J Quant Criminol
DOI 10.1007/s10940-011-9137-7

ORIGINAL PAPER

A Comparison of Logistic Regression, Classification and Regression Tree, and Neural Networks Models in Predicting Violent Re-Offending

Yuan Y. Liu · Min Yang · Malcolm Ramsay · Xiao S. Li · Jeremy W. Coid

Abstract Previous studies that have compared logistic regression (LR), classification and regression tree (CART), and neural networks (NNs) models for their predictive validity have shown inconsistent results in demonstrating superiority of any one model. The three models were tested in a prospective sample of 1225 UK male prisoners followed up for a mean of 3.31 years after release. Items in a widely-used risk assessment instrument (the Historical, Clinical, Risk Management-20, or HCR-20) were used as predictors and violent reconvictions as outcome. Multi-validation procedure was used to reduce sampling error in reporting the predictive accuracy. The low base rate was controlled by using different measures in the three models to minimize prediction error and achieve a more balanced classification. Overall accuracy of the three models varied between 0.59 and 0.67, with an overall AUC range of 0.65–0.72. Although the performance of NNs was slightly better than that of LR and CART models, it did not demonstrate a significant improvement.

Keywords Violence reconviction · Risk assessment · Neural networks · Classification and regression tree · HCR-20

Y. Y. Liu · M. Yang · X. S. Li (✉