Am J Crim Just (2015) 40:47–74
DOI 10.1007/s12103-014-9246-6

Assessing the Predictive Utility of Logistic Regression, Classification and Regression Tree, Chi-Squared Automatic Interaction Detection, and Neural Network Models in Predicting Inmate Misconduct

Fawn T. Ngo · Ramakrishna Govindu · Anurag Agarwal

Received: 3 February 2014 / Accepted: 7 May 2014 /
Published online: 23 May 2014

Abstract This study assesses the relative utility of a traditional regression approach - logistic regression (LR) - and three classification techniques - classification and regression tree (CART), chi-squared automatic interaction detection (CHAID), and multi-layer perceptron neural network (MLPNN)—in predicting inmate misconduct. The four models were tested using a sample of inmates held in state and federal prisons and predictors derived from the importation model on inmate adaptation. Multi-validation procedure and multiple evaluation indicators were used to evaluate and report the predictive accuracy. The overall accuracy of the four models varied between 0.60 and 0.66 with an overall AUC range of 0.60–0.70. The LR and MLPNN methods performed significantly better than the CART and CHAID techniques at identifying misbehaving inmates and the CHAID method outperformed the CART approach in classifying defied inmates. The MLPNN method performed significantly better than the LR technique in predicting inmate misconduct among the training samples.

Keywords Actuarial risk assessment techniques · Comparative statistical techniques · Logistic regression · Classification and regression tree · Chi-squared automatic interaction detection · Neural networks · Importation model · Inmate misconduct

Introduction

Developing effective models to predict future risks has been a challenging task confronting criminal justice researchers and practitioners. Generally, there are two

F. T. Ngo (✉)
College of Arts and Sciences, University of South Florida Sarasota-Manatee, 8350 N. Tamiami Trail, Sarasota, FL 34243, USA
e-mail: fawnngo@sar.usf.edu

R. Govindu · A. Agarwal
College of Business, University of South Florida Sarasota-Manatee, 8350 N. Tamiami Trail, Sarasota, FL 34243, USA

types of risk assessment instruments employed in criminal justice settings: those based on clinical judgment and those based on actuarial practices (Gottfredson and Gottfredson, 1986). Clinical methods, also known as first-generation assessments or subjective assessments (Bonta, 1996), involves interviews with offenders using unstandardized questions or checkoff lists formulated by professionals to gauge behavioral indicators.[1] Actuarial techniques, also known as second-generation assessments or statistically-based instruments, are grounded in theory and research and factors identified from empirical studies to be related to criminal behavior are employed to develop a predictive model (Bonta, 1996).

Compared to clinical methods, actuarial risk assessment instruments have been demonstrated to be more accurate in focasting future risks (Dawes, Faust, and Meehl, 1989; Gottfredson and Moriarty, 2006; Hanson and Morton-Bourgon, 2007; 2009; Hilton, Harris, and Rice, 2006; Jones, 1996; BUT see Singh, Grann, and Fazel, 2011). However, actuarial risk assessment methods are not without limitations. In particular, critics of these methods have contended that since they are generally based on generalized linear models (i.e., linear or logistic regression), they tend to assume a "one size fits all" approach, with no individual considerations. This assumption essentially ignores the possibility that different factors might apply to different subgroups of individuals when predicting risks (Steadman et al., 2000). Critics of actuarial methods have also noted the loss of accuracy resulting from applying the actuarial rules to an offender population different from the population used to develop the rules—a sampling problem common to any statistical analysis (Gendreau, Goggin, and Smith, 2002; Glover et al., 2002; Gottfredson and Gottfredson, 1986; Grove and Meehl, 1996) - as well as although the overall accuracy rates of these instruments represent an improvement over chance, the magnitude of the improvement is small and not deemed to be clinically significant (Menzies et al., 1994; Steadman et al., 2000). Recently, critics of conventional actuarial techniques, particularly logistic regression, have demonstrated that other approaches such as machine learning methods are superior in forecasting future risks especially in situations where the best decision boundary is complex (Berk and Bleich, 2013).

To date, researchers and scholars have attempted to assess different statistical methods and approaches for risk assessment instruments to identify the best method for the purpose of testing and developing new instruments. Within the arena of criminal justice, prior studies have compared logistic regression, classification tree methods, and neural networks models for their relative accuracy in predicting violence and criminal recidivism. In this paper, we seek to contribute to the scholarship on the identification of the most effective statistical methods to use when predicting future risks by assessing the relative predictive utility of logistic regression, classification and regression tree (CART), chi-squared automatic interaction detection (CHAID), and neural networks in predicting an outcome that has not been examined by previous scholars and researchers - inmate misconduct. To the best of our knowledge, our paper is the first to compare

[1] Clinical approaches to risk assessment can be further dichotomized into unstructured and structured clinical judgment. With unstructured clinical judgment, a clinician relies solely on his/her professional experience for accuracy in predicting an individual's risk. With structured clinical judgment, the clinician utilizes empirically-based risk factors to guide his/her prediction of an individual's risk (for further descriptions of these two types of risk assessment methods, see Aegisdottir et al., 2006; Hanson, 2005; Singh and Fazel, 2010; Singh, Grann, and Fazel, 2011).

these four techniques together. We also draw from one of the leading theoretical perspectives on inmate misconduct, the importation model, for our predictor variables.

The remainder of our paper is organized as follows. First, we provide a brief description of logistic regression, CART, CHAID, and neural networks. Next, we present a review of the literature on prior comparison studies involving the above statistical techniques. We also discuss the importation model and describe our methods and data. Finally, we discuss our findings and their implications.

Logistic Regression, Cart, Chaid, and Neural Networks

Logistic Regression

Logistic regression (LR) is a statistical technique for classification based on the logistic function. LR involves the estimation of the probability of a binary event occurring (e.g., whether or not an inmate will recidivate within 12 months of release from prison). Since the application of the Ordinary Least Squared (OLS) regression to a binary outcome variable would violate the assumption of homoscedasticity and normality of the error term, as well as yield predicted probabilities that fall outside of the range 0 to 1, the employment of LR remedies those problems by transforming the binary outcome variable into the natural log of the odds ("the log odds") of the event of interest's occurrence (Aldrich and Nelson, 1984).

The general logistic regression model is,

$$\mathrm{Log}\left[p\Big/(1-p)\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_i x_i$$

where p is the estimated probability of the outcome of interest, β_0 is the intercept term, and $\beta_{1,\ \ldots,}\ \beta_{i,}$ are the logistic regression coefficients associated with the predictors $x_{1,}$ …., x_i. Further, to classify all cases into the defined category groups, LR applies a classification cut-off to the estimated probability. For the classification of two groups or categories (e.g., recidivated or did not recidivate), the LR default cut-off probability is set at 0.50 under the assumption of equal misclassification costs for the two types of misclassification, false positives and false negatives.[2] However, in situations involving disparate initial proportions (e.g., the proportion of sex offenders who recidivate is 5 % and this proportion is much lower than the proportion of sex offenders who do not recidivate—95 %) and in which the classification will favor the category with the greater overall proportion or prior probabilities, researchers would need to change the cut-off probability to achieve a better balance prediction. Due to its lenient requirements (LR does not require normally distributed variables or assume homoscedasticity; Hosmer and Lemeshow, 1989; Thomas et al., 2005), LR emerges as the prevailing technique of choice among actuarial approaches for a dichotomous classification. However, because LR gives a probability value for outcome variable estimated in an aggregate manner using all the factor values, it does not help the user to formulate an

[2] A false positive is defined as a positive result on a diagnostic test for a condition in an individual who actually does *not* have that condition and a false negative is defined as a negative result on a diagnostic test for a condition in an individual who actually *does* have that condition.

interview strategy. Critics of the LR method also contend that risk assessment tools *should* reflect actual clinical thinking and that some other classification approaches such as classification trees may be more suitable (Gardner et al., 1996; Rosenfeld and Lewis, 2005; Steadman et al., 2000; Thomas et al., 2005). We next discuss the classification tree approaches.

Classification Trees (CART and CHAID)

Both CART and CHAID are methods that build what is called a classification tree. The classification tree approach has been advocated as an alternative to regression models because it resembles a clinician type methodology of using a series of inquiries for the purpose of classifying subjects into defined category groups (e.g., recidivists and non-recidivists; Gardner et al., 1996; Rosenfeld and Lewis, 2005; Steadman et al., 2000; Thomas et al., 2005). Through a sequential process (or a set of logical *if-then* conditions), the sample is divided into "branches"[3] and within each of these branches, the best predictor is determined until no more variance can be explained with the remaining variables or some other criterion such as a minimum group size, has been reached. The resulting category groups represent subgroups of the original sample that differ in terms of the probability of the outcome variable.

There is an array of classification tree models including Classification and Regression Tree (CART; Breiman et al., 1984; Ripley, 1996), Chi-squared Automatic Interaction Detection (CHAID; Kass, 1980), Quick, Unbiased, Efficient Statistical Trees (QUEST; Loh and Shih, 1997), Decision Tree Forests (Breiman, 2001), Boosting Trees (Friedman, 1999), and Iterative Classification Tree (ICT; Steadman et al., 2000). Among the classification tree models, the CART and CHAID approaches appear to be both clinically feasible and effectual in predicting future risks, particularly future violence (Silver, Smith, and Banks, 2000; Thomas and Leese, 2003; Steadman et al., 2000). Both CART and CHAID methods produce tree graphs that present actuarial data in a manner that is simple and straight forward to use in a clinical settings (see Figure 1 and Appendix 1).

CART was developed and popularized by Breiman and colleagues (Breiman et al. 1984; see also, Ripley, 1996) while CHAID was originally proposed by Kass (1980). Both CART and CHAID will produce classification trees if the outcome variable is categorical and regression trees if the outcome variable is numerical or a combination. Further, the differences between the various tree-based methods lie in the strategies for splitting the trees into branches and sub branches. To split a tree into branches, CART employs the Gini measure of impurity[4] for categorical outcome variables and the least-squared deviation (LSD) measure of impurity for numerical outcome variables (Breiman et al., 1984). The Gini index is given by,

[3] A classification tree starts with the top decision branch, sometimes called the *root* or *parent* node, and the top branch is split into subsequent branches known as *child nodes Terminal nodes* are branches on the tree beyond which no further decisions are made.

