Prevention Science
https://doi.org/10.1007/s11121-019-01040-2

Can Machine Learning Improve Screening for Targeted Delinquency Prevention Programs?

William E. Pelham III[1] · Hanno Petras[2] · Dustin A. Pardini[3]

Abstract
The cost-effectiveness of targeted delinquency prevention programs for children depends on the accuracy of the screening process. However, screening accuracy is often poor, resulting in wasted resources and missed opportunities to avert negative outcomes. This study examined whether screening approaches based on logistic regression or machine learning algorithms can enhance accuracy relative to traditional sum-score approaches when identifying boys in the 5th grade ($N = 1012$) who would be repeatedly arrested for violent and serious crimes from ages 13 to 30. Screening algorithms were developed that incorporated facets of teacher-reported externalizing problems and other known risk factors (e.g., peer rejection). The predictive performance of these algorithms was evaluated and compared in holdout (i.e., test) data using the area under the receiver operating curve (AUROC) and Brier score. Both the logistic and machine learning methods yielded AUROC superior to traditional sum-score screening approaches when a broad set of risk factors for future delinquency was considered. However, this improvement was modest and was not present when using item-level information from a composite scale assessing externalizing problems. Contrary to expectations, machine learning algorithms performed no better than simple logistic models. There was a large apparent advantage of machine learning that disappeared after appropriate cross-validation, underscoring the importance of careful evaluation of these methods. Results suggest that screening using logistic regression could improve the cost-effectiveness of targeted delinquency prevention programs in some cases, but screening using machine learning would confer no marginal benefit under currently realistic conditions.

Keywords This data is mandatory Please provide.

Introduction

Violence and other forms of serious criminal behavior are major public health problems that convey substantial economic and emotional costs to society (Bureau of Justice Statistics 2015; Federal Bureau of Investigation 2017). To address this situation, there has been a proliferation of interventions for elementary-age children designed to prevent the emergence of severe and chronic delinquency during adolescence and young adulthood (O'Connell et al. 2009; Wilson and Lipsey 2007). These interventions are typically implemented in the school setting using an indicated or selective approach in which a subsample of children with known early risk factors for later delinquency receives the intervention. Ideally, scarce intervention resources are allocated to those children at highest risk for the chronic and severe pattern of criminal offending that conveys substantial emotional and financial burden (Foster and Jones 2006), but this requires accurately identifying the target children. The goal of this report is to evaluate whether novel screening methods using (a) logistic regression or (b) machine learning can improve accuracy in identifying children at risk for chronic or severe criminal offending.

The ultimate success of targeted delinquency prevention programs is contingent upon the use of accurate screening procedures. In the case of delinquency prevention, the

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11121-019-01040-2) contains supplementary material, which is available to authorized users.

✉ William E. Pelham, III
wp8000@gmail.com

[1] Department of Psychology, Arizona State University, Tempe, AZ, USA

[2] American Institutes for Research, Washington, District of Columbia, USA

[3] School of Criminology & Criminal Justice, Arizona State University, Phoenix, AZ, USA

traditional screening approach is to compute a risk index (i.e., a single variable measuring risk for later delinquency) and then use one of two methods to choose which children should receive the intervention. The first method is to enroll only those children that exceed some threshold on a nationally normed scale assessing early behavior problems, such as those with a teacher-report *t*-score at or above the 85th percentile (e.g., Lochman and Wells 2004). The second method is to enroll children that exceed a sample-specific cut-score on a summative risk index, such as children whose teacher-reported behavior problem score exceeds the 80th percentile among youth within the particular school district (e.g., Q6 Lochman et al. 2015). This method can be implemented using a scale assessing behavior problems or a screen score representing the cumulative number of risk factors present in a child's life (e.g., living in a single-parent household and attending an underfunded school). These traditional screening approaches can be conceptualized as "sum-score" methods because the risk score used for selecting children is calculated by adding up equally weighted items on a behavior problem scale or counting the number of risk factors present.

Unfortunately, sum-score screening methods have been shown to have relatively poor accuracy in identifying those children that will go on to exhibit serious and persistent delinquency (Hill et al. 2004; Loeber et al. 2005; Petras et al. 2013). For example, one longitudinal study found that a summative risk score based on teacher-rated conduct problems accurately identified about half of all boys in grades 1 to 5 (44–67% across grades) who would be arrested for violence during adolescence when equally weighting false Q7 positives and false negatives (Petras et al. 2004). This screening method also produced a large number of false positives across grades (48–71%). False positive predictions result in finite resources being wasted on children who would not have gone on to exhibit delinquency, and false negative predictions result in the missed opportunity to avert costly delinquency acts from occurring during adolescence and adulthood. Both types of errors undermine the utility of targeted interventions, underscoring the need for improved screening approaches that result in superior accuracy and thus return on investment.

Moreover, the performance of these screening procedures is likely worse than published literature suggests. Most prior work has developed the sum-score approach (e.g., selected the risk factors to count and the screening cutpoint) in the same data that was used to evaluate the approach (e.g., to calculate screening accuracy). This procedure results in a screening method that is "overfit" to the data, produces positively biased estimates of performance and is unlikely to perform well on new cases to be screened (Babyak 2004). Thus, it is important to find screening approaches that perform well in data not used to develop the method.

Closer consideration of the traditional sum-score approach to screening reveals two key limitations that may contribute to poor accuracy. First, sum-score approaches assume that each risk factor should be given equal weight when generating an overall risk score. For example, targeted delinquency prevention programs for children have used summative risk scores that equally weight minor covert (e.g., lying) and more serious overt (e.g., physical fighting) behavior problems, even though the latter are more strongly associated with future criminal offending (Loeber et al. 2005). Others have used multi-domain risk scores that equally weight ancillary (e.g., low family income) and primary (e.g., childhood conduct problems) risk factors when identifying children for enrollment in delinquency prevention programs (Dishion et al. 2008).

A second limitation of sum-score approaches is that they treat risk factors as having linear associations with future delinquency and do not account for potential interactive effects between different risk factors. A failure to account for these more complex associations may limit classification accuracy. For example, deviant peer group affiliation in childhood may only confer risk for future criminal offending when present at high, but not moderate, levels (Loeber et al. 2008). Similarly, higher cognitive control abilities may protect youth with persistent anger problems from engaging in criminal offending (Hawes et al. 2016). Q8 Risk factors may also exhibit a combination of non-linear and interactive associations, although these types of complex relations are rarely tested.

