

Risk and Protective Factors of Current Opioid Use Among Youth Living on or Near American Indian Reservations: An Application of Machine Learning

Mark A. Prince^{1, 2}, Bradley T. Conner^{1, 2}, Samuel R. Davis¹, Randall C. Swaim^{1, 2},
and Linda R. Stanley²

¹Department of Psychology, Colorado State University

²Tri-Ethnic Center for Prevention Research, Colorado State University

Opioid use among youth, particularly among American Indian (AI) youth, is rising, resulting in a large number of accidental overdoses and deaths. In order to develop effective prevention strategies, we need to use exploratory data analysis to identify previously unknown predictors of opioid use among youth living on or near reservations. The present study is an application of Machine Learning, a type of exploratory data analysis, to the Our Youth, Our Future epidemiological survey ($N = 6482$) to determine salient risk and protective factors for past 30-day opioid use. The Machine Learning algorithm identified 11 salient risk and protective factors. Importantly, highest risk was conferred for those reporting recent cocaine use, having ever tried a narcotic other than heroin, and identifying as American Indian. Protective factors included never having tried opioids other than heroin, infrequent binge drinking, having fewer friends pressuring you to use illicit drugs, initiating alcohol use at a later age, and being older. This model explained 61% of the variance in the training sample and, on average, 24% of the variance in the bootstrapped samples. Taken together, this model identifies known predictors of 30-day opioid use, for example, recent substance use, as well as unknown predictors including being AI, Snapchat use, and peer encouragement for use. Notably, recent cocaine use was a more salient predictor of recent opioid use than lifetime opioid use.

What is the significance of this article for the general public?

This study identified risk and protective factors for opioid use among reservation-based youth. Moreover, the highest risk was conferred for those reporting recent cocaine use, having ever tried a narcotic other than heroin, and identifying as American Indian. By using Machine Learning we were able to identify predictors of opioid use including and beyond those that would be predicted from theory. These risk and protective factors can help inform prevention and intervention efforts targeting youth living on or near American Indian reservations.

Keywords: American Indian, Machine Learning, opioid use, reservation-based youth

Supplemental materials: <https://doi.org/10.1037/tps0000236.supp>

This article was published Online First February 11, 2021.
Mark A. Prince  <https://orcid.org/0000-0002-8848-0568>
Bradley T. Conner  <https://orcid.org/0000-0001-6069-4401>

Samuel R. Davis  <https://orcid.org/0000-0002-2352-6422>
Linda R. Stanley  <https://orcid.org/0000-0002-5306-7394>

This research was supported by the National Institute on Drug Abuse (National Institute on Drug Abuse) Grant R01-DA003371.

Correspondence concerning this article should be addressed to Mark A. Prince, Department of Psychology, Colorado State University, 1876 Campus Delivery, Fort Collins, CO 80523, United States. Email: mark.prince@colostate.edu

“The opioid epidemic is quietly ravaging Indian Country” (Tipps et al., 2018, p. 422). Data over the past 20 years indicate that rates of opioid misuse and opioid-related deaths have rapidly increased among American Indians (AI)/Alaska Natives (AN). For example, the Centers for Disease Control and Prevention (CDC) data indicate that 2018 age-adjusted opioid-related overdose deaths among AI/AN were nearly five times the comparable rate in 1999 (2.9 per 100,000 in 1999; 14.2 per 100,000 in 2018; Tipps et al. 2018; Wilson et al. 2020). Moreover, these numbers are likely undercounted by as much as 40% due to misidentification of race on AI/AN death certificates (Arias et al., 2016). Adjusting for misclassification reveals rates equal to or greater than those of non-Hispanic Whites, the population hardest hit by the opioid epidemic (Martins et al., 2015). With respect to opioid misuse, 2018 data from the National Survey on Drug Use and Health (NSDUH; Substance Abuse and Mental Health Services Administration [SAMHSA], 2019) indicate that 5.8% of AI/ANs age 12 and older reported pain reliever misuse, compared to 3.6% of the general population, while AI/AN opioid misuse rates were 5.8%, compared to 3.7% nationally.

Recent research shows that particularly high rates of opioid misuse are seen among AI adolescents who live on or near reservations. Annual misuse of OxyContin and heroin among AI reservation adolescents in 2018 was significantly greater than among the comparable national population (Stanley et al., in press). For example, among 10th graders, past year OxyContin use was 2.9 times higher than national 10th Grade rates, while 10th Grade heroin rates were 20.8 times higher than national rates.

Two recent studies provide some context to the current study (Barnett et al., 2019; Nalven et al., 2020). Barnett and colleagues (2019) examined risk factors for nonprescription opioid use (not including heroin) among U.S. adolescents using the 2017 Youth Risk Behavior Survey. While the number of potential risk factors was limited to other substance use, mood, amount of sleep, school, academic performance, and race/ethnicity, they found that identification as AI/AN was associated with more nonprescription opioid use. One limitation of this finding is the few number of AI/AN in the sample and the failure to distinguish between reservation and urban AI youth. In contrast, Nalven and colleagues (2020) limited their study to a population-based sample of reservation

AI youth. With predictors limited to peer, family, and school factors, they found that variables from each of these domains were associated with higher nonprescription opioid and heroin use.