[4] The three measures of impurity generally used for classification problems are the Gini measure, the generalized Chi-square measure, and the generalized G-square measure. The Chi-square measure is similar to the standard Chi-square value computed for the expected and observed classifications, and the G-square measure is similar to the maximum-likelihood Chi-square. The Gini measure is the index most often used for measuring purity in the context of classification problems and the method advocated by the developers of CART (Breiman et al., 1984).

$$g(t) = \Sigma_{j \neq i} C(i|j) p(j|t) p(i|t)$$

Where $p(j|t)$ is the estimated probability that an observation belonging to group j given that it is in subgroup t, $p(i|t)$ is the estimated probability that an observation belonging to group i given that it is in subgroup t, and $C(i|j)$ is the probability of misclassifying a category j case as category i. On the other hand, the LSD measure is computed as,

$$R(t) = 1 \Big/ N_w(t) \Sigma w_i f_i \left(y_i - \overline{\overline{y}}(t) \right)^2$$

where $N_w(t)$ is the weighted number of cases in group t, w_i is the value of the weighting variable for case i, f_i is the value of the frequency variable, y_i is the value of the outcome variable, and $\overline{y}(t)$ is the weighted mean for group t. Both the Gini index and the LSD value are measures of goodness of fit. The CART procedure attempts to split a tree to maximize the goodness of fit.

CHAID uses a different approach to split a tree into "branches." CHAID uses tests of statistical significance. Similar to CART, the statistical test employed by CHAID is determined by the outcome variable. If the outcome variable is continuous, an F test is used. If the outcome variable is ordinal, a likelihood-ratio test is used and if the outcome variable is nominal, a Pearson chi-squared test is used (Rokach and Maimon, 2008). For each input attribute a_i, CHAID finds the pair of values in V_i that is least significantly different with respect to the outcome variable. Further, for each selected pair of values, CHAID assesses if the p value obtained is greater than a certain merge threshold. If it is, then CHAID merges the values and searches for an additional potential pair to be merged. The process is repeated until no more significant pairs are found or a pre-specified condition is met (e.g., the maximum tree depth is reached; for further description of CHAID, see Hill and Lewicki, 2006; Rokach and Maimon, 2008).

Similar to LR, the classification tree method is based on a probability value that is estimated for each case and the cutoff for equal misclassification cost is assumed to be 0.5. Different cutoff values can also be used for different misclassification costs. Further, to select the "right sized" tree[5] or to determine if the classification tree computed from the "training" sample in which the outcomes are already known will perform equally well in predicting outcomes in a second, independent "test" sample,[6] it is recommended that researchers perform cross-validation. Cross validation is essentially an empirical approach to the problem of obtaining an unbiased estimate of predictive accuracy (Gottfredson and Moriarty, 2006). The two proposed options for performing cross-validation by the developers of CART (Breiman et al., 1984) are the test sample cross-validation and the global cross-validation. In the test sample cross-validation, the tree is computed from the "training" sample, and its predictive accuracy

[5] The size of a tree in the classification and regression trees analysis is an important issue since an unreasonably big tree can only make the interpretation of results more difficult. There are two recommended strategies for selecting the "right-sized" tree. One strategy is to grow the tree to just the right size, where the right size is determined by the researcher, based on the knowledge from previous research, diagnostic information from previous analyses, or even intuition. The other strategy is to use a set of well-documented, structured procedures developed by Breiman et al. (1984) for selecting the "right-sized" tree.

[6] A classification tree model encompasses at least two samples - training and testing—and the training sample is used to build the model and the testing sample is employed to validate its performance.

is applied to predict the group membership in another independent "test" sample. Alternatively, if a large "training" sample is available, a randomly selected proportion of the cases (e.g., a third or a half) could be reserved and used as the "test" sample. In the global cross-validation, the entire analysis is replicated a specified number of times holding out a proportion of the "training" sample equal to 1 over the specified number of times, and using each hold-out sample in turn as the "test" sample.

Neural Networks

Neural network (NN) models, also known as artificial neural networks, have emerged from research in artificial intelligence. NN models are *adaptive* statistical models based on an analogy with the structure of the brain. They are adaptive in that they can *learn* to estimate the parameters of some population such as the values of "*a"* and "*b"* in the regression equation $y=a+bx$ (Abdi, Valentin, and Edelman, 1999). A neural network is built from simple units called processing elements (PEs; or neurons by analogy). The PEs are interlinked by a set of connection weights that loosely correspond to regression coefficients in regression models. The processing ability of the neural network is stored in these connection weights obtained through the process of training (or *learning*). In other words, the goal of the network is to *learn,* or to discover, some patterns between input and output values. The *learning* is accomplished through the modification of the connection weights between PEs. The learning process also specifies the *algorithm* used to estimate the parameters or the values of the connections between PEs (Abdi et al., 1999; Caulkins et al., 1996; Gurney, 1997; Ripley, 1996).

There is a wide variety of neural network models and architectures.[7] However, the most commonly used design, and perhaps the most studied, is the multi-layer perceptron neural network (MLPNN; Rumelhart and McClelland, 1986; see also, Minsky and Papert, 1969). The MLPNN consists of an input layer (the first layer), an output layer (the last layer), and a hidden layer (the intermediate layer). Each PE in the input layer corresponds to a feature or characteristic that the researcher is interested in using as a predictor (or independent variable). The MLPNN is also characterized by a feed forward structure in that the information to be analyzed is input into each PE, processed, and then passed on to each PE in the next layer. Also, information out of the output PE is used to compute the estimates of errors, which are used to modify the weights (see Figure 2 and Appendix 2).

The purpose of the MLPNN design is to map the input units to a desired output similar to the way in which the dependent variable is a function of the independent variables in regression analysis. This goal is accomplished by updating or adjusting the weights on the connections between PEs, iteratively, in response to error signals transmitted back through the network. That is, when the network is presented with an input pattern, it computes the activation of the output unit(s) using the current network weight structure (the weights are initialized randomly prior to training). The difference between the output of the network and the desired output constitutes the error signal and this signal is then propagated back through the network (via the PEs) for the

[7] For a detailed description of the different neural network models see Abdi et al. (1999), Carpenter and Grossberg (1991), Gurney (1997), and Rumelhart and McClelland (1988).

purpose of updating or adjusting the connection weights. The connection weights are continually updated until the sum of all error signals is minimized (White, 1989).

Similar to CART or CHAID, the MLPNN uses a training dataset to build a classification model, in this case an artificial neural network, which can then be used to classify cases in the testing dataset. Within the MLPNN design, the training is conducted using the following steps (Gurney, 1997; Price et al., 2000):

1) Patterns or "signals" are presented at the input layers. The signals are sent from the input layers to the hidden layers,

$$v_{pj} = \sum_{i=1}^{N} W_{ij} X_{pi}$$

where x_{pi} are the input values of the PEs in the input layer, p is the observation number, i is index for the PE in the input layer, w_{ij} is the weight between the i^{th} PE in the input layer and the j^{th} PE in the hidden layer, and v_{pj} is the input to the j^{th} PE in the hidden layer. N is the total number of independent variables. Further,

$$x_{pj} = f\left(v_{pj}\right) + B_{inp,j}$$

where $f(v)$ is the activation function in the hidden layer (with the most commonly used function being the logistic function), $B_{inp,j}$ is the input of the bias unit between the input and hidden layers (note that this term corresponds with the intercept term of regression models; Bishop, 1995);

2) The patterns or "signals" are sent from the hidden layer to the output layer,

$$v_{pk} = \Sigma_{j=1}^{M} w_{jk} x_{pj}$$

where w_{jk} is the weight between the j^{th} hidden unit and the k^{th} output unit (given that there are k output units), M is the number of hidden PEs, and v_{pk} is the middle set of "signals" in this process. Further,

$$y_{pk} = f\left(v_{\text{pk}}\right) + B_{\text{outj}}$$

where $f(v)$ is the activation function in the output layer (which is often set as softmax function for the purpose of classification), B_{outj} is the output of the bias unit between hidden and output layers, and y_{pk} is the output of the output PEs.

3) The difference between the predictions (v_{pk}) of the network and the target values (or desired output) is determined using an error function, with the sum of squared error and cross entropy being the two most popular kinds of error functions (Liu et al., 2011).
4) The training algorithm is used to adjust the weights (w_{ij} and w_{jk}) of the network. The most common training algorithm includes back propagation, gradient decent, conjugate descent, and BFGS (Abdi et al., 1999; Bishop, 1995; Gurney, 1997).
5) Steps 1 through 4 are repeated through a number of training cycles until the network arrives at satisfactory outputs.

Table 1 Prior comparative studies examining the predictive utility of logistic regression, classification tree (CART and CHAID), and neural networks

Study	Statistical techniques examined	Sample	Outcome variable	Overall conclusion
Brodzinski et al. (1994)	Discriminant classification & neural networks	Juvenile probationer cases between 1985 and 1986	Recidivism	Neural networks outperformed discriminant classification
Caulkins et al. (1996)	Multiple regression & neural networks	Offenders released from prisons in the U.S. from 1970 to 1972	Recidivism	Comparable performance between multiple regression & neural networks
Gardner et al. (1996)	Negative binomial regression & CART	Psychiatric patients in a prospective longitudinal study	Incidents of community violence	Comparable performance between negative binomial regression & CART
Grann & Langstrom (2007)	Logistic regression & neural networks	Forensic patients in Sweden	Violent reconviction	Comparable performance between logistic regression & neural networks
Liu et al. (2011)	Logistic regression, CART & neural networks	Male prisoners released from UK prisons	Violent reconviction	The overall predictive accuracy of all three models was comparable
Neuilly et al. (2011)	Logistic regression & CTA (Classification Tree Analysis) optimization using random forest	Homicide offenders released from New Jersey Department of Corrections between 1999 and 2000	Parole violation & reincarceration	CTA optimization through random forest outperformed logistic regression
Palocsay et al. (2000)	Logistic regression & neural networks	Offenders released from prisons in North Carolina in 1978 & 1980	Recidivism	Neural networks outperformed logistic regression
Rosenfeld & Lewis (2005)	Logistic regression & CART	Stalking offenders referred for psychiatric evaluation in New York City	Violent reconviction	Comparable performance between logistic regression & CART but there was evidence of shrinkage in the CART model
Silver et al. (2000)	Burgess's method, logistic regression, CART & ICT[a] model	Sentencing data in the State of New Jersey	Recidivism	The ICT model outperformed the Burgess's method, logistic regression, and CART but *only* in terms of the percentage of cases classified as high or low risk
Stalans et al. (2004)	Logistic regression & CTA (Classification Tree Analysis)	Violent offenders on probation from October 30, 2000, to November 30, 2000, in the State of Illinois	Violent recidivism	CTA outperformed logistic regression

