J Exp Criminol (2015) 11:299–318
DOI 10.1007/s11292-014-9221-8

Isolating modeling effects in offender risk assessment

Zachary Hamilton · Melanie-Angela Neuilly · Stephen Lee · Robert Barnoski

Published online: 13 November 2014

Abstract
Objectives Recent evolutions in actuarial research have revealed the potential increased utility of machine learning and data-mining strategies to develop statistical models such as classification/decision-tree analysis and neural networks, which are said to mimic the decision-making of practitioners. The current article compares such actuarial modeling methods with a traditional logistic regression risk-assessment development approach.
Methods Utilizing a large purposive sample of Washington State offenders (N= 297,600), the current study examines and compares the predictive validity of the currently used Washington State Static Risk Assessment (SRA) instrument to classification tree analysis/random forest and neural network models.
Results Overall findings varied, being dependent on the outcome of interest, with the best model for each method resulting in AUCs ranging from 0.732 to 0.762. Findings reveal some predictive performance improvements with advanced machine-learning methodologies, yet the logistic regression models demonstrate comparable predictive performance.
Conclusions The study concluded that while data-mining techniques hold potential for improvements over traditional methods, regression-based models demonstrate comparable, and often improved, prediction performance with noted parsimony and greater interpretability.

Z. Hamilton (✉)
Department of Criminal Justice and Criminology, Washington State University, SAC 403K, Spokane, WA, USA
e-mail: zachary.hamilton@wsu.edu

M.-A. Neuilly
Department of Criminal Justice and Criminology, Washington State University, 701 Johnson Tower, Pullman, WA, USA

S. Lee
Bioinformatics and Computational Biology, University of Idaho, Brink Hall 412, Moscow, ID, USA

R. Barnoski
Department of Criminal Justice and Criminology, Washington State University, Spokane, WA, USA

Keywords Random forest · Neural network · Recidivism · Risk assessment

Introduction

With their seminal articles in 1990, Andrews and colleagues established the correctional Principles of Risk, Need, and Responsivity (Andrews et al. 1990a, b). As the discussion on predictors of risks and needs has grown, tools created to assess and forecast offender recidivism have progressed. Although most correctional agencies utilize some form of actuarial assessment to assist with supervision practices, nearly all general[1] offender assessments use some form of regression-based methodology in their development.

Recently, more specialized applications of offender assessments have been developed making use of data-mining techniques (Berk et al. 2009; Schaffer et al. 2011; Steadman et al. 2000). While some have claimed the superiority of prognostic methods (Breiman 2001a), supportive examples within criminal justice settings are rare (Berk et al. 2009; Liu et al. 2011) and as such, the comparative performance within the field is relatively unexplored. Generally, when the concurrent validity of assessments has been examined, efforts have surrounded improvements of one instrument over another with multiple methodological variations existing as possible causes for predictive improvements, such as: use of dynamic items, analytic versus Burgess weighting, and/or outcome type and duration (see Austin et al. 2003; Barnoski and Aos 2003).

Unfortunately, the development of offender risk assessment instruments rarely provides for experimental methodologies, making for a somewhat clouded understanding of optimal methods. The current study seeks to reach experimental rigor by isolating the effects of one prevalent risk assessment scale creation method (logistic regression) and two contemporary data-mining methods (neural networks and random forests). This provision of a controlled condition in which we isolate the impact of varying statistical methods is used to determine the adequacy of an aforementioned debate centered on examining hypotheses of incremental performance improvement found in non-regression data-mining methods.

Review of the literature

The field's notable foray into offender assessments is typically identified as Don Andrew's creation of the Level of Service Inventory (LSI). Developed to guide probation case managers, the LSI consisted of a "single-sheet inventory with 62 'zero—one' items which would fit in officers' case-books," (Andrews 1982, p. 3). Since this time, a variety of offender-assessment instruments have been developed, with the vast majority of U.S. jurisdictions making use of one of the following assessments: the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) (Brennan et al. 2009a, b), the Static Risk Assessment (SRA) (Barnoski and Drake

[1] We use the term "general" offender recidivism assessments to draw a distinction between those used for a correctional offender population and those used for specific populations, namely: sex offenders, psychopaths, and the mentally ill.

2007), the Ohio Risk Assessment System (ORAS) (Latessa et al. 2009), and the Women's Risk Need Assessment (WRNA) (Van Voorhis et al. 2010), along with updated versions of LSI, namely the LSI-R and LS/CMI (Andrews 1995; Andrews and Bonta 1995). While each provide wide variations in item selection and assessments of validity, all utilize some form of regression-based modeling for instrument development.

One of the primary goals of risk prediction is to create an instrument with high predictive performance. The predictive powers of assessments are based on the ability to identify true positives and true negatives—recidivism versus no recidivism. Some have discussed the difficulty with improving our predictions and reduced optimism exists for identifying new ways to "build a better mouse trap" (Gottfredson and Moriarty 2006). While a fair amount of attention has been paid to four generations of assessments via the adoption of additional item types (i.e., static, dynamic, and responsivity) (Andrews et al. 2006), others have described a possible "fifth generation" utilizing non-regression methodologies (Schaffer et al. 2011).

Machine learning and prediction

Machine learning is a subfield of computer science and devises algorithms to identify data patterns (Breiman 2001a; Wasserman 2014). Contrary to classic statistics approaches, machine learning is focused on prediction rather than interpretation. Without assuming a data distribution, algorithms seek out the best predictors (Breiman 2001a). While these approaches have been slowly embraced by statisticians, there are still very much "two cultures" (Breiman 2001a; Wasserman 2014).