There are two clear strategies for overcoming these fundamental limitations of sum-score screening methods. The first is to differentially weight risk factors using weights determined via regression. In the case of a binary outcome (e.g., delinquent vs. non-delinquent), this can be accomplished by fitting a logistic regression and weighting each risk factor by the regression coefficient indicating its relation to the outcome to be predicted. Thus, primary risk factors for delinquency (e.g., aggression) can contribute more to the overall screening score than more ancillary risk factors (e.g., academic achievement). However, logistic regression methods are limited in their ability to comprehensively account for non-linear and interactive effects of multiple risk factors. Although it is possible to add non-linear and interactive predictors into logistic models, one often encounters estimation and separation issues when the number of predictors becomes large (Peduzzi et al. 1996). In addition, it is difficult to satisfy the standard recommendation that logistic regressions be fit to datasets with a minimum of 10 events per predictor variable when evaluating potential interactive and non-linear effects (Peduzzi et al. 1996). For example, considering only 10 risk factors for delinquency, there are 45 potential two-way interactions and 120 potential three-way interactions, which would necessitate a sample that contains 1750 youth who exhibited the targeted delinquency outcome.

Machine Learning as an Untapped Approach

A more flexible and effective strategy that can be used to account for non-linear and interactive associations is machine learning. This class of techniques is commonly used in statistics, computer science, and engineering to build data-driven predictive algorithms (Hastie et al. 2009). Although these methods have improved prediction in diverse contexts, they have not yet been applied to screening for delinquency prevention. Relative to both sum-score approaches and to logistic regression, machine learning techniques are better able to reproduce complicated causal structures including higher-order interactions, are more accommodating of non-linear relationships between predictors and outcome, and are capable of using a much greater number of predictor variables (Hastie et al. 2009). For example, the popular random forest algorithm aggregates the results of hundreds of classification "trees," each of which recursively partitions the sample into subgroups that are maximally different in the outcome. This modeling strategy allows for highly discontinuous effects (i.e., cutpoints can occur anywhere within the range of a risk factor), permits many-way interactions (i.e., five recursive partitions would indicate a five-way interaction), and enforces no restriction on the number of predictors. Thus, machine learning may be able to address both the key limitations of existing screening methods identified above and thereby improve screening accuracy (Yarkoni and Westfall 2017).

We are aware of only two published applications in which machine learning was compared to logistic regression in the prospective prediction of delinquency outcomes. Neuilly et al. (2011) found that a classification tree algorithm produced accuracy superior to logistic regression (88% vs. 82%) when predicting recidivism among 320 convicted adult homicide offenders. Kleinberg et al. (2018) compared a gradient boosted trees algorithm with logistic regression when predicting re-offense among more than 20k adult defendants awaiting trial, finding that gradient boosted trees better identified offenders in the highest range of the risk continuum (e.g., positive predictive value of 56% vs. 46% in the uppermost 1% of risk). However, both studies focused on adults who had already offended, leaving it unclear whether machine learning can similarly improve the prospective prediction of which children will go on to display serious and persistent delinquency.

Present Study

The current study used longitudinal data collected on a school-based sample of 1012 boys to investigate whether logistic regression and/or machine learning algorithms can improve screening for targeted delinquency prevention programs, relative to traditional sum-score methods. Risk factors for delinquency were collected via teacher-report in the 5th grade using a well-validated and nationally normed rating scale, similar to those used to screen children for targeted delinquency prevention programs (e.g., Coping Power; Lochman et al. 2010). Serious and persistent delinquency outcomes from adolescence through early adulthood were derived from official records. Models predicting delinquency outcomes were developed using logistic regression and machine learning algorithms with a portion of the study sample. The performance of these models was then evaluated on an independent holdout sample and compared to traditional sum-score screening methods.

Methods

Sample

This study used longitudinal data collected from boys in the youngest and middle cohorts of the Pittsburgh Youth Study (PYS). Boys were selected for the study following a screening assessment conducted with a random sample of 1st grade (youngest cohort) and 4th grade (middle cohort) students enrolled in the Pittsburgh public schools (youngest $N = 849$; middle $N = 868$). At the screening, the boys' conduct problems were assessed via measures given to parents, teachers, and the boys themselves. Boys who scored in the upper 30% of risk on the screener and a roughly equal number of boys randomly selected from the remainder participated in the follow-up (youngest total $N = 503$; middle total $N = 508$). Across both cohorts, boys were predominately Black (54%) or White (42%). At the screening, most boys were living with their biological mothers (93%), and just under half had a father figure in the home (42%). Boys in the follow-up sample did not differ from those screened in terms of race, family configuration, or level of parental education (for details see Loeber et al. 1998).

The current study focused on evaluating the accuracy of methods designed to identify youth for targeting delinquency prevention programs during late elementary school, so all predictors were drawn from teacher-report data collected on both cohorts during the spring of the 5th grade (90% retention). This was the first assessment after screening at which teacher-report data was collected on both cohorts at the same grade level. This grade-equivalent assessment was used to combine data from both cohorts to achieve a sample large enough for cross-validation.

Measures

Teacher-Report Form

Predictors of delinquency outcomes were drawn from an expanded version of the teacher-report form (TRF) (Achenbach 1991) that included supplemental items added by the PYS investigators. The TRF has well-established reliability and validity in predicting later criminal offending (e.g., Pardini et al. 2018; Verhulst et al. 1994). The TRF instructed teachers to rate how true a series of statements were about the participant using a three-point scale: *not true* (0), *somewhat or sometimes true* (1), *very true or often true* (2). Items assessed adaptive functioning in multiple domains: internalizing and externalizing problems, hyperactivity/impulsivity, inattention, social difficulties, and academic motivation. In addition, teachers provided information about a child's academic performance in the subjects of reading and math on a 5-option scale (1 = *far below grade* to 5 = *far above grade*).

Analyses were repeated using two different sets of predictors derived from the TRF. Each set had different advantages. In addition, since our primary research question was the relative performance of sum-score, logistic regression, and machine learning methods, we wished to probe whether relative performance depended on the nature of the predictors used.