Beyond the study by Nalven et al. (2020); there is only limited data on opioid use among AI/AN youth living on or near reservations and none that included a comprehensive set of potential risk and protective factors. As such, there is a need to explore a broad range of contextual factors that may portend risk for opioid use among AI youth that live on or near reservations. Further, AI/AN youth often attend schools that are attended by non-AI/AN youth and live in communities that are not exclusively AI/AN as well (Stanley & Swaim, 2018). In studies that collect data from school aged youth that live on or near reservations, non-AI/AN students are often excluded from analyses to provide clearer focus on AI/AN youth (e.g., Stanley & Swaim, 2018). However, this may be a missed opportunity to study all youth attending these schools. Indeed, while there are likely factors that are not controlled, (e.g., neighborhood factors), there are many contextual factors that are shared among youth that attend the same schools (poverty, reduced access to prevention and treatment services). This creates a natural experiment to begin to distill out potential contributions of ethnic/racial identity influences, for example, historical trauma, that may affect likelihood of engagement in high-risk behaviors, including opioid use.

High rates of adolescent opioid use on and near reservations underscore the need for culturally congruent and efficacious prevention interventions for opioid misuse. Such interventions must be based on sound evidence about the risk and protective factors underlying opioid use. However, little research has specifically examined factors that contribute to opioid use among reservation-based adolescents. Additionally, efforts to reduce opioid use over the past two decades have resulted in limited success as evidenced by relatively higher rates of use (Nalven et al., 2020; Stanley et al., in press). One potential explanation for this lack of improvement despite efforts is a failure to identify meaningful predictors that explain opioid use in this population that are beyond the scope of typically invoked theories (Hellemann et al., 2009). For example, factors associated with adolescent substance use in general are numerous and broad in scope (Stone et al., 2012). However, the list of risk and protective factors is both long and likely includes variables that fail to

explain significant amounts of variance. Additionally, an exhaustive list of protective and risk factors does not exist. This leads to a need to both organize and select factors that have already been identified and use methodology that will identify risk and protective factors that have yet to be studied.

Bronfenbrenner's ecological theory (1977) may be useful in selecting among known predictors as it provides a framework for the selection of potentially useful variables that cuts across individual, social, and community factors. This wide selection of variables is also supported in the review of risk and protective factors for adolescent substance use presented by Hawkins and colleagues (1992). Machine Learning as a data analytic tool may be useful in identifying new risk and protective factors (Hellemann et al., 2009). Machine Learning is essentially an empirical atheoretical approach to data analysis that relies on feature selection, or salience, in identifying salient predictors (Adkins, 2017; Bleidorn & Hopwood, 2019; Dipnall et al., 2016; Witten & Frank, 2002). Taking an atheoretical approach using exploratory data analysis techniques, such as Machine Learning, allows for the identification of previously unknown salient predictors (Lemon et al., 2003). With replication, these newly identified predictors can be incorporated into existing models to increase explanatory power and improve the effectiveness of prevention and intervention strategies.

Data from the Our Youth, Our Future (OYOF) study, a large ongoing epidemiological study of adolescents living on or near reservations (Swaim & Stanley, 2018b), affords an opportunity to use Machine Learning to examine a large group of such measures in order to identify the most predictive risk and protective factors for reservation-based youth. One of the advantages of Machine Learning is the capitalization of the immense amount of data collected in large epidemiological studies such as OYOF (Dipnall et al., 2016). In addition, these data offer a unique opportunity to compare whether risk and protective factors for opioid use operate similarly between AI youth and non-AI youth attending the same schools. Because reservation AI youth sometimes attend off-reservation schools, due to no grade-appropriate on-reservation school, the OYOF study samples schools on or within 25 miles of a reservation with at least 20% AI enrollment. Moreover, some on-reservation schools have significant numbers of non-AI students due to large non-AI populations residing on reservations. Because the survey

is given to all students attending a school, a substantial number of non-AI students are surveyed as well. This comparison is important for several reasons. Prevention programming and substance use policy are often school- or community-based. For these to be effective, commonalities and differences between AI and non-AI students must be ascertained prior to implementation. In addition, it is often asserted that there are important differences in how risk and protective factors operate for reservation AI youth versus other youth, in part due to the unique environment of the reservation and due to cultural and ethnicity factors (Szlemko et al., 2006; Zeledon et al., 2020). Using a large number of predictors in a Machine Learning study with both AI and non-AI youth can add evidence toward answering whether this is true. Unlike studies that compare reservation AI adolescents to national adolescent samples, our sample of AI and non-AI youth have some similarities in terms of shared schools and, to a lesser degree, living environment. Therefore, it provides a unique point of comparison unlike that typically seen in the literature where reservation AI adolescents are compared to national adolescent samples.

This study sought to use Machine Learning to identify previously unknown salient predictors of past 30-day opioid use from a large population-based sample of youth (both AI and non-AI) living on or near reservations. Given that Machine Learning is an atheoretical approach to data analysis (Witten & Frank, 2002), no directional a priori directional hypotheses were tested. Instead, the primary hypothesis for this study was that Machine Learning would identify both previously known and unknown salient predictors of opioid use among reservation-based youth. While a pressing public health concern, opioid use is a rare event that raises methodological challenges. Machine Learning methods are well suited for making predictions about factors that portend risk or that can be protective for/against rare events like opioid use among youth (cf. Weiss & Hirsh, 2000).