It is often without a good understanding of this epistemological divide that criminologists have become interested in the applicability of these methods for recidivism prediction. Two broad types of techniques have been of particular interest: decision trees (i.e., classification tree analysis, as well as random forests), and artificial neural networks (Berk et al. 2009; Brodzinski et al. 1994; Caulkins et al. 1996; Gardner et al. 1996; Grann and Langstrom 2007; Monahan et al. 2000; Palocsay et al. 2000; Silver et al. 2000; Stalans et al. 2004; Steadman et al. 2000; Thomas et al. 2005; Tollenaar and van der Heijden 2013). Although additional methods exist, our current interest is focused on examining those techniques that have, at the time of writing this piece, garnered the most attention—decision trees/random forests, and neural networks, and comparing them to the traditional logistic regression.

Decision trees and random forests

Decision trees are non-parametric question-decision models, which use predictors to split the data into homogeneous groups through a series of conditioning answers (Breiman et al. 1984; Liu, et al. 2011). The data are first split into two groups using the best possible predictor, then further split using the next best predictor for that individual grouping. This process is repeated, leading to more and more homogeneous groups, until a stopping rule is reached (Harper 2005). Decision trees are limited in that they do not provide assessments of statistical

significance or strength of association and they are unstable due to overfitting[2] (Berk et al. 2009; Breiman 1996).

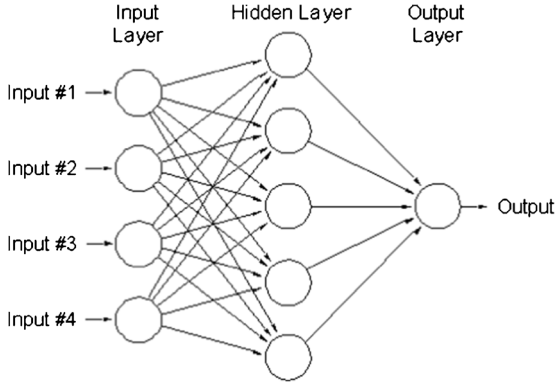
Random forests (RF) (Breiman 2001b) are used to stabilize decision trees (Berk et al. 2009; Neuilly et al. 2011) and consist of the amalgamation of multiple trees, randomly drawn from the same data using bootstrapping (Breiman 2001b). The forest leads to identifying an "average" tree, used for prediction purposes. The forest is first fitted on a training subset of the data, and its performance is tested on a validation subset (Liu et al. 2011; Tollenaar and van der Heijden 2013).

Studies have consistently shown no real difference in predictive performance between regression models of risk assessment and decision trees and RFs (see Appendix Table 3). A variety of study limitations, such as sample size, statistical power, and validation procedures, however, have provided rationales for underwhelming findings (Banks et al. 2004; Gardner et al. 1996; Monahan et al. 2006; Neuilly et al. 2011; Rosenfeld and Lewis 2005; Stalans et al. 2004; Thomas et al. 2005). Nonetheless, it is important to note that RFs do possess some very attractive qualities, including overfit reduction (Breiman 2001b), ability to identify interactions, and tuning for relative costs of false positives/false negatives (Berk et al. 2009).

Neural networks

NNs are a family of models that provide a flexible way to generalize linear regression and logistic regressions (Bishop 1995; Haykin 1999; Hertz et al. 1990; Ripley 1996; Smith 1993; Wasserman 1993). As described, a NN consists of interconnected hidden levels of artificial neurons, and it processes information using a connectionist approach to computation. In most cases, a NN is an adaptive system that changes its structure during a learning phase. NNs are used to model complex relationships between inputs and outputs or to find patterns in data.

To illustrate, we describe a simple but most common form with one hidden layer and one output unit as shown below.



[2] Overfitting is a term used to indicate that a model is trained too closely to the development (construction) sample and loses predictive accuracy when applied to additional (validation) samples.

Image retrieved from http://www.cs.bgu.ac.il/~icbv061/StudentProjects/ICBV061/ICBV-2006-1-TorIvry-ShaharMichal/index.php on 01/31/2013

The input units distribute their information to the hidden neurons (units) in the second layer. These hidden units (h =1 to 5) sum the input units (i =1 to 4) after suitable weights w_{ih}, add a constant w_{ih}, (the bias) and take a fixed activation function φ_h of the result. The output unit is the same form, but with output activation function φ_0. Thus

$$f(x) = \varphi_0\left(w_0 + \sum_{h=1}^{5} w_h * \varphi_h\left(w_{0h} + \sum_{i=1}^{4} w_{ih}x_i\right)\right)$$

The activation function $\varphi's$ of the hidden layer is almost always the logistic function

$$\varphi(z) = \frac{\exp(z)}{1+\exp(z)}$$

and the activation function φ^' s of the output is also logistic when y is binary, and a linear function when y is continuous. The weights w's are the model parameters, which are estimated by minimizing an appropriate fitting criterion between the NN output f(x) and the response variable y, for example, the total deviance when y is binary and the least squares when y is continuous. The number of parameters equals the number of connecting arcs between all units plus the bias terms. The number of weights increases rapidly with the number of hidden layers. It has been shown that NNs models can approximate any continuous function f(x) uniformly in high-dimensional space simply by increasing the number of hidden layers (Ripley 1996).

Much like RFs, NNs present advantages and disadvantages. NNs do not require distribution assumptions, fit variables at any measurement level and identify complex nonlinear relationships. In terms of model flexibility, NNs are an extension to logistic regression. The price to pay for model flexibility is model interpretability. The model parameters in NNs are numerous and findings may lack face validity. For this reason, NNs are considered as a 'black-box' prediction model. Few studies have used NNs to forecast recidivism but overall, results are positive, yet underwhelming (see Appendix I). Comparisons of NNs and other methodologies are mixed, demonstrating negligible (Caulkins et al. 1996) to small performance improvements (Brodzinski et al. 1994; Liu et al. 2011; Palocsay et al. 2000). Similar to RFs, proper validation strategies were sometimes lacking.