Predictor Set #1: Externalizing Problem Items

The first predictor set comprised all 34 items from the TRF externalizing composite scale subscale, plus the age of the child in years. The items indicate a broad array of childhood conduct problems, including aggression, oppositional and defiant behaviors, rule-breaking, anger outbursts, destruction of property, and truancy. This predictor set had the advantage of being easy to replicate or use in future studies that have collected the TRF. In addition, by including more variables than would typically be used in a sum-score or logistic regression (i.e., $n = 34$), it probed the possibility that machine learning would manifest advantage over the other methods when the number of predictors was larger.

Predictor Set #2: Risk Factor Subscales

The second predictor set consisted of subscales measuring nine different risk factors for delinquent behavior (see Table S1 for descriptive statistics). The subscales measured subdomains of conduct problems (e.g., aggression and oppositionality/defiance) and other risk factors (e.g., academic achievement and peer rejection) that have been associated with severe delinquent behavior in prior research (Loeber et al. 2008). Multiple TRF items were averaged to create each subscale (Cronbach alphas ranged from 0.84 to 0.94). The nine subscales were as follows: aggression (3 items), oppositionality/defiance (4 items), hyperactivity/impulsivity (5 items), inattention (6 items), dysregulated anger (4 items), interpersonal callousness (8 items), peer rejection (4 items), poor academic achievement (2 items), and negative attitude toward school (3 items). Relative to the TRF externalizing items, this predictor set had the advantages of (a) tapping into other domains of risk not captured in the externalizing items and (b) measuring risk factors with greater reliability than do the individual items. See supplement for citations to past work validating these subscales and confirmatory factor analysis of their structure in these data.

Outcomes: Serious and Persistent Criminal Offending

Criminal offending was measured using official records of criminal charges received between the 5th grade assessment and age 30. Juvenile criminal charges were collected from the Allegheny County Juvenile Court and the Pennsylvania (PA) Juvenile Court Judges' Commission. Adult criminal charges were collected by searching records managed by the PA State Police, PA Clerk of Courts, and the Federal Bureau of Investigation. Official criminal record searches were conducted on all study participants.

In order to investigate whether our conclusions were consistent across different specifications of the target outcome, a series of different criteria were used to delineate individuals who exhibited a pattern of serious and persistent criminal behavior. Three target outcomes were created based on the number of violent charges (i.e., simple assault, aggravated assault, rape, robbery, murder, and kidnapping), number of serious charges (i.e., felony violence or theft), and total number of charges received. For each charge outcome, a cutpoint was chosen that identified (as close as possible) the uppermost 25% of the sample. This cutpoint was chosen to identify those boys that exhibit a costly, chronic pattern of offending rather than isolated acts of delinquent behavior, which are relatively common among urban males living in impoverished environments. (Sensitivity analyses found conclusions were unchanged when using more liberal cutpoints [see supplement]). Using this approach, the target groups were as follows: (1) individuals with 3 or more violent charges (27%), (2) individuals with 3 or more serious charges (27%), and (3) individuals with 23 or more total charges (28%). A fourth group was also created consisting of individuals who met one or more of the three criteria outlined above (37%). Membership in each of these four groups was predicted as a binary outcome.

Data Analysis

Q9 All analyses were conducted in R (v3.5.2) (R Core Team 2018).

Handling of Missing Data

Prior to analysis, 34 cases were eliminated because they died prior to the last criminal record data collection, and 99 cases were eliminated because families did not participate in the 5th grade assessment. An additional 15 cases were eliminated because teachers failed to complete enough items assessing the targeted predictors. Remaining missing data was minimal (0.37% of all item response values) and each missing value was replaced with the median on that variable (see supplement for discussion of why median imputation was preferable to, e.g., multiple imputation). The final dataset included 864 (85%) of the original 1012 participants.

Creation of Training and Test Datasets

Prior to analysis, participants were randomly assigned to one of two mutually exclusive datasets, referred to as "training" and "test" datasets. Developing and evaluating a screening method using the same data results in overfitting, and thus an overestimate of predictive performance in future data (Hastie et al. 2009). To avoid this problem, we randomly assigned participants to the training set (probability = 0.70) or test set (probability = 0.30). The choice of a 70/30% split balanced (a) the desire to have as many cases as possible in the training set to increase the precision of the predictive model and (b) the need to have sufficient number of cases remaining for the test set to produce credible estimates of performance on holdout data (Hastie et al. 2009). The training set was used to develop and select the predictive models using repeated 10-fold cross-validation; the test set was used to evaluate the predictive performance of the final models in holdout data.

Sum-Score Approach

The predictive performance of logistic and machine learning models was contrasted with traditional sum-score screening approaches. For the predictor set comprising the TRF externalizing items, the sum-score risk score was the TRF total externalizing problems *t*-score, which is a function of the summed responses to each item. For the predictor set comprising the risk factor subscales, the sum-score risk score was calculated as the sum of the standardized values (i.e., *z*-scores) on each of the nine subscales.

Logistic Regression

Logistic regression models were used to examine whether screening performance was improved when components of the risk score were differentially weighted. We fit logistic models (a) using all items from the TRF externalizing scale as separate predictors and (b) using the risk factor subscales as separate predictors. These logistic models were fit to the training data for each delinquency outcome. Coefficients were saved and then applied to the test dataset to evaluate the predictive performance of the logistic models.

Machine Learning

Analyses also examined the performance of five different machine learning algorithms: lasso, random forest, gradient boosted trees, neural networks, and support vector machines. These methods have demonstrated success in numerous machine learning applications and represent different algorithmic approaches (see supplement, and also James et al. 2013). Best-practice entails trying multiple approaches to the same problem and selecting as the final model that which produces the best performance in the training dataset (Hastie et al. 2009; Kuhn and Johnson 2013). Thus, although five different machine algorithms were developed in the training data, only the one that produced the best cross-validated performance was compared to the sum-score and logistic methods in the test data.

