating substance use at a younger age also contributes to more severe use later in life (Kim et al., 2017; Tillson et al., 2019; Hawke et al., 2020). Several

Abbreviations: CPQ-A, Adolescent Cannabis Problems Questionnaire; HONC, Hooked on Nicotine Checklist; LOOCV, Leave-one-out cross-validation; MCGLM, Multivariate Covariance Generalized Linear Model; RAPI, Rutgers Alcohol Problems Index; RMSE, Root mean squared error; SD, Standard deviation; SE, Standard error.

* Corresponding authors at: 800 W Campbell Rd, FO 35, Richardson, TX 75025, USA.
E-mail addresses: swati.biswas@utdallas.edu (S. Biswas), pankaj@utdallas.edu (P.K. Choudhary).

https://doi.org/10.1016/j.pmedr.2021.101674
Received 3 August 2021; Received in revised form 23 November 2021; Accepted 12 December 2021
Available online 13 December 2021

studies have also explored risk factors that lead to co-use of substances by adolescents including individual, familial, and sociodemographic factors (White et al., 2019; D'Amico et al., 2020).

Given the potential short- and long-term harms associated with hazardous substance use in adolescence, it is imperative to identify adolescents at high risk of developing hazardous substance use. Hazardous substance use is use of substances that increases the future risk or likelihood of health consequences; this does not include use that has already led to health consequences (Saitz et al., 2021). A few tools are available for individual substance risk prediction. In particular, Hayatbakhsh et al. (2009) developed a risk score for cannabis use and disorder in early adulthood based on early life risk factors. More recently, a simple cumulative risk index was developed to classify which adolescents are at risk for developing persistent substance disorders in adulthood using risk factors from childhood and adolescence (Meier et al., 2016). Another recent study built a model for predicting quantitative risk of developing cannabis use disorder in adults based on personal risk factors using statistical and machine learning approaches (Rajapaksha et al., 2020). Yet another recent study (Jing et al., 2020) built models for predicting risk of developing substance use disorder by thirty years of age using separate sets of predictors from late childhood to 22 years of age.

There is a substantial literature on exploring factors that lead to co-use of multiple substances (White et al., 2019; D'Amico et al., 2020). However, to our knowledge, modeling of hazardous use of multiple substances jointly has not been considered especially in the context of risk prediction modeling. More specifically, the key differences between these previous studies and our present work are that they (1) do not provide a measure of quantitative risk or score (they are, in fact, not intended to be used as risk prediction models), (2) model concurrent use rather than *hazardous use* of multiple substances, and (3) focus on a selected number of risk factors (hypothesis-driven or domain-specific) rather than a comprehensive set of potential risk factors. As such, there is a need to develop risk prediction models for hazardous use of multiple substances based on personal risk factors of adolescent users.

This study aims to fulfill this need by developing preliminary models for jointly predicting hazardous use of alcohol, cannabis, and tobacco for adolescents who have used all three substances in their lifetime. Joint statistical modeling of multiple outcomes utilizes the correlation between them, which can lead to higher power for detecting association between risk factors and outcomes and can additionally provide insight into the shared underlying mechanisms. As our goal is statistical risk prediction rather than hypothesis testing, we consider a set of potential risk factors as suggested by the literature.

The classical multivariate regression is a standard method for modeling multiple outcomes jointly. However, it assumes a common set of predictors for all outcomes, which limits its applicability in our context of risk prediction. This is because if a variable is predictive of one outcome but not another, model parsimony dictates that the variable should be included only in the model for the former but not the latter. Adding unimportant variables to a model adversely affects its ability to predict accurately for new (future) participants that are not included in building the model. Moreover, regularization of regression coefficients in the model can protect against overfitting of the model especially when sample sizes are not large. An overfitted model is suboptimal for the purpose of predicting for new participants (James et al., 2013). However, regularization is not available in the classical approach. Therefore, we apply two relatively new statistical and machine learning methods, each of which addresses one of these limitations. Specifically, we utilize multivariate covariance generalized linear models (MCGLM; Bonat and Jørgensen, 2016) and penalized multivariate regression with a lasso penalty (Friedman et al., 2010). These methods have not been used to model multiple outcomes in the substance use literature perhaps because the development of joint risk prediction models has not yet been considered in a formal way.