Logistic regression

Logistic regression is a standard technique of choice for binary-dependent variables. Probability of occurrence P(y=1) is modeled given the p variables x_i using an S-shaped curve. The classical logistic model uses the form

$$\log\left(\frac{P(y=1)}{P(y=0)}\right) = \beta_0 + \sum_{i=1}^{p} (\beta_0 + \beta_i x_i)$$

where model parameters β's are estimated by minimizing differences between the observed versus the predicted y, that is, P(y=1). As logistic regression is considered a generalized model, the link function uses a natural log. The β estimates the change in the log odds of Y for one-unit changes in X.

Logistic regression covariates identify the unique contribution of each measure included in the model. In prediction models, beta estimates are used to create weights, from which instrument scoring is derived. The anticipated disadvantage of logistic regression is that each measure is taken in isolation, or controlling for the other measures in the model. By only examining the unique contribution of each item, the shared variance and interactive combinations of predictors are muted (Steadman et al. 2000).

Present study

The current study sought to identify the incremental predictive performance provided by non-regression modeling techniques. We hypothesize a substantive variation in predictive performance when comparing modeling strengths of logistic regression, RF, and NN estimations. Specifically, we hypothesize that NN will perform better than RF, which, in turn, will perform better than logistic regression. Additionally, we make said comparisons of predictive validity while providing methodological advances from previous studies and focusing on increased criterion validity (Liu et al. 2011; Tollenaar and van der Heijden 2013), specifically using one of the largest samples ever assembled to examine risk model prediction (N=297,600), a fixed 3-year follow-up, previously validated predictor items with known strength of specificity and predictive power (Barnoski and Drake 2007), a heterogeneous sample (including females), and comparison including one broadband (felony) and three narrowband models (violent, drug, and sex reconvictions). Additionally, models were fine-tuned to prevent over-fitting, a common characteristic of data-mining procedures.

Although their implementation has the potential to touch nearly every aspect of the correctional system, the creation, techniques, and utility discussions of offender risk assessments have been restricted to a select group of instrument developers (e.g., Andrews and Bonta 1995; Baird 1981; Barnoski and Drake 2007; Brennan and Oliver 2000; Duwe 2013; Hare 1991; Latessa et al. 2009). Due to variations such as base rates, item types, jurisdiction distinctions, item selection, weighting/scoring, and validation procedures, meta-analytic and head-to-head instrument comparisons have focused discussions on which instrument is better (Barnoski and Drake 2007; Brennan et al. 2009a, b; Skeem and Louden 2007; Smith et al. 2009) and rarely provide empirical findings as to predictive performance of instrument development methods. This study provides a methodologically rigorous examination of criterion validity, comparing non-regression recidivism prediction methods. Our focus here is not on item selection, but rather on modeling, our aim being to isolate modeling effects from item-selection effects, in order to promote further experimentation based on instrument variation.

Methodology

Sample frame

In Washington State, the Department of Corrections (WADOC) oversees sentences for felony offenders and gross misdemeanants. We selected a purposive sample of offenders reentering the community following a WADOC jurisdiction conviction and having been assessed using Washington State's Static Risk Assessment (SRA). The SRA is generated automatically using 23 post-sentence static risk factors. Although many have argued that the inclusion of dynamic items may improve predictive performance (Cottle et al. 2001; Jung and Rawana 1999; Loeber and Farrington 1998), and we do not disagree, the use of a static items assists in the isolation of methodological performance differences, eliminating issues related to staff training, fidelity, and interrater reliability. Although implemented in 2008, the WADOC automated SRA calculations, providing back-dated assessments including all reentering offenders, begin in July of 1986 through January of 2008 (N=297,600).

Measures

The 23 SRA item weights are publicly available (see Barnoski and Drake 2007). To provide a direct comparison between prediction models, the original operationalization of items was utilized. Items used to compute the SRA include two demographic (age group and male gender) and 21 criminal history measures. Univariate sample descriptives are provided in Table 1. It is noteworthy that the sample's Race/Ethnicity frequencies are included in the table but not used in the calculation of risk. The primary outcome predicted is "felony conviction," indicating any felony conviction in the three years following reentry. Additional specified felony predictions for violent, drug, and sex convictions were also assessed. All outcome measures were dichotomously coded (0/1).

Analytic plan

To examine predictive performance, a validation approach was used for each of the four recidivism models—any felony, violent felony, felony drug, and felony sex convictions—and for each of the three prediction methods, binary logistic regression, RFs, and NNs. Logistic regression models were computed using the predefined items weights outlined for the SRA (see Barnoski 2010). The number of hidden units for neural networks and the number of trees for random forest model were optimized for each of the four recidivism models.

The bootstrapping validation procedures were completed using a modified process first described by Harrell et al. (1996). First, 100 bootstrap samples are drawn, where eligible cases are selected with replacement until the bootstrap sample reaches the original sample size. Subjects selected are considered the "in boot" sample while those not selected are the "out of boot" sample. The "in boot" represents the construction sample while the "out of boot" subjects

Table 1 Univariate sample descriptives

Item	Frequency
Race/ethnicity (not used in model calculation)	
White	79.7
Black	17.0
Other	3.3
Age	
13 to 17	0.2
18 to 19	7.6
20 to 29	40.4
30 to 39	30.4
40 to 49	16.0
50 to 59	4.2
60 or older	1.3
Male	81.3
Prior juvenile convictions	
0	85.2
1	6.8
2+	3.4
3	2.0
4	1.2
5+	1.5
Prior juvenile non-sex violent conviction	
0	95.2
1	3.8
2+	0.9
3	2.0
4	1.2
5+	1.5
Total felonies	
0	7.9
1	46.4
2	19.0
3	10.4
4	6.2
5+	10.1
Felony homicide offense: Murder/Manslaughter	1.3
Felony sex offense	
0	86.0
1	12.6
2	1.2
3+	0.2

Table 1 (continued)