Table 1 (continued)

Study	Statistical techniques examined	Sample	Outcome variable	Overall conclusion
Steadman et al. (2000)	Logistic regression, CHAID & ICT model	Acute psychiatric patients discharged from three hospitals[b]	Rate of violence	The ICT model outperformed CHAID & logistic regression
Thomas et al. (2005)	Logistic regression & CART	Psychotic patients in four inner-city mental health services in the UK	Rate of violence	CART outperformed logistic regression but the robustness of CART is questionable
Yang et al. (2010)	Logistic regression, CART & neural networks	Male & female offenders released from UK prisons between November 2002 & October 2005	Violent recidivism	The overall predictive accuracy of all three models was comparable; logistic regression appeared to be a more robust technique

[a] Iterative Classification Tree method developed in the MacArthur Violence Risk Assessment Study (Steadman et al., 2000)

[b] The three facilities were: 1) the Western Psychiatric Institute and Clinic (Pittsburg, PA), 2) the Western Missouri Mental Health Center (Kansas City, MO), and 3) the Worcester Hospital and the University of Massachusetts medical Center (Worcester, MA)

Prior Comparison Studies

Prior comparative studies on the best statistical approaches to construct risk assessment instruments have examined the relative predictive performance among conventional regression techniques (multiple regression, logistic regression, negative binomial regression, discrimination classification), classification tree methods (CART, CHAID), and NN models. These studies have generally focused on the outcome of future violent behavior or recidivism as well as drawn from samples of psychiatric patients and criminal offenders. To the best of our knowledge, to date, there are seven studies assessing the relative predictive utility of conventional regression models with classification tree models, four studies comparing conventional regression models with NN models, and two studies contrasting all three methods.[8] Overall, the results generated from prior comparison studies are mixed and inconclusive (a description and overall conclusion for the above studies are provided in Table 1).

Of the seven studies that examined conventional regression models with classification tree models, two studies reported similar accuracy performance between the two techniques in predicting future violence among psychiatric patients (Gardner et al., 1996) and stalking offenders (Rosenfeld and Lewis, 2005; see Table 1). The remaining five studies reported better performance by the classification models compared to the conventional regression methods in predicting violence or recidivism among psychiatric patients, probationers, and homicide offenders (Neuilly et al., 2011; Silver et al., 2000; Stalans et al., 2004; Steadman et al., 2000; Thomas et al., 2005; see Table 1).

Among the four studies that compared conventional regression models with NN models, two studies reported that NN models performed better than conventional regression methods in predicting recidivism among juvenile probationers (Brodzinski, Crable, and Scherer, 1994) and inmates released from prisons (Palocsay et al., 2000). The remaining two studies showed that NN models were neither superior nor worse than the conventional methods in predicting recidivism among offenders released from prisons (Caulkins et al., 1996) and violent reconviction among psychiatric patients (Grann and Langstrom, 2007). Pertaining to the two studies that compared LR, CART, and NN models, the first study found that while the performance of NNs was slightly better than that of LR and CART models in predicting violent reconviction from a sample of male offenders released from UK prisons, the differences did not reach significance. The authors concluded that the three techniques exhibited similar accuracy performance (Liu et al., 2011). In the second study, the authors also reported similar accuracy performance among the three models in predicting violent reconviction from a UK sample of men and women prisoners. However, relative to NNs and CART, LR appeared to be a more robust model (Yang et al., 2010; see Table 1).

With the exception of the two studies that examined LR, CART, and NN models together (Liu et al., 2011; Yang et al., 2010), prior comparison studies were limited due to various issues including small sample size, inadequate cross-

[8] Our review of prior research only includes studies that explicitly compare conventional regression models with classification tree models and/or neural network models. There were prior studies that attempt to validate the actuarial model developed in the MacArthur Violence Risk Assessment Study or examine the development of the Classification of Violence Risk (COVR) software and these studies are not included in our review of prior literature.

validation strategies, and disregard for low base rates. Additionally, the inconsistency in findings reported by previous comparison studies may have been attributed to factors such as specificity of predictors, homogeneity of samples, definition criteria of outcomes, computational approach for model validation, handling of over-fitting, tuning procedures to improve forecasting accuracy, and improper implementation of forecasting procedures (Berk and Bleich, 2013; Liu et al., 2011)

In this study, we seek to extend the literature on comparative studies by assessing the relative merits of four statistical techniques - LR, CART, CHAID, and NN models—for their predictive utility on an outcome that has not been examined in prior research—inmate misconduct. Inmate misconduct is a salient area of inquiry because inmate misconduct reflects inmates' adjustment to prison, affects the prison order and safety of correctional staff and other inmates, and is closely related to prison classification in that not only are inmate disciplinary infractions one of the measures of classification effectiveness, they are also a necessary element for reclassification in prison (Gendreau, Goggin, and Law, 1997; Jiang and Fisher-Giorlando, 2002). We also incorporate a wide range of predictors drawn from one of the leading theoretical perspectives on inmate misconduct - the importation model[9] - and employ large training and testing samples. We also draw data from a nationally representative sample of incarcerated offenders in state and federal prisons. Before we present our data, we describe the importation model on inmates' adjustment to prison.

The Importation Model on Inmate Adaption

One of the leading theoretical perspectives used to account for inmate adjustment to prison is the importation model. According to this perspective, inmates' behavior in confinement is determined by their distinctive traits and social background prior to incarceration. That is, inmates import their roles from outside of prison into the prison culture (Irwin, 1981; Irwin and Cressey, 1962) and if an inmate was violent outside prison, it is very likely that the inmate will also be violent while incarcerated. Further, the importation model argues that adaption to prison will depend on the inmate's ability to find a "niche" that meets his needs (Seymour, 1977; Toch, 1977) as well as since inmates come from different subgroups with different belief systems and norms, they do not represent a solitary group in prison as some prison scholars have suggested (Irwin and Cressey, 1962; Paterline and Petersen, 1999; Toch and Adams, 1986; Wooldredge, 1991).

Given the focus of the importation model on the effect of pre-prison factors on prison adjustment, prior empirical studies examining the efficacy of this perspective in explaining inmate behavior have examined factors including race, gender, age, social class, marital status, education, employment, offense type, criminal history, gang membership, drug use, and personality variables (Byrne and

[9] The four leading theoretical perspectives on inmates' adjustment to prison are the deprivation, importation, situational, and administrative control models (Clemmer, 1940; Irwin and Cressey, 1962; DiIulio 1987; Sykes, 1958; Steinke, 1991). In this paper, we chose to focus on the importation model and defer the examination of the other models in subsequent papers since including measures from all four models would prove too cumbersome.

Table 2 Descriptive statistics of variables for the five samples[a]

	Sample 1 training/testing	Sample 2 training/testing	Sample 3 training/testing	Sample 4 training/testing	Sample 5 training/testing
Outcome variable					
Found guilty of breaking any rules					
1 = Yes	0.53/0.53	0.53/0.54	0.53/0.53	0.53/0.52	0.53/0.52
0 = No	0.47/0.47	0.47/0.46	0.47/0.47	0.47/0.48	0.47/0.48
Predictor variables					
Gender					
1 = Male	0.81/0.80	0.81/0.80	0.81/0.8	0.81/0.81	0.80/0.83
0 = Female	0.19/0.20	0.19/0.20	0.19/0.19	0.19/0.19	0.20/0.17
Age					
0 = Less than 20	0.04/0.03	0.04/0.03	0.04/0.04	0.04/0.03	0.03/0.03
1 = 20–35	0.49/0.52	0.49/0.51	0.49/0.47	0.50/0.47	0.50/0.50
2 = 36 or Older	0.47/0.45	0.47/0.46	0.47/0.49	0.46/0.50	0.47/0.47
Race					
1 = African American	0.45/0.45	0.45/0.46	0.45/0.45	0.45/0.45	0.45/0.45
0 = Other race	0.55/0.55	0.55/0.54	0.55/0.55	0.55/0.55	0.55/0.55
Marital status					
1 = Married	0.16/0.15	0.16/0.15	0.16/0.15	0.15/0.17	0.16/0.16
0 = Not married	0.84/0.85	0.84/0.85	0.84/0.85	0.85/0.83	0.84/0.84
Number of prior arrests					
0 = none	0.01/0.01	0.01/0.01	0.01/0.01	0.01/0.01	0.01/0.01
1 = One arrest	0.22/0.23	0.23/0.23	0.22/0.24	0.23/0.21	0.23/0.22
2 = Two to five arrests	0.48/0.48	0.48/0.46	0.48/0.47	0.47/0.49	0.48/0.48
3 = Five or more arrests	0.29/0.28	0.28/0.30	0.29/0.28	0.29/0.29	0.28/0.29
Age at first arrest					
0 = Less than 13	0.09/0.09	0.09/0.10	0.09/0.09	0.10/0.08	0.09/0.10
1 = 13–20	0.64/0.65	0.64/0.65	0.64/0.64	0.64/0.64	0.64/0.63
2 = 21 or older	0.27/0.26	0.27/0.25	0.27/0.27	0.26/28	0.27/0.27
Employment prior to incarceration					
1 = Yes	0.69/0.70	0.70/0.67	0.69/0.68	0.69/0.70	0.69/0.70
0 = No	0.31/0.30	0.30/0.33	0.31/0.32	0.31/0.30	0.31/0.30

[a] Each of the training samples consists of 8,000 cases and each of the testing samples includes 2,000 cases

Hummer, 2007; Goodstein and Wright, 1989; Paterline and Petersen, 1999; Wooldredge, 1991; Wright, 1991). Overall, the results generated from prior assessments of the importation model have demonstrated support for it even in competing criminological models. For example, Harer and Steffensmeier (1996)

examined the applicability of the importation and deprivation models in accounting for inmate violent misconduct. The authors likened the deprivation versus importation models to the structural versus cultural theories of violence (respectively) and hypothesized that the patterns of inmate violence in prisons would parallel those in the larger society.