2. Methods

2.1. Participants

This study was a preliminary analysis of the baseline data from a larger randomized controlled trial (RCT) intervening in adolescent alcohol and cannabis use (Feldstein Ewing et al., 2021). The original study consisted of 506 alcohol and/or cannabis using adolescents (ages 13–18) who were justice-involved and recruited from the southwestern US. It was conducted with local institutional review board (IRB) approval and a federal certificate of confidentiality. Youth aged 18 or older provided consent and parent consent/adolescent assent was obtained for youth under 18 (Feldstein Ewing et al., 2021). Eligible youth used alcohol and/or cannabis regularly (i.e., at least once per month for the past 6 months). Since the present preliminary analysis involves identifying risk factors for hazardous substance use, only baseline assessment data were used, as they were collected prior to the introduction of a behavioral intervention in the original study. Thus, these baseline data serve as cross-sectional data for the present study.

2.2. Outcome variables

Following recent calls in the literature (Silvers et al., 2019), this study focused on hazardous use as the central outcome variables of interest and treated them as quantitative response variables. Thus, we had a total of three outcome variables, consisting of scores of hazardous alcohol use, hazardous cannabis use, and hazardous tobacco use. These were respectively measured using Rutgers Alcohol Problems Index (RAPI) (White and Labouvie, 1989), Adolescent Cannabis Problems Questionnaire (CPQ-A) (Martin et al., 2006), and Hooked on Nicotine Checklist (HONC) (DiFranza et al., 2002). These measures are designed specifically for adolescents. They are described further in the supplement.

2.3. Data preparation

The analyzed data were restricted to polysubstance users of all three target substances. We utilized the following screening process to identify potential risk factors. We began by selecting only the variables that either remain unchanged (e.g., gender and race) or remain relatively stable over time. Given the cross-sectional nature of the data, restricting attention to such variables protected against using variables that may actually be an effect of hazardous substance use. Some new variables were derived by combining similar constructs (e.g., mother's and father's past substance use were combined to form parental past substance use) to reduce the amount of missing data. This screening process resulted in 18 risk factors. The final dataset consisted of $n = 270$ participants with complete data on the selected risk factors.

2.4. Risk factors

Supplementary Table 1 summarizes the 18 risk factors. They include participant demographics (e.g., age, gender, and cultural identification); general environmental factors (e.g., early life stress, maternal education, level of parental monitoring, level of parental attachment, and level of peer attachment); their own substance use (e.g., age of first use for each substance and lifetime use of substances other than alcohol, cannabis, and tobacco); and family substance use (e.g., history of parental substance use, family's cigarette use, and family history of hazardous alcohol and cannabis use).

Not all participants used substances other than alcohol, cannabis, and tobacco, and hence for them there was no corresponding age of first use of other substances. Therefore, to include this age variable in the model, a binary indicator of lifetime use of other substances was added to the model together with its interaction with the age of first use. This way, the interaction term had a non-zero value only for the users of other

substances, representing their age of first use. All ordinal variables in the model were treated as continuous.

2.5. Data analysis

We used two multivariate statistical modeling frameworks for joint modeling of the three outcome variables: MCGLM and multivariate lasso.

MCGLM: It is a novel generalization of the classical multivariate regression (Bonat and Jørgensen, 2016) allowing modeling of the mean structure, variance function, and within-response covariance structure. MCGLM allows outcome-specific predictors, i.e., the predictors need not be shared across all outcomes. Its predictive accuracy was measured by root mean square error (RMSE), computed using leave-one-out cross-validation (LOOCV) (James et al., 2013). This measure differs from the ordinary RMSE and allows assessment of model performance on future unseen data more accurately by protecting against overfitting. RMSE of a model indicates average size of errors in its predictions. Hence a lower RMSE implies a better model.

Multivariate Lasso (penalized multivariate regression with a lasso penalty): In multivariate regression with lasso, the three-dimensional vectors of regression coefficients are regularized by a group lasso penalty (Friedman et al., 2010), with tuning parameter selected optimally via LOOCV. A risk factor retained in the model (i.e., having non-zero coefficient) is shared across all responses. This modeling approach is directed towards predictive accuracy. So it does not provide standard errors or p-values, but instead measures performance using RMSE computed via LOOCV.

Further details on these methods can be found in Supplementary Materials. The methods were fitted using mcglm (Bonat, 2018) and glmnet (Hastie et al., 2021) packages of statistical software system R (R Core Team 2020). They do not allow missing data.