Item	Frequency
Felony violent property	
0	91.5
1	7.5
2+	0.9
Felony assault—Not domestic violence related	
0	86.0
1	12.6
2	1.2
3+	0.2
Felony domestic violence assault or related offense	
0	98.4
1	1.5
2+	0.2
Felony weapon	
0	95.9
1	3.7
2+	0.3
Felony property offense	
0	52.7
1	27.9
2	9.5
3	4.6
4	2.4
5+	2.9
Felony drug offense	
0	58.5
1	26.9
2	8.4
3+	6.3
Felony Escape	8.5
Misdemeanor assault offense—Not domestic violence related	
0	84.1
1	11.5
2	2.9
3	1.0
4	0.3
5+	0.2
Misdemeanor domestic violence assault or violation	
0	85.6
1	8.2

Table 1 (continued)

Item	Frequency
2+	6.2
Misdemeanor sex offense	
0	97.4
1	1.8
2+	0.8
Misdemeanor other domestic violence	1.2
Misdemeanor weapons offense	3.7
Misdemeanor property offense	
0	70.3
1	15.6
2	6.0
3+	8.1
Misdemeanor drug offense	
0	86.7
1	9.6
2+	3.8
Misdemeanor escape	0.8
Misdemeanor alcohol offense	16.8
Outcomes (Three-year follow-up)	
Any felony conviction	31.5
Felony property conviction	12.2
Felony drug conviction	9.7
Violent felony conviction	8.6

represent the validation sample. Model performance criteria are computed on the "out of boot" samples for a given model.

Performance criteria

To compare methods, we relied on multiple criteria to provide a comprehensive examination of predictive performance. There are three areas of performance used for evaluation purposes—discrimination, calibration, and accuracy. To evaluate *discrimination* we utilize the common AUC metric, which is the area under the receiver operating characteristic (ROC) curve. The ROC curve plots the fraction of true positives out of the observed positives (i.e., true-positive rate) versus the fraction of false positives out of the total negatives (i.e., false-positive rate), at various threshold settings. True-positive rate is also known as *Sensitivity* or *recall* in machine learning. The false-positive rate can be calculated as one minus the *Specificity*. An alternate measure on discrimination is Youden's J, which is *Sensitivity* plus *Specificity* minus one. Lastly, the cross-

entropy (CE) is a discrimination measure of 'distance' between the predicted probabilities distribution versus the observed outcome's distribution; it ranges from zero-to-one, where smaller values indicate improved prediction.

Calibration examines the relationship between the predicted probabilities and the observed outcomes. We examine the overall calibration error (CALerr), also called the 'slope,' which is the difference between the expected probability and the proportion of the observed outcome. Calibration is also measured by the root mean squared error (RMSE), which for the current study represents the square root of the average squared discrepancies between the predicted probabilities and observed values.

Accuracy (ACC) is the proportion of cases correctly classified. A threshold, or cut point, is required and for methods comparison, we used the common base rate. Two additional accuracy statistics for binary classification are the F_1-score, which is the harmonic mean of recall and precision, and Matthews correlation coefficient (MCC), which is the correlation of the predicted probabilities and observed values. A combined measure of discrimination, calibration, and accuracy, the SAR, is computed as: (AUC + ACC +1 – RMSE)/3. Finally, 95% confidence intervals (CI) were computed for each of the aforementioned statistics and are used to determine statistical significance between model estimates.[3]

Results

Results of the comparisons are provided in Table 2. As smaller values are indicators of better performance for some criteria, the best value for each method is bolded within each of the four outcome types. Overall, the logistic regression models provided the best (or tied for the best) performance in 33 of the 36 comparison. With regard to discrimination, the differences appear to be smaller for broadband felony (AUC difference =0.01) versus any of the three narrowband offense types, with AUC difference ranging from 5 to 15 %. For measures of accuracy, the differences are much smaller, finding ACC differences of 0.01 or none at all. Calibration differences were also negligible, with logistic regression models taking top position across all four models, while RF models identified slightly greater slope values. Additional measures (maximum entropy, F_1-score, Matthews coefficient, and Youden's J) all identified logistic regression models to possess greater performance over NNs and RFs. Finally, the combined measure of discrimination, accuracy, and calibration (SAR) also identified logistic regression models to have better performance, although, again, the differences were minor.

[3] It should be noted that with 100 bootstrap draws, the 95 % CI for each performance measure is calculated as (m-1.96*SD/ √(100), m+1.96*SD/ √(100)). For presentation purposes, CIs are not included with model results but may be obtained by contacting the corresponding author.

Table 2 Comparison of logistic regression, neural network, and random forest models (N=297,600)