Method

Sample and Recruitment

The present study used survey data from 37 schools that participated in the Our Youth, Our

Future (OYOF) survey during the spring semester of 2015 and the 2016–2017 academic year. Schools were stratified into seven regions (Northeast, Northwest, Southeast, Southwest, Northern Plains, Southern Plains, and Upper Great Lakes) based on cultural and other similarities among AI groups as described in Snipp (2005). OYOF is an ongoing epidemiologic investigation of substance use rates of adolescents living on or near reservations and related risk and protective factors. Its methods are described in detail in Swaim and Stanley (2018b). Randomly selected schools were contacted via mail, e-mail, and phone, and if schools agreed to participate, appropriate tribal and/or school board approval was obtained. Each school received a report of their survey findings and compensation for resources used to complete the survey process, with median payment of \$1,500.

Procedure

Several weeks prior to the survey, parents of enrolled middle and high school students (grades 7–12) received letters describing the survey and instructions for opting their child out, if desired. Students not opted out were administered the OYOF survey online via Qualtrics software and were instructed to skip questions they did not want to answer. Participants were anonymous, and all procedures were approved by the Colorado State University Institutional Review Board.

Participants

The sample consisted of 6,482 youth in grades 7–12. Fewer than 1% of the participants were opted out or refused to take the survey. Of the 6,482 participants, 48% identified as AI including those who identified as AI and another race ($n = 3,098$), while 52% identified as non-Native ($n = 3,384$). Among those who identified as non-Native: 60% identified as White, 11% identified as Black/African American, 33% identified as Hispanic, and 17% identified with another racial/ethnic identity (note, race and ethnicity categories within non-Native are not mutually exclusive). Approximately 50% of participants self-identified as male. The average age was 14.78 ($SD = 1.79$).

Measures

Drug use items on the survey mirror those on the Monitoring the Future survey, a nationally representative annual survey given to middle and high school students. The outcome variable was a dichotomized sum of endorsement of any opioid use in the past 30 days, coded as 1 for any use and 0 for no use and calculated from two items measuring (a) “How many times (if any) have you taken narcotics other than heroin on your own—that is, without a doctor telling you to take them—during the last 30 days?,” where other narcotics were defined as a list of opioids: methadone, opium, morphine, codeine, Demerol, Vicodin, OxyContin, and Percocet; and (b) “how many times (if any) have you taken heroin during the last 30 days.” Summed scores across these two items greater than 1 were reduced to 1, creating a binary 30-day opioid use outcome variable. Three percent ($n = 175$) of the sample endorsed any opioid use within 30 days of completing the survey, while 97% ($n = 6,185$) reported no opioid use within the past 30 days. A greater percentage of participants in the Southeast reported use within the past 30 days (6%, $n = 36$) compared to the other regions.

Other OYOF survey items measure risk and protective factors deriving from the social development model, a general theory of human behavior that incorporates predictors in the individual and in the family, school, community, and peer group environments into a coherent development model predicting antisocial and prosocial behaviors (see Cambron et al. [2019] for an overview of this model).

In particular, the survey contains cultural, religious, family, personality, and peer measures (cf. Swaim & Stanley, 2018b). Three hundred fourteen items were used as predictors of past 30-day opioid use. These predictors can be generally categorized as: demographics (e.g., age, gender, ethnicity), substance use (e.g., alcohol, cannabis, cocaine), consequences of substance use, cultural and religious identities, family structure and values, personality, peer norms and influence regarding substance use, and attitudes/beliefs about substance use. Predictors were single items rather than aggregates of multiitem measures. Example predictors are “When (if ever) did you first do narcotics other than heroin?,” “how often do you use Snapchat,” and “how often have your friends asked you to use other illegal drugs?” A complete list of the predictors and the outcome along with

the response options are included as an [online supplemental table](#).

Analyses

All analyses were conducted in R (R Core Team, 2019). Data were cleaned and analyzed for missingness. After preliminary analyses of univariate descriptive statistics, data were randomly split into two samples - the training (in this case 80% of the data) and testing (20% of the data) samples. The research question was addressed using recursive partitioning, a Type of decision tree-based Machine Learning (Zhang & Singer, 2013) as implemented through the rpart package (Therneau et al., 2019) on the training sample and testing misclassification in the testing sample.

Recursive partitioning is a nonparametric type of Machine Learning that builds a “tree” that identifies homogeneous orthogonal subgroups of a population, called “leaves” or “nodes,” whose members share common characteristics that predict the outcome variable (Lemon et al., 2003); in this study, presence or absence of past 30 day opioid use. Model construction is binary and recursive in that, at each step of the model building process, every predictor variable is used to predict the outcome variable and split the sample into a more homogeneous subsample, even if that predictor variable has already been used in model construction (Zhang & Singer, 2013). The advantage of recursive partitioning is that a large number of predictors can be evaluated *a theoretically* to determine salient predictors of the outcome variable. This process can be used to build a robust set of predictor variables and to determine intricate relations among predictor variables in predicting the outcome of interest. The first variable in the model is considered the most salient predictor of the outcome variable. All predictor variables retained in the model after the first step are predicting the outcome variable through the preceding steps; in other words, they are interacting with all predictor variables at all previous steps to predict the outcome variable. In decision tree analyses, chains of predictor variables are referred to as branches of the tree. Branches on the left represent the protective side of the model, in this study, no opioid use in the past 30 days, while branches on the right represent the risk side of the model, or any opioid use in the past 30 days. Model construction continues until each branch of the tree ends in a “node” that is perfectly homogeneous (i.e., all members of the sample have exactly the

same values on all variables; the most homogeneous sample is typically a sample of one person). Model fit is evaluated by examining the relative error of the final model. The lower the relative error, the better the model fits the data, while $1 - \text{the relative error}$ is an estimate of the variance explained by the model.

