Document Title:	Development and Validation of an Actuarial Risk Assessment Tool for Juveniles with a History of Sexual Offending
Author(s):	KiDeuk Kim, Grant Duwe, Emily Tiry, Ashlin Oglesby-Neal, Cathy Hu, Ryan Shields, Elizabeth Letourneau, Michael Caldwell
Document Number:	253444
Date Received:	September 2019
Award Number:	2013-AW-BX-0053

Summary Report

Development and Validation of an Actuarial Risk Assessment Tool for Juveniles with a History of Sexual Offending [1]

June 2019

KiDeuk Kim
Urban Institute
kkim@urban.org
202-261-5346

Emily Tiry
Urban Institute
etiry@urban.org
202-261-5630

Cathy Hu
Urban Institute
chu@urban.org
202-261-5978

Elizabeth Letourneau
Johns Hopkins University
elizabethletourneau@jhu.edu
410-955-9913

Grant Duwe
Minnesota Department of Corrections
grant.duwe@state.mn.us
(651) 361-7377

Ashlin Oglesby-Neal
Urban Institute
aoglesby@urban.org
202-261-5411

Ryan Shields
University of Massachusetts Lowell
Ryan_Shields@uml.edu
978-934-4335

Michael Caldwell
University of Wisconsin-Madison
mfcaldwell@wisc.edu
608-347-6764

[1] This project was supported by Award No. 2013-AW-BX-0053, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect those of the Department of Justice. Direct all correspondence to KiDeuk Kim (email: kkim@urban.org / mailing address: Urban Institute, 500 L'Enfant Plaza SW, Washington, District of Columbia, 20024).

Abstract

For juvenile justice practitioners working with youth who have sexually offended, an accurate risk assessment instrument can help guide important decisions about placement, supervision, and treatment. However, current knowledge and practice in assessing the risk of sexual recidivism for youth is limited, as there are few existing tools that are empirically valid and reliable. The current project thus examined current practice and policy in the assessment, treatment, and management of juveniles with a history of sexual offending across multiple jurisdictions (Florida, New York, Oregon, Pennsylvania, and Virginia) and developed a prototype assessment tool, state-specific risk assessment models, and practical guidance for building a risk assessment for sexual recidivism in juvenile justice settings.

Specifically, the project team collected case-level information on more than 8,000 juveniles who sexually offended between 2009 and 2013. The project team analyzed the database to develop numerous prediction models for sexual recidivism, defined as being re-arrested for a sex crime. The overall approach to model development focused on maximizing the utility of existing information to yield the most effective and stable prediction results. Given that extant research has persistently suffered from small sample sizes and inconsistencies across studies and jurisdictions, it was critical to develop and evaluate the models under various testing conditions and demonstrate reliable performance. Therefore, compiling extensive historical data on youth who have sexually offended was logical and necessary as a primary mode of data collection.

Similarly, instead of relying on one method of classification as typically pursued in prior research, the project team examined the performance of numerous classifiers, including traditional regression and machine learning (ML) algorithms, and tested the predictive validity of those models in multiple ways, including traditional hold-out validation, k-fold cross-validation, and bootstrapping. The project team also simulated the validation results from multiple models over 1,000 times. From hundreds of models developed and tested individually for each participating jurisdiction and jointly for all of them, in addition to interviews with practitioners in each jurisdiction, the project team distilled several lessons to inform current practice in risk assessment for sexual recidivism.

The key findings/conclusions from the project highlight that predicting sexual recidivism among youth entails numerous inherent challenges due to the low frequency of occurrence, chief among which is the lack of reliability in risk prediction. Our simulation analysis reveals that a risk prediction model that performs adequately in one setting often reversely classifies individuals in another setting (i.e., high risk to be identified as low risk and vice versa). Adopting an off-the-shelf assessment tool, either public or commercial, should be avoided without extensive customization to local settings and testing, which involves updating the list of predictors as well as their weights.

In the current project, each jurisdiction-specific model with its own set of predictors and weights has shown better performance in terms of predictive validity and reliability than to a multi-state model. The Area Under the Curve (AUC) for the jurisdiction-specific models range from high 0.70s to low 0.80s whereas the AUC for the multi-state model is in the 0.60s. Despite its moderate performance, the project team developed a prototype risk assessment tool out of the multi-state model because it has features more widely applicable than the jurisdiction-specific models. As an

example of how ML applications of risk assessment can be implemented in practice, the prototype provides a platform for improving current practice in sex offense risk assessment through the use of advanced technology and existing administrative data.

URBAN INSTITUTE

Final Summary Report

Development and Validation of an Actuarial Risk Assessment Tool for Juveniles with a History of Sexual Offending

For juvenile justice practitioners working with youth who have sexually offended, an accurate risk assessment instrument can help guide important decisions about placement, supervision, and treatment. However, current knowledge and practice in assessing the risk of sexual recidivism for youth is limited, as there are few existing tools that are empirically valid and reliable. Due to various challenges to risk prediction, these tools often over-predict the risk of sexual recidivism.

Recognizing the gap in the field, this project sought to develop an actuarial risk assessment instrument that would effectively predict the risk of sexual recidivism among youth. In partnership with leading experts and five jurisdictions across the country (Florida, New York, Oregon, Pennsylvania, and Virginia), the Urban Institute has developed a prototype assessment tool, state-specific risk assessment models, and practical guidance for conducting risk assessment for sexual recidivism in juvenile justice settings.

This report provides a summary of how those accomplishments were achieved and what lessons/implications were learned from the project. Before turning to the summary, some of the key findings and conclusions from the project are worth highlighting.

1) Due to the low frequency of occurrence, predicting sexual recidivism among youth yields results that are highly sensitive to the research settings in which the models are developed. Adopting an off-the-shelf assessment tool, either public or commercial, should be avoided without extensive customization to local settings, which entails updating the list of predictors as well as their weights. In this project, each jurisdiction-specific model has different predictors and weights, reflecting the differences in population characteristics, as well as data availability and quality in each place.

2) All of our partner jurisdictions have an assessment process for youth who come to the attention of the juvenile justice system to inform two types of decision-making: 1) developing case management and treatment plans and 2) determining placements or sentence length. Despite this similarity, how they actually conduct risk assessment and inform decisions based on the assessed risk level varies considerably across the jurisdictions. This is probably the case for the juvenile justice system in the United States at large.

3) As such, the accuracy and reliability of prediction models can markedly improve when customized to a particular setting and population to which the models are to be applied.

The Area Under the Curve (AUC) for one of our best and reliably performing models for an individual jurisdiction exceeded 0.80 (Oregon). All other models for single jurisdictions achieved similar prediction results (high 0.70s). This is notably higher than the performance for the multi-jurisdiction model (mid 0.60s), demonstrating how prediction models can benefit from being customized to the specific setting and population of interest.

4) When predicting sexual recidivism among youth, prior criminal history had the greatest predictive power. However, several dynamic factors, such as the extent of delinquent or positive peer association, impulsivity, school attendance and performance, remorseful feelings, mental health issues, and substance use, were also predictive of sexual recidivism even after controlling for prior criminal history. Although how those dynamic factors predicted sexual recidivism varied across hundreds of models tested in the current project, there was a sufficient empirical basis to suggest that any future work to improve the performance of risk assessment tools should further consider those dynamic factors.

5) The use of ML algorithms holds high promise for improving our capacity to make data-driven, risk-based decisions for youth with a history of sexual offending. Throughout the current project, prediction models based on ML algorithms notably outperformed traditional prediction models. However, ML algorithms generally require a large volume of data to be optimally effective. They are also subject to over-fitting, which requires more rigorous testing and updating.

6) How to determine cut points for risk levels has important implications, especially for sex offense risk assessment for youth. Because sexual recidivism rates are typically low (5% in this sample), without extensive tool customization, strategic planning, and consensus building among key stakeholders, it is highly likely to have a risk classification system that identifies someone unlikely to recidivate as "high-risk." For example, if individuals in our sample with predicted probabilities in the 75th percentile or above are classified as high-risk, that would not be out of keeping with current practice. However, on average, only 6.5% of them were rearrested for a sex offense within two years. In other words, 93.5% of them did not recidivate sexually.

   It begs the question of whether that 75th percentile should be used as a threshold to separate high-risk individuals from the rest. The definition of "high-risk" should not be derived solely on the basis of statistical properties. Criminal and juvenile justice stakeholders must have an open conversation about how much "risk" is tolerable given their system capacity to effectively manage youth with a history of sexual offending. This also gives rise to the need to evaluate prediction models for their absolute risk estimates because current practice in risk assessment focuses primarily on how to rank order individuals by risk (i.e., statistical discrimination) without necessarily estimating their chance of recidivism more precisely (i.e., calibration).

Background

There has long been a general consensus that actuarial assessments are superior to clinical judgments (Meehl, 1965). Actuarial assessments also offer additional benefits of objectivity and accountability in decision-making although such benefits have yet to be widely recognized or discussed. In the era of evidence-based practice and shrinking resources, juvenile and criminal justice agencies have increasingly adopted actuarial risk assessment instruments to determine how to allocate limited resources.

While there have been steady improvements in the development, validation, and implementation of risk assessments (Brennan et al., 2009), the management of youth with a history of sexual offending still relies heavily on clinical or subjective judgement to determine the level of supervision and the type of rehabilitative programming. The most commonly used juvenile sexual recidivism risk assessment instruments, the Juvenile Sex Offender Assessment Protocol-II (J-SOAP-II; Prentky & Righthand, 2003) and the Estimate of Risk of Adolescent Sexual Offense Recidivism (ERASOR; Worling & Curwen, 2001), were not designed for actuarial calculations of risk, but rather for clinical assessment of treatment needs. The other routinely used instrument, the Juvenile Sexual Offense Recidivism Risk Assessment Tool-II (JSORRAT-II; Epperson, Ralson, Fowers, DeWitt, & Gore, 2006), includes only static risk factors and has not been widely validated. In addition, according to a recent study on JSORRAT-II (Epperson & Ralston, 2015), the tool only shows moderate prediction accuracy (AUC=0.65), leaving much room for improvement.

Clearly, the field of sex offense risk assessment for youth lags behind current research on actuarial decision-making. There are a number of factors that make it difficult to accurately and reliably assess the risk of sexual recidivism among youth. First and foremost, adolescents are not fully mature in their judgement, problem-solving, and decision-making capabilities. Therefore, predicting their future behavior is inherently challenging. Second, it is difficult to capture the extensive developmental change that occurs during adolescence using a limited set of risk and protective factors. Conversely, it is also difficult to collect information on a youth sample large enough to afford some credibility to empirical findings. Nearly all of the empirical studies that exist today are based on small, non-probability samples, thereby limiting their applicability outside the research setting in which they were conducted. Lastly, sexual recidivism does not occur frequently. Low base rates of sexual recidivism among youth, along with other data challenges such as outliers and missing data, impede the use of traditional regression-based approaches to risk modeling.

Despite these inherent challenges, the current project sought to overcome the prediction problems prevalent in existing risk assessment tools for youth with a history of sexual offending, and the project team accomplished that in two important ways. First, the project compiled a large, multi-state dataset on youth with a history of sexual offending and considered a gamut of predictors, including static and dynamic risk factors as well as protective factors. Our general findings and implications for policy and practice are more reliable and widely applicable to the field than those from small, convenient samples that are often collected from a single juvenile justice or mental health institution, which are predominant in extant research. Second, the project explored various methodological approaches to risk prediction, including traditional regression-based modeling and machine learning algorithms and compared their performance in various test settings. Machine learning algorithms have been widely used in other fields, including healthcare, financial services, manufacturing and retail, and their potential to improve predictive performance is increasingly recognized in the field of criminal justice as well (Berk & Bleich, 2013; Bushway, 2013; Ridgeway, 2013b).

Design & Methods

In response to the NIJ's solicitation, this project was initially conceived as an opportunity to develop risk models for predicting short-term sexual reoffending among youth using a prospective data collection approach. The project team proposed a 36-month research design that included prospective data collection of potential risk and protective factors for approximately 1,200 youth in four partner sites and up to 18 months of follow-up for recidivism analysis.

However, after our initial observations of and conversations with the sites, the project design was modified to replace the prospective data collection with a retrospective approach. This shift largely resulted from suggestions from the partner agencies that the project team should consider a longer follow-up period for recidivism analysis. In addition, this shift was a logical and necessary decision to better support our proposed analytic approach using machine learning (ML) algorithms and to overcome one of the most critical limitations in extant research, which is the limited validity of empirical findings due largely to skewed, small, and non-representative samples used.

It was critical for the project to ensure that prediction models be developed and evaluated under various testing conditions and demonstrate reliable performance. Therefore, compiling as much historical data as possible on youth who have sexually offended was necessary as a primary mode of data collection. Tracking sexual recidivism prospectively is likely to result in a smaller youth cohort to begin with and high rates of data attrition, leading to another limited sample in terms of size and generalizability, not to mention a shorter length of follow-up.

A longer follow-up period would provide results better suited to inform the management of youth who have sexually offended for two reasons. First, the rates of sex offense recidivism are generally low, thus requiring a longer follow-up to observe an adequate number of failures. This helps reliably predict the risk of sexual recidivism. Second, public interest in sex offense recidivism would not be limited to the first 12 or 18 months of youth's release from state custody. In addition, the retrospective data collection approach would unlock the potential of rich administrative data maintained by the site partners for all of their juvenile cases.

Thus, our revised research plan included two data collection components: 1) collecting historical administrative data from the partner sites, and 2) conducting site visits with each partner agency to collect information on current policy and practice in sex offender treatment and management through semi-structured interviews and focus groups.

Administrative Data Collection

During the initial site visits with each of the partner sites, the project team gathered information about the types of administrative data that each site maintains. Though the availability of specific items varied across sites, the sites typically collected data about prior juvenile and criminal justice involvement, history of delinquency, or referrals to child protection agencies; social history; family and other social supports; substance use/abuse; employment history; school performance and conduct; personality traits; and conditions of supervision and sex offender treatment history. Using these domains as well as prior research as a guide, the project team worked with each site to develop a list of data elements to request, secure research approval, and clarify questions about data interpretation.

Semi-structured Interviews and Focus Groups

In shifting our focus to administrative data, the project team also needed to develop an understanding of how current policy and practice in sex offender treatment and management are reflected in those administrative data, as well as how policy and practice vary across sites. The project team was particularly interested in identifying how the sites manage youth who have sexually offended, how risk assessments are administered, what treatment programs are available, and what kind of supervision they receive. During our subsequent visits to each site, the project team interviewed key personnel within that jurisdiction, including juvenile justice agency staff (e.g., treatment directors, probation and/or correctional officers), as well as residential outpatient treatment facility staff (e.g., program directors and clinical staff), to learn about current practice in the assessment, management, and treatment of youth who have sexually offended.

Multi-State Data Set

With administrative data from five different states[2], we created a combined data set to be used for risk model development. This data set included youth from each state whose current disposition was for a sex offense or who had previously been adjudicated delinquent of a sex offense. The data set provides information on demographic characteristics, current offense, past delinquency and criminal justice involvement, school performance, peer associations, personality traits, and other factors that are traditionally known in criminological research as relevant to criminal behaviors.

The administrative data of each state was standardized so that they all could be combined into one dataset. Summary statistics of the data file are provided in Table 1. The combined dataset has 8,035 unique individuals, approximately 25% of whom were used in model development. The main reason for data loss is that not all of the predictors used in the multi-state model were consistently available from all of the partner jurisdictions. In particular, a significantly large number of cases from Florida did not make it to the core sample (n=1,835) because those attrited cases were not given a full assessment while under the supervision of the Florida Department of Juvenile Justice that included several dynamic factors included in our multi-state risk model. This core sample consists of predominantly males charged with a sex offense. The representation of each partnering agency is different between the whole sample and core sample, and so is the racial/ethnic composition. However, the whole sample and core sample are fairly comparable in terms of the extent of criminal history and recidivism. The re-arrest rates for a sex offense are five percent for both samples.

[2] Originally, four counties in Pennsylvania had agreed to participate in the project (Allegheny, Berks, Bucks, and Montgomery). However, eligible cases from all counties in Pennsylvania were subsequently identified from the Pennsylvania Juvenile Case Management System and extracted for data analysis.

TABLE 1

Summary Statistics of Study Sample

	Whole Sample (n=8,035)		Model Development Sample (n=1,835)	
Characteristic	Percentage/Mean	Frequency	Percentage/Mean	Frequency
Male	94%	7,556	97%	1,789
Female	6%	469	3%	46
White	47%	3,275	61%	999
Black	36%	2,552	25%	406
Hispanic	15%	1,075	9%	153
Other race	2%	139	4%	71
Age at qualifying offense	14.7	8,023	15.0	1,834
Most serious current charge is sex offense	91%	6,889	90%	1,369
Sex offense re-arrest within two years	5%	364	5%	83
New York	31%	2,520	28%	518
Pennsylvania	5%	438	17%	318
Virginia	4%	315	15%	276
Oregon	10%	801	33%	604
Florida	49%	3,961	7%	120

Risk Model Development

The overall approach to model development focuses on maximizing the utility of existing information to yield most effective and stable prediction results. Given that extant research has persistently suffered from small-sample studies whose results were inconsistent from study to study and from jurisdiction to jurisdiction, it was critical for the project to ensure that prediction models be developed and evaluated under various testing conditions and demonstrate reliable performance. Therefore, compiling as much historical data on youth who have sexually committed as possible was desirable as a primary mode of data collection. This was also logical because all of our partner agencies already tracked and maintained a considerable amount of information on their juvenile population as part of their assessment/classification process. They all used general risk assessment instruments that collectively covered a broad range of areas, including prior juvenile and criminal justice involvement, history of delinquency, or referrals to child protection agencies; social history; family and other social supports; substance use/abuse; employment history; school performance and conduct; personality traits; and conditions of supervision and sex offender treatment history. As those data elements have been broadly examined in decades of criminological research as correlates of delinquency or criminality, the first step in our modeling work involved examining their empirical association with sexual recidivism.

With respect to the outcome predicted, sexual recidivism was defined as being re-arrested for sex crime within two years. The definition of sex crime used in our analytic models follows how each individual state's statutes classify sexual offenses. The definition thus varies from jurisdiction to jurisdiction, but generally refers to a broad range of acts that are considered sexually deviant in nature, including but is not limited to rape, sexual abuse, sexual battery, child molestation, child pornography, prostitution, and indecent exposure.

Using the rich dataset from multiple jurisdictions, the project team tested various combinations of predictors and their interactions individually for each jurisdiction and jointly for all of them. The modeling processes for the state-specific and combined models were largely similar. To develop the multi-state model, the project team first randomly split the data into a training set and a test set. The training set contained two-thirds of cases, while the test set contained one-third of cases.[3] Both sets had similar rates of sex offense recidivism. The state-specific data sets were split temporally. The model was created using the training set, and then the performance was evaluated on the test set. Data splitting is common in predictive modeling, and helps prevent the model from over-fitting the data by allowing it to be tested on an "out-of-sample" set of observations. This method ensures that the model does not pick up on specific patterns in the training set, and instead generalizes to the whole sample. We repeated the random data splitting process 1,000 times for the multi-state model to ensure the model was stable across all samples of the data. In addition, we tested k-fold cross-validation and bootstrapping approaches to ensure that our test results were not particularly sensitive to the type of validation methods used.

With respect to predictive modeling, the project team tested many different machine learning and more traditional algorithms to identify the best-performing model. The traditional approaches included logistic regression and decision tress, while the machine learning approaches included regularized logistic regression, stochastic gradient descent, artificial neural networks, and support vector machines. For each model, the project team used a five-fold cross validation and tested many different parameters on the training set. The project team then evaluated the performance on the test set by calculating the area under the curve (AUC), overall accuracy, sensitivity, and specificity.

Risk Factor Identification

The process of selecting predictors in the machine learning literature is known as feature selection and is fundamentally statistical in nature. Common practice for "feature selection" relies on a statistical evaluation on how each predictor or a combination of predictors performs in the prediction model. Depending on the type of machine learning algorithms used, this process can be obscure, which is one of the major criticisms for machine learning approaches to classification. The algorithm can be a black box, providing little or no insights into the process of how each predictor contributes to the prediction, let alone which predictors are used.

It is therefore important to emphasize that the project team selected an initial set of predictors to be used in model building on the basis of their theoretical salience and availability across the jurisdictions. As listed above, those data domains cover a wide range of elements that have been examined in decades of criminological research, such as prior and current involvement in the juvenile and criminal justice systems, family characteristics, substance use/abuse, school

[3] The project team experimented with numerous approaches to partitioning the data, varying the proportion of training and test sets as well as the mode of data partitioning (e.g., random split, temporal split). No one approach performed uniformly well across various prediction models and datasets. In the absence of a theoretical basis to assume one approach is necessary better than another, we followed conventional practice in portioning the data (i.e., 2/3 for model building and 1/3 for validation).

performance, employment history, personality traits, and conditions of supervision and sex offender treatment history. The project team conducted a qualitative review on each of the common data fields across the jurisdictions to determine which predictors should be used in the process of model building. The machine learning algorithm then exploited the potential of those theoretically relevant predictors for separating recidivists from non-recidivists.

Still, the output from machine learning algorithms is not as intuitive as that of regression analysis. Some algorithms (e.g., support vector machine or neural networks) are particularly prone to interpretability issues. Regardless of the type of machine learning algorithms used, however, it is simple and straightforward to reverse-engineer that machine learning process and distill insights by regressing the predicted probabilities of recidivism on the predictors used. To identify factors associated with sex offense recidivism, the project team performed this regression analysis using only the set of youth included in the final risk model (n = 1,835) (see Table 4).

Risk Assessment Prototype

One of the machine learning algorithms that performed reliably across numerous test settings, involving different subsets of predictors as well as the study sample, was regularized logistic regression. Regularization is a statistical process that helps avoid over-fitting especially when the sample size is small or the number of predictors is large. The project team thus used the regularized logistic regression model to predict sexual recidivism, and the prediction algorithm was prototyped to demonstrate how it would work in practice.[4] The prototype tool was built using R Shiny, a platform that allows for the development of web applications (see Appendix B for the layout).[5] Just like any risk assessment instrument, the prototype tool allows a practitioner to submit information about an individual and then provides an output of the individual's predicted probability of sexual recidivism.

As mentioned above, it is important to note that the accuracy and reliability of our prediction models improved markedly when customized to each jurisdiction, meaning that each state-specific model has different predictors and weights. The performance of our multi-jurisdiction model was moderate. However, the multi-jurisdiction model has broad applicability than the state-specific models in terms of the prediction algorithm and predictors used. As a way of demonstrating how ML applications of risk assessment can be implemented in practice, the project team thus developed the prototype tool out of the multi-jurisdiction model.

The project team cautions against implementing our prototype tool for widespread use without adjusting it to the local setting. Practitioners often select an off-the-shelf assessment tool, and implement it in their jurisdiction. The project team is of the opinion that no risk assessment

[4] Unless otherwise noted, "assessment tools," "prediction models," and "algorithms" are loosely and interchangeably used in this report. These terms are often referenced differently in the literature, depending on the context used. However, the use of these terms typically refers to an actuarial process by which to quantify the risk of recidivism whether that is turned into a trade-marked product (i.e., assessment instruments) or an statistical script (i.e., algorithms), or remain in the summary table of an academic publication (i.e., prediction models).

[5] The prototype tool is available from the following website: https://www.urban.org/assessing-risk-sexual-offense-recidivism-among-youth

instruments that exist today can reliably be adopted without considerable adjustments to local settings, followed by local validation. Our assessment tool, which is sensitive to specifics of the data used for risk prediction and validation, should be no exception, hence the name, Risk Assessment Prototype. A later section of this report further discusses the sensitivity of our prediction models.

Findings

Site Visit Findings

In addition to collecting administrative data from each site, the project team conducted site visits with each partner site to collect information on current policy and practice in sex offender treatment and management through semi-structured interviews and focus groups. The project team conducted interviews with staff within secure youth facilities (e.g., treatment directors, probation and/or correctional officers) and residential and outpatient treatment facilities (e.g., program directors, clinical staff). The project team aimed to identify how each jurisdiction manages youth who have been adjudicated for a sex offense, how risk assessments are administered, what treatment programs are available, and what kind of supervision they receive.

Across all sites, risk assessment is conducted both at intake and throughout the time that a youth is under supervision to guide decisions about their supervision conditions and treatment plans. The states use the results of these assessments for two types of decision making: 1) for developing treatment and case management plans, and 2) for determining placements or sentence length. They utilize a number of assessment tools, relying primarily on general youth risk assessment instruments that are used across all justice-involved youth (e.g. the Youth Assessment Screening Instrument (YASI), Youth Level of Service (YLS), or the Positive Assessment Change Tool (PACT), but also conducting some sex offense-specific assessments (e.g. JSOAP or ERASOR). Reportedly, a screening tool is commonly used at intake, and only those assessed to be medium and high risk receive the full assessment that informs sex offender treatment and programming. These assessments are conducted either by licensed psychologists or counselors, or trained probation officers. This practice varies widely across the jurisdictions.

A primary purpose of conducting risk assessments with these youth is to guide the creation of individualized treatment plans. Sometimes these plans are developed in conjunction with other stakeholders that participate in the youth's treatment, such as therapists, case managers, and juvenile correctional officers. There are often mandatory sex offense treatment programs that are offered either in a community-based or residential setting based on risk level. Treatment plans involve multi-systemic, cognitive behavioral, and sexual trauma approaches, and can be updated to reflect changes in risks and needs as youth are re-assessed over time.

For the most part, the states employ evidence-based approaches such as CBT and MST. Following ATSA guidelines and the RNR principles, most states also made an effort to individualize treatment based on a youth's risk level and criminogenic needs. In some cases, they individualize treatment by tailoring treatment goals within a single treatment program; in others, they first place youth in different treatment tracks based on risk and needs, and then further individualize their treatment goals. It should also be noted, however, that the process of informing treatment plans based on the assessed risk and criminogenic needs was largely clinical and varies considerably across the jurisdictions and individual treatment providers/officers.

The states also vary in the extent to which they require youth who have been adjudicated of a sex offense to register with a sex offender registry. Most states have some mechanism for youth registration, whether through set criteria or judicial discretion. However, juvenile sex offender registration is a fast-changing landscape, with numerous recent legal challenges and state policy changes generally moving states toward fewer registration requirements for youth and bringing them more in line with ATSA's recommendation not to subject youth to registration.

Multi-state Risk Model

As mentioned earlier, the multi-state model has features more widely applicable than the jurisdiction-specific models and therefore deserves some discussion below. However, the multi-state model should not be interpreted as our recommendation for how exactly to develop risk estimates for general use by individual clinicians or juvenile justice agencies.

The project team used several performance metrics to evaluate the predictive performance of the risk model. Across all of them, the performance of the risk model was moderately acceptable. As displayed in Table 2, the AUC was above 0.6 for models using a few different algorithms, showing a moderate level of performance (Hosmer and Lemeshow 2000). In other words, there is, at least, a 60% chance that a randomly selected youth who did go on to commit another sex offense was scored as higher risk than a randomly selected youth who did not go on to commit another sex offense using each of these models. The highest performing model with an AUC of 0.658 was a regularized lasso logistic regression; other high performing models included: generalized regression (GBM), artificial neural networks, and traditional logistic regression.

TABLE 2

Predictive Performance of Risk Models

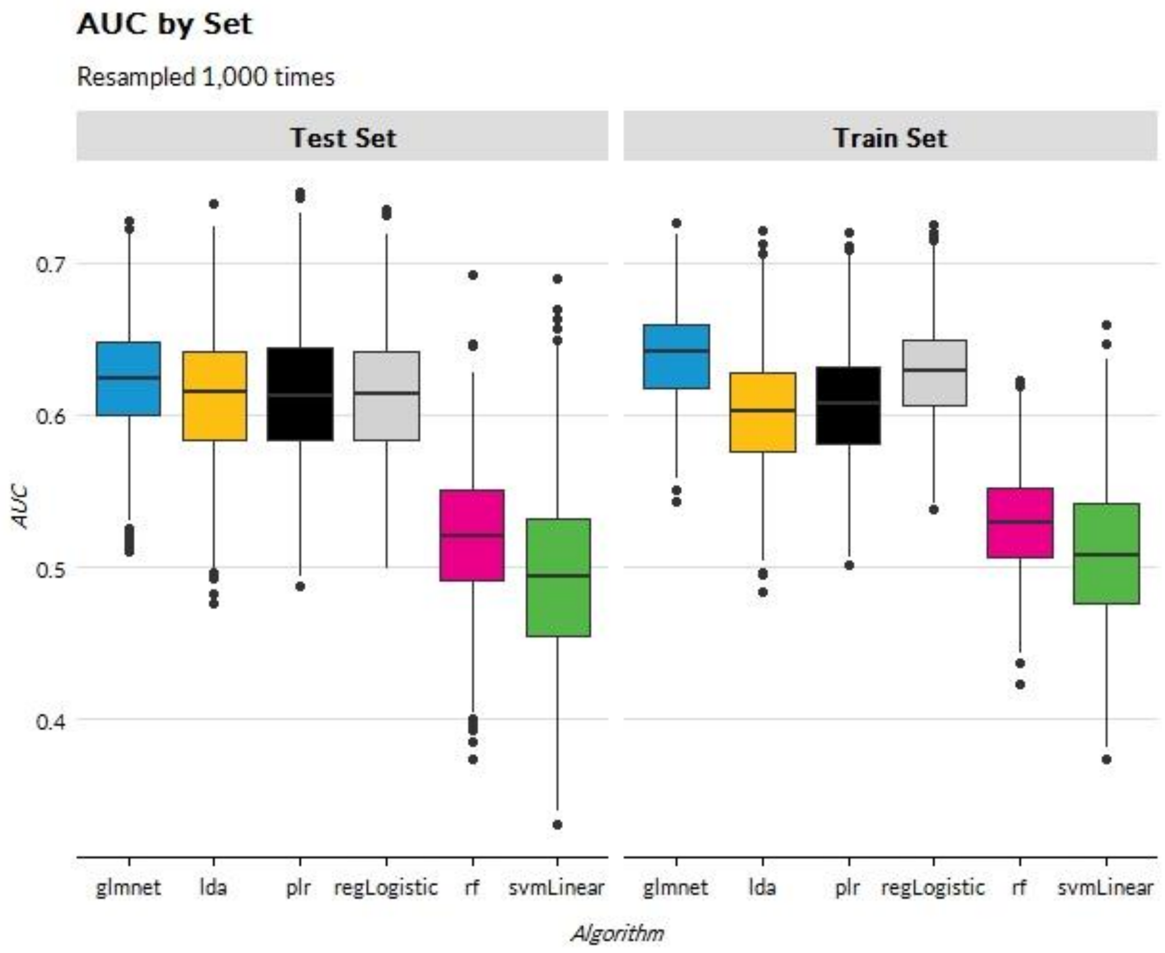
Algorithm	Train AUC	Test AUC
Regularized Logistic Regression	0.664	0.658
Generalized Regression	0.661	0.601
Artificial Neural Network	0.620	0.604
Logistic Regression	0.597	0.654

Notes: An AUC score represents the probability that a randomly selected person with a failure event (e.g. re-arrest) would obtain a higher score on the risk scale than a randomly selected person without a failure event. An AUC of 0.5 would be expected due to chance; 1.0 would represent perfect classification.

The project team estimated numerous models to arrive at this set of findings. Some of our tests resulted in prediction models with a markedly higher AUC than those reported above and in the literature. However, results from those models were sensitive to how the data were split for model building and validation. The project team learned that randomness in data splitting alone can have a great influence on the performance of the risk assessment model. The most common approach to validation is to randomly split the data set into a training set and a test set. Typically, the proportion of observations to use for model building ranges from 50% to 80%, and tool developers would select the highest performing model in terms of separating recidivists from non-recidivists. This practice of randomly splitting the data has not been of much interest in the literature. However, the notion of random split should be understood with a great deal of caution for sex offense risk assessment because of the low base rates.

FIGURE 3

Simulation Results on AUC across Randomly Split Datasets[6]



The goal of random split is to create two or more datasets that are similar to each other. Because there are a very few individuals who would reoffend sexually, however, the random split approach may or may not achieve balance in the characteristics of recidivists between the datasets created. In other words, when tossing a coin twice, there is no guarantee that you will get one head and one tail. However, if you repeat coin tossing for an infinite number of times, the proportion of heads vs tails will be closely to 50:50. In the real world application of risk assessment, there are simply not enough sexual recidivists to expect a high degree of similarity across randomly split datasets.

[6] glmnet (Lasso and Elastic-Net Regularized Generalized Linear Model); lda (linear discriminant analysis); plr (penalized logistic regression); regLogistic (regularized logistic regression); rf (random forest); svmLinear (support vector machine)

The project team thus simulated the influence of the random split on model performance across several machine learning algorithms by replicating the random splitting process 1,000 times. The AUC of the training and test sets varies widely based on the way the observations are split between them. As shown in Figure 1, the model using penalized logistic regression has an AUC that ranges from 0.5 to 0.73 for the training set. The range for the training set is from just below 0.5 to nearly 0.75.

It should also be noted that the performance of other models, especially support vector machine (svmLinear), is alarming in that they would perform adequately in one setting but would classify individuals completely reversely in another setting (i.e., high risk to be identified as low risk and vice versa). The performance of risk models can be highly unreliable due partly to the low frequency of sexual recidivism.

Although our simulation only reviews the performance of machine learning algorithms, the issues raised above are applicable to sex offense risk assessment in general and have thus far received little or no recognition in the literature. Interpreting the validation results from existing sex offense risk assessment tools thus requires special attention as tool developers would cherry-pick their best model without a systematic assessment on the model's stability.

State-specific Risk Models

Risk models were also developed for each state that participated in the study – Florida, New York, Oregon, Pennsylvania, and Virginia. Following an iterative process of building prediction models and evaluating their performance in different test settings, Table 3 reports the top-preforming model in each state by predictive validity. These models rely on a different set of predictors (see Appendix A for details).

TABLE 4

Predictive Performance of Jurisdiction-Specific Risk Models

State	AUC (Validation)	Algorithm
Florida	0.777	Regularized Logistic Regression
New York	0.797	Regularized Logistic Regression
Oregon	0.810	Support Vector Machine
Pennsylvania	0.776	Support Vector Machine
Virginia	0.772	Bagged Trees

Note. All models predict re-arrest for a sex offense within two years, with the exception of the model for Pennsylvania, which predicts one-year sex offense re-arrest.

Table 3 provides the validation results of each state in terms of AUC. Three models had an AUC above 0.7, indicating an acceptable level of performance, and two models had an AUC that rounded to or was above 0.8, indicating a strong performance. The best-performing algorithms across the five states were regularized logistic regression, support vector machines, and bagged trees, but no one machine learning algorithm performed universally well across the jurisdictions. Regardless of the type of machine learning algorithms used in each jurisdiction, these models are far better in terms of predictive performance to existing risk assessment tools for sexual recidivism in terms of identifying who would recidivate sexually.

The improved performance of the state-specific models compared to the multi-state model is potentially due to two factors. First, the state-specific models were each able to include more variables than the multi-state model, which was constrained to a subset of predictors commonly available across the jurisdictions. Second, populations are likely more similar within states than across them – generalizing across multiple states likely leads to the multi-state model performing at a lower level.

Risk Factors of Sexual Offending

It is important to reiterate that some of the machine learning algorithms developed in this project do not provide item-level details as to how each predictor contributes to the estimated probability of sexual recidivism. However, results from a series of bivariate regression models are shown in Table 4 to provide a sense of what predictors were commonly used in our modeling building and how influential they were. These analyses cannot be used or interpreted beyond that exploratory purpose.

First, as shown consistently in the literature, criminal history is the most important factor in risk prediction. Several factors regarding prior arrests for a sex offense and against-person felonies have the greatest impact on sexual recidivism, as well as offending experience and delinquent peer associations. Second, several dynamic factors, including school performance, peer association, remorseful feelings, impulsivity, mental health, and substance issues, are predictive of sexual recidivism. Across numerous models tested in our state-specific and multi-state datasets, the impact of those dynamic factors was not always consistent. However, there was a sufficient empirical basis to suggest that any future work to improve the performance of risk assessment tools should further consider those dynamic factors. Third, some of these bivariate correlations may seem counterintuitive, but they do not account for other important predictors of sexual recidivism. In particular, without considering sentencing or treatment conditions to which high-risk individuals may be subject, it would be difficult to interpret some of the risk factors having a protective effect.

TABLE 4

Bivariate Regression Results

Cases Included in Model Development

Variable	Coef	P		Beta
Diversity in offenses against persons	0.050	0.000	***	0.551
Delinquent peers	0.024	0.000	***	0.268
Prior commitments	0.004	0.000	***	0.164
Enrollment in special education	0.008	0.000	***	0.114
School attendance issues	0.010	0.000	***	0.112
Feelings of guilt about misbehavior	0.008	0.000	***	0.086
History of family member being incarcerated	0.005	0.020	**	0.075

Low academic performance	0.006	0.002	***	0.073
Impulsive	0.004	0.085	*	0.040
School behavior issues	0.002	0.288		0.025
Mental health issues	0.001	0.798		0.006
Substance use	0.000	1.000		0.000
Total prior arrests where most serious misdemeanor offense was a sex offense	0.000	0.999		0.000
Positive peers	-0.006	0.021	**	-0.054
Age at qualifying event	-0.002	0.000	***	-0.089
History of physical abuse victimization	-0.010	0.000	***	-0.123
History of sexual abuse victimization	-0.011	0.000	***	-0.139
History of family substance use issues	-0.013	0.000	***	-0.169
History of neglect	-0.015	0.000	***	-0.206
Most serious current charge is sex offense	-0.030	0.000	***	-0.248
Total prior arrests where most serious offense was felony	-0.013	0.000	***	-0.367
Total prior arrests for against-person felonies	-0.022	0.000	***	-0.442
Total prior arrests where most serious felony offense was a sex offense	-0.028	0.000	***	-0.509

Notes: * p < 0.1, ** p < 0.05, *** p < 0.01, **bolded, italicized variables** are dynamic risk factors Diversity in offenses against persons is a variable that indicates how specialized an individual's offense history is in offenses against persons. A higher coefficient represents greater diversity in types of felony offenses (against persons and not against persons felonies); a lower coefficient represents less diversity in types of felony offenses (a higher percentage of against persons felonies).

Implications for Practice and Research

Can Machine Learning Help Identify Youth at High-Risk of Sexual Recidivism?

Typically, actuarial risk assessment tools are developed by iteratively testing a different combination of predictors until a model with satisfactory performance is found. From each unsatisfactory model tested, something is learned and applied to the next model. This process is the same for any tool developers, outcomes predicted, or methodological approaches used.

Machine learning techniques allow this learning process to be repeated a large number of times without requiring much human input. They also follow clear rules about what to learn from unsuccessful models whereas the traditional or theory-driven approach typically pays little systematic attention to unsuccessful models. In other words, the machine learning approach to risk prediction is designed to maximize the potential of existing knowledge (Kim and Duwe, 2016).

As such, if prediction is the game, machine learning techniques should be the high draft picks (Bushway, 2013, p. 565).

By way of example, our risk models based on machine learning algorithms outperform the traditional model based on logistic regression. The state-specific models, in particular, show markedly better performance than existing risk assessment tools for sexual recidivism in terms of separating recidivsts from non-recidivists. Clearly, the data-driven appraoch holds promise in leveraging our limited knowledge and data on sexual offending to inform how to effectively manage youth at risk of sexual recidivism.

What are the Challenges to using Machine Learning?

There are a growing number of studies that argue over one methodological approach versus another (Berk & Bleich, 2013; Caruana & Niculescu-Mizil, 2006; Duwe & Kim, 2015; Hamilton et al., 2015; ; Hess & Turner, 2013; Liu et al., 2011; Stalans, Yarnold, Seng, Olson, & Repp, 2004; Tollenaar & van der Heijden, 2013). Since all prediction models could be expressed or understood in the machine learning language (Ridgeway, 2013), however, debating on which statistical approach should be used or should not be used on the basis of their predictive accuracy does not seem particularly useful. If tuned properly, the machine learning approach can, at least, match the performance of traditional models (Berk & Bleich, 2013). As such, it would be more practical and productive to discuss when and how to use machine learning algorithms, as opposed to whether or not to use them. There are a few important considerations to note.

First, actuarial assessment tools are used in a variety of human service contexts and reflect various disciplinary orientations and approaches. What is unique about risk assessment for sexual recidivism compared to other applications of actuarial decision-making in the criminal justice system (e.g., pretrial risk assessment, inmate classification system) is that there are a lot of individual clinicians with specialized training – and sometimes a license and even research experience in developing or evaluating assessment tools – who routinely assess the risk and needs of youth with histories of sexual offending. As they may be accustomed to conduct these assessments in various ways deemed clinically suitable, there is also a relatively stronger demand from them to learn about how to apply new advances, such as the machine learning approach to risk prediction, to their practice.

However, the development and implementation of a machine learning application to predict sexual recidivism is not best-suited for an individual clinician to engage for his or her own cases. Machine learning models can be most effectively developed and maintained when applied to an enterprise level data infrastructure. Ideally, a local or state government agency responsible for the management and treatment of at-risk youths would be better suited to maintain such a system and provide individual clinicians with the most up-to-date and accurate information about the risk and needs of justice-involved individuals.

Second and relatedly, machine learning models tend to be sensitive to specifics of the data used for risk prediction and validation. They typically require large data (i.e., a large number of observations and predictors), as well as routine updates and re-validation to maintain effectiveness in predicting sexual recidivism. As discussed earlier, a common challenge in predicting sexual recidivism is the low base rate. In our sample, only 5 percent were re-arrested for a sex offense within two years. This challenge is not just about whether the algorithm can predict such a rare event, but also about how much confidence should be placed on our decision-making for at-risk youth based on a rare phenomenon that only happens to 1 in 20 individuals.

Suppose there is a study with the sample size of 400, which by no means is a small sample in the current literature on sex offense risk assessment. A prediction model would be built on the characteristics of 20 recidivists, assuming the 5 percent base rate. If the sample is split for validation, even fewer individuals would be available to contribute to the prediction model. It is simply unrealistic to develop reliable prediction estimates from no more than a dozen individuals. When developing risk assessment tools that predict outcomes with the low base rates, including sexual recidivism, as many cases as possible should be included. Furthermore, machine learning tools should be updated routinely as they are subject to over-fitting.[7] All told, there are all the more reasons to suggest that the implementation of machine learning applications is more suitable at the jurisdiction level than at the individual institution or clinician level.

Third, a great deal of skepticism about the utility of machine learning applications comes from the lack of transparency or interpretability inherent in machine learning algorithms. Some algorithms (e.g., random forest) are more intuitive than others (e.g., support vector machine), but still they are generally less transparent about how each of the predictors explain recidivism individually and jointly. When it comes to the utility of machine learning algorithms, there is a trade-off between predictive performance and transparency. For those who conduct risk assessments, it is important to think about how much of an improvement in predictive accuracy should be considered acceptable at the expense of transparency because a clear communication about their expectations can help calibrate the performance of machine learning applications.

Fourth, there seems to be a slight misalignment between risk assessment and case management. The terms, “risk assessment,” and “risk and needs assessment,” are loosely and interchangeably used in the field, as well as in this report. However, it is important to clarify that risk assessment tools usually provide little information that can be used to directly inform treatment planning whereas needs assessment involves measuring the extent and nature of the needs of a given target population so that services can respond to them.

It is a popular idea that we can mitigate someone’s risk by addressing criminogenic factors that are identified in the process of risk assessment, but strictly speaking, the causal effect of criminogenic factors on recidivism cannot be learned from risk assessment because it is not an analytic apparatus that can help us understand causality. Scientific knowledge has traditionally been pursued with the two primary goals – explanation and prediction. In many scientific fields, especially social sciences, statistical models are predominantly used for causal explanation as opposed to empirical prediction. Risk assessment is primarily about predicting one’s risk of recidivism based on the distributional relationships between the outcome and predictors. The impact of each individual predictor on recidivism should not necessarily be interpreted as causal.

[7] An overfit model is one that is too closely fit to a limited set of data points, which simply means that the model is not generalizable outside the original dataset.

That said, whether or not machine learning algorithms provide much information about the impact of each predictor should not be much of a concern for service planning purposes.[8] Instead, the use of machine learning applications can be explored as a way to monitor someone's progress. If implemented properly, the machine learning application of risk assessment can be automated to produce risk estimates with minimal effort and can provide a standardized metric of assessing progress or detecting change.

What Else Have We Learned About Risk Assessment Development?

This study also provides some general recommendations for how to develop a risk assessment for at-risk youth with histories of sexual offending and how to successfully validate and implement risk assessments in the juvenile justice setting.

First, one other common challenge in risk assessment development and validation is the inability to account for treatment or management that may suppress the level of risk expected among the high-risk group. In this retrospective study, the project team could not control for any treatment, placements, or supervision that the youth received because not all jurisdictions collected this information or were able to share it. It is possible that youth who had more serious offenses or demonstrated treatment needs received more intense services or treatment that may have suppressed their anticipated levels of risk. Ideally, services and treatment received by each youth should be accounted for in the risk model development process.

This data challenge is also one of the reasons why risk assessment should not be interpreted in the causal framework. Not to mention statistical assumptions behind any prediction work, there are important factors that can influence recidivism but typically are not accounted for in the process of risk assessment – treatment, supervision, and management practices, for example. Without understanding how those factors affect the outcome of interest, it is difficult to interpret the effects of other predictors included in the model.

Second, we conducted traditional cross-validation in which a prediction model was constructed on a subset of the data and tested against on the unseen remaining data. However, one of the questions this project raised during the early stage of development extended this traditional validation framework to see whether or not an agency should implement an assessment tool that has been validated in other jurisdictions but not in their own. To what extent can they expect that the tool to perform reasonably well? Throughout numerous models tested in various settings, the project offers no assurance that adopting an off-the-shelf assessment tool, either public or commercial, can work without extensive customization to local settings, which entails updating the list of predictors as well as their weights. Local validation is a must-do, especially, for risk models that are computation-heavy.

A final important consideration of risk assessment development is choosing cut points for risk classification. Across the multiple algorithm types for the overall risk model, the mean predicted probability of recidivism was 4.92% on the test set, which approximates the observed recidivism rate of 4.95% in the test set. A convenient option is to classify cases with predicted probabilities higher than the mean as high risk and those below the mean as low risk because relative to the

[8] However, transparency can be rightfully required or desired for other purposes in the juvenile and criminal justice systems (e.g., accountability).

mean they are low or high risk. However, in absolute terms, a case manager may not perceive a youth with a predicted probability of recidivism of 7% as high risk because his or her risk of recidivism is highly unlikely. When implementing a risk assessment, jurisdictions should consider both the relative and absolute risk of recidivism when deciding risk classification levels.

Conclusions

Across our five study sites, juvenile justice agencies conduct general and sexual recidivism risk assessments on youth with histories of sexual offending to inform individualized supervision and treatment plans. However, there are limitations to the predictive accuracy of tools currently in use that require careful attention. Predicting risk of sexual reoffending among youth comes with a number of unique challenges, including accounting for extensive developmental change during adolescence, low base rates of recidivism, and treatment intervention effects.

The current project also encountered these challenges and addressed them to the extent possible through innovative modeling methodologies. Moving beyond the traditional logistic regression approach to classification, the project team tested several machine learning algorithms to examine whether they could adapt to complex interactions in the data and produce more accurate predictions of recidivism. The highest performing models in both the multi-state dataset and the state-specific datasets were machine learning models such as regularized logistic regression. These findings indicate that there is promise in further exploring advanced modeling techniques in the justice field.

In addition to equipping juvenile justice agencies with a framework for developing and validating a risk assessment instrument for their own youth population, the findings of this study provide insights into the nature of sexual offending among youth more broadly. Sexual recidivism among youth with histories of sexual offending is very low, despite their high level of needs. Practitioners should keep this into account when determining the level and type of supervision and services for these youth – recognizing that imposing greater restrictions and requiring more programming may not have the intended effect on youth that have a low likelihood of reoffending. Furthermore, there are many different pathways to sexual offending behavior, including histories of victimization and abuse, that require individualized attention and treatment plans.

Ultimately, more accurate predictions of risk and more nuanced understandings of risk and protective factors can better inform practitioners' decisions about how to manage and treat youth with histories of sexual offending. Lessons learned from this study serve to advance the field of sexual offense risk assessment and management in juvenile justice settings, and present considerations for jurisdictions seeking to develop and implement this type of risk assessment effectively and efficiently.

References

Berk, R. A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative assessment. Criminology & Public Policy, 12, 513.

Brennan, T., Dieterich, W., & Ehret, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. Criminal Justice and Behavior, 36(1), 21-40.

Bushway, S. (2013). Is there any logic to using Logit: Finding the right tool for the increasingly important job of risk prediction. Criminology and Public Policy, 12(3), 563-567.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms using different performance metrics. Paper presented at the The 23rd International Conference on Machine Learning, New York, NY.

Duwe, G., & Kim, K. (2015). Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism. Criminal Justice Policy Review, Forthcoming. doi:10.1177/0887403415604899

Epperson, D. L., & Ralston, C. A. (2015). Development and validation of the Juvenile Sexual Offense Recidivism Risk Assessment Tool–II. Sexual Abuse, 27(6), 529-558.

Epperson, D., Ralston, C., Fowers, D., DeWitt, J. & Gore, K. (2006). Actuarial risk assessment with juveniles who offend sexually: Development of the Juvenile Sexual Offense Recidivism Risk Assessment Tool – II (JSORRAT – II). In Prescott, D. (Ed.) Risk Assessment of Youth Who Have Sexually Abused, Oklahoma City, Oklahoma: Woods & Barnes.

Hosmer, D.W. & Lemeshow, S. (2000). Applied Logistic Regression (2nd Ed.). New York, NY: A Wiley-Interscience Publication.

Hamilton, Z., Neuilly, M.-A., Lee, S., & Barnoski, R. (2015). Isolating modeling effects in offender risk assessment. Journal of Experimental Criminology, 11(2), 299-318.

Hess, J., & Turner, S. (2013). Risk assessment accuracy in corrections population management: Testing the promise of tree based ensemble predictions. Retrieved from Irvine, CA.

Kim, K. & Duwe, G. (2016). Improving the performance of risk assessments: A case study on the prediction of sexual offending among juvenile offenders. In F. S. Taxman (Ed.), Handbook on risk and need assessment: Theory and practice. ASC Division on Corrections & Sentencing Handbook series. NY: Routledge.

Liu, Y. Y., Yang, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural network models in predicting violent re-offending. Journal of Quantitative Criminology, 27, 547-573.

Meehl, P. E. (1965). Seer over sign: The first good example. Journal of Experimental Research in Personality, 1, 27–32.

Prentky, R. A., & Righthand, S. (2003). Juvenile Sex Offender Assessment Protocol II (J-SOAP-II) manual. Washington, DC: U.S. Department of Justice, Office of Justice Programs, Office of Juvenile Justice and Delinquency Prevention.

Ridgeway, G. (2013). The Pitfalls of Prediction. National Institute of Justice Journal, Issue No.271.

Stalans, L. J., Yarnold, P. R., Seng, M., Olson, D. E., & Repp, M. (2004). Identifying three types of violent offenders and predicting violent recidivism while on probation: A classification tree analysis. Law and Human Behavior, 28(3), 253-271.

Tollenaar, N., & van der Heijden, P. (2013). Which method predicts recidivism best? A comparison of statistical, machine learning and data mining predictive methods. Journal of the Royal Statistical Society: Series A, 176(2), 565-584.

Worling, J. R., & Curwen, T. (2001). Estimate of Risk of Adolescent Sexual Offense Recidivism (Version 2.0: The "ERASOR"). In M. C. Calder, Juveniles and children who sexually abuse: Frameworks for assessment (pp. 372-397). Lyme Regis, Dorset, UK: Russell House Publishing.

Acknowledgements

We are thankful for contributions from juvenile justice professionals in Florida, New York, Oregon, Pennsylvania, and Virginia. The data analyzed was provided by the Florida Department of Juvenile Justice; the New York State Division of Criminal Justice Services (DCJS); the Oregon Youth Authority; the Pennsylvania Juvenile Court Judges' Commission; and the Virginia Department of Juvenile Justice. The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, its funders, DCJS, or the aforementioned agencies. Neither New York State nor DCJS, nor any aforementioned state, agency, or department assumes liability for its contents or use thereof. The project team also extends gratitude to Erika Parks, Daniel Lawrence, Jeremy Welsh-Loveman, Cybele Kotonias, Margaret Goff, Sino Esthappan, and Kyle Ueyama of the Urban Institute for the research support provided and Marie Garcia of NIJ for her support and guidance.

Appendix A. Risk Model Details

TABLE X

Summary of Top-performing Risk Models

	Combined	FL	NY	OR	PA	VA
Sex offense rearrest	2 yr	2 yr	2 yr	2 yr	1 yr	2 yr
Model type	RLR	RLR	RLR	SVM	SVM	Bagged trees
Data split method	Random	Temporal	Temporal	Temporal	Temporal	Temporal
Train cases	1231	1415	434	1910	281	648
Test cases	605	679	201	1125	157	515
Recidivism rate train set	4.5%	4.7%	6.5%	1.8%	5.7%	1.9%
Recidivism rate test set	4.5%	4.4%	6.0%	1.1%	7.0%	1.9%
AUC on test set	0.658	0.777	0.797	0.810	0.776	0.772
Risk Items	17	16	15	31	23	11
Risk item categories						
Delinquency history	X	X	X	X	X	X
Family history		X	X	X	X	
Prosocial factors	X	X		X	X	
Antisocial factors	X		X	X	X	X
Needs	X		X	X	X	X
Demographics		X	X	X		

Risk items in each model

Combined Model

- Total prior arrests with felony allegations
- Total prior arrests with misdemeanor allegations
- Total prior arrests where most serious offense was a felony
- Total prior arrests where most serious offense was a misdemeanor
- Total prior arrests for against-person felonies
- Total prior arrests for against-person misdemeanors
- Total prior arrests for sex offense felonies
- Total prior arrests for sex offense misdemeanors
- Academic performance
- School behavior
- School attendance

- Positive peers
- Delinquent peers
- Impulsivity
- Feelings of guilt
- Mental health
- Substance use

Florida

- Total violent offenses in index referral
- Total prior arrests with felony allegations
- Total prior arrests with felony against-person offenses
- Total prior arrests with misdemeanor against-person offenses
- Total prior arrests where most serious felony was weapons offense
- Total prior arrests where most serious felony was burglary
- 3 or more misdemeanor referrals
- Specialization/Diversity in drug and property offending (index referrals)
- Sexual misconduct misdemeanor referrals
- Age at first offense
- Attitude toward pro-social rules and conventions in society
- History of being victim of emotional abuse or neglect
- Family member(s) youth feels close to/has good relationship with
- Living with at least one of biological parents
- Older sibling currently/ever in jail or prison
- Hispanic

New York

- PreScreen Legal Risk
- Criminal record of mom in household
- No Problems with sibling
- Criminal record of mom in primary environment
- Close to dad/male
- Close to male sibling
- ADHD
- Mental health issues - dad
- Past mental health medication
- History of abuse by parent

- Sexual aggression
- Substance use - mom
- Age at risk assessment
- Total prior arrests where most serious offense was a felony
- Thinking responses

Oregon

- Age at assessment
- Revocation in 12 months
- Most severe offense is sex crime
- Total prior sex offense referrals
- Total prior referrals where most serious felony was crim. mischief
- Total prior referrals where most serious felony was sex offense
- Total prior referrals where most serious felony was theft
- Total prior referrals where most serious felony was assault
- Total prior referrals where most serious felony was < ounce
- Most serious felony theft count
- Most serious violation < ounce count
- Oregon Youth Authority Sex Crime score
- Most serious felony theft score
- Not a sex offender
- Sex offender plus
- Felony referrals
- Against person felony referrals
- Special education student
- History of successful employment
- Current pro-social community ties
- Out-of-home/shelter care placements exceeding 30 days
- History of dependency petitions filed
- Incarceration of household members for 3+ months
- Substance use treatment participation
- Total alcohol use past 4 weeks
- Total drug use past 4 weeks
- History of sexual abuse
- History of ADHD
- History of mental health problems

- Impulsivity
- Problem with sexual aggression not in criminal history

Pennsylvania

- Total prior arrests where most serious felony offense was a sex offense
- Any total prior arrests with sex offense allegations
- Total prior arrests where most serious misdemeanor offense was stolen property
- Disruptive behavior on school property
- Peer relations risk
- Substance abuse interferes with life
- No personal interests
- Not seeking help
- Total prior arrests for against-person felonies
- Total prior arrests where most serious misdemeanor offense was assault
- Total prior arrests where most serious felony offense was robbery
- Total prior arrests where most serious misdemeanor offense was theft
- Inadequate supervision
- Limited organized activities
- Verbally aggressive, impudent
- Short attention span
- Substance abuse risk score
- No/few positive acquaintances
- Total prior arrests with misdemeanor allegations
- Leisure/recreation risk score
- Total prior arrests for against-person misdemeanors
- Inconsistent parenting
- Attitudes/orientation risk

Virginia

- Academic performance
- Total prior arrests with felony allegations
- Total prior arrests with misdemeanor allegations
- Total prior arrests with weapons allegations
- Total prior arrests for against-person misdemeanors
- Total prior arrests where most serious offense was a misdemeanor
- Total prior arrests where most serious offense was runaway

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

- Juvenile classified as special education
- Cognitive distortions
- Delinquent peer affiliation
- Second most severe offense in recent referral

Appendix B. Prototype Risk Assessment Tool

Instructions: Please use your best judgment from conversations with the youth and a review of their records to assess each item.

Assessment Items

Total prior arrests with felony allegations

1 50

○

0 5 10 15 20 25 30 35 40 45 50

Total prior arrests with misdemeanor allegations

0 2 50

○

0 5 10 15 20 25 30 35 40 45 50

Total prior arrests where most serious offense was a felony

2 50

○

0 5 10 15 20 25 30 35 40 45 50

Total prior arrests where most serious offense was a misdemeanor

1 50

○

0 5 10 15 20 25 30 35 40 45 50

Total prior arrests for against-person felonies

1 50

○

0 5 10 15 20 25 30 35 40 45 50

Total prior arrests for against-person misdemeanors

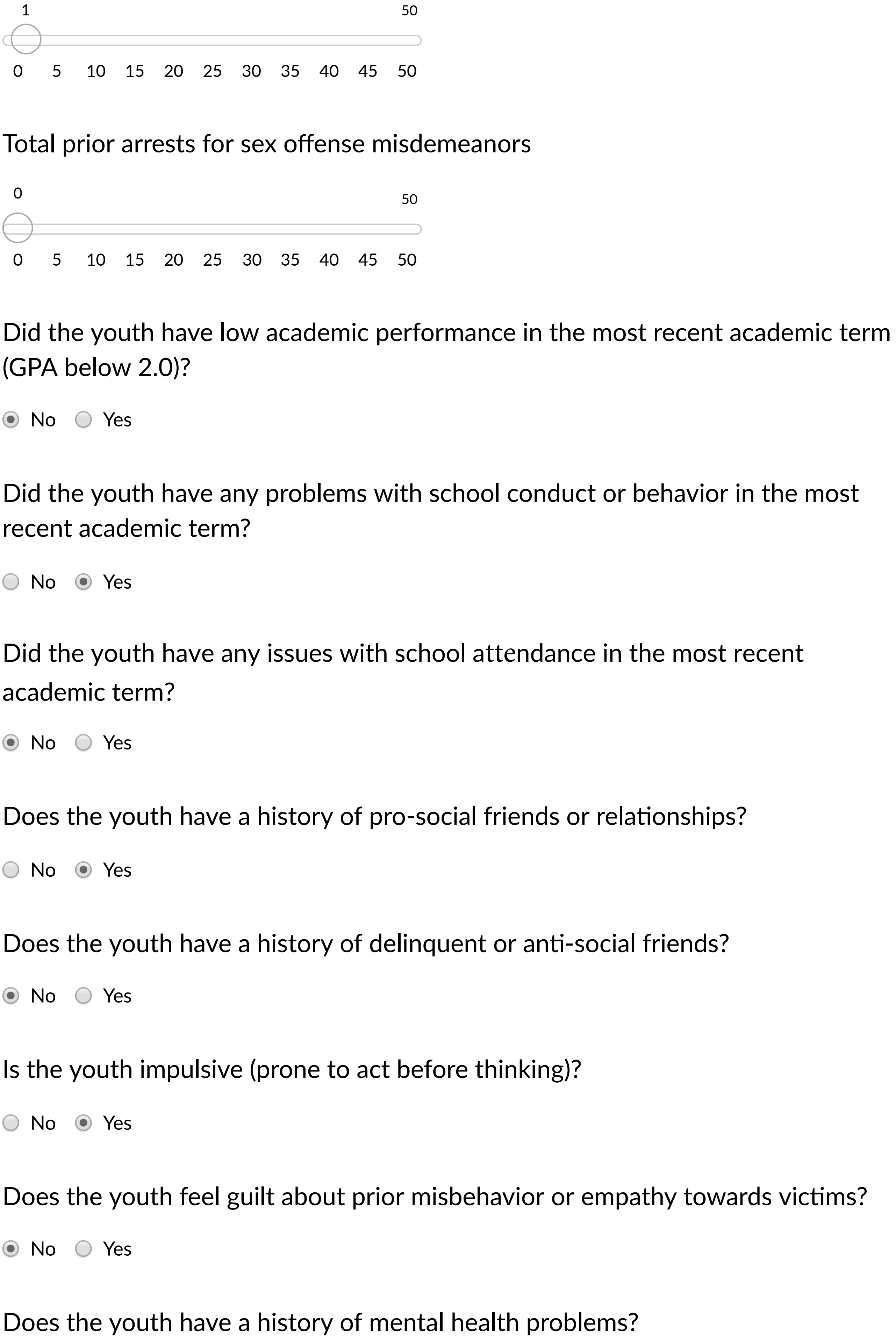
0 50

○

0 5 10 15 20 25 30 35 40 45 50

Total prior arrests for sex offense felonies

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.



1 50

0 5 10 15 20 25 30 35 40 45 50

Total prior arrests for sex offense misdemeanors

0 50

0 5 10 15 20 25 30 35 40 45 50

Did the youth have low academic performance in the most recent academic term (GPA below 2.0)?

● No ○ Yes

Did the youth have any problems with school conduct or behavior in the most recent academic term?

○ No ● Yes

Did the youth have any issues with school attendance in the most recent academic term?

● No ○ Yes

Does the youth have a history of pro-social friends or relationships?

○ No ● Yes

Does the youth have a history of delinquent or anti-social friends?

● No ○ Yes

Is the youth impulsive (prone to act before thinking)?

○ No ● Yes

Does the youth feel guilt about prior misbehavior or empathy towards victims?

● No ○ Yes

Does the youth have a history of mental health problems?

○ No ● Yes

Does the youth have a history of drug or alcohol use?

● No ○ Yes

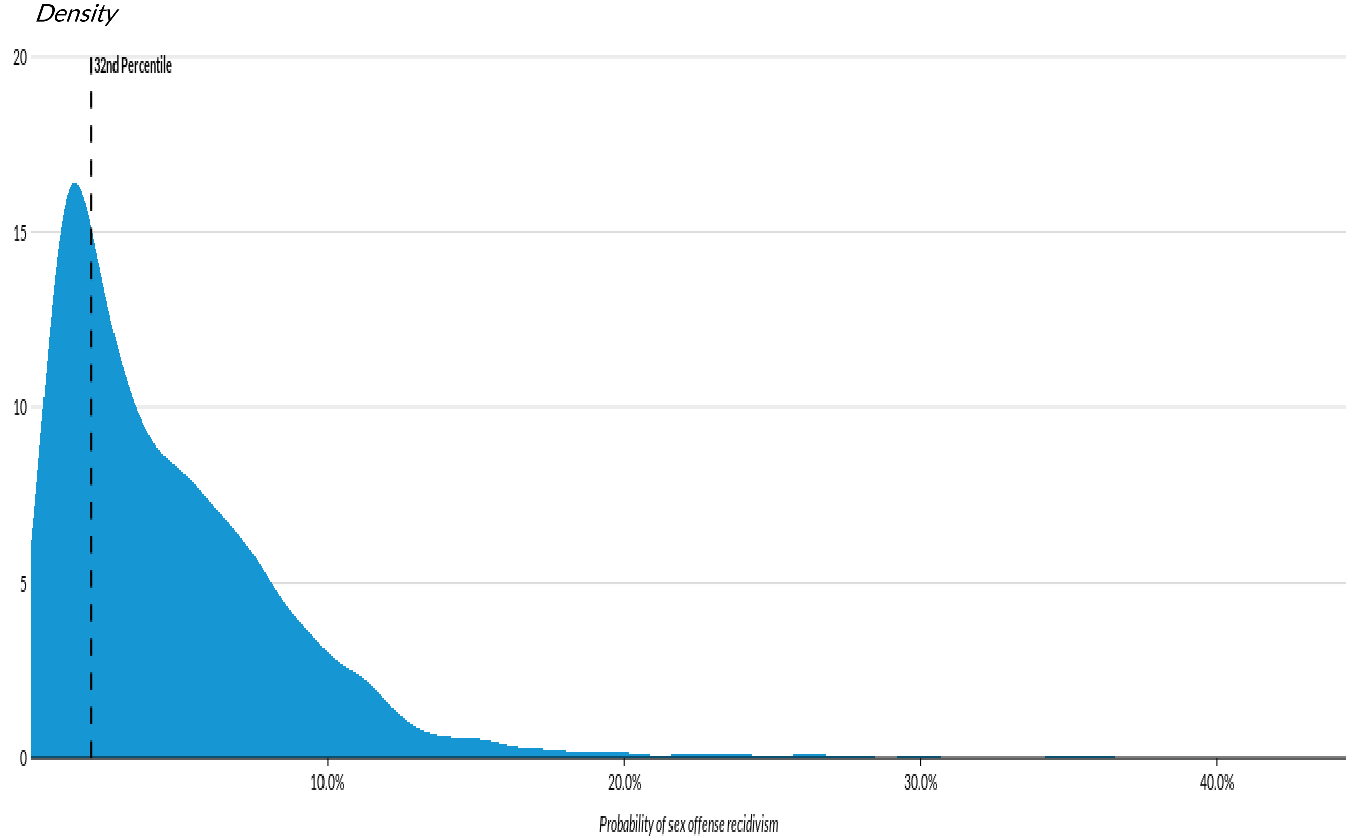
SUBMIT

Sex Offense Risk Prediction

This individual's likelihood of rearrest for a sex offense within two years is 2.04%.

A risk score of 2.04% is in the 32nd Percentile.

Distribution of Probabilities of Sex Offense Recidivism among Youth in Study Sample



This density plot shows the distribution of sex offense risk predictions among the youth in this study's sample. The dashed vertical line identifies where the current prediction falls within this distribution.

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.