	Felony recidivism			Drug recidivism			Violent recidivism			Sex recidivism		
Performance	LR (SD)	NN (SD)	RF (SD)	LR (SD)	NN (SD)	RF (SD)	LR (SD)	NN (SD)	RF (SD)	LR (SD)	NN (SD)	RF (SD)
AUC	0.74 (0.00)	0.73 (0.01)	0.73 (0.00)	0.75 (0.00)	0.73 (0.02)	0.70 (0.00)	0.76 (0.00)	0.73 (0.03)	0.71 (0.00)	0.77 (0.00)	0.70 (0.04)	0.62 (0.01)
ACC	0.73 (0.00)	0.73 (0.01)	0.72 (0.00)	0.92 (0.00)	0.93 (0.01)	0.92 (0.00)	0.92 (0.00)	0.92 (0.01)	0.92 (0.00)	0.99 (0.00)	0.99 (0.00)	0.99 (0.00)
RMSE	0.42 0.00)	0.42 (0.01)	0.43 (0.00)	0.21 (0.00)	0.22 (0.02)	0.22 (0.00)	0.21 (0.00)	0.21 (0.00)	0.22 (0.00)	0.07 (0.00)	0.07 (0.00)	0.07 (0.00)
CALerr (slope)	0.15 (0.00)	0.16 (0.02)	0.17 (0.00)	0.06 (0.00)	0.05 (0.01)	0.06 (0.00)	0.06 (0.00)	0.05 (0.01)	0.07 (0.00)	0.01 (0.00)	0.01 (0.00)	0.01 (0.00)
MXE	0.53 (0.00)	0.72 (0.86)	0.61 (0.01)	0.18 (0.00)	0.45 (0.51)	0.65 (0.02)	0.18 (0.00)	0.29 (033)	0.64 (0.02)	0.03 (0.00)	0.08 (0.10)	0.32 (0.01)
F_1-score	0.54 (0.00)	0.53 (0.05)	0.52 (0.00)	0.22 (0.00)	0.16 (0.04)	0.17 (0.00)	0.22 (0.00)	0.19 (0.03)	0.18 (0.00)	0.05 (0.00)	0.04 (0.01)	0.03 (0.00)
MCC	0.35 (0.00)	0.34 (0.02)	0.32 (0.00)	0.18 (0.00)	0.15 (0.04)	0.13 (0.00)	0.18 (0.00)	0.16 (0.02)	0.13 (0.00)	0.05 (0.00)	0.04 (0.01)	0.02 (0.00)
Youden's J	0.35 (0.00)	0.33 (0.05)	0.32 (0.00)	0.18 (0.00)	0.11 (0.04)	0.13 (0.00)	0.18 (0.00)	0.15 (0.03)	0.13 (0.00)	0.05 (0.00)	0.04 (0.01)	0.02 (0.00)
SAR	0.68 (0.00)	0.68 (0.00)	0.67 (0.00)	0.82 (0.00)	0.68 (0.00)	0.80 (0.00)	0.82 (0.00)	0.82 (0.01)	0.80 (0.00)	0.89 (0.00)	0.88 (0.01)	0.85 (0.00)

With those comparisons in mind, the difference between logistic regression, RF, and NN models is estimated to be moderate to small in magnitude. Statistical significance was also identified when one model's performance demonstrated a non-overlapping CI improvement over the two comparison models. Examining the nine model comparison statistics, logistic regression demonstrated a significantly improved performance over NNs and RFs in five of estimates of the general felony recidivism model, seven estimates in the drug recidivism model, six estimates in the violent recidivism model, and six estimates in the sex recidivism model. Thus, although not demonstrating universal advantages across all models and measures, the vast majority of comparisons and tests indicated logistic regression methods provided improved predictive performance.

Discussion

In a time when quantitative methods are viewed as potential solutions to complex problems, it is understandable to infer that greater sophistication will lead to improved prediction. The current study sought to isolate the effects of various risk-assessment models and test the hypothesis that non-regression methods would lead to improved predictive validity. As a result, we *do not* find evidence to make that claim, and in contrast, find regression methods performed better by comparison on the majority of performance criteria. However, there are far more performance similarities among the three methods, where many performance indicators revealed less than substantive differences between the three model types. Furthermore, we do not claim that a "fifth-generation" tool cannot eventually be achieved through more advanced methodologies. We only contend that the application of non-regressive data-mining techniques for offender assessment is not 'limitless' and greater development and understanding of their impact is needed.

Due to the relative similarities in performance, at this time we do not believe there is a substantive reason to choose more complex prediction models over logistic regression for offender populations. Although some argue that interactions and non-additive model solutions are ignored (Steadman et al. 2000), in terms of performance, we find logistic regression models to be comparable, and more often provide improved performance, than NN and RF. Furthermore, as the advantage of model interpretability is often paramount, we do not recommend adding unnecessary levels of modeling complexity associated with data-mining techniques that may inhibit practitioner trustworthiness and use of the tool.

Previously we described several substantial methodological advances of the current study. These stated sample and measurement-related strengths made it possible to avoid pitfalls to which many of the previous studies have fallen prey and, in turn, provide a stronger test of the incremental improvement of prediction methods. Specifically, we were not faced with the need to collapse categories of offenses or make statistical adjustments to further increase the base rate of offending as done in previous comparisons (Brodzinski et al. 1994; Caulkins et al. 1996; Palocsay et al. 2000; Silver and Chow-Martin 2002;

Silver et al. 2000). We were also able to validate each of our models while more adequately controlling for shrinkage, which has often been problematic (Banks et al. 2004; Brodzinski et al. 1994; Monahan et al. 2000, 2005, 2006; Silver and Chow-Martin 2002; Silver et al. 2000; Steadman et al. 2000). Although not a requirement of future analyses, the utility of a static risk assessment created a large study sample with *fewer moving parts*, allowing for a better experimental isolation of the compared methodological variations. As a greater variety of offender assessments have been created in the last decade, arguments are being waged as to which instrument uses the *correct set of methods* to maximize prediction. Rather than head-to-head instruments comparisons, we argue for the application of experimental isolation of potential methodological variations.

Limitations

The current study is not without shortcomings. First, we did not select predictor items with the comparison of methods in mind. A rigorous investigation was previously conducted, selecting measures that collectively contribute to the prediction of felony convictions using logistic regression models (see Barnoski and Drake 2007). In this way, the measures used for the methods comparisons were optimized for a binary logistic regression and the item operationalizations are thought to provide a slight advantage for said modeling procedure. Thus, measures utilized may have contributed to relatively comparable performances of the three model types. Instead it may be advantageous for subsequent method comparisons to start with item selection from a pool of measures to potentially maximize each method's performance. Finally, these data-mining procedures may be better apt at identifying patterns among a more homogeneous and/or rarer base rate population, where patterns of subject responses can be established among known categorizations of offender specializations.

Conclusions

Much of the discussion surrounding risk assessment improvements revolves around identifying better items, improving measurement scaling, and properly assigning weights. While all of this is necessary, we also believe it worthwhile to sometimes take a step back and examine the more mechanical underpinnings of the risk assessment machinery, and run confirmatory experimental tests which examine statistical methods as stimuli, focusing their relative performance. As a result, a case can be made for the continued use of regression-based modeling techniques for offender risk-scale development, and yet the door is left open to further research aimed at the continuing investigation of avenues for improvement. Using the current performance of each prediction method as a base, future research should indeed focus on adding elements of variation and investigating whether each method's performance can be increased by modifying item selection, methodological tuning, etc. The present report can thus serve as springboard for future experimentation.