3. Results

3.1. Sample characteristics

Descriptive statistics for the three outcome variables are as follows: hazardous alcohol use (RAPI; mean = 11.84, SD = 11.01, range = 0–72), hazardous cannabis use (CPQ-A; mean = 7.84, SD = 5.03, range = 0–23), and hazardous tobacco use (HONC; mean = 4.18, SD = 3.67, range = 0–10).

Supplementary Table 2 provides a summary of categorical and continuous risk factors computed from all 270 participants. A majority of the participants were male (74.4%), lived with their family (93%), identified with Hispanic culture (57.4%), and reported elevated early life stress (72.2%). In addition, most of the youth had used other substances (88.1%) at least once in the lifetime. A large proportion also reported having parents with previous substance use (99.3%), cigarette smokers among family members (74.4%), and a family history of hazardous alcohol use (63.3%). However, family history of hazardous cannabis use was less prevalent (40%). On average, a participant was 16 years of age, started to use alcohol, cannabis, and tobacco before age 13, and started to use other substances between ages 13 and 14. In addition, on average, a participant reported less than two years of college for maternal education and moderate levels of parent monitoring and attachment with parents and peers.

3.2. Results from multivariate models

A square root transformation was applied to RAPI and CPQ-A scores for adherence with model assumptions. The means and SDs of the transformed variables are: 3.05 and 1.59 (RAPI) and 2.62 and 0.98 (CPQ-A). The three outcomes are moderately correlated with correlation coefficients of 0.30 (square-roots of RAPI and CPQ-A), 0.22 (HONC and square-root of RAPI), and 0.27 (HONC and square-root of CPQ-A). For age of first tobacco use, quadratic and cubic effects were included in the models in addition to the linear effect as suggested by an initial exploratory data analysis.

3.2.1. Results from MCGLM

The final model consisted of eleven risk factors as presented in Table 1 (age of first tobacco use is considered as one variable even though it had three terms). There was some overlap among the risk factors across the outcome-specific components of the model. The RAPI and CPQ-A components had six risk factors each, whereas the HONC component had seven. Only age of first tobacco use is common to all three components. Further, this age had a linear effect on the transformed RAPI and CPQ-A scores and a nonlinear (cubic) effect on HONC score.

The RAPI component included age, lifetime use and age of first use of other substances, family histories of hazardous cannabis and alcohol use, and age of first tobacco use. Age, lifetime use of other substances, and family history of hazardous cannabis use were positively associated with the outcome, whereas age of first use of other substances, age of first tobacco use, and family history of hazardous alcohol use were negatively associated.

The CPQ-A component included maternal education, parental attachment, early life stress, lifetime use of other substances, and ages of first cannabis and tobacco use. Early life stress, lifetime use of other substances, and age of first cannabis use had a positive association with the outcome, whereas maternal education, parental attachment, and age of first tobacco use had a negative association.

The HONC component included age, parental attachment, early life stress, family use of cigarette, family history of hazardous cannabis use, age of first cannabis use, and cubic effect of age of first tobacco use. Age, early life stress, and family cigarette use were positively associated with the outcome, whereas parental attachment, age of first cannabis use, and family history of hazardous cannabis use were negatively associated. The non-linear (cubic) effect of age of first tobacco use on average HONC score can be described as a slight decrease until age 8, followed by a slight increase between 8 and 12 years, and then a sharp decrease with age.

The overall RMSE of the joint model computed via LOOCV was 2.15. The outcome-specific RMSEs were: 1.53 (RAPI), 0.91 (CPQ-A), and 3.26 (HONC). For reference, the range of these measures in data are 0–8.49 for RAPI (in square root scale), 0–4.80 for CPQ-A (in square root scale) and 0–10 for HONC. This suggests that predictions are most accurate for CPQ-A and least accurate for HONC.

3.2.2. Results from multivariate lasso

Multivariate lasso identified a total of nine risk factors. These are presented in Table 2. As multivariate lasso does not allow risk factors to vary by outcome, all of these risk factors have non-zero coefficients for all three outcomes. All nine risk factors were also identified by MCGLM, which additionally included maternal education and family history of hazardous alcohol use. Both models had similar directions for effects of the selected risk factors except the age of first use of other substances, whose effect on RAPI score was estimated to be practically zero by multivariate lasso while it was negative and significant by MCGLM. Although age of first tobacco use had a cubic effect on HONC score, the effect is much smaller compared to that in MCGLM. The cubic term was included in RAPI and CPQ-A components as well, however, its coefficients were practically zero. Overall, results from this model were consistent with those from MCGLM. Finally, the RMSE for the joint multivariate lasso model was 2.18, with the outcome-specific RMSEs of 1.55 (RAPI), 0.94 (CPQ-A), and 3.32 (HONC).