Machine learning models were developed using a series of steps performed using the training data (see supplement for technical detail). For each machine learning model, optimal values on that algorithm's "tuning parameters" were selected based on the results of a repeated 10-fold cross-validation procedure (10 repeats). Each algorithm has different tuning parameters that control its functioning. For example, the lasso algorithm has one tuning parameter: a value for λ, a penalty factor that shrinks the estimated regression coefficients toward zero. Thus, we evaluated the performance of a model (i.e., combination of algorithm and potential tuning values) in the following way. First, participants in the training dataset were randomly divided into ten equal-size blocks. Next, the model was estimated using data from 9 of the 10 blocks, and then tested on the 10th. This was repeated ten times, each time holding out a different one of the 10 blocks, and then the results were averaged together. Thus, the performance of the models was always evaluated using data from an independent group of participants, reducing model overfitting.

Predictive performance in training data was evaluated using the Brier score, which is the mean squared difference between each participant's model predicted the probability of experiencing the outcome and participant's observed outcome (i.e., 0 = did not experience the outcome; 1 = did experience the outcome). A lower Brier score indicates that the predicted probabilities are closer to the participants' true probabilities. For each algorithm, the tuning parameter specification that produced the lowest Brier score metric was used to generate the final predictive model based on the entire training dataset. Finally, we chose the algorithm whose final predictive model produced the lower Brier score in the training data as the machine learning method to apply in the test data.

Comparing the Performance of Screening Approaches in Test (i.e., Holdout) Data

Next, we compared the performance of the three predictive approaches (i.e., sum-score method, logistic regression, and machine learning) in the test dataset. For the logistic and machine learning methods, each participant was assigned a predicted probability score by inputting their observed values on the predictors to the final model selected. These probability values were treated as continuous risk scores for evaluating performance (predicted probabilities can range 0–1).

Screening methods were compared using two performance metrics: (1) the area under the receiver operating characteristic curve (AUROC) and (2) the Brier score. The AUROC indexes the ability of a test to correctly classify those with and without the outcome, with a higher AUROC indicating better accuracy in terms of discriminating positive from negative cases (AUROC = 0.50 indicates discrimination at the level of random guessing; AUROC = 1 indicates perfect discrimination). To verify that our models were discriminating delinquents from non-delinquents at a rate better than chance, we tested the null hypothesis that the AUROC was equal to 0.50 using the Mason and Graham (2002) method. The Brier score is the mean squared difference between the predicted probability of the delinquency outcome and the actual outcome. Because Brier scores are based on predicted probabilities, they cannot be calculated for sum-score approaches.

To address the primary research question, we tested whether the AUROC and Brier score values generated by each of the three screening methods were significantly different. For each combination of predictor set and outcome, the performance of the sum-score approach was compared with the logistic regression and machine learning models, and then the performance of the logistic regression and machine learning models were compared. The Delong et al. (1988) method was used to compare AUROCs, and the percentile bootstrap (Davison and Hinkley 1997) was used to compare Brier scores. There were 8 comparisons of AUROCs from sum-score vs. logistic regression models, 8 comparisons of AUROCs from sum-score vs. machine learning models, 8 comparisons of AUROCs from logistic regression vs. machine learning models, and 8 comparisons of Brier scores from logistic regression vs. machine learning models. To reduce concerns about multiple testing, we focus interpretation on the pattern of results across conditions (i.e., combinations of predictors and outcome) rather than any specific statistical contrast.

Performance Across Screening Cutpoints

Performance measures that utilize the predicted probabilities or rank order (e.g., AUROC and Brier score) are more efficient and robust than metrics that utilize cutpoints to make discrete predictions (e.g., accuracy, sensitivity, and specificity). To complement our statistical comparison of the screening methods using the AUROC and Brier score, we also descriptively compared their positive predictive value (PPV) and negative predictive value (NPV) across a range of screening cutpoints. PPV indicates how often the children identified as positives (i.e., those to be enrolled in the intervention) go on to manifest the delinquency outcome. NPV indicates how often the children identified as negatives (i.e., those to be excluded from the intervention) go on to *not* manifest the delinquency outcome. PPV and NPV of screening algorithms were estimated via repeated 10-fold cross-validation in the training data since estimates from the test data would have been too unstable (e.g., screening 10% of test data into intervention comprises only 26 youth for the PPV calculation).

Results

See supplement for complete reporting of model performance (Table S2), model comparisons (Table S4), and the final machine learning algorithm selected under each condition.

Overall Predictive Performance

Within the test data, AUROC analyses confirmed that every predictive model discriminated which children would later manifest the delinquency outcome significantly better than chance (all $ps < 0.001$). AUROCs ranged from 0.68 to 0.78, with a median value of 0.74. Brier Scores ranged from 0.161 to 0.208, with a median value of 0.168.

Comparing Sum-Score Methods to Logistic Regression and Machine Learning

Table 1 reports the AUROC values for the sum-score, logistic regression, and machine learning models for each combination of predictor set and outcome. When the predictor set was comprised of the TRF externalizing problem items, there were no statistically significant differences between the sum-score method and the logistic regression or machine learning methods (*ns*). Averaging across outcomes, mean AUROC was 0.73 for sum-

Table 1 Predictive performance in test data for risk scores produced by sum-score, logistic, and machine learning methods

Predictor set	Outcome Variable	Sum-score method [95% CI]	Logistic regression [95% CI]	Machine learning [95% CI]
Area under the receive operating curve (AUROC)				
Risk factor subscales	3+ violent charges	0.75 [0.69, 0.81][a]	0.77 [0.71, 0.84][a]	0.78 [0.71, 0.84][a]
	3+ serious charges	0.73 [0.67, 0.79][a]	0.78 [0.72, 0.84][b]	0.78 [0.72, 0.84][b]
	23+ total charges	0.68 [0.61, 0.75][a]	0.74 [0.67, 0.80][b]	0.73 [0.67, 0.80][b]
	Violent, serious, or total charges	0.73 [0.67, 0.79][a]	0.75 [0.69, 0.82][a]	0.75 [0.69, 0.82][a]
Externalizing problem items	3+ violent charges	0.75 [0.68, 0.81][a]	0.73 [0.66, 0.80][a]	0.75 [0.69, 0.82][a]
	3+ serious charges	0.72 [0.66, 0.79][a]	0.73 [0.66, 0.81][a]	0.75 [0.69, 0.81][a]
	23+ total charges	0.69 [0.63, 0.76][a,b]	0.74 [0.67, 0.80][a]	0.68 [0.61, 0.76][b]
	Violent, serious, or total charges	0.74 [0.68, 0.80][a,b]	0.69 [0.62, 0.76][a]	0.74 [0.67, 0.80][b]
Brier scores				
Risk factor subscales	3+ violent charges	–	0.161 [0.138, 0.184][a]	0.166 [0.145, 0.190][b]
	3+ serious charges	–	0.161 [0.137, 0.185][a]	0.163 [0.140, 0.187][a]
	23+ total charges	–	0.177 [0.150, 0.204][a]	0.174 [0.149, 0.200][a]
	Violent, serious, or total charges	–	0.193 [0.168, 0.218][a]	0.193 [0.171, 0.217][a]
Externalizing problem items	3+ violent charges	–	0.167 [0.141, 0.195][a]	0.167 [0.144, 0.190][a]
	3+ serious charges	–	0.166 [0.138, 0.196][a]	0.167 [0.143, 0.192][a]
	23+ total charges	–	0.168 [0.142, 0.198][a]	0.184 [0.159, 0.211][b]
	Violent, serious, or total charges	–	0.208 [0.179, 0.238][a]	0.194 [0.170, 0.219][a]