Appendix I

Table 3 Recapitulation of previous research comparing linear and non-linear risk-assessment tools

Study	Methods compared	Sample size	Population	Dependent variable	Follow-up period	Predictors	Risk categories	Performance measures	Validation strategy
Berk et al. 2009	Random forests/ Logistic regression	Over 66,000	Adult probation and parole	Homicide	2 years	Age Gender Race Prior record Current offense Neighborhood information Age at first charge	Homicide versus no homicide	Training sample Error rate Random forest 0.07 LR 0.01	Validation sample Error rate Random forest 0.07
Brodzinski et al. 1994	NN/ Discriminant analysis	778	Juvenile probation cases	Recidivism	Unknown	Age at first adjudication School history Substance abuse Family criminal history Peer relations Prior adjudications for unruliness Delinquency Probation violations	Recidivist versus Non recidivist	Construction sample Accuracy DA 64.36 %	Validation sample Accuracy NN 99.48 % DA 62.63 %
Caulkins et al. 1996	NN/ Multiple regression/ Association analysis/ Predictive attribute analysis	3,389	Offenders released from prison	Recidivism	2 years	Offense Criminal history Social history Institutional adjustment	Recidivist versus Non-recidivist	Construction sample MCR values NN 0.460 Other models 0.338–0.440	Validation sample MCR values NN 0.416 Other models 0.328–0.436
Gardner et al. 1996	CART/ Negative binomial regression	784	Psychiatric patients	Count of community violence	4 months	Demographics Clinical diagnosis Compliance with medication Brief symptom inventory	Cutpoints	Sensitivity CART 7.7 % NBR 9.3 % Specificity CART 99.2 % NBR 99.1 %	Bootstrapping

Table 3 (continued)

Study	Methods compared	Sample size	Population	Dependent variable	Follow-up period	Predictors	Risk categories	Performance measures	Validation strategy
Liu et al. 2011	CART/ NN/ Logistic regression	1,225	Offenders released from prison	Violent re-conviction	From 1.31 to 4.24 years	HCR-20	Violent versus non-violent reoffense	Training and testing samples AUC values Accuracy LR 0.58–0.68 CART 0.57–0.70 NN 0.59–0.66 Sensitivity LR 0.57–0.73 CART 0.40–0.81 NN 0.59–0.83	Multiple validation samples AUC values Accuracy LR from 0.59 to 0.64 CART 0.57–0.65 NN 0.60–0.67 Sensitivity LR 0.64–0.69 CART 0.33–0.72 NN 0.63–0.73
Neuilly et al. 2011	CTA/ Random forest/ Logistic regression	320	Released homicide offenders	Recidivism	Up to 5 years	Demographics Incident characteristics Prior criminal history Lifestyle	Reoffense versus Parole violation	Classification error rate CTA 12–20 % LR 18 %	Bootstrapping Prediction error rate Random forest 25–50 %
Palocsay et al. 2000	NN/ Logistic regression	10,357	Offenders released from prison	Recidivism	Unknown	Gender Race Age at release Drug and alcohol history Prior records Prior incarcerations Sentence served	Recidivist versus non-recidivist	Multiple training and testing samples Accuracy NN 68.71–71.50 % LR 68.05–68.15 %	Multiple validation samples Accuracy NN 65.96–69.23 % LR 64.29–66.73 %
Rosenfeld and Lewis 2005	CART/ Logistic regression	204	Psychiatric patients	Violent reconviction	Up to 5 years	Demographics Clinical variables Offense-related variables	Low High	Three models each AUC values CART 0.79–0.85 LR 0.78–0.80	Jack-knifed cross-validation AUC values CART 0.64–0.66 LR 0.71–0.74

Table 3 (continued)

Study	Methods compared	Sample size	Population	Dependent variable	Follow-up period	Predictors	Risk categories	Performance measures	Validation strategy
Stalans et al. 2004	CTA/ Logistic regression	1,344	Violent probationers	Violent recidivism	4 weeks	Demographics Types of violent offenders Type of prior violence Frequency of prior arrests Offense characteristics Substance use Mental health Probation conditions Dynamic predictors	Low Medium High	AUC values CTA 0.67 LR 0.71 Sensitivity CTA 35.2 % LR 9.8 % Specificity CTA 88.4 % LR 98.7 %	Bootstrapping for CTA only
Steadman et al. 2000	ICT using CHAID/ Logistic regression	939	Psychiatric patients	Triangulated violence	20 weeks	Clinical factors Historical risk factors	Low High	AUC values Probability ICT 0.82 CHAID 0.79 LR 0.81	Bootstrapping
Thomas et al. 2005	CART/ Logistic regression	708	Psychiatric patients	Counts of physical assault	2 years	Sociodemographic risk factors Historical risk factors Clinical factors	Unknown	Sensitivity CART 29 % LR 19 % Specificity CART 98 % LR 96 % Accuracy CART 82 % LR 78 %	Tenfold cross-validation Sensitivity CART 14 % LR 19 % Specificity CART 93 % LR 94 % Accuracy CART 75 % LR 77 %

References

Andrews, D. A. (1982). *The level of supervision inventory (LSI)*. Ontario Ministry of Correctional Services.

Andrews, D. A. (1995). *No title*. Cited in LS/CMI Manual pp. 117–144.

Andrews, D. A., & Bonta, J. (1995). *The LSI-R: the level of service inventory–revised*. Toronto: Multi-Health Systems.

Andrews, D., Bonta, J., & Hoge, R. (1990a). Classification for effective rehabilitation: rediscovering psychology. *Criminal Justice and Behavior, 17*(1), 19–52.

Andrews, D., Zinger, I., Hodge, R., Bonta, J., Gendreau, P., & Cullen, F. (1990b). Does correctional treatment work? a clinically relevant and psychologically informed meta-analysis. *Criminology, 28*(3), 369–392.