Perfect homogeneity means that the model is overfit to the particular sample of the population under study (Zhang & Singer, 2013). Thus, two processes are used to maximize the amount of variance explained by the model while also maximizing the generalizability of the model to other samples from the population of interest. The first, bagging (Strobl et al., 2009); is to cross-validate the model using a bootstrapping procedure where 80% of the sample is extracted n number of times and, using replacement, built up to the same n as the original sample. In the present study, 100 bootstrapped samples were extracted for cross-validation. The model is then reestimated in each of the bootstrapped samples. Each reestimated model has a relative error associated with it. All of these relative errors are then used to calculate a mean relative error and a standard error of that distribution of relative errors. The second process is to use a cost-complexity pruning rule to select the model that most parsimoniously explains both the most variance in the sample and, on average, the most variance from the cross-validation, based on the model relative error and the mean and standard error of the relative errors (for Hellemann et al., 2009). The goal is to select the model that uses the fewest splits to explain the most variance across all samples. This pruning process uses parsimony to reduce model overfit, which typically results in less variance explained in the original sample, but more variance explained across all of the bootstrapped samples. A good-fitting tree is one with the fewest splits that has smaller cross-validation error than a model with no splits.

Model results are also evaluated with a confusion matrix and by computing accuracy, error rate, specificity, and sensitivity of the model. The confusion matrix assesses the accuracy of the classification algorithm learned in the training sample on the data in the testing sample (Clarke et al., 2009). Data are presented in a 2×2 table so that predicted class, learned in the training data, is compared to actual class in the testing data. Cells in the table represent the true positive (predicted positive, actually positive), true negative (predicted negative, actually negative), false

positive (predicted positive, actually negative), and false negative (predicted negative, actually positive). Accuracy, error rate, specificity and sensitivity are calculated from these values. Values for accuracy, specificity, and sensitivity are better the closer they get to 1, while error rate is better the closer is gets to 0.

Model construction used the “class” method in part as the outcome variable was binary (Therneau et al., 2019). Splitting of the sample continued as long as there were five observations remaining in the previous node and splitting stopped when there were at least 2 observations remaining in any node. Past 30-day opioid use was rare in this sample; 175 out of 6,482 (2.7%) participants reported any opioid use in the past 30 days. The cost-complexity parsimony rule that governs model selection stipulates that the model with the fewest number of splits that explains the most variance in cross-validation should be selected to maximize generalizability while minimizing overfit to the sample. However, research indicates that this pruning method often overprunes the model (Esposito et al., 1997). Thus, with rare outcomes, as is the case herein, pruning the model one step past parsimony maximizes the variance explained while still minimizing overfitting, assuming the more complex model satisfies all other model fit criteria (e.g., explains more variance in the boot-strapped samples on average compared to a model with no splits).

Results

Results from recursive partitioning identified a good-fitting model to the training sample data that was stable in the cross-validation procedure and replicated in the testing data with adequate classification statistics. Model fit statistics are presented in Table 1, the confusion matrix is presented in Table 2, and classification indicators are presented in Table 3. The model, before pruning, had 5 steps, each with an increasing number of splits, up to 25 splits. Evaluation of the fit statistics from the model construction and cross-validation procedures indicated that a model with 3 splits was the most parsimonious (i.e., Step 2). However, given the rareness of the outcome variable, the model with 12 splits (i.e., Step 4) maximized explanatory power of the model while still minimizing overfit to the sample. Thus, the Step 4 model was retained as the final model.

Table 1
Recursive Partitioning Tree Fit Statistics

Step	CP	# of splits	Rel error	X error	X STD
1	0.201987	0	1.00000	1.00000	0.080160
2	0.039735	2	0.59603	0.81457	0.072552
3	0.016556	3	0.55629	0.60927	0.062943
4	0.015000	12	0.39073	0.75497	0.069911
5	0.01000	21	0.26490	0.76159	0.070210

Note. CP = model complexity parameter; Rel Error = relative error; X Error = mean of the relative error of all cross validations; X STD = standard error of the distribution of cross validation relative errors. Bolded text denotes the step where pruning occurred.

The final pruned model is presented in Figure 1. This model explained 61% of the variance in the training sample and, on average, 24% (18%–31%) of the variance in the bootstrapped samples. The most salient predictor of any 30-day opioid use was the number of times an individual used cocaine in the past 30 days. Lowest risk of any 30-day opioid use was predicted by having used cocaine less than twice in the past 30 days and never having tried opioids other than heroin. Conversely, highest risk of any past 30-day opioid use was predicted by two or more instances of cocaine use in the past 30 days, ever having tried opioids other than heroin, and self-reporting as an American Indian. Intermediate risk for any opioid use in the past 30 days was predicted by other substance use (i.e., binge drinking, amphetamine, cannabis, and huffing glue), peer pressure to use illicit substances, age and social media use as depicted in Figure 1.