3.2.3. The final model

MCGLM had slightly smaller RMSEs and hence its predictions were slightly more accurate than multivariate lasso for all three components. Due to this reason and the fact that MCGLM allows outcome-specific risk

Table 1
Summary of results for the joint MCGLM model.

Variable	Hazardous alcohol use (RAPI)		Hazardous cannabis use (CPQ-A)		Hazardous tobacco use (HONC)	
	Estimate (SE)	P-value	Estimate (SE)	P-value	Estimate (SE)	P-value
Intercept	0.196 (1.26)	0.876	2.406 (0.41)	<0.001	−0.051 (9.77)	0.996
Age	0.214 (0.08)	0.007			0.914 (0.17)	<0.001
Maternal education			−0.068 (0.02)	0.002		
Parental attachment			−0.021 (0.01)	0.083	−0.074 (0.04)	0.096
Early life stress			0.354 (0.13)	0.005	1.178 (0.46)	0.010
Lifetime use of other substances	2.557 (0.76)	0.001	0.701 (0.17)	<0.001		
Age of first use of other substances	−0.138 (0.05)	0.008				
Age of first use of cannabis			0.051 (0.02)	0.034	−0.174 (0.09)	0.051
Age of first use of tobacco	−0.095 (0.03)	0.005	−0.065 (0.02)	0.002	−3.212 (2.80)	0.251
(Age of first use of tobacco)^2					0.375 (0.26)	0.144
(Age of first use of tobacco)^3					−0.014 (0.01)	0.067
Family use of cigarette					1.748 (0.47)	<0.001
Family history of hazardous alcohol use	−0.402 (0.19)	0.039				
Family history of hazardous cannabis use	0.589 (0.19)	0.002			−0.925 (0.40)	0.022

Table 2
Estimated coefficients for the variables retained by the multivariate lasso model.

Variable	Hazardous alcohol use (RAPI)	Hazardous cannabis use (CPQ-A)	Hazardous tobacco use (HONC)
Intercept	1.138	2.428	−2.968
Age	0.108	−0.023	0.450
Parental attachment	−0.024	−0.016	−0.045
Early life stress	0.124	0.263	0.949
Lifetime use of other substances	0.500	0.431	1.048
Age of first use of other substances	0.000†	0.010	0.037
Age of first use of cannabis	−0.002	0.016	−0.082
Age of first use of tobacco	0.000	0.000	0.000
(Age of first use of tobacco)^2	0.000	0.000	0.000
(Age of first use of tobacco)^3	0.000†	0.000†	−0.001
Family use of cigarette	0.323	0.141	1.315
Family history of hazardous cannabis use	0.283	−0.005	−0.506

†Non-zero coefficient < 0.001.

factors, the MCGLM model is our final proposed model. The distributions of the predicted scores from this model are displayed in Fig. 1. To illustrate the application of this model, we consider an average participant from the study — the one whose quantitative risk factors are equal to their average values and qualitative risk factors equal their most common categories (as per Supplementary Table 2). The predicted RAPI, CPQ-A, and HONC scores (on original scales) for this participant are 7.56, 7.88, and 6.23, respectively. Using Fig. 1, these scores fall between 1st and 2nd quartiles for RAPI, 2nd and 3rd quartiles for CPQ-A, and above 3rd quartile for HONC.

4. Discussion

Hazardous substance use in adolescence can increase risk for substance use disorders in adulthood, which remain a major public health issue in the United States. So, it is essential to develop new tools for identifying adolescents at risk of developing hazardous substance use among co-users in order to possibly increase the impact of prevention and intervention programs. This study took a preliminary step by determining risk factors associated with hazardous use of alcohol, cannabis, and tobacco — the three of the most commonly used substances by adolescents — and employing them to build joint models for predicting risk of hazardous substance use. To our knowledge, this is the first study that attempted to jointly model hazardous use of multiple substances by adolescents specifically for the purpose of quantitative risk prediction. With information about salient combinations of individual, family and sociodemographic factors, the proposed model can be used to predict scores on quantitative measures of hazardous use for an adolescent user of alcohol, cannabis, and tobacco. A predicted score can be compared with its distribution (e.g., the quartiles of predicted scores provided in Fig. 1) to assess the relative level of risk for the user.