Andrews, D., Bonta, J., & Wormith, S. (2006). The recent past and near future of risk/need assessment. *Crime and Delinquency, 52*(1), 7–27.

Austin, J., Coleman, D., Peyton, J., & Johnson, K.D. (2003). *Reliability and validity study of the LSI-R risk assessment instrument*. Washington, D.C. Institute on Crime, Justice, and Corrections at The George Washington University.

Baird, C. S. (1981). Probation and parole classification: the Wisconsin model. *Corrections Today, 43*, 36–41.

Banks, S., Robbins, P. C., Silver, E., Vesselinov, R., Steadman, H. J., Monahan, J., Mulvey, E. P., Appelbaum, P. S., Grisso, T., & Roth, L. H. (2004). A multiple-models approach to violence risk assessment among people with mental disorder. *Criminal Justice and Behavior, 31*(3), 324–340.

Barnoski, R. (2010) *Washington State static risk assessment—version 2.0*. [Modified to improve reliability and validity, requested by Washington State Center for Court Research].

Barnoski, R., Aos, S. (2003). *Washington's Offender Accountability Act: An Analysis of the Department of Corrections' Risk Assessment*. #03-12-1202.

Barnoski, R., Drake, E. (2007). *Washington's Offender Accountability Act: Department of Corrections' Static Risk Instrument*. #07-03-1201.

Berk, R., Sherman, L., Barnes, G., Kurtz, E., & Ahlman, L. (2009). Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning. *Journal of the Royal Statistical Society A, 172*(Part 1), 191–211.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford: Oxford University Press.

Breiman, L. (1996). Bagging predictors. *Machine Learning Journal, 26*, 123–140.

Breiman, L. (2001a). Statistical modeling: the two cultures. *Statistical Science, 16*(3), 199–215.

Breiman, L. (2001b). Random forests. *Machine Learning Journal, 45*(1), 5–32.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Monterey: Wadsworth and Brooks/Cole.

Brennan, T., & Oliver, W. L. (2000). *Evaluation of reliability and validity of COMPAS scales: national aggregate sample*. Traverse City: Northpointe Institute for Public Management.

Brennan, T., Dieterich, B., Breitenbach, M., & Mattson, B. (2009a). Commentary on NCCD "A questions of evidence: A critique of risk assessment models used in the justice system". Northpointe Institute for Public Management, Inc. http://www.northpointeinc.com/files/whitepapers/Baird_Response_060409.pdf

Brennan, T., Dieterich, W., & Ehret, B. (2009b). Evaluation the predictive validity of the comps risk and needs assessment system. *Criminal Justice and Behavior, 36*(1), 21–40.

Brodzinski, J. D., Crable, E. A., & Scherer, R. F. (1994). Using artificial intelligence to model juvenile recidivism patterns. *Computers in Human Services, 10*(4), 1–18.

Caulkins, J., Cohen, J., Gorr, W., & Wei, J. (1996). Predicting criminal recidivism: a comparison of neural networks with statistical models. *Journal of Criminal Justice, 24*(3), 227–240.

Cottle, C. C., Lee, R. J., & Heilbrun, K. (2001). The prediction of criminal recidivism in juveniles: a meta-analysis. *Criminal Justice and Behavior, 28*(3), 367–394.

Duwe, G. (2013). The development, validity, and reliability of the Minnesota screening tool assessing recidivism risk (MnSTARR). *Criminal Justice Policy Review, XX*, 1–35. doi:10.1177/0887403413478821.

Gardner, W., Lidz, C., Mulvey, E., & Shaw, E. (1996). A comparison of actuarial methods of identifying repetitively violent patients with mental illness. *Law and Human Behavior, 20*(1), 35–48.

Gottfredson, S., & Moriarty, L. (2006). Statistical risk assessment: Old problems and New applications. *Crime and Delinquency, 52*(1), 178–200.

Grann, M., & Langstrom, N. (2007). Actuarial assessment of violent risk: to weigh or Not to weigh. *Criminal Justice and Behavior, 34*(1), 22–36.

Hare, R. (1991). *The revised psychopathy checklist*. Toronto: Multi-Health Systems.

Harper, P. R. (2005). A review and comparison of classification algorithms for medical decision making. *Health Policy, 71*(3), 315–331.

Harrell, F., Lee, K., & Mark, D. (1996). Multivariate prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15*, 361–387.

Haykin, S. (1999). *Neural networks: a comprehensive foundation*. Upper Saddle River: Prentice Hall.

Hertz, J., Palmer, R. G., & Krogh, A. S. (1990). *Introduction to the theory of neural computation*. New York: Perseus Books.

Jung, S., & Rawana, E. P. (1999). Risk-need assessment of juvenile offenders. *Criminal Justice and Behavior, 26*, 69–89.

Latessa, E., Smith, P., Lemke, R., Markarios, M., Lowenkamp, C. (2009) *Creation and Validation of the Ohio Risk Assessment System – Final Report*. Ohio Department of Rehabilitation and Correction. # 2005-JG-E0R-6269 and 2005-JG-C01-T8.

Liu, Y. Y., Yang, M., Ramsay, M., Li, X. S., & Coid, J. W. (2011). A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent Re-offending. *Journal of Quantitative Criminology, 27*(4), 547–573.

Loeber, R., & Farrington, D. (1998). *Serious and violent juvenile offenders: risk factors and successful interventions*. Thousand Oaks: Sage.

Monahan, J., Steadman, H. J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., Silver, E., Roth, L. H., & Grisso, T. (2000). Developing a clinically useful actuarial tool for assessing violence risk. *British Journal of Psychiatry, 176*(4), 312–319.

Monahan, J., Steadman, H. J., Robbins, P. C., Appelbaum, P., Banks, S., & Grisso, T. (2005). An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric Services, 56*(7), 810–815.