Within each row, values that do *not* share a superscript differ significantly, $p < 0.05$

score method, 0.72 for logistic regression, and 0.73 for machine learning algorithms.

In contrast, when the predictor set comprised the risk factor subscales, there were four (of 8 possible) statistically significant differences between the sum-score method and the other two approaches. Both logistic regression and machine learning produced higher AUROCs than did the sum-score method ($p < 0.05$) when predicting outcomes involving repeated serious charges or total charges. Averaging across all models, mean AUROC was 0.72 for sum-score method, 0.76 for logistic regression, and 0.75 for machine learning algorithms, indicating a modest advantage of the more complex approaches.

Comparing Logistic Regression to Machine Learning

Table 1 reports the AUROC and Brier scores of logistic regression and machine learning models for each combination of predictor set and outcome. Both the statistical tests (i.e., the p values) and the descriptive results (i.e., the means) suggested that there was no consistent difference in the performance of the two methods. Logistic regression performed significantly better (i.e., higher AUROC and lower Brier score) when predicting the total charges criterion using the externalizing problem items ($p < 0.05$). In contrast, machine learning performed significantly better (i.e., higher AUROC) when predicting which participants would meet criteria for one or more of the delinquency outcomes using the externalizing problem items ($p < 0.05$). Averaging across all models, the mean performance of logistic regression was almost identical to that of machine learning on both AUROC (0.741 vs. 0.747) and Brier score (0.175 vs. 0.176).

Performance Across Screening Cutpoints

Figure 1 shows the positive predictive value (PPV) obtained by each of the three methods—sum-score, logistic regression, and machine learning—when between 0 and 50% of children are screened into the preventive intervention (i.e., across cutpoints that 0 to 50% of children exceed). When predicting the violent, serious, and total charges outcomes, PPVs generally ranged between 40 and 60% and were (as expected) higher when a smaller proportion of children were screened into the intervention. When predicting the outcome of meeting any of the three charges criteria, PPVs generally ranged from 50 to 70% and again were higher when a smaller proportion of children were screened into the intervention. Advantages of (a) machine learning and logistic regression over (b) sum-score methods were most apparent when predicting the outcome of meeting any of the three charges criteria, where PPVs were approximately 10 to 20 percentage points higher when between 10 and 20% of children were screened into intervention.

Figure 2 shows the same for negative predictive value (NPV). When predicting the violent, serious, and total charges outcomes, NPVs generally ranged between 75 and 85% and were (as expected) higher when a larger proportion of children were screened into the intervention. When predicting the outcome of meeting any of the three charges criteria, PPVs generally ranged