Employing data from a sample of only white and black inmates from 58 male federal correctional institutions, Harer and Steffensmeier (1996) uncovered that net of deprivation or structural measures (crowding, length of time served, prison turnover rate, staff to inmate ratios, furloughs, security level of prison, and staff perception of their ability to communicate and work constructively with inmates), black inmates had significantly higher levels of prison violence than white inmates but lower levels of alcohol/drug misconduct. The authors attributed their findings as evidence supporting the cultural or importational view of prison behavior since both inside and outside prison, blacks have higher rates of violence than whites but whites have as high or higher rates of alcohol and drug abuse than blacks (Wallace and Bachman, 1991).

In a recent study, Jiang and Fisher-Giorlando (2002) compared the importation, deprivation, and situational models for their efficacy in explaining inmate misconduct and violent incidents against staff and other inmates. Drawing from official data collected from 186 inmates in a single correctional facility, the authors found that the situational and deprivation models helped explain violent incidents against staff while the situational and importation models helped explain violent incidents against inmates (see also, Cao, Zhao, and Dine, 1997; Dhami, Ayton, and Loewenstein, 2007; Sorensen, Wrinkle, and Gutierrez, 1998; Paterline and Petersen, 1999; Wooldredge 1991).[10]

In this study, we employ a wide range of importation variables to evaluate the relative utility of LR, CART, CHAID, and MLPNN models in predicting inmate misconduct. In the next section, we describe our data and methods.

Data and Methods

Data for the current study come from the 2004 Survey of Inmates in State and Federal Correctional Facilities (SISFCF) conducted for the Bureau of Justice Statistics (BJS) by the Bureau of the Census (ICPSR #4572). Data collection for SISFCF involved a two-stage stratified sample design with correctional facilities chosen at the first stage and inmates within facilities chosen at the second stage. SISFCF provides nationally representative data on inmates held in state and federal prisons with personal interviews with the inmates occurring between October 2003 and May 2004. Inmates participated in SISFCF provided information about their current offense and sentence, criminal history, family background and characteristics, prior drug and alcohol use, medical and

[10] Since reviews of the importation model are readily available elsewhere (see for example, Byrne and Hummer, 2007; Cao, Zhao, and Dine 1997; Goodstein and Wright, 1989; Paterline and Petersen, 1999; Wooldredge, 1991; Wright, 1991) and due to page limitation, we forego a thorough review of the literature on this perspective.

mental health conditions, participation in treatment programs, gun possession and use, and prison activities, programs, and services.

Samples

A total of 14,499 inmates participated in the 2004 SISFCF and after accounting for missing data and nonresponses, the sample size was reduced to 10,328. Since we were interested in the global cross-validation method, we randomly selected 10,000 cases from the original dataset and partitioned it into five sub-datasets of 2,000 cases each labeled A, B, C, D, and E. We tested all four techniques (LR, CART, CHAID, and MLPNN) five times using each of the five subsets as testing sample and the remaining four subsets together as training sample. For example, in the first set, subsets B, C, D, and E were combined and used as the training sample while subset A was used as the testing sample. In the second set, subsets C, D, E, and A were combined and used as the training sample while subset B was used as the testing sample[11] and so on. We believe this multi-validation method yields a more reliable classification accuracy than a single-sample validation as the latter approach may result in high model fit values (Grann and Langstrom, 2007).

Table 2 shows the demographic and other characteristics of the five sub-datasets. As shown in Table 2, all five sub-datasets demonstrated similar demographic and other characteristics. In particular, in all five sub-datasets, the majority of the respondents were males (80–83 %), almost half were African Americans (45–46 %), and many of the respondents (47–52 %) were in the 20 to 35 years age range. Additionally, almost one-fifth of the respondents were married (15–17 %) and many were employed (68–70 %) as well as attended high school (73–76 %) before being incarcerated. On average, respondents in all five sub-datasets reported having between two and five prior arrests and the age of their first arrest was between the ages of 13 and 20. Many respondents (47–52 %) in each of the five sub-datasets indicated that they consumed alcohol or used drugs previously and approximately one-third of the respondents (28–30 %) in each sub-dataset reported that they have been diagnosed with a mental or personality disorder. Finally, almost half of the respondents in each of the five sub-datasets were serving their current sentences for a violent offense (46–48 %) and slightly over half of the respondents (52–54 %) reported that they had violated at least one type of prison rules or regulations (Table 2).

Measures

Outcome Variable Our outcome variable in this study is whether the inmate was cited or found guilty of any prison violations.[12] More specifically, this variable was measured using the question, "Since your admission, have you been written up for or

[11] The combinations of the five sub-datasets are as followed with the letter on the left side represents the testing sample and the letters in the right side represent the training sample: Sample 1 = A/BCDE; Sample 2 = B/CDEA; Sample 3 = C/DEAB; Sample 4 = D/EABC; and Sample 5=E/ABCD.

[12] We elected to employ a general measure of any prison misconduct because prior comparative research has utilized similar measures such as delinquency, recidivism, or violence (see for example, Caulkins et al., 1996; Rosenfeld and Lewis, 2005; Thomas et al., 2005).

Table 3 The comparison among LR, CART, CHAID, and MLPNN by sensitivity, specificity, accuracy, and 95 % C.I. of accuracy

Sub-sets combination[a]		LR			CART			CHAID			MLPNN		
		Sen.	Spe.	Acc.(95 % CI)	Sen.	Spe.	Acc.(95 % CI)	Sen.	Spe.	Acc.(95 % CI)	Sen.	Spe.	Acc.(95 % CI)
A/BCDE	Train	0.71	0.57	0.64 (0.63, 0.65)	0.61	0.65	0.63 (0.62, 0.64)	0.57	0.69	0.63 (0.62, 0.63)	0.68	0.61	0.64 (0.63, 0.65)
	Test	0.68	0.55	0.62 (0.60, 0.64)	0.58	0.66	0.62 (0.60, 0.64)	0.58	0.66	0.62 (0.60, 0.64)	0.66	0.60	0.64 (0.61, 0.66)
	Total	0.70	0.56	0.64 (0.63, 0.65)	0.61	0.65	0.63 (0.62, 0.64)	0.58	0.69	0.63 (0.62, 0.64)	0.67	0.60	0.64 (0.63, 0.65)
B/CDEA	Train	0.69	0.56	0.63 (0.62, 0.64)	0.76	0.47	0.62 (0.61, 0.64)	0.67	0.58	0.63 (0.62, 0.64)	0.69	0.60	0.64 (0.63, 0.65)
	Test	0.72	0.57	0.65 (0.63, 0.68)	59	0.68	0.63 (0.61, 0.65)	0.59	0.68	0.63 (0.61, 0.65)	0.72	0.58	0.66 (0.64, 0.68)
	Total	0.70	0.57	0.64 (0.63, 0.64)	0.73	0.51	0.63 (0.62, 0.63)	0.65	0.60	0.63 (0.62, 0.64)	0.69	0.59	0.65 (0.64, 0.66)
C/DEAB	Train	0.71	0.56	0.64 (0.63, 0.65)	0.72	0.54	0.63 (0.62, 0.64)	0.78	0.47	0.63 (0.62, 0.64)	0.69	0.59	0.64 (0.63, 0.65)
	Test	0.69	0.57	0.63 (0.61, 0.66)	0.55	0.64	0.60 (0.57, 0.62)	0.55	0.64	0.60 (0.57, 0.62)	0.66	0.59	0.63 (0.61, 0.65)
	Total	0.71	0.56	0.64 (0.63, 0.65)	0.68	0.56	0.63 (0.62, 0.64)	0.73	0.51	0.63 (0.62, 0.64)	0.68	0.59	0.64 (0.63, 0.65)
D/EABC	Train	0.70	0.55	0.64 (0.63, 0.65)	0.76	0.49	0.63 (0.62, 0.64)	0.67	0.59	0.63 (0.62, 0.64)	0.68	0.60	0.64 (0.63, 0.65)
	Test	0.70	0.56	0.64 (0.62, 0.66)	0.58	0.64	0.61 (0.59, 0.63)	0.58	0.64	0.61 (0.59, 0.63)	0.68	0.60	0.64 (0.62, 0.66)
	Total	0.70	0.56	0.64 (0.63, 0.65)	0.72	0.52	0.63 (0.62, 0.64)	0.65	0.60	0.63 (0.62, 0.64)	0.69	0.60	0.64 (0.63, 0.65)
D/EABC	Train	0.70	0.56	0.64 (0.63, 0.65)	0.77	0.47	0.63 (0.62, 0.64)	0.77	0.47	0.63 (0.62, 0.64)	0.69	0.58	0.64 (0.63, 0.65)
	Test	0.71	0.56	0.64 (0.62, 0.65)	0.56	0.66	0.61 (0.59, 0.63)	0.56	0.66	0.61 (0.59, 0.63)	0.69	0.58	0.64 (0.62, 0.66)
	Total	0.70	0.56	0.64 (0.63, 0.65)	0.73	0.51	0.63 (0.61, 0.65)	0.73	0.51	0.63 (0.62, 0.64)	0.69	0.58	0.64 (0.63, 0.65)

[a] The letter on the left side represents the testing sample (N=2,000) and the letters on the right side represent the training sample (N=8,000)

found guilty of breaking any prison rules," and the response options were "Yes" and "No." This variable was then coded as a dichotomous variable with 1 = the inmate was cited/found guilty of at least one prison misconduct and 0 = the inmate was not cited/found guilty of any prison misconduct. The descriptive statistics for this outcome variable are shown in Table 2.

Predictor Variables Eleven measures derived from the importation model were included in the study as predictor variables. *Gender* was coded as a dichotomous variable with 1 = Male and 0 = Female and *Race* was also coded as a dichotomous variable with 1 = African American and 0 = Other Race.[13] *Marital Status* was measured using the question, "Are you now married, widowed, divorced, separated, or have you never been married," and the responses were recoded with 1 = Married and 0 = Not Married (i.e., widowed, divorced, separated or never been married). *Employment Prior to Incarceration* was measured using the question, "During the month before your arrest, did you have a job or a business," and this variable was also coded as a dichotomous variable with 1 = the inmate had a job/business before incarceration and 0 = the inmate did not have a job/business before incarceration.

For the variable *Age,* respondents were asked to report their age in years and the responses were collapsed into three categories with 0 = Less than 21 years old, 1 = 21 years old to 35 years old, and 2 = 36 years or older. Likewise, for the variable *Age at First Arrest*, inmates' responses to the question, "How old were you the first time you were arrested for a crime," were collapsed into three categories with 0 = Less than 13 years old, 1 = 13 years old to 20 years old, and 2 = 21 years or older. The variable, *Education Prior to Incarceration*, was measured using the question, "Before your admission, what was the highest grade of school that you had attended," and the responses were collapsed into four categories with 0 = Less Than High School (i.e., kindergarten through 8th grade), 1 = High School (i.e., 9th grade through 12th grade), 2 = Some College (i.e., freshman through senior in college), and 3 = College or Graduate Degree (i.e., Bachelor Degree or Higher).