An important novelty of this work is the application of relatively new statistical and machine learning methods for analyzing multiple outcomes. To the best of our knowledge, these methods have not been used so far in the substance use literature. They are better suited for risk prediction than the classical multivariate regression because one (MCGLM) allows outcome-specific predictors, thereby providing flexibility for joint modeling that may lead to higher predictive accuracy; and the other (multivariate lasso) allows regularization of regression coefficients to protect against overfitting, thereby helping the model to generalize well on future unseen data.

For comparison, we also fitted the classical multivariate regression model, whose results are in Supplementary Table 3. This model identified a total of ten risk factors, all of which were also identified by MCGLM, but the latter additionally identified family history of hazardous alcohol use and had higher predictive accuracy. We also fitted three separate linear regression models for the three outcomes (results not shown). None of the individual models identified age of first cannabis use, age of first use of other substances, and family history of hazardous alcohol use, which were identified by MCGLM. Of these, the age of first cannabis use, an important risk factor as per the literature, was identified by MCGLM to be associated with both CPQ-A and HONC scores. Although the individual models identified guardian, parental monitoring, and parental past substance use, which were not identified by MCGLM, the effects of these variables could have been captured partially by similar variables included in the joint model, e.g., parental attachment and family history variables.

One of our key findings is that early age of onset of tobacco use is associated with hazardous use of all three substances. In line with previous literature, one study found that, among adolescent drinkers, past-year smokers were at a higher risk for alcohol use disorders than non-smokers (Grucza and Bierut, 2006). Another study evaluating the association between e-cigarette use and the use of other substances showed that early onset of e-cigarette use was associated with increased use of alcohol and cannabis (McCabe et al., 2018). Tobacco has also been found to be a partial driver of cannabis dependence in young people who use tobacco and cannabis (Hindocha et al., 2015). Another key finding is that lower levels of parental attachment are associated with higher scores of hazardous cannabis and tobacco use. Lower levels of parental monitoring and attachment are also known to be associated with

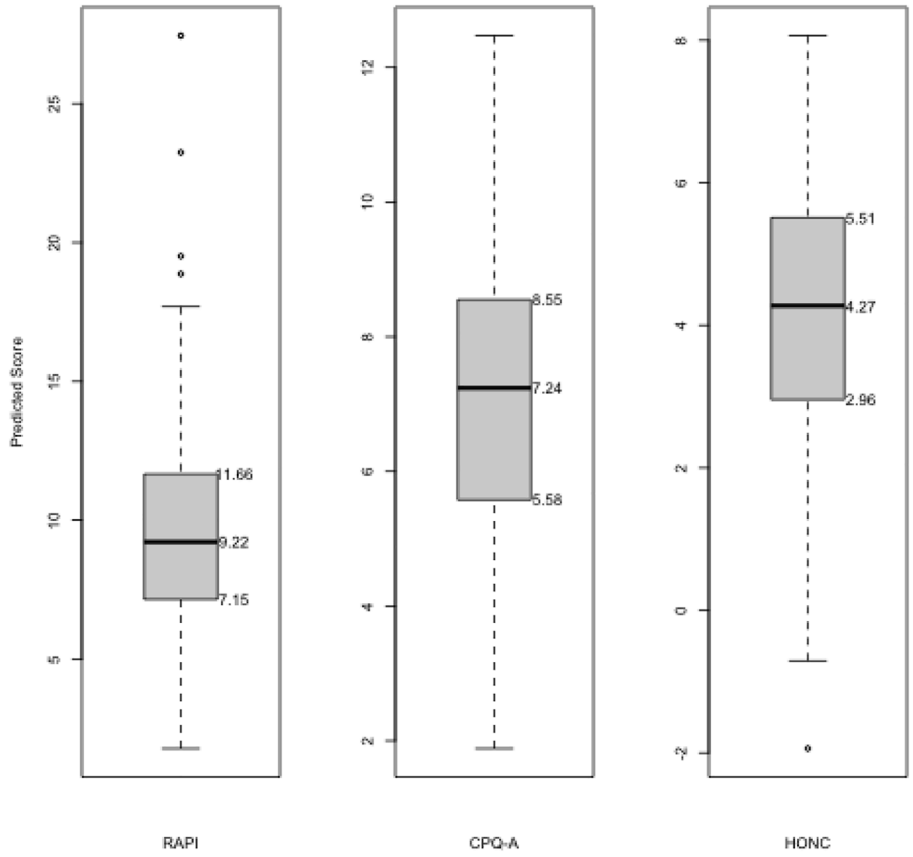

Fig. 1. Distributions of predicted score estimated using MCGLM. The corresponding mean (SD) are 9.64 (3.58) for RAPI, 7.03 (2.00) for CPQ-A, and 4.18 (1.88) for HONC.

adolescent substance use disorders (Van Ryzin et al., 2012; Whitesell et al., 2013; Rusby et al., 2018).