Discussion

Opioid use is a major public health concern among youth, particularly AI youth (Stanley et al., in press), yet little is known about the risk and protective factors affecting its use. The present study used Machine Learning to identify risk and protective factors of past 30-day opioid use among a large representative sample of youth living on or near AI reservations. Results demonstrated that Machine Learning was able to predict current opioid use accurately, and thus it can be an effective tool for identifying salient predictors of such use. One key finding from this study is that youth who report recent use of substances other than

Table 2*Confusion Matrix of the Final Tree*

Prediction	Testing data	
	No opioid use	Opioid use
No opioid use	1,252 (TN - A)	13 (FN - B)
Opioid use	7 (FP - C)	11 (TP - D)

Note. The confusion matrix was run on the Testing Data, which is comprised of 20% of the available data.

opioids are more likely to report past 30-day opioid use. Thus, efforts focused on preventing and treating opioid use should focus on substance use more broadly. In particular, the most salient predictor of recent opioid use was 30-day use of cocaine. This finding is consistent with observations of growing concurrent use of stimulants and opioids and of users intentionally switching back and forth between stimulants and opioids (Enos, 2019).

In addition to ever using opioids other than heroin, a critical risk factor for past 30-day opioid use was self-identifying as AI. Identifying as AI appeared as one of the most important predictors of past 30-day opioid use among the over 300 predictors tested (i.e., it was the third most salient predictor on the high-risk side of the model). This implies that, among youth living on or near reservations, AI youth are more likely than their non-AI peers to use opioids. Applying Bronfenbrenner's theoretical framework for identifying predictors, our modeling included measuring the most common risk and protective factors among adolescents for substance use, items that measure key constructs within the family, peer, school, and individual domains. These items have been found to be predictive of substance misuse in all populations of adolescents, including AIs. In addition, measures of cultural identity, ethnic pride, and perceived discrimination are included in the survey and thus in the Machine Learning model. However, even with many culturally relevant items included in the set of possible predictors, identifying as Native was a salient predictor of last 30-day opioid use.

There are several reasons why this might be the case. First, there are variables that differentiate the AI and non-AI groups and that predict substance use but that are not measured in the survey. These include socioeconomic variables, such as poverty (Henry et al., 2011); and variables specific to AI populations, for example, historical trauma (Skewes & Blume, 2019). Historical trauma results

from the Indigenous history of colonization, genocide, forced removal from lands, and attempts to deny and abolish cultural traditions, including language and healing practices. Historical trauma is transmitted intergenerationally through diminished family functioning (Cromer et al., 2018); and is hypothesized to impair parenting practices. Moreover, Indigenous children are often victims of or witnesses to high rates of trauma and loss, often associated with suicide, accidents, violence, and substance abuse (Brookie et al., 2013; Sarche & Spicer, 2008); and this trauma has been linked to substance use. This history is both an important contributor to a host of Indigenous health inequities (McOliver et al., 2015). In the present study, this history may be the reason identifying as Native, an Indigenous identity, was a salient predictor in the model. This is consistent with Davis et al. (2020) who found that AI adolescents were more likely to endorse coping motives for marijuana use than their non-AI counterparts attending the same schools. Motives for opioid use was not measured in that survey, but it could help explain why being AI is a risk factor for opioid use.

For those reporting recent cocaine use and never having tried opioids other than heroin, the next most salient predictor of past 30-day opioid use was past 30-day inhalant use, another serious substance misuse behavior (Manikandan & Balasubramanian, 2020). This finding is similar to that found for a national sample of adolescents age 12–17 years showing that persistent inhalant use was a strong predictor for nonmedical opioid use (Nakawaki & Crano, 2012).

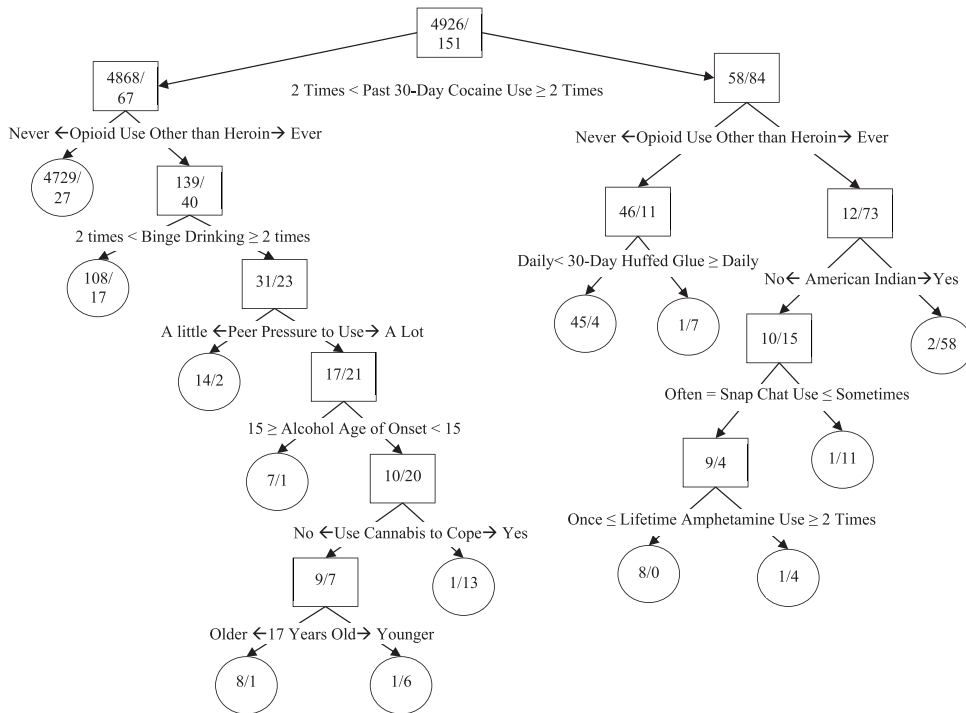
Staying on the high-risk side of the model, the Machine Learning algorithm identified Snapchat use as a protective factor, with those using Snapchat often being less likely to endorse recent opioid use. Snapchat use was a salient predictor for individuals not identifying as AI. One interpretation of this finding is that Snapchat use is a proxy for social engagement. Those non-AI youth who are not using

Table 3*Classification Indicators*

	Testing data
Error rate	0.02
Accuracy 95% CI	0.98 (0.98, 0.99)
Sensitivity	0.46
Specificity	0.99

Note. Error Rate closer to .00 is best, all others closer to 1.00 is best.

Figure 1
Final Regression Tree



Note: Squares represent intermediate nodes and circles represent final nodes. Numbers in the squares and circles denote the number of participants in the node, those on the left of the diagonal report no opioid use in the past 30 days and those on the right denote those reporting 30-day opioid use.

Snapchat may be more socially isolated which may be a risk factor for engagement in substance use. Some research has shown that social media use is linked to benefits including social capital (e.g., Ellison et al., 2007); especially among marginalized youth (Jarrett, 2003; Stevens et al., 2017). Future research is needed to better understand risky versus protective engagement in social media with respect to substance use. The final variable on the high-risk side of the model differentiating those that reported more Snapchat use was a lifetime history of amphetamine use, with those reporting use more likely to report past 30-day opioid use.

In contrast to risk factors, the Machine Learning algorithm also identified a host of protective factors. Among the protective factors were less engagement with substance use broadly with the exception of using marijuana to cope. It is possible that using marijuana to cope is protective because youth who use marijuana to cope are not choosing to use opioids. However, using

marijuana to cope may be a harbinger of problems that are yet to develop. Longitudinal research can help determine if using marijuana to cope has lasting protective benefits against engaging in opioid use.

Beyond personal substance use, those who reported less peer pressure to use illegal drugs were less likely to report current opioid use. This finding, along with the social media engagement described above, point to the importance of social factors in youth's decisions to engage in opioid use. Prevention and intervention efforts should include both individual and social strategies to help defend against the risks associated with opioid use.

One surprising finding was that geographic region did not emerge as either a risk or protective factor in the classification tree. Examining the prevalence of opioid use across regions indicates that there are regions of higher risk (e.g., Southeast, 6% reporting opioid use) and regions

with less risk (e.g., Southern Plains, 0% reporting opioid use); however, region did not emerge as a salient predictor among those studied. This indicates that other risk factors, that also vary by region, explain these differences in use.

Overall, the model identified predictors that are known. For instance, substance use of any kind was predictive of past 30-day opioid use. However, lifetime use of opioids other than heroin was not the most salient predictor of the outcome. Instead, more recent cocaine use was a more salient predictor. Beyond substance use, the model identified nuanced predictors of risk. For instance, while we know in general that AI adolescents are at higher risk for substance use (e.g., Swaim & Stanley, 2018a), the algorithm identified Snapchat use as a salient predictor. More research is needed to understand this relation. Finally, the algorithm identified being younger as portending more risk for past 30-day adolescent use among individuals reporting increased alcohol misuse, more peer influences to use drugs, and initiating alcohol use at a younger age. This finding is surprising and important in differentiating who might be at high risk for opioid use from a young age.

Strengths and Limitations

This study had both strengths and limitations. Major strengths included the large sample of both AI and non-AI youth living on or near reservations, an understudied, vulnerable population, and the application of Machine Learning to a major public health concern in order to effectively identify a list of salient risk and protective factors for current opioid use among youth. Additionally, an ensemble approach to the recursive partitioning was used by training the algorithm in one data set and testing it in another and by employing the cross-validation and cost-complexity pruning methods. This allowed for the assessment of the model by a number of indices and minimized overfit while maximizing generalizability. Limitations included cross-sectional data that precludes causal inferences and some degree of overfitting due to the rarity of opioid use among youth. Given the rarity of past 30-day opioid use in absolute terms, it would be helpful to replicate the current findings in a new sample or to test the predictive validity of the final model in a new dataset. We did not ask whether students were using substances simultaneously, a parti-

cularly dangerous form of use (Ochoa et al., 2001); but further research should examine whether reservation students are using heroin and cocaine together, the so-called speedball, that can be quite lethal (Leri et al., 2003).

Implications

There are a number of implications of the current study. First, findings from these exploratory analyses identified risk and prevention strategies that can be used to inform prevention and intervention efforts, including factors (e.g., social media engagement) that have not been part of the national discussion around the opioid epidemic. These findings can be shared with regional and tribal treatment and healing centers to be used as screening tools or to help providers have a richer understanding of factors that may portend risk for opioid use or that may be protective. Some of these findings may be surprising (e.g., the potentially helpful use of marijuana to cope) and may provide some context for providers in elucidating risky substance use (e.g., cocaine use, binge drinking) and less risky substance use. Second, this study demonstrates that Machine Learning, and more specifically, classification tree analysis with recursive partitioning, is an effective tool for predicting rare events.

References

- Adkins, D. E. (2017). Machine learning and electronic health records: a paradigm shift. *American Journal of Psychiatry*, 174(2), 93–94.
- Arias, E., Heron, M., & Hakes, J. (2016). The validity of race and Hispanic-origin reporting on death certificates in the United States: An update. *Vital and Health Statistics*, 2(172), 1–21.
- Barnett, T. E., Thompson, E. L., Litt, D. M., & Lewis, M. A. (2019). Correlates of nonmedical prescription opioid use among U.S. adolescents. *American Journal of Preventive Medicine*, 57(5), e175–e179. <https://doi.org/10.1016/j.amepre.2019.05.006>
- Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, 23(2), 190–203. <https://doi.org/10.1177/1088868318772990>
- Brockie, T. N., Heinzelmann, M., & Gill, J. (2013). A framework to examine the role of epigenetics in health disparities among Native Americans. *Health Disparities in Genomics and Genetics*, 2013, 410395–410399. <https://doi.org/10.1155/2013/410395>

- Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, 32(7), 513–531. <https://doi.org/10.1037/0003-066X.32.7.513>
- Cambron, C., Catalano, R. F., & Hawkins, J. D. (2019). The social development model in The Oxford Handbook of Developmental and Life-Course Criminology. In D. P. Farrington, L. Kazemian, & A. R. Piquero (Eds.), *The Oxford handbook of development and life-course criminology* (pp. 224–247). Oxford University Press.
- Clarke, B., Fokoue, E., & Zhang, H. H. (2009). *Principles and theory for data mining and machine learning*. Springer Science & Business Media.
- Cromer, L., Gray, M., Vasquez, L., & Freyd, J. (2018). The relationship of acculturation to historical loss awareness, institutional betrayal, and the intergenerational transmission of trauma in the American Indian experience. *Journal of Cross-Cultural Psychology*, 49(1), 99–114. <https://doi.org/10.1177/0022022117738749>
- Davis, S. R., Prince, M. P., Swaim, R. C., & Stanley, L. R. (2020). Comparing cannabis use motive item performance between American Indian and White youth. *Drug and Alcohol Dependence*, 213, 108086. <https://doi.org/10.1016/j.drugalcdep.2020.108086>
- Dipnall, J. F., Pasco, J. A., Berk, M., Williams, L. J., Dodd, S., Jacka, F. N., & Meyer, D. (2016). Fusing data mining, machine learning and traditional statistics to detect biomarkers associated with depression. *PLoS ONE*, 11(2), Article e0148195.
- Ellison, N. B., Steinfield, C., & Lampe, C. (2007). The benefits of Facebook “friends”: Social capital and college students’ use of online social network sites. *Journal of Computer-Mediated Communication*, 12(4), 1143–1168. <https://doi.org/10.1111/j.1083-6101.2007.00367.x>
- Enos, G. A. (2019, October 29). Concurrent use of stimulants, opioids complicates policy initiatives. *Addiction Professional*. <https://www.psychcongress.com/article/stim-spotlight-concurrent-use-stimulants-opioids-defies-past-explanations>
- Esposito, F., Malerba, D., Semeraro, G., & Kay, J. (1997). A comparative analysis of methods for pruning decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5), 476–491. <https://doi.org/10.1109/34.589207>
- Hawkins, J. D., Catalano, R. F., & Miller, J. Y. (1992). Risk and protective factors for alcohol and other drug problems in adolescence and early adulthood; Implications for substance abuse prevention. *Psychological Bulletin*, 112(1), 64–105. <https://doi.org/10.1037/0033-2909.112.1.64>
- Helleman, G., Conner, B. T., Anglin, M. D., & Longshore, D. (2009). Seeing the trees despite the forest: Applying recursive partitioning to the evaluation of drug treatment retention. *Journal of Substance Abuse Treatment*, 36(1), 59–64. <https://doi.org/10.1016/j.jsat.2008.03.005>
- Henry, K. L., McDonald, J. N., Oetting, E. R., Silk Walker, P., Walker, R. D., & Beauvais, F. (2011). Age of onset of first alcohol intoxication and subsequent alcohol use among urban American Indian adolescents. *Psychology of Addictive Behaviors*, 25(1), 48–56. <https://doi.org/10.1037/a0021710>
- Jarrett, R. L. (2003). Worlds of development: The experiences of low-income, African-American youth. *Journal of Children & Poverty*, 9(2), 157–188. <https://doi.org/10.1080/10796120305437>
- Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D., & Rakowski, W. (2003). Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *Annals of Behavioral Medicine*, 26(3), 172–181. https://doi.org/10.1207/S15324796ABM2603_02
- Leri, F., Bruneau, J., & Stewart, J. (2003). Understanding polydrug use: Review of heroin and cocaine co-use. *Addiction*, 98(1), 7–22. <https://doi.org/10.1046/j.1360-0443.2003.00236.x>
- Manikandan, K., & Balasubramanian, S. (2020). An uprising health concern of inhalant abuse among adolescents: A mini review. *Journal of Pharmaceutical Sciences and Research*, 12(1), 151–155.
- Martins, S. S., Santaella-Tenorio, J., Marshall, B. D., Maldonado, A., & Cerdá, M. (2015). Racial/ethnic differences in trends in heroin use and heroin-related risk behaviors among nonmedical prescription opioid users. *Drug and Alcohol Dependence*, 151, 278–283. <https://doi.org/10.1016/j.drugalcdep.2015.03.020>
- McOliver, C. A., Camper, A. K., Doyle, J. T., Eggers, M. J., Ford, T. E., Lila, M. A., Berner, J., Campbell, L., & Donatuto, J. (2015). Community-based research as a mechanism to reduce environmental health disparities in American Indian and Alaska Native communities. *International Journal of Environmental Research and Public Health*, 12(4), 4076–4100. <https://doi.org/10.3390/ijerph120404076>
- Nakawaki, B., & Crano, W. D. (2012). Predicting adolescents’ persistence, non-persistence, and recent onset of nonmedical use of opioids and stimulants. *Addictive Behaviors*, 37(6), 716–721. <https://doi.org/10.1016/j.addbeh.2012.02.011>
- Nalven, T., Spillane, N. S., & Schick, M. R. (2020). Risk and protective factors for opioid misuse in American Indian adolescents. *Drug and Alcohol Dependence*, 206, 107736. <https://doi.org/10.1016/j.drugalcdep.2019.107736>
- Ochoa, K. C., Hahn, J. A., Seal, K. H., & Moss, A. R. (2001). Overdosing among young injection drug users in San Francisco. *Addictive Behaviors*, 26(3), 453–460. [https://doi.org/10.1016/S0306-4603\(00\)00115-5](https://doi.org/10.1016/S0306-4603(00)00115-5)
- R Core Team. (2019). R: A language and environment for statistical computing (Version 3.5.3) [Computer software]. R Foundation for Statistical Computing. <http://www.R-project.org/>

- Sarche, M., & Spicer, P. (2008). Poverty and health disparities for American Indian and Alaska Native children: Current knowledge and future prospects. *Annals of the New York Academy of Sciences*, 1136, 126–136. <https://doi.org/10.1196/annals.1425.017>
- Skewes, M. C., & Blume, A. W. (2019). Understanding the link between racial trauma and substance use among American Indians. *American Psychologist*, 74(1), 88–100. <https://doi.org/10.1037/amp0000331>
- Snipp, C. M. (2005). *American Indian and Alaska native children: Results from the 2000 Census*. Population Reference Bureau. <https://www.prb.org/pdf05/AmericanIndianAlaskaChildren.pdf>
- Stanley, L. R., Crabtree, M. A., & Swaim, R. C. (in press). Opioid misuse among American Indian adolescents. *American Journal of Public Health*.
- Stanley, L. R., & Swaim, R. C. (2018). Latent classes of substance use among American Indian and white youth living on or near reservations. *Public Health Reports*, 133(4), 432–441. <https://doi.org/10.1177/0033354918772053>
- Stevens, R., Gilliard-Matthews, S., Dunaev, J., Woods, M. K., & Brawner, B. M. (2017). The digital hood: Social media use among youth in disadvantaged neighborhoods. *New Media & Society*, 19(6), 950–967. <https://doi.org/10.1177/1461444815625941>
- Stone, A. L., Becker, L. G., Huber, A. M., & Catalano, R. F. (2012). Review of risk and protective factors of substance use and problem use in emerging adulthood. *Addictive Behaviors*, 37(7), 747–775. <https://doi.org/10.1016/j.addbeh.2012.02.014>
- Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, 14(4), 323–348. <https://doi.org/10.1037/a0016973>
- 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 (HHS Publication No. PEP19-5068, NSDUH Series H-54). Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration <https://www.samhsa.gov/data/>
- Swaim, R. C., & Stanley, L. R. (2018a). Effects of family conflict and anger on alcohol use among American Indian students: Mediating effects of outcome expectancies. *Journal of Studies on Alcohol and Drugs*, 79(1), 102–110. <https://doi.org/10.15288/jsad.2018.79.102>
- Swaim, R. C., & Stanley, L. R. (2018b). Substance use among American Indian youths on reservations compared with a national sample of U.S. adolescents. *JAMA Network Open*, 1(1), e180382–e180382. <https://doi.org/10.1001/jamanetworkopen.2018.0382>
- Szlemko, W. J., Wood, J. W., & Thurman, P. J. (2006). Native Americans and alcohol: Past, present, and future. *The Journal of General Psychology*, 133(4), 435–451. <https://doi.org/10.3200/GENP.133.4.435-451>
- Therneau, T., Atkinson, B., & Ripley, B. (2019). Package “rpart” [Computer software]. <https://cran.r-project.org/web/packages/rpart/rpart.pdf>
- Tipps, R. T., Buzzard, G. T., & McDougall, J. A. (2018). The opioid epidemic in Indian country. *The Journal of Law, Medicine & Ethics*, 46(2), 422–436. <https://doi.org/10.1177/1073110518782950>
- Weiss, G. M., & Hirsh, H. (2000). July. Learning to predict extremely rare events. [Workshop] In *AAAI, Learning from imbalanced data sets* (pp. 64–68). AAAI Press.
- Wilson, N., Kariisa, M., Seth, P., Smith, I. V. H., & Davis, N. L. (2020). Drug and opioid-involved overdose deaths—United States, 2017–2018. *Morbidity and Mortality Weekly Report*, 69(11), 290–297. <https://doi.org/10.15585/mmwr.mm6911a4>
- Witten, I. H., & Frank, E. (2002). Data mining: Practical machine learning tools and techniques with Java implementations. *SIGMOD Record*, 31(1), 76–77. <https://doi.org/10.1145/507338.507355>
- Zeledon, I., West, A., Antony, V., Telles, V., Begay, C., Henderson, B., ... Soto, C. (2020). Statewide collaborative partnerships among American Indian and Alaska Native (AI/AN) communities in California to target the opioid epidemic: Preliminary results of the Tribal Medication Assisted Treatment (MAT) key informant needs assessment. *Journal of Substance Abuse Treatment*, 108, 9–19. <https://doi.org/10.1016/j.jsat.2019.04.003>
- Zhang, H., & Singer, B. H. (2013). *Recursive partitioning in the health sciences*. Springer Science & Business Media.

Received April 7, 2020

Revision received December 1, 2020

Accepted December 11, 2020 ■