The variable, *Number of Prior Arrests*, was measured using the question, "How many times have you ever been arrested, as an adult or a juvenile, before your arrest (for the current offense)" and the responses were collapsed into four categories with 0 = None, 1 = One Arrest, 2 = Two to Five Arrests, and 3 = Five or More Arrests. On the other hand, the variable, *Current Offense*, was measured using inmates' responses to the question, "Are you currently sentenced to serve time for any offense," and the responses included four categories with 0 = Public Disorder, 1 = Drug Offense, 2 = Property Offense, and 3 = Violent Offense.

Finally, respondents were asked if they have ever used any of the following: 1) heroin, 2) other opiates, 3) methamphetamine, 4) other amphetamine, 5) methaqualone,

[13] It is noteworthy that prior research on the importation model tends to involve selective coding of the variable race. Some studies compare Black inmates with non-Black inmates, other studies compare White inmates with non-White inmates, and some studies even encompass several dichotomous measures of race (i.e., Black vs. other racial groups, White vs. other racial groups, Hispanic vs. other racial groups, etc.). We elected to code our race variable as Black vs. Non-Black because the importation model emphasizes the effect of pre-prison characteristics on prison adjustment and there is evidence that outside prison, Blacks have higher crime rates than Whites and other racial groups (see for example, Snyder, 2011).

Table 4 The comparison among LR, CART, CHAID, and MLPNN by AUC and 95 % C.I. of AUC

Sub-sets combination[a]		(1) LR AUC (95 % CI of AUC)	(2) CART AUC (95 % CI of AUC)	(3) CHAID AUC (95 % CI of AUC)	(4) MLPNN AUC (95 % CI of AUC)
A/BCDE	Train	0.68 (0.67, 0.69)	0.66 (0.65, 0.67)	0.68 (0.67, 0.69)	0.69 (0.68, 0.70)
	Test	0.66 (0.64, 0.68)	0.62 (0.60, 0.64)	0.65 (0.63, 0.67)	0.67 (0.65, 0.69)
	Total	0.68 (0.67, 0.69)	0.65 (0.64, 0.66)	0.67 (0.66, 0.68)	0.69 (0.68, 0.70)
B/CDEA	Train	0.68 (0.67, 0.69)	0.64 (0.63, 0.65)	0.67 (0.66, 0.68)	0.69 (0.68, 0.70)
	Test	0.70 (0.67, 0.72)	0.63 (0.61, 0.65)	0.65 (0.63, 0.67)	0.69 (0.67, 0.71)
	Total	0.68 (0.67, 0.69)	0.64 (0.63, 0.65)	0.67 (0.66, 0.68)	0.69 (0.68, 0.70)
C/DEAB	Train	0.68 (0.67, 0.69)	0.65 (0.64, 0.66)	0.68 (0.67, 0.69)	0.69 (0.68, 0.70)
	Test	0.67 (0.65, 0.69)	0.60 (0.58, 0.62)	0.62 (0.60, 0.64)	0.67 (0.65, 0.69)
	Total	0.68 (0.67, 0.69)	0.64 (0.63, 0.65)	0.67 (0.66, 0.68)	0.69 (0.68, 0.70)
D/EABC	Train	0.68 (0.67, 0.69)	0.65 (0.64, 0.66)	0.67 (0.66, 0.68)	0.69 (0.68, 0.70)
	Test	0.67 (0.65, 0.69)	0.65 (0.64, 0.66)	0.64 (0.62, 0.66)	0.68 (0.66, 0.70)
	Total	0.68 (0.67, 0.69)	0.64 (0.63, 0.65)	0.67 (0.66, 0.68)	0.68 (0.67, 0.69)
E/ABCD	Train	0.68 (0.67, 0.69)	0.65 (0.64, 0.66)	0.67 (0.66, 0.68)	0.69 (0.68, 0.70)
	Test	0.68 (0.66, 0.70)	0.61 (0.59, 0.63)	0.63 (0.61, 0.65)	0.67 (0.65, 0.69)
	Total	0.68 (0.67, 0.69)	0.64 (0.63, 0.65)	0.67 (0.66, 0.68)	0.69 (0.68, 0.70)

[a] The letter on the left side represents the testing sample (N=2,000) and the letters on the right side represent the training sample (N=8,000)

6) barbiturates, 7) tranquilizers, 8) crack, 9) cocaine, 10) PCP, 11) ecstasy, 12) LSD, 13) marijuana/hashish, and 14) any other drugs. The responses were combined to create the variable *Prior Substance Use* and this variable was coded as a dichotomous variable with 1 = the inmate used at least one drug previously or 0 = the inmate did not use any drugs previously. Likewise, respondents were asked if they have ever been diagnosed by a mental health professional such as a psychiatrist or a psychologist with any of the following disorders: 1) a depressive disorder, 2) manic- depression, bipolar disorder, or mania, 3) schizophrenia or other psychotic disorders, 4) post-traumatic disorder, 5) other anxiety disorders such as panic disorder, 6) a personality disorder, and 7) other mental or emotional condition. The responses were combined to create the variable *Prior Mental/Personality Disorders* and this variable was coded as a dichotomous variable with 1 = the inmate was diagnosed with at least one mental or personality disorder and 0 = the inmate was not diagnosed with any of the mental or personality disorders. The descriptive statistics for all of the predictor variables are shown in Table 2.

Evaluation Indicators

Measures commonly used in prior comparative studies to gauge the "predictive accuracy" of a risk assessment instrument include "percent correctly classified," sensitivity and specificity, false positive and false negative rates, proportionate reduction in error (PRE), and Mean Cost Rating (MCR; Gottfredson and Moriarty, 2006). It is noteworthy that the above measures have been criticized for their

instability with varying predictor base rates as well as biases in favor of certain outcomes (Gottfredson and Moriarty, 2006; Rice and Harris, 1995). In recent years, the area under the curve (AUC) of the Receiver Operating Characteristic (ROC) has been advocated as an effective and appropriate index of accuracy because it is unaffected by differential base rates (Mossman, 1994; Rice and Harris, 1995).

The ROC curves originated in signal detection theory (Egan, 1975) and over the past several years, have been extended to machine learning and data mining applications for model evaluation and selection (Perlich, Provost, and Simonof, 2003; Yan et al., 2003). The ROC curve for a binary classification problem plots the true positive rate[14] as a function of the false positive rate [15] for all observed predictor values. Essentially, the ROC curve depicts the tradeoff in the false positive rate that occurs as the true positive rate increases with lower cutoff scores and vice versa (see Figure 3 and Appendix 3). The AUC of the ROC represents the effect size estimate derived from the ROC analysis and ranges from 0.0 (perfect negative prediction) to 1.0 (perfect positive prediction).

Ensuing prior comparative studies, we rely on multiple evaluation indicators in assessing the predictive accuracy of the four statistical techniques included in our study. In addition to reporting the AUC value and 95 % C.I. of AUC, we also report the sensitivity, specificity, overall accuracy, and 95 % C.I. of the overall accuracy. The sensitivity indicator represents the proportion of true positives that are correctly identified by a statistical technique[16] and the specificity indicator denotes the proportion of true negatives classified by a statistical method.[17] The overall accuracy indicator signifies the proportion of true positives and true negatives identified by a statistical method.[18] Finally, to build LR, CART, CHAID, and MLPNN models, we employ the STATISTICA 11.0 software package (StatSoft 2008) and our cut-off probability is set at 0.5.[19]

Results

Table 3 shows the sensitivity, specificity, accuracy, and a 95 % C.I.[20] of the overall accuracy for LR, CART, CHAID, and MLPNN models using the 0.5 cut-off probability. As shown in Table 3, the sensitivity values generated by the LR model for the five training samples range from 0.69 to 0.71, and for the corresponding testing samples, range from 0.68 to 0.72. The results on Table 3 also reveal that compared to the sensitivity values, the specificity values generated by the LR model for both training and testing samples were lower (the specificity values range from 0.55 to 0.57). However, there appears to be no evidence of reduce accuracy or shrinkage[21] in the

[14] The true positive rate = (the # of true positives)/(the # of all positives)
[15] The false positive rate = 1—[(the # of true negatives)/(the # of all negatives)]
[16] Sensitivity = (the # of true positives)/(the # of all positives)
[17] Specificity = (the # of true negatives)/(the # of all negatives)
[18] Overall Accuracy = (the # of true positives and true negatives)/(the # of all positives and all negatives)
[19] We selected 0.50 as the cut-off probability based on the fact that between 52–54 % of the inmates in the five sub-samples were found guilty of breaking any rules.
[20] Given that the overall accuracy is a proportion, we constructed the confidence intervals using standard methods for proportions (see for example, Gardner and Altman, 1989).
[21] Shrinkage or over-fitting occurs when a statistical model demonstrates poor predictive performance or when the predictive accuracy of a model decreases from the training sample to the testing sample.

testing samples pertaining to the sensitivity and specificity accuracy (compare the values of the training samples with the values of the testing samples under the columns labeled "Sen" and "Spe" for the LR model on Table 3). Given the higher sensitivity values generated by the LR model relative to the specificity values, it appears that the LR model performs better at predicting inmate misconduct than at identifying inmate obedience.

The LR model also generated an accuracy value from 0.63 to 0.64 for the training samples (*95 % CI*: 0.62–0.65) and an accuracy value from 0.62 to 0.065 for the testing samples (*95 % CI*: 0.60–0.68; see the column labeled "Acc" under the LR model on Table 3). Further, the overall performance of the LR model (i.e., the combined performance of training and testing samples) for the five sets of data is 0.64 (*95 % CI*: 0.63–0.65; see the rows labeled "Total" under the LR model on Table 3).

Unlike the LR model, the sensitivity values generated by the CART and CHAID models for the training samples varied widely (from 0.61 to 0.77 and from 0.57 to 0.78, respectively; see Table 3). The specificity values generated by both models for the training samples follow a similar pattern (they range from 0.47 to 0.65 and from 0.47 to 0.69, respectively; see Table 3). On the other hand, the sensitivity and specificity values generated by both models for the testing samples were less dispersed (the sensitivity values generated by both models range from 0.55 to 0.59 and the specificity values range from 0.64 to 0.68; see Table 3). Additionally, while there was evidence of shrinkage in the testing sample pertaining to the sensitivity accuracy for both models, there was no evidence of shrinkage pertaining to the specificity accuracy (compare the values of the training samples with the values of the testing samples under the columns labeled "Sen" and "Spe" for the CART and CHAID models on Table 3).