Other findings in this study are also generally consistent with the literature. For example, increased age is associated with higher RAPI and HONC scores. Indeed, there is a higher percentage of young adults with alcohol use disorders compared to adolescents, and cigarette use tends to increase with age (Substance Abuse and Mental Health Services Administration, 2019, Johnston et al., 2019). The observed reduction in CPQ-A scores with an increase in maternal education is supported by literature showing lower socio-economic status is associated with hazardous substance use (Andrabi et al., 2017; Lee et al., 2018). The finding that increased CPQ-A and HONC scores are associated with greater early life stress is also consistent with literature demonstrating that various forms of early life stress such as physical and/or sexual abuse, negligence, parental divorce, and domestic violence tend to increase hazardous use of alcohol, cannabis, and tobacco (Afifi et al., 2020; Kirsch et al., 2020). Our result indicating that family cigarette use is associated with higher HONC scores is consistent with other studies reporting on the association between family smoking habits and increased adolescent smoking (Joung et al., 2016; Rozi et al., 2016).

This study had some limitations including (1) cross-sectional nature of the data used for model building, (2) participant recruitment from the southwestern US among justice-involved youth who use alcohol and cannabis regularly, and (3) a limited sample size after restricting to users of all three substances. To partially address (1), we only used risk factors that would remain relatively stable over time. For (2), we note that regular alcohol and/or cannabis use was defined as at least once per month for the past 6 months, so the study inclusion criteria is rather representative of many adolescents. To mitigate (3), we used leave-one-out cross-validation to evaluate the predictive accuracy of our models, which guards against overfitting. In light of these limitations, we consider the proposed model to be preliminary.

In summary, this study is a novel attempt to build a risk prediction model for adolescent users of multiple substances. It serves as an important step towards our ultimate goal of building a comprehensive risk prediction model based on a large nationally representative longitudinal sample of adolescent users. Such a model can help us identify risk factors for hazardous substance use in adolescence and ideally flag factors that might help us understand the future transition into substance use disorders in adulthood.

5. Role of Funding Sources

This work was funded by the University of Texas at Dallas SPIRe seed grant and supported by the National Institute on Alcohol Abuse and Alcoholism (R01 AA017878-01A2; K24 AA026876-01 to SFE). The sponsors had no role in the study design, collection, analysis or interpretation of data, writing the manuscript and the decision to submit this manuscript for publication.

CRediT authorship contribution statement

Thanthirige Lakshika Maduwanthi Ruberu: Formal analysis, Investigation, Methodology, Writing – original draft. **Emily A. Kenyon:** Writing – review & editing. **Karen A. Hudson:** Data curation. **Francesca Filbey:** Conceptualization, Funding acquisition, Investigation, Writing – review & editing. **Sarah W. Feldstein Ewing:** Conceptualization, Investigation, Writing – review & editing. **Swati Biswas:** Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Writing – original draft. **Pankaj K. Choudhary:** Conceptualization, Funding acquisition, Investigation, Methodology, Supervision, Writing – original draft.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.pmedr.2021.101674.

References

Afifi, T.O., Taillieu, T., Salmon, S., Davila, I.G., Stewart-Tufescu, A., Fortier, J., Struck, S., Asmundson, G.J., Sareen, J., MacMillan, H.L., 2020. Adverse childhood experiences (ACEs), peer victimization, and substance use among adolescents. Child Abuse Neglect 106, 104504.

Andrabi, N., Khoddam, R., Leventhal, A.M., 2017. Socioeconomic disparities in adolescent substance use: Role of enjoyable alternative substance-free activities. Soc. Sci. Med. 176, 175–182.

Bonat, W., 2018. Multiple response variables regression models in R: The mcglm package. J. Stat. Softw. 84 (4), 1–30.

Bonat, W.H., Jørgensen, B., 2016. Multivariate covariance generalized linear models. J. Roy. Stat. Soc.: Ser. C (Appl. Stat.) 65 (5), 649–675.