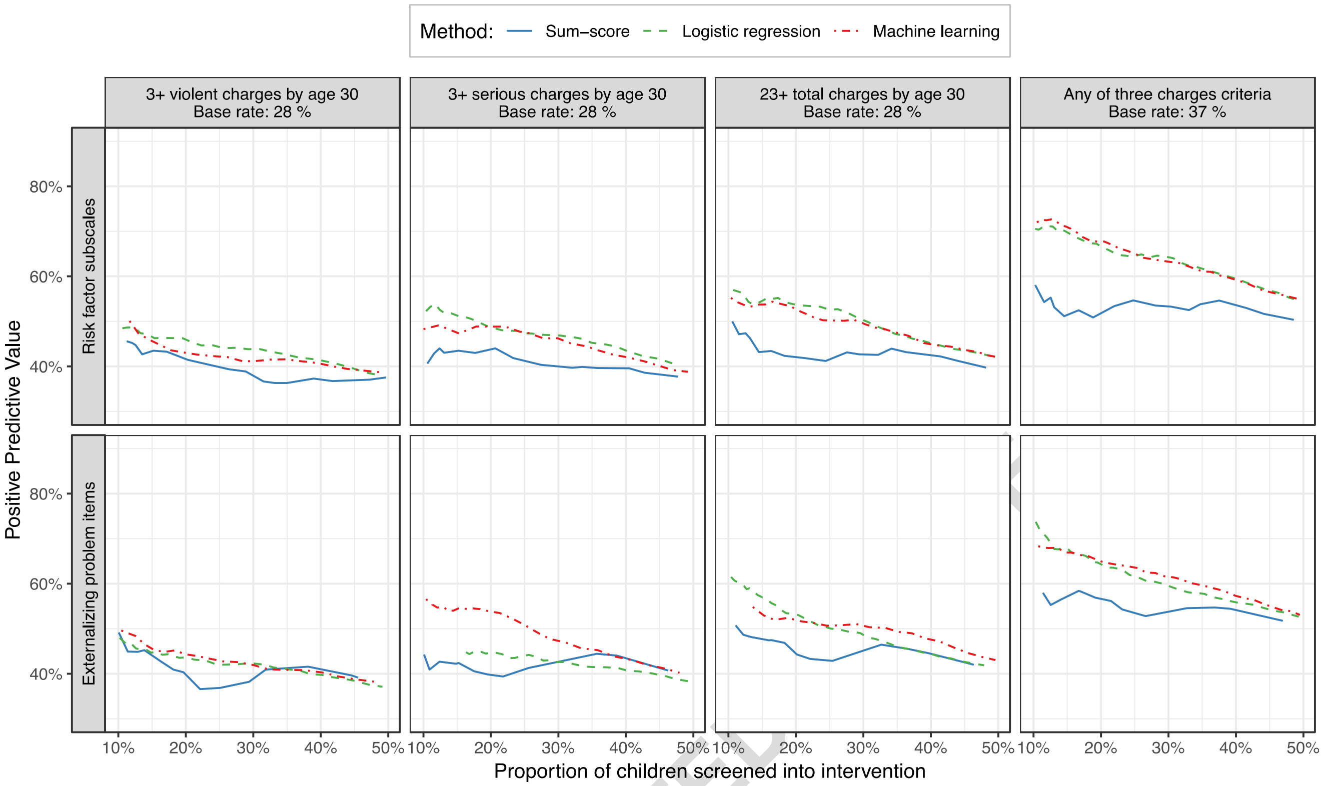

Fig. 1 Positive predictive value (PPV) across screening thresholds, per repeated 10-fold cross-validation in training data. Values calculated via repeated 10-fold cross-validation in training data. Positive predictive value is the probability that when the model predicts a child will go on to exhibit the delinquency outcome, the child will in fact exhibit that outcome. Graph shows positive predictive value achieved by each screening method across range of proportion of kids screened into intervention. Results separated by predictor set (rows) and outcome to be predicted (columns). Each line is constructed by calculating the positive predictive value and proportion predicted to be positive across a range of possible cutpoints in the risk score produced by the method (i.e., the sum score or the predicted probability)

from 65 to 80% and again were higher when a larger proportion of children were screened into the intervention. While the NPV curves for machine learning and logistic regression were higher than those for sum-score methods in almost all cases, differences in NPV were very small in magnitude (5% at maximum).

Discussion

Data from a prospective, longitudinal study was used to evaluate whether logistic regression and/or machine learning algorithms improved screening for targeted delinquency prevention programs, relative to traditional sum-score methods. To protect against overfitting, the performance of logistic and machine learning methods was tested on an independent holdout sample using eight different combinations of predictor set and outcome. Results indicated that both the logistic and machine learning methods could improve on traditional sum-score screening approaches when multiple-domain risk factors were used to predict repeated criminal offending. However, there was no evidence that the complex machine learning algorithms provided better predictive performance than simpler logistic models.

All screening approaches obtained AUROCs of between 0.68 and 0.78, indicating that they would correctly classify a randomly selected pair of delinquent and non-delinquent boys 68–78% of the time. These AUROC values were comparable to those obtained in prior studies predicting violent arrests (Petras et al. 2004; AUROCs up to 0.74) or diagnoses of antisocial personality disorder (Petras et al. 2013; AUROCs from 0.62 to 0.71) from teacher-reported aggression during elementary school. However, our AUROCs were calculated on holdout data not used to estimate the risk score, so they will be lower (and more accurate) than the values obtained in past work that has not maintained this distinction.

More complex methods—logistic regression and machine learning—performed better than the sum-score approach only when the risk score was based on multiple risk factor subscales. One potential explanation for this discrepancy is a key limitation of sum-score approaches raised earlier: Ancillary risk factors receive equal weighting to primary risk factors when they are incorporated into the risk score. The risk factors used in the current study included ancillary risks (e.g., school motivation and peer rejection) that were weighted equally to primary risks