Monahan, J., Steadman, H. J., Appelbaum, P. S., Grisso, T., Mulvey, E. P., & Roth, L. H. (2006). The classification of violence risk. *Behavioral Sciences and the Law, 24*(6), 721–730.

Neuilly, M.-A., Zgoba, K. M., Tita, G. E., & Lee, S. (2011). Predicting recidivism in homicide offenders using classification tree analysis. *Homicide Studies, 15*(2), 154–176.

Palocsay, S. W., Wang, P., & Brookshire, R. G. (2000). Predicting criminal recidivism using neural networks. *Socio-Economic Planning Sciences, 34*(4), 271–284.

Ripley, B. D. (1996). *Pattern recognition and neural networks*. New York: Cambridge University Press.

Rosenfeld, B., & Lewis, C. (2005). Assessing violence risk in stalking cases: a regression tree approach. *Law and Human Behavior, 29*(3), 343–357.

Schaffer, D., Kelly, B., & Lieberman, J. (2011). An exemplar-based approach to risk assessment: validating the risk management systems instrument. *Criminal Justice Policy Review, 22*(2), 167–186.

Silver, E., & Chow-Martin, L. (2002). A multiple-models approach to assessing recidivism risk: implications for judicial decision making. *Criminal Justice and Behavior, 29*(5), 538–568.

Silver, E., Smith, W. R., & Banks, S. (2000). Constructing actuarial devices for predicting recidivism: a comparison of methods. *Criminal Justice and Behavior, 27*(6), 733–764.

Skeem, J., & Louden, J. (2007). *Assessment of Evidence on the quality of the correctional offender management profiling for alternative sanctions (COMPAS). Prepared for the California department of corrections and rehabilitation (CDCR)*. CA: Davis.

Smith, M. (1993). *Neural networks for statistical modeling*. New York: Van Nostrand Reinhold.

Smith, P., Cullen, F., & Latessa, E. (2009). Can 14,737 women be wrong? a meta-analysis of the LSI-R and recidivism for female offenders. *Criminology and Public Policy, 8*(1), 183–208.

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), 253–262.

Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Clark Robbins, P., Mulvey, E. P., Grisso, T., Roth, L. H., & Banks, S. (2000). A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior, 24*(1), 83–100.

Thomas, S., Leese, M., Walsh, E., McCrone, P., Moran, P., Burns, T., Creed, F., Tirer, P., & Fahy, T. (2005). A comparison of statistical models in predicting violence in psychotic illness. *Comprehensive Psychiatry, 46*, 296–303.

Tollenaar, N., & van der Heijden, P. G. M. (2013). Which method predicts recidivism best?: a comparison of statistical, machine learning, and data mining predictive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 176*(2), 565–584.

Van Voorhis, P., Wright, E. M., Salisbury, E., & Bauman, A. (2010). Women's risk factors and their contributions to the existing risk/needs assessment: the current status of a gender-responsive supplement. *Criminal Justice and Behavior, 37*, 261–288.

Wasserman, P. (1993). *Advanced methods in neural computing*. New York: Van Nostrand Reinhold.

Wasserman, L. (2014). Rise of the machines. In X. Lin, D. L. Banks, C. Genest, G. Molenberghs, D. W. Scott, & J.-L. Wang (Eds.), *Past, present, and future of statistical science* (pp. 1–12). Boca Raton: CRC Press. Chapter 1.

Zachary K. Hamilton PhD, is a professor of criminal justice and criminology and director of theWashington State Institute of Criminal Justice (WSICJ) at Washington State University. He received his PhD in criminal justice and criminology from Rutgers University. His recent work has focused on offender assessment, supervision, and treatment and has appeared in such journals as Substance Abuse Treatment,Experimental Criminology, and Justice Quarterly. His research interests includequantitative methods, risk assessment, reentry, specialty courts, and offender change programming.

Melanie-Angela Neuilly received a Ph.D. in Criminal Justice from Rutgers University, School of Criminal Justice in May 2007, and a Ph.D. in Psychology from the University of Rennes, France, in December 2008. Since 2006, Dr. Neuilly is employed as an Assistant Professor in the Department of Sociology and Anthropology at the University of Idaho. Her research focuses on violent crime and methodological issues of crime and public health data collection in an international, comparative context. Her primary interest is in homicide and violent death research. Previouspublications have looked at the social construction of the crime of pedophilia in France and in the United States, issues of theoretical foundations of crime scene analysis and criminal profiling, questions of understanding homicide and criminal violence within the broader context of violent deaths and non-lethal injuries, as well as on international drug trafficking.

Stephen Sauchi Lee received his Ph.D. in Statistics from Florida State University, a M.A. in Mathematics from the University of West Florida, and a B.S. in Mathematics from the University of Hong Kong. He joined the Department of Statistics at the University of Idaho in 1993. His research interests are in Multivariate Analysis,Computational Methods, Classification, and Statistical Modeling. They include: Integrating models and methods from statistics, neural networks, machine learning, and data mining communities to discover relationships and recognize patterns in databases; modeling for interpretations and predictions; extracting information and patterns; developing computational algorithms to increase efficiency and prediction accuracy; applying in Analytics, Bioinformatics, and Genomics.

Robert "Barney" Barnoski received his Ph.D. in psychometrics in 1976 from Temple University. Dr. Barnoski has 30 years of experience in applying statistics and research methods to aid in decision-making and policy analysis. For the majority of his career he was employed as a Senior Researcher for the Washington State Institute for Public Policy (WSIPP). During his tenure at WSIPP he developed the Static Risk Assessment(SRA) for adult offenders, the Positive Achievement Change Tool (PACT) for juveniles, and an assessment for family reconciliation services for the Washington State Children's Administration. He is currently an Affiliate Faculty member for Washington State University in the Department of Criminal Justice and Criminology. Visit the Institute's website, www.wsipp.wa.gov, to read the reports written by Barney while at the Institute.