CDC WONDER, 2020. Underlying Cause of Death 1999-2019 on CDC WONDER Online Database, released in 2020. Data are from the Multiple Cause of Death Files, 1999-2019, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Retrieved October 28, 2021 from http://wonder.cdc.gov/ucd-icd10.html.

Choi, H.J., Lu, Y., Schulte, M., Temple, J.R., 2018. Adolescent substance use: Latent class and transition analysis. Addict. Behav. 77, 160–165.

D'Amico, E.J., Rodriguez, A., Tucker, J.S., Dunbar, M.S., Pedersen, E.R., Shih, R.A., Davis, J.P., Seelam, R., 2020. Early and late adolescent factors that predict co-use of cannabis with alcohol and tobacco in young adulthood. Prev. Sci. 21 (4), 530–544.

DiFranza, J.R., Savageau, J.A., Fletcher, K., Ockene, J.K., Rigotti, N.A., McNeill, A.D., Coleman, M., Wood, C., 2002. Measuring the loss of autonomy over nicotine use in adolescents: the DANDY (development and assessment of nicotine dependence in youths) study. Arch. Pediatr. Adolesc. Med. 156 (4), 397–403.

Feldstein Ewing, S.W., Bryan, A.D., Dash, G.F., Lovejoy, T.I., Borsari, B., Schmiege, S.J., 2021. Randomized controlled trial of motivational interviewing for alcohol and cannabis use within a predominantly Hispanic adolescent sample. Experiment. Clin. Psychopharmacol. In press.

Friedman, J., Hastie, T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. J. Stat. Softw. 33 (1), 1–22.

Grucza, R.A., Bierut, L.J., 2006. Cigarette smoking and the risk for alcohol use disorders among adolescent drinkers. Alcohol. Clin. Exp. Res. 30 (12), 2046–2054.

Hastie, T., Qian, J., Tay, K., 2021. An introduction to glmnet. Retrieved January 10 from https://cran.r-project.org/web/packages/glmnet/vignettes/glmnet.pdf.

Hawke, L.D., Wilkins, L., Henderson, J., 2020. Early cannabis initiation: Substance use and mental health profiles of service-seeking youth. J. Adolesc. 83, 112–121.

Hayatbakhsh, M.R., Najman, J.M., Bor, W., O'Callaghan, M.J., Williams, G.M., 2009. Multiple risk factor model predicting cannabis use and use disorders: A longitudinal study. Am. J. Drug Alcohol Abuse 35 (6), 399–407.

Hindocha, C., Shaban, N.D., Freeman, T.P., Das, R.K., Gale, G., Schafer, G., Falconer, C.J., Morgan, C.J., Curran, H.V., 2015. Associations between cigarette smoking and cannabis dependence: A longitudinal study of young cannabis users in the United Kingdom. Drug Alcohol Depend. 148, 165–171.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer.

Jing, Y., Hu, Z., Fan, P., Xue, Y., Wang, L., Tarter, R.E., Kirisci, L., Wang, J., Vanyukov, M., Xie, X.-Q., 2020. Analysis of substance use and its outcomes by machine learning I. Childhood evaluation of liability to substance use disorder. Drug Alcohol Depend. 206, 107605.

Johnston, L.D., Miech, R.A., O'Malley, P.M., Bachman, J.G., Schulenberg, J.E., Patrick, M.E., 2019. Monitoring the future national survey results on drug use, 1975–2018: Overview, key findings on adolescent drug use. Institute for Social Research, University of Michigan, Ann Arbor.

Joung, M.J., Han, M.A., Park, J., Ryu, S.Y., 2016. Association between family and friend smoking status and adolescent smoking behavior and e-cigarette use in Korea. Int. J. Environ. Res. Public Health 13 (12), 1183.

Kim, M.J., Mason, W.A., Herrenkohl, T.I., Catalano, R.F., Toumbourou, J.W., Hemphill, S.A., 2017. Influence of early onset of alcohol use on the development of adolescent alcohol problems: A longitudinal binational study. Prev. Sci. 18 (1), 1–11.

Kirsch, D., Nemeroff, C.M., Lippard, E.T., 2020. Early life stress and substance use disorders: Underlying neurobiology and pathways to adverse outcomes. Advers. Resilience Sci. 1, 29–47.

Lee, J.O., Cho, J., Yoon, Y., Bello, M.S., Khoddam, R., Leventhal, A.M., 2018. Developmental pathways from parental socioeconomic status to adolescent substance use: Alternative and complementary reinforcement. J. Youth Adolesc. 47 (2), 334–348.

Martin, G., Copeland, J., Gilmour, S., Gates, P., Swift, W., 2006. The adolescent cannabis problems questionnaire (CPQ-A): Psychometric properties. Addict. Behav. 31 (12), 2238–2248.

McCabe, S.E., West, B.T., McCabe, V.V., 2018. Associations between early onset of e-cigarette use and cigarette smoking and other substance use among us adolescents: A national study. Nicotine Tob. Res. 20 (8), 923–930.

Meier, M.H., Hall, W., Caspi, A., Belsky, D.W., Cerdá, M., Harrington, H., Houts, R., Poulton, R., Moffitt, T.E., 2016. Which adolescents develop persistent substance dependence in adulthood? Using population-representative longitudinal data to inform universal risk assessment. Psychol. Med. 46 (4), 877–889.

Moss, H.B., Chen, C.M., Yi, H.-Y., 2014. Early adolescent patterns of alcohol, cigarettes, and marijuana polysubstance use and young adult substance use outcomes in a nationally representative sample. Drug Alcohol Depend. 136, 51–62.

NIDA, 2020. Principles of adolescent substance use disorder treatment: A research-based guide. Retrieved January 10 from. https://www.drugabuse.gov/publications/principles-adolescent-substance-use-disorder-treatment-research-based-guide/principles-adolescent-substance-use-disorder-treatment.

R Core Team, 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, https://www.R-project.org/.

Rajapaksha, R.M.D.S., Hammonds, R., Filbey, F., Choudhary, P.K., Biswas, S., 2020. A preliminary risk prediction model for cannabis use disorder. Prev. Med. Rep. 20, 101228.

Rozi, S., Mahmud, S., Lancaster, G., Zahid, N., 2016. Peer pressure and family smoking habits influence smoking uptake in teenage boys attending school: Multilevel modeling of survey data. Open J. Epidemiol. 6 (3), 167–172.

Rusby, J.C., Light, J.M., Crowley, R., Westling, E., 2018. Influence of parent–youth relationship, parental monitoring, and parent substance use on adolescent substance use onset. J. Fam. Psychol. 32 (3), 310–320.

Saitz, R., Miller, S.C., Fiellin, D.A., Rosenthal, R.N., 2021. Recommended use of terminology in addiction medicine. J. Addict. Med. 15 (1), 3–7.

Silvers, J.A., Squeglia, L.M., Rømer Thomsen, K., Hudson, K.A., Feldstein Ewing, S.W., 2019. Hunting for what works: Adolescents in addiction treatment. Alcohol. Clin. Exp. Res. 43 (4), 578–592.

Substance Abuse and Mental Health Services Administration, 2019. Key substance use and mental health indicators in the United States: Results from the 2018 national survey on drug use and health. Retrieved January 20, from. https://www.samhsa.gov/data/.

Tillson, M., Staton, M., Strickland, J.C., Pangburn, K., 2019. An examination of the age of substance use onset and adult severity of use among offenders entering treatment. J. Drug Issues 49 (2), 238–252.

Van Ryzin, M.J., Fosco, G.M., Dishion, T.J., 2012. Family and peer predictors of substance use from early adolescence to early adulthood: An 11-year prospective analysis. Addict. Behav. 37 (12), 1314–1324.

White, H.R., Kilmer, J.R., Fossos-Wong, N., Hayes, K., Sokolovsky, A.W., Jackson, K.M., 2019. Simultaneous alcohol and marijuana use among college students: Patterns, correlates, norms, and consequences. Alcohol. Clin. Exp. Res. 43 (7), 1545–1555.

White, H.R., Labouvie, E.W., 1989. Towards the assessment of adolescent problem drinking. J. Stud. Alcohol 50 (1), 30–37.

Whitesell, M., Bachand, A., Peel, J., Brown, M., 2013. Familial, social, and individual factors contributing to risk for adolescent substance use. J. Addict. 2013, 579310.

Yule, A.M., Wilens, T.E., Martelon, M., Rosenthal, L., Biederman, J., 2018. Does exposure to parental substance use disorders increase offspring risk for a substance use disorder? A longitudinal follow-up study into young adulthood. Drug Alcohol Depend. 186, 154–158.