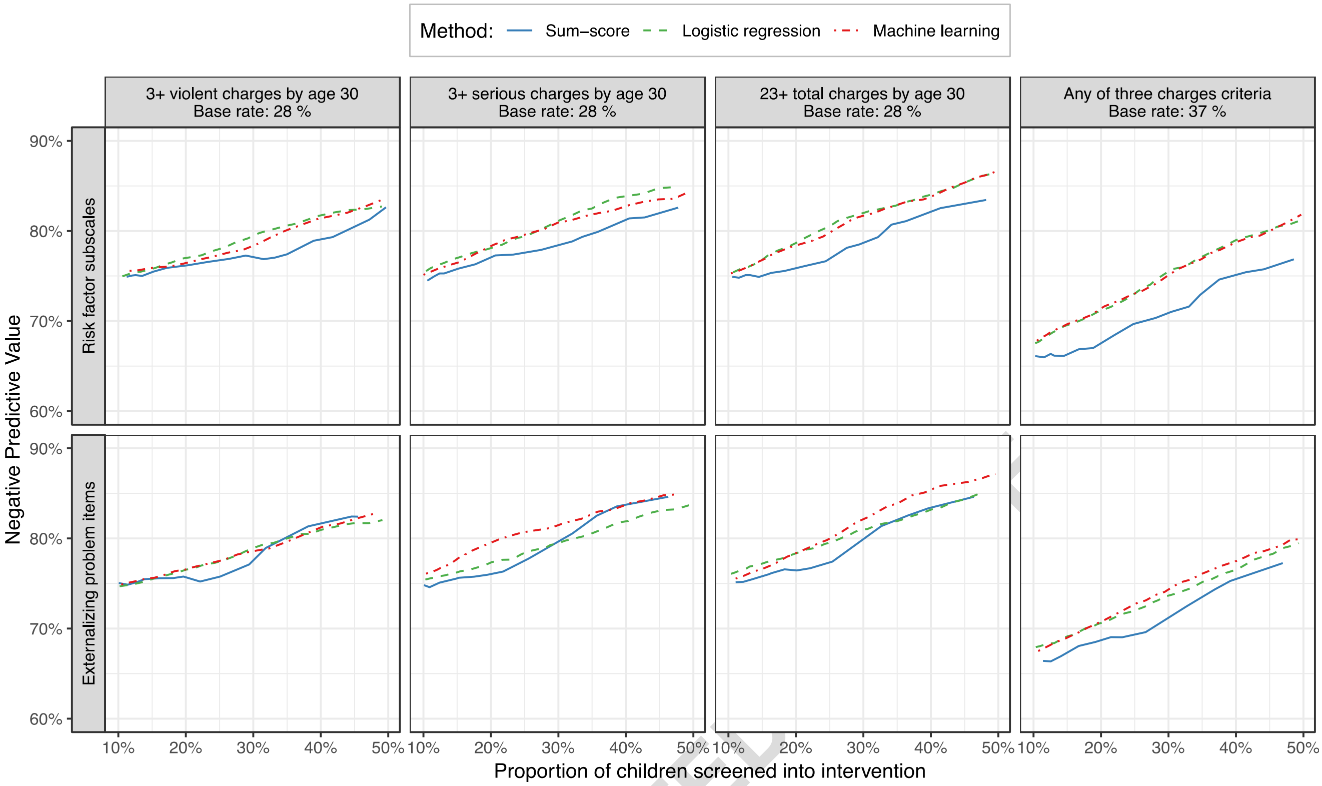

Fig. 2 Negative predictive value (NPV) across screening thresholds, per repeated 10-fold cross-validation in training data. Values calculated via repeated 10-fold cross-validation in training data. Negative predictive value is the probability that when the model predicts a child will not go on to exhibit the delinquency outcome, the child will in fact not exhibit that outcome. Graph shows negative predictive value achieved by each screening method across range of proportion of kids screened into intervention. Results separated by predictor set (rows) and outcome to be predicted (columns). Each line is constructed by calculating the positive predictive value and proportion predicted to be positive across a range of possible cutpoints in the risk score produced by the method (i.e., the sum score or the predicted probability)

(e.g., aggression and interpersonal callousness) when forming a risk score with the sum-score method. In contrast, the TRF externalizing problem items were all assessing the primary risk domain of disruptive behavior problems, so permitting differential weighting may have had less impact on screening performance. Thus, logistic regression or machine learning methods for screening may confer benefit beyond sum-score approaches when using subscales measuring multiple different risk factors, but not when using item-level information from a composite scale assessing externalizing problems.

However, the advantage of the more complex methods was modest and not universally present. The mean increase in AUROC was approximately 0.04 (Table 1). The increase in positive predictive value was in some cases substantial (e.g., of 15–20% in upper, rightmost panel of Fig. 1), but only across part of the range of potential cutpoints, and only with certain combinations of predictor set and outcome to be predicted. There was no substantial increase in negative predictive value. Thus, whether these methods' improved screening performance justifies their increased difficulty of implementation would depend on the specific screening situation at hand. Important factors would include the relative cost of false positives and false negatives and whether the data needed to develop such an algorithm (i.e., to find the regression coefficients) are in existence or being routinely collected.

Machine Learning vs. Logistic Regression

Although logistic regression provided differential weighting to risk factors, only machine learning permitted complex combinations of non-linear associations and interactive effects between risk factors. Nonetheless, the machine learning algorithms we evaluated did not perform any better than logistic models. This finding is consistent with a recent systematic review of 71 studies comparing clinical prediction models developed in many different fields of medicine (Christodoulou et al. 2019). The authors found that when pooling comparisons at low risk of bias, the mean difference in AUC between logistic regression and the machine learning algorithm was almost exactly zero.

In our study, perhaps the number of risk factors (i.e., predictors) was insufficient to realize the benefit of machine learning. The number of constructs assessed in the current study was limited by reliance on teacher report, and the performance of machine learning may prove superior to logistic methods when considering a broader set of risk factors (e.g., family functioning, neighborhood crime) assessed via multiple informants (e.g., parents, youth). Similarly, it is possible that machine learning would be superior to other methods when predicting offending outcomes measured in a different way (Jo et al. 2018) or at a different point in development.

Perhaps the dataset used in the current study was too small to benefit from machine learning (van der Ploeg et al. 2014). Our effective sample size was 864 cases, and after placing 30% of the cases in the test set, this left approximately 605 cases to fit the model in the training set. This is a large sample size relative to most psychological research, but it is small when compared to the many successful applications of machine learning in technology or administrative databases (e.g., datasets with 50k images, 80k insurance claims). There may have been an insufficient number of cases to reliably recover the non-linear, interactive relationships that machine learning algorithms would (in theory) be better able to model.

Perhaps the simplest explanation for our findings is that additive, weighted effects across risk factors, captures most of the predictable variance in delinquency outcomes assessed via criminal records. In other words, there may not be many strong non-linear, interactive relationships for the machine learning algorithms to recover (van der Ploeg et al. 2016). The fact that simpler models often achieve nearly as good performance as more complicated ones has been documented in a variety of contexts (Hand 2006; Holte 1993; Jamain and Hand 2008). In fact, in this study, a simple sum-score performed as well as both logistic regression and the far more complicated machine learning methods when predicting offending outcomes using the externalizing problem items (Dawes 1979; Wainer 1976).

Although two previous studies have found that machine learning outperformed logistic regression in predicting criminal offending outcomes, these studies differed from the current investigation in several notable ways. Both were samples of adults (mean age > 30 years) that had already been arrested for or convicted of a criminal offense, whereas the current sample was a community sample of children in the 5th grade. Machine learning may be more potent when predicting re-offense among active offenders than predicting later offense among children. Moreover, Kleinberg et al. (2018) developed their machine learning models using a dataset that included more than 200k cases. Thus, the observed advantage may have been explained by the fact that machine learning was able to recover complex interactions than was logistic regression. Neuilly et al. (2011) calculated the predictive performance of methods using the same data used to fit the model, compared with our use of estimates in holdout data. Thus, the observed advantage of machine learning may have resulted from the algorithm overfitting the data more than did the logistic regression. Consistent with this notion, machine learning outperformed logistic regression in the present study when data from the same participants were used to fit and evaluate the model, but this discrepancy disappeared when the models were appropriately cross-validated (see Figs. S1 and S2). This underscores the importance of evaluating the performance of machine learning algorithms using estimates from either internal cross-validation procedures (e.g., 10-fold cross-validation) or a holdout sample (Arlot and Celisse 2010).

Limitations

In addition to the limitations already described, the nature of the PYS sample limits the generalizability of our conclusions. Oversampling for children with conduct problems increased the base rate of delinquency outcomes in this sample. However, the metric we used to compare screening approaches (AUROC) is in theory independent of the base rate, and sensitivity analyses confirmed that our screening algorithms performed similarly within the lower risk and higher risk portions of the sample (see supplement). In addition, we have no reason to believe that the sampling scheme would differentially impact the sum-score, logistic regression, or machine learning methods, leaving their *relative* screening performance unaffected.