With regard to the accuracy values, the CART model generated an accuracy value from 0.62 to 0.63 (*95 % CI*: 0.61–0.65) for the five training samples and an accuracy value from 0.60 to 0.63 (*95 % CI*: 0.57–0.65) for the five testing samples (see the column labeled "Acc" under the CART model on Table 3). On the other hand, the CHAID model generated an accuracy value of 0.63 for all five training samples and an accuracy value from 0.60 to 0.63 for the five testing samples (Table 3). The LR model (accuracy=0.64; *95 % CI*: 0.63–0.65) seems to demonstrate slightly greater predictive utility relative to the CART and CHAID models in predicting inmate misconduct (the accuracy value for both CART and CHAID was 0.63; *95 % CI*: 0.61–0.65; see Table 3).

Pertaining to the MLPNN model, the sensitivity values generated by this technique for the five training samples range from 0.68 to 0.69 and for the five testing samples range from 0.66 to 0.72 (Table 3). On the other hand, the specificity values generated by the MLPNN model for the five training samples range from 0.58 to 0.61 while those for the five testing samples range from 0.58 to 0.60. Furthermore, there was minimal evidence of shrinkage in the testing sample pertaining to the sensitivity and specificity accuracy (compare the values of the training samples with the values of the testing samples under the columns labeled "Sen" and "Spe" for the MLPNN model on Table 3) and similar to the LR model where the generated specificity values for all five sets of data were lower than the sensitivity values, it appears that the MLPNN technique perform better at detecting inmate misconduct than at identifying inmate obedience.

The MLPNN model also yielded an accuracy value of 0.64 (*95 % CI*: 0.63–0.65) for the five training samples and from 0.63 to 0.66 (*95 % CI*: 0.61–0.68) for the five testing samples (see the column labeled "Acc" under the MLPNN model on Table 3).

Additionally, the MLPNN model (accuracy=0.64–0.65; *95 % CI*: 0.63–0.66) seems to demonstrate slightly better predictive power relative to the LR (accuracy=0.64; *95 % CI*: 0.63–0.65), CART and CHAID models in predicting inmate misconduct (accuracy=0.63; *95 % CI*: 0.61–0.65; see Table 3).

Table 4 presents AUC under ROC values for LR, CART, CHAID, and MLPNN models and also the 95 % C.I. of AUC values. It should be noted that the "Total" AUC values recorded in Table 4 represent the weighted averages of the AUCs for the training and testing samples and the weights are based on the size of the training and testing samples.[22] The results on Table 4 reveal that the LR model generated an AUC value of 0.68 (*95 % CI*: 0.67–0.69) for each of the five training samples and an AUC value from 0.66 to 0.70 (*95 % CI*: 0.64–0.72) for the five testing samples (Table 4). Table 4 also reveals that the Total AUC value yielded by the LR model was 0.68 (*95 % CI*: 0.67–0.69; see the rows labeled "Total" under the LR model on Table 4) and there was minimal evidence of shrinkage in the testing sample.

The CART model generated an AUC value of 0.64 to 0.66 (*95 % CI*: 0.63–0.67) for the training samples and an AUC value of 0.60 to 0.63 (*95 % CI*: 0.58–0.65) for the testing samples (Table 4). The Total AUC value generated by the CART model for the five sets of data ranges from 0.64 to 0.65 (*95 % CI*: 0.63–0.66; see the rows labeled "Total" under the CART model on Table 4). With regard to the CHAID model, it generated an AUC value from 0.67 to 0.68 (*95 % CI*: 0.66–0.69) for the training samples and an AUC value from 0.62 to 0.65 (*95 % CI*: 0.60–0.67) for the testing samples (Table 4). The CHAID model also yielded a Total AUC value of 0.67 (*95 % CI*: 0.66–0.68; see the rows labeled "Total" under the CHAID model on Table 4) and for both CART and CHAID models, there was evidence of shrinkage in the testing sample.

Table 4 reveals that the MLPNN model generated an AUC value of 0.69 (*95 % CI*: 0.68–0.70) for each of the training samples and an AUC value from 0.67 to 0.69 (*95 % CI*: 0.65–0.71) for the testing samples (Table 4). The MLPNN model also generated a Total AUC value of 0.68 to 0.69 (*95 % CI*: 0.67–0.70) for the five sets of data (see the rows labeled "Total" under the MLPNN model on Table 4) and similar to the LR model, there was minimal evidence of shrinkage in the testing sample. Accordingly, it appears that the predictive utility of the MLPNN model in predicting inmate misconduct (AUC value=0.68 to 0.69; *95 % CI*: 0.67–0.70) is slightly superior than the predictive utility of the LR model (AUC value=0.68; *95 % CI*: 0.67–0.69), and the predictive powers of both LR and MLPNN models are greater than the predictive utilities of both CART and CHAID models (AUC value=0.64 to 0.65; *95 % CI*: 0.63–0.65 and AUC value=0.67; *95 % CI*: 0.66–0.68, respectively; Table 4). The results also reveal that the predictive utility of the CHAID model (AUC value=0.67; *95 % CI*: 0.66–0.68) is better than the predictive power of the CART model (AUC value=0.64 to 0.65; *95 % CI*: 0.63–0.65; Table 4).[23]

[22] Total value = [(8,000 X AUC of the training sample + 2000 X AUC the testing sample)/10,000].

[23] We also performed Analysis of Variance (ANOVA) test for testing the differences in the means of classification accuracy and conduct pairwise *t*-test for the various pairs formed between the four classification methods. We found the MLPNN and the LR techniques performed significantly better than the CART and CHAID methods in predicting inmate misconduct (p-value<0.001), the CHAID technique outperformed the CART method (p-value<0.001) in classifying disobeyed inmates, and the MLPNN approach performed significantly better than the LR method (p-value<0.01) in predicting inmate misconduct (results are not shown but available upon request).

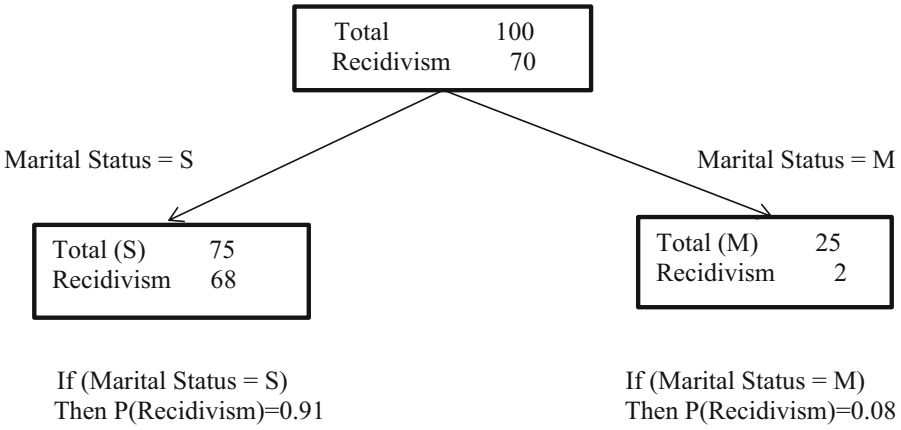

Fig. 1 A one-factor decision tree diagrama

Discussion and Conclusion

In this paper, we seek to contribute to the scholarship on the identification of the best statistical methods to use for the purpose of testing and developing risk instruments by comparing the predictive power of a traditional regression approach, LR, with three classification techniques, CART, CHAID, and MLPNN, in predicting inmate misconduct. Given the sparse applicability of CART, CHAID, and NN models in the field of criminal justice as well as their equivocal comparability with the LR model, we feel the undertaking of such comparison study is both essential and warranted. To the best of our knowledge, no study to date has focused on the outcome of inmate misconduct as well as examined the predictive utility of the above four models together.

We found several findings that parallel the conclusions reached by previous researchers as well as discovered several stimulating areas for future research. First, in line with prior research, we found evidence of shrinkage (or over-fitting) in the testing sample among the CART and CHAID methods (Liu et al., 2011; Rosenfeld and Lewis,

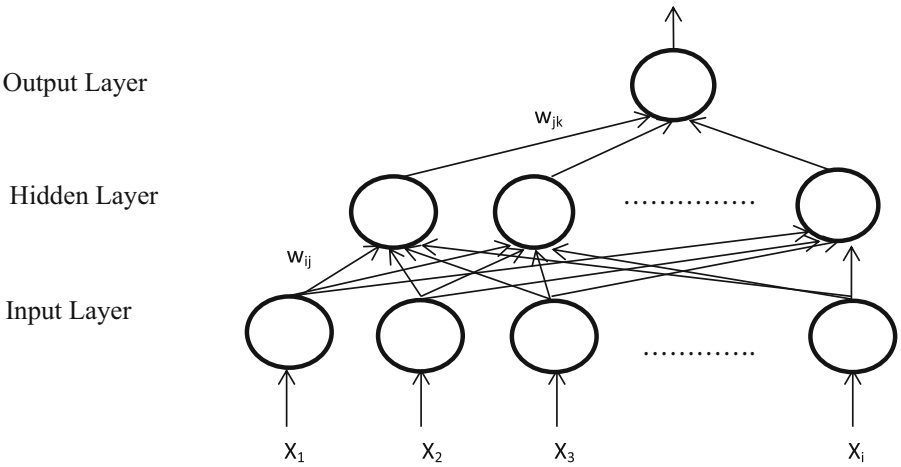

Fig. 2 A Simple Multilayer Feed Forward Neural Network Architecture with Backpropagation

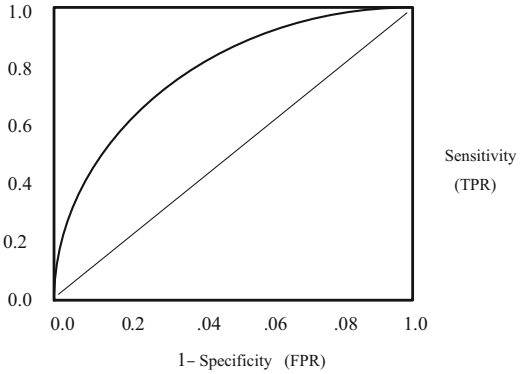

Fig. 3 The ROC curve

2005; Thomas et al., 2005; Yang et al., 2010). Second, the results generated from our study provide further support for the stability and robustness of the LR model relative to the CART and CHAID models (Gardner et al., 1996; Liu et al., 2011; Rosenfeld and Lewis, 2005; Stalans et al., 2004; Thomas et al., 2005). Third, although there was slight evidence of over-fitting with the MLPNN approach, it nevertheless demonstrated comparable predictive accuracy to the LR approach (Liu et al., 2011; Palocsay et al., 2000; Yang et al., 2010). Fourth, the AUCs under the ROC generated from our study appear to correspond with findings reported in prior research in that they did not exceed the 0.75 threshold (i.e., 0.75 is the maximum AUC value reported in previous research; Kroner and Mills, 2001; Coid et al., 2007).

In addition to the above findings, we also discovered that all four techniques examined in our study performed better at predicting misbehaving inmates than at identifying obedient inmates (i.e., all four approaches were more sensitive than specific in predicting inmate misconduct; see Table 3). Further, we found that overall, the CHAID technique performed better than the CART method at both predicting misbehaving inmates as well as identifying obedient inmates (see Table 3). Given the neglect of the CHAID technique in prior comparison studies,[24] we hope this finding will promote the inclusion of this classification tree technique in future comparison studies and projects. We also uncovered that the MLPNN method demonstrated slightly better performance at predicting obedient inmates relative to the LR technique (i.e., the specificity values of the MLPNN method were slightly better than those of the LR approach; see Table 3).

We encourage researchers and scholars to further validate the predictive performance of the above four techniques using different predictors, for different outcomes, and employing different populations. In the current study, we elected to draw solely from the importation model on inmate adaptation for our predictors because the inclusion of factors from the other leading perspectives (i.e., the deprivation, situational, and administrative control models) would prove too cumbersome. We inspire future studies to examine the predictive utility of the above four approaches utilizing predictors derived from the other leading models on inmate adaptation.

[24] To the best of our knowledge, to date, only one study has investigated the predictive utility between CHAID and LR (see Steadman et al., 2000).

Relatedly, given the lack of specification with regard to the outcome variable included in our study (i.e., inmates who were written up for or found guilty of breaking *any* prison rules versus inmates who did not break *any* prison rules), we encourage researchers to examine model performance with different outcome specifications such as comparing "true" obedient inmates with those found guilty of violent misconduct or exclude inmates found guilty of non-violent misconduct from the obedient inmate group. The above strategies would result in more homogeneous or more specific outcome category. Similarly, given the heterogeneity of the sample included in our study (i.e., almost 90 % of the inmates in our sample reported that they had used drugs previously; see Table 2), we recommend that future comparison studies consider exploring the effect of heterogeneousness on the predictive validity of all included models.

We also questioned whether our results corroborated with the importation model on inmate adjustment or whether there is any support to the theory that inmates bring with them distinct traits and characteristics leading to misconduct in prison. The sensitivity values or the rates of correctly predicting inmate misconduct generated by the LR and MLPNN models in our study ranged from 0.66 to 0.72. The overall accuracy for these two techniques ranged from 0.64 to 0.65 (see Table 3). Given that these accuracies were significantly greater than our base rate of 50 %, we can say that the results provided evidence in support of the importation perspective. If these importation variables were not good predictors, we would get a classification accuracy of 50 % which is our base rate. Nonetheless, we encourage future research to include predictor variables from the importation model and other theoretical perspectives on inmate adaptation to see if higher prediction accuracies are possible.

Finally, given the recent advocacy for wider use of statistical techniques other than conventional regression-based procedures (see Berk and Bleich, 2013; Bushway, 2013; Ridgeway, 2013) and given the evidence generated from our study and from previous research that the MLPNN model possesses comparable predictive accuracy as the LR method, we encourage researchers and scholars to further investigate the predictive performance and utility of NN models. In particular, notwithstanding the potential issues of over-fitting and the "black box"[25] nature associated with NNs, there is evidence that the NN technique is particularly effective when the primary goal is outcome prediction (i.e., forecasting) and when complex nonlinearities exist in a dataset (see Tu, 1996). Additionally, since NNs are "trained" to solve a particular type of problem (i.e., they are adaptive statistical models), this "learning" ability enables NNs to tackle a wide range of problems, some of which have proved taxing using conventional computing methods (Florio, Einsfeld, and Levy, 1994). There is also evidence that NN models are more sensitive than other types of models (i.e., LR and CART) with the inclusion of dynamic variables (Yang et al., 2010), and thus, NN models might be useful and amenable to the prediction of changing behavior and responsiveness.

[25] One issue inherent in all NN models is model transparency. Unlike the LR method, it is not possible to determine which variables contribute mainly to a particular output in a NN model (for further discussion on the issue of model transparency, see Bigi et al., 2005; Grann and Langstrom, 2007; Guerriere and Detsky, 1991; Ning et al., 2006)

Appendix 1

Decision tree diagrams are a common way to visually display classification schemes. A decision tree consists of nodes which further split or into two or more branches, creating more nodes. The diagram starts with a *root node*, which is split into two or more nodes based on some splitting rule. The splitting rule is based on the values of a certain variable. The node that splits into multiple nodes is called the *parent node* and the split nodes are called *child nodes*. The child nodes, in turn become parent nodes when they are split based on another splitting rule. When a node does not split any further, we call that node a *leaf node* or a *terminal node*. A branch ends with a *terminal node*. The terminal node shows the probability of the class in which a case belongs.

The decision tree diagram shown above displays the probability of an offender recidivating based on marital status. In this example, the root node shows the percentage of cases involving "recidivism". The root node is split into two branches based on the value of the variable "marital status". The first child node includes all the cases with a marital status of single and the second node includes all the cases with a marital status of married. The corresponding data indicate that from the sample of 100 offenders, 70 recidivated and 30 did not. Further, among the 75 offenders who were single, 68 recidivated and among the 25 offenders who were married, only 2 recidivated. Accordingly, the probability of an offender who is single recidivating is 91 % and the probability of an offender who is married recidivating is 8 %.

Appendix 2

The simple multilayer neural network architecture shown above has three layers: input, output and hidden. Each layer consists of a number of processing elements (PEs or neurons). In feed-forward neural network architectures, information is input into each PE, processed, and then passed on to each PE in the layer above. In the case of the output PE, information is simply passed out of the network.

Each PE in the input layer corresponds to a feature or characteristic that the researcher is interested in using as an independent variable. The goal of the network is to map the input units to a desired output similar to the way in which the dependent variable is a function of the independent variables in regression analysis.

The PEs are interlinked by a set of connections which are characterized by weights. In feed-forward networks with backpropagation, the networks "learn" to map the input units to the output units by adjusting the weights on the connections in response to error signals transmitted back through the network. The difference between the output of the network and the target mapping constitutes the error signal. The error signal is propagated back through the network via the PEs and their connections and the weights are updated. This process continues until the sum of all error signals is minimized.

Appendix 3

A Receiver operating characteristic (ROC) curve is a graphical plot which illustrates the performance of a binary classifier system as its discriminant threshold is varied. It is created by plotting the fraction of true positives out of all the positives (Sensitivity) on the y-axis and false positives out of the negatives (1-specificity) on the x-axis at various threshold settings. The location of a point in the ROC space depicts the classification accuracy of a classification instrument. For example, the point at coordinate of (0,1) indicates that the classification instrument has a sensitivity of 100 % and specificity of 100 % (i.e., perfect classification). Classification instruments with 50 % sensitivity and 50 % specificity can be visualized on the diagonal determined by coordinate (0,0) and coordinate (1,0) and a point predicted by a classification instrument that falls into the area *above* the diagonal represents a good prediction. Conversely, a point predicted by a classification method that falls into the area *below* the diagonal represents a bad prediction. Theoretically, a random guess would give a point on the diagonal.

The ROC curve depicts the tradeoff between the true positive rate (TPR) and false positive rate (FPR) for different cut-points of a classification instrument. The interpretation of the ROC curve is similar to the single point in the ROC space in that the closer the points on the ROC curve to the ideal coordinate (0,1) the more accurate the classification instrument is. On the other hand, the closer the points on the ROC curve to the diagonal, the less accurate the classification instrument is. In the above diagram, a typical ROC curve looks like the curved line and the area under that curve is called the AUC under ROC. Higher the AUC values, better the classifier.

References

Abdi, H., Valendin, D., & Edelman, B. (1999). *Neural networks* (Vol. 124). Newbury Park: Sage.

Aegisdottir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2006). The meta-analysis of clinical judgment project: fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist, 34*, 341–382.

Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit, and probit models* (Vol. 45). Newbury Park: Sage.

Berk, R. A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: a comparative assessment. *Criminology and Public Policy, 12*, 513–544.

Bigi, R., Gregori, D., Cortigiani, L., Desideri, A., Chiarotto, F. A., & Toffolo, G. M. (2005). Artificial neural networks and robust Bayesian classifiers for risk stratification following uncomplicated myocardial infartion. *International Journal of Cardiology, 101*, 481–487.

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

Bonta, J. (1996). Risk-needs assessment and treatment. In A. T. Harland (Ed.), *Choosing correctional options that work: defining the demand and evaluating the supply* (pp. 18–22). Thousand Oaks: Sage.

Breiman, L. (2001). Decision tree forest. *Machine Learning, 45*, 5–32.

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

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

Bushway, S. D. (2013). Is there any logic to using logit: finding the right tool for the increasingly important job of risk prediction. *Criminology and Public Policy, 12*, 563–567.

Byrne, J. M., & Hummer, D. (2007). Myths and realities of prison violence: a review of the evidence. *Victims and Offenders An International Journal of Evidence-based Research Policy and Practice, 2*, 77–90.
Cao, L., Zhao, J., & Van Dine, S. (1997). Prison disciplinary tickets: a test of the deprivation and importation models. *Journal of Crime and Justice, 25*, 103–113.
Carpenter, & Grossberg. (1991). Causal attributions in expert parole decisions. *Journal of Personality and Social Psychology, 36*, 1501–1511.
Caulkins, J., Cohen, J., Gorr, W., & Wei, J. (1996). Predicting criminal recidivism: a comparison of neural network models with statistical methods. *Journal of Crime and Justice, 24*, 227–240.
Clemmer, D. (1940). *The prison community*. Boston: Christopher.
Coid, J., Yang, M., & Ullrich, S., et al. (2007). Predicting and understanding risk of reoffending: The prisoner cohort study. *Research Summary,* Ministry of Justice 6.
Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674.
Dhami, M. K., Ayton, P., & Lowenstein, G. (2007). Adaption to imprisonment: indigenous or imported? *Criminal Justice and Behavior, 34*, 1085–1100.
DiIulio, J. J., Jr. (1987). *Governing prisons: a comparative study of correctional management*. New York: Free Press.
Egan, J. P. (1975). *Signal detection theory and ROC analysis*. New York: Academic.
Florio, T., Einfeld, S., & Levy, F. (1994). Neural network and psychiatry: candidate applications in clinical decision making. *Australian and New Zealander Psychiatry, 28*, 651–666.
Friedman, J. H. (1999). *Stochastic gradient boosting*. Stanford: Stanford University.
Gardner, M. J., & Altman, D. G. (1989). Estimating with confidence. In M. J. Gardner & D. G. Altman (Eds.), *Statistics with confidence* (pp. 6–19). London: British Medical Journal.
Gardner, W., Lidz, C. W., Mulvey, E. P., & Shaw, E. C. (1996). A comparison of actuarial methods for identifying repetitively violent patients with mental illnesses. *Law and Human Behavior, 20*, 35–48.
Gendreau, P., Goggin, C. E., & Law, M. A. (1997). Predicting prison misconducts. *Criminal Justice and Behavior, 24*, 414–431.
Gendreau, P., Goggin, C. E., & Smith, P. (2002). Is the PCL-R really the "unparelleled" measure of offender-risk? A lesson in knowledge accumulation. *Criminal Justice and Behavior, 29*, 397–426.
Glover, A., Nicholson, D., Hemmati, T., Bernfeld, G., & Quinsey, V. (2002). A comparison of predictors of general and violent recidivism among high risk federal offenders. *Criminal Justice & Behavior, 29*, 235–249.
Goodstein, L., & Wright, K. N. (1989). Inmate adjustment to prison. In L. Goodstein & D. L. MacKenzie (Eds.), *The American prison: issues in research and policy* (pp. 229–251). NY: Plenum.
Gottfredson, S. D., & Gottfredson, D. M. (1986). Accuracy of prediction models. In A. Blumstein, J. Cohen, J. Roth, & C. A. Visher (Eds.), *Criminal careers and "Career Criminals"* (pp. 212–290). Washington: National Academy of Sciences Press.
Gottfredson, S. D., & Moriarty, L. J. (2006). Statistical risk assessment: old problems and new applications. *Crime and Delinquency, 52*, 178–200.
Grann, M., & Langstrom, N. (2007). Actuarial assessment of violence risk: to weigh or not to weigh? *Criminal Justice and Behavior, 34*, 22–36.
Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithm) prediction procedures: the clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*, 293–323.
Guerriere, M. R., & Detsky, A. S. (1991). Neural networks: what are they? *Annals of Internal Medicine, 115*, 906–907.
Gurney, K. (1997). *An Introduction to neural networks*. New York: UCL Press.
Hanson, R. K. (2005). Twenty years of progress in violence risk assessment. *Journal of Interpersonal Violence, 20*, 212–217.
Hanson, R.K. & Morton-Bourgon, K.E (2007) The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis. Public Safety and Emergency Preparedness Canada.
Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: a meta-analysis of 118 prediction studies. *Psychological Assessment, 21*, 1–21.
Harer, M. D., & Steffensmeier, D. J. (1996). Race and prison violence. *Criminology, 34*, 323–355.
Hill, T., & Lewicki, P. (2006). *Statistics, methods and application: a comprehensive reference for science, industry, and data mining*. Tulsa: StatSoft, Inc.
Hilton, N. Z., Harris, G. T., & Rice, M. E. (2006). Sixty-six years of research on the clinical versus actuarial prediction of violence. *The Counseling Psychologist, 34*, 400–409.
Hosmer, D. W., & Lemeshow, S. (1989). *Applied logistic regression*. New York: Wiley.

Irwin, J. K. (1981). Sociological studies of the impact of long term confinement. In D. A. Ward & K. F. Schoen (Eds.), *Confinement in maximum custody* (pp. 33–68). Lexington: D.C. Health.

Irwin, J. K., & Cressey, D. (1962). Thieves, convicts, and the inmate culture. *Social Problems, 10*, 142–155.

Jiang, S., & Fisher-Giorlando, M. (2002). Inmate misconduct: a test of the deprivation, importation, and situational models. *The Prison Journal, 82*, 335–358.

Jones, P. R. (1996). Risk prediction in criminal justice. In A. T. Harland (Ed.), *Choosing correctional options that work: defining the demand and evaluating the supply* (pp. 33–68). Thousand Oaks: Sage.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*, 119–127.

Kroner, D. G., & Mills, J. F. (2001). The accuracy of five appraisal risk instruments in predicting institutional misconduct and new convictions. *Criminal Justice and Behavior, 28*, 471–489.

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*, 547–573.

Loh, W. Y., & Shih, Y. S. (1997). Split selection methods for classification trees. *Statistica Sinica, 7*, 815–840.

Menzies, R., Webster, S. D., McMain, S., Staley, S., & Scaglione, R. (1994). The dimensions of dangerousness revisited. *Law and Human Behavior, 18*, 1–28.

Minsky, M., & Papert, S. (1969). *Perceptrons*. Cambridge: MIT Press.

Mossman, D. (1994). Assessing prediction of violence: being accurate about accuracy. *Journal of Consulting and Clinical Psychology, 62*, 783–792.

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

Ning, G. M., Su, J., Li, Y. Q., Wang, X. Y., Li, C. H., & Yan, W. M. (2006). Artificial neural network base model for cardiovascular risk stratification in hypertension. *Medical and Biological Engineering and Computing, 44*, 202–208.

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

Paterline, B. A., & Petersen, D. M. (1999). Structural and social psychological determinants of prisonization. *Journal of Crime and Justice, 27*, 427–441.

Perlich, C., Provost, F., & Simonof, J. (2003). Tree induction vs. logistic regression: a learning curve analysis. *Journal of Machine Learning Research, 4*, 211–255.

Price, R. K., Spitznagel, E. L., Downey, T. J., Meyer, D. J., Risk, N. K., & El-Ghazzawy, O. G. (2000). Applying artificial neural network models to clinical decision making. *Psychological Assessment, 12*, 40–51.

Rice, M. E., & Harris, G. T. (1995). Violent recidivism: assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63*, 737–748.

Ridgeway, G. (2013). Linking prediction and prevention. *Criminology and Public Policy, 12*, 545–550.

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

Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: theory and application*. Hackensack: World Scientific Publishing.

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

Rumelhart, D. E., & McClelland, J. L. (1986). *Parallel distributed processing* (Vol. 1). Cambridge: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1988). *Parallel distributed processing* (Vol. 1 and 2). Cambridge: MIT Press.

Seymour, J. (1977). Niches in prisons. In H. Toch (Ed.), *Living in prison: the ecology of survival* (pp. 18–22). New York: Free Press.

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

Singh, J. P., & Fazel, S. (2010). Forensic risk assessment: a metareview. *Criminal Justice and Behavior, 37*, 965–988.

Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: a systematic review and metaregression analysis of 68 studies and 25, 980 participants. *Clinical Psychology Review*. doi:doi:10.1016/j.cpr.2010.11.009.

Snyder, H. N. (2011). *Patterns & Trends: Arrests in the United States, 1980–2009*. Bureau of Justice Statistics: U.S. Department of Justice.

Sorensen, J., Wrinkle, R., & Gutierrez, A. (1998). Patterns of rule-violating behaviors and adjustment to incarceration among murderers. *The Prison Journal, 78*, 222–231.

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*, 253–271.

StatSoft Inc (2008) Data mining, predictive analytics, statistics, StatSoft electronic textbook. http://www.statsoft.com/textbook/.

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

Steinke, P. (1991). Using situational factors to predict types of prison violence. *Journal of Offender Rehabilitation, 17*, 119–132.

Sykes, G. M. (1958). *The society of captives*. Princeton: Princeton University Press.

Thomas, S., & Leese, M. (2003). A green-fingered approach can improve the clinical utility of violence risk assessment tools. *Criminal Behavior and Mental Health, 13*, 153–158.

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

Toch, H. (1977). *Living in prison: the ecology of survival*. New York: Free Press.

Toch, H., & Adams, K. (1986). Pathology and disruptiveness among prison inmates. *Journal of Research in Crime and Delinquency, 23*, 7–21.

Tu, J. V. (1996). Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *Journal of Clinical Epidemiology, 9*, 1225–1231.

Wallace, J. M., & Bachman, J. G. (1991). Explaining racial/ethnic differences in adolescent drug use: the impact of background and lifestyles. *Social Forces, 38*, 333–354.

White, H. (1989). Some asymptotic results for learning in single hidden-layer feedforward network models. *Journal of the American Statistical Society, 84*, 1003–1013.

Wooldredge, J. D. (1991). Correlates of deviant behavior among inmates of U.S. correctional facilities. *Journal of Crime and Justice, 14*, 1–25.

Wright, K. N. (1991). A study of individual, environmental, and interactive effects in explaining adjustment to prison. *Justice Quarterly, 8*, 217–242.

Yan, L., Dodier, R., Mozer, M.C., & Wolniewicz, R (2003) Optimizing classifier performance via the Wilcoxon-Mann–Whitney statistics. *Proceedings of the International Conference on Machine Learning.*

Yang, M., Liu, Y.Y. & Coid, J.W (2010) Applying neural networks and classification tree models to the classification of serious offenders and the prediction of recidivism. Research summary, Ministry of Justice, UK, available online at www.justice.gov.uk/publications/research.htm.

Fawn T. Ngo is an Assistant Professor in the College of Arts & Sciences at the University of South Florida Sarasota-Manatee. Her research interests include criminological theory, evaluative research, and quantitative methods in criminology.

Ramakrishna Govindu is an Instructor in the College of Business at the University of South Florida, Sarasota. He got his Ph.D. in Industrial Engineering from Wayne State University. His research interests include Advanced Analytics, Supply Chain Modeling, Simulation, and Optimization.

Anurag Agarwal is a professor in the College of Business at the University of South Florida, Sarasota. His Ph.D. is from The Ohio State University. His research interests include data mining, artificial intelligence, neural networks and optimization.