The PYS sample consisted entirely of boys, and perhaps machine learning would outperform other screening methods in a mixed-gender sample (e.g., by permitting gender by risk factor interactions). There was only one assessment during elementary school that overlapped for both cohorts (5th grade), and the relative performance of screening methods may vary at different developmental stages (e.g., during early childhood or adolescence). Finally, the sample was drawn from boys attending school in one urban city, which might affect the distributions on risk factors in such a way as to attenuate or accentuate differences among screening methods.

Conclusions

How can logistic regression or machine learning approaches contribute to screening for targeted delinquency prevention? Our results suggest that both approaches may improve screening when a broader set of risk factors are used to generate an overall risk score, but the improvements are modest and situation-dependent. None of the complex machine learning methods we evaluated was superior to simple logistic regression, suggesting the latter is preferred. However, the field needs more studies applying these algorithms in diverse contexts to fully evaluate the potential benefits of machine learning (Dwyer et al. 2018). It will be critical to compare the performance of machine learning models to other methods using appropriate cross-validation procedures as these methods may otherwise produce misleading estimates of predictive accuracy. There remains a clear need for strategies that can improve screening for targeted delinquency prevention, and more work is necessary to determine if machine learning will ultimately be one of those strategies.

Funding This research was supported with NICHD grant R01 HD092094; NIDA grants T32 DA039772, R37 DA009757, and UH2 DA041713; and NIAAA grant F31 AA026768.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Informed consent/assent was obtained from all participants in this study.

References

Achenbach, T. M. (1991). *Manual for the teacher's report form and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40–79.

Babyak, M. A. (2004). What you see may not be what you get: A brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine, 66*, 411–421.

Bureau of Justice Statistics. (2015). *Justice expenditure and exployment extracts, 2012 - Preliminary (no. NCJ 248628)*. U.S. Department of Justice.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., et al. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology, 110*, 12–22.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582.

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

Dishion, T. J., Shaw, D., Connell, A., Gardner, F., et al. (2008). The family check-up with high-risk indigent families: Preventing problem behavior by increasing parents' positive behavior support in early childhood. *Child Development, 79*, 1395–1414.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology, 14*, 91–118.

Federal Bureau of Investigation. (2017). *Uniform crime report: Crime in the United States, 2016*. Department of Justice: Washington D.C..

Foster, E. M., & Jones, D. (2006). Can a costly intervention be cost-effective?: An analysis of violence prevention. *Archives of General Psychiatry, 63*, 1284–1291.

Hand, D. J. (2006). Classifier technology and the illusion. *Statistical Science, 21*, 1–14.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer Science & Business Media.

Hill, L. G., Coie, J. D., Lochman, J. E., & Greenberg, M. T. (2004). Effectiveness of early screening for externalizing problems: Issues of screening accuracy and utility. *Journal of Consulting and Clinical Psychology, 72*, 809–820.

Holte, R. C. (1993). Very simple classification rules perform well on most commonly used datasets. *Machine Learning, 11*, 63–90.

Jamain, A., & Hand, D. J. (2008). Mining supervised classification performance studies: A meta-analytic investigation. *Journal of Classification, 25*, 87–112.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in R*. New York: Springer Science & Business Media.

Jo, B., Findling, R. L., Hastie, T. J., Youngstrom, E. A., Wang, C.-P., Arnold, L. E., et al. (2018). Construction of longitudinal prediction targets using semisupervised learning. *Statistical Methods in Medical Research, 27*, 2674–2693.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics, 133*, 237–293.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. New York: Springer Science & Business Media.

Lochman, J. E., & Wells, K. C. (2004). The coping power program for preadolescent aggressive boys and their parents: Outcome effects at the 1-year follow-up. *Journal of Consulting and Clinical Psychology, 72*, 571–578.

Lochman, J. E., Boxmeyer, C. L., Powell, N. P., Barry, T. D., & Pardini, D. A. (2010). Anger control training for aggressive youths. In J. R. Weisz & A. E. Kazdin (Eds.), *Evidence based psychotherapies for children and adolescents* (2nd ed., pp. 227–242).

Loeber, R., Farrington, D. P., Stouthamer-Loeber, M., & Van Kammen, W. B. (1998). *Antisocial behavior and mental health problems: Explanatory factors in childhood and adolescence*. Mahwah: Lawrence Erlbaum Associates.

Loeber, R., Pardini, D., Homish, D. L., Wei, E. H., Crawford, A. M., Farrington, D. P., et al. (2005). The prediction of violence and homicide in young men. *Journal of Consulting and Clinical Psychology, 73*, 1074.

Loeber, R., Farrington, D. P., Stouthamer-Loeber, M., & White, H. R. (2008). *Violence and serious theft: Development and prediction from childhood to adulthood*. New York: Routledge.

Neuilly, M.-A., Zgoba, K. M., Tita, G. E., & Lee, S. S. (2011). Predicting recidivism in homicide offenders using classification tree analysis. *Homicide Studies, 15*, 154–176.

Pardini, D. A., Byrd, A. L., Hawes, S. W., & Docherty, M. (2018). Unique dispositional precursors to early-onset conduct problems and criminal offending in adulthood. *Journal of the American Academy of Child & Adolescent Psychiatry, 57*, 583–592.e3.

Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. *Journal of Clinical Epidemiology, 49*, 1373–1379.

Petras, H., Buckley, J. A., Leoutsakos, J.-M. S., Stuart, E. A., & Ialongo, N. S. (2013). The use of multiple versus single assessment time points to improve screening accuracy in identifying children at risk for later serious antisocial behavior. *Prevention Science, 14*, 423–436.

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology, 14*, 137.

van der Ploeg, T., Nieboer, D., & Steyerberg, E. W. (2016). Modern modeling techniques had limited external validity in predicting mortality from traumatic brain injury. *Journal of Clinical Epidemiology, 78*, 83–89.

Verhulst, F. C., Koot, H. M., & Van der Ende, J. (1994). Differential predictive value of parents' and teachers' reports of children's problem behaviors: A longitudinal study. *Journal of Abnormal Child Psychology, 22*, 531–546.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*, 213–217.

Wilson, S. J., & Lipsey, M. W. (2007). School-based interventions for aggressive and disruptive behavior: Update of a meta-analysis. *American Journal of Preventive Medicine, 33*, S130–S143.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691617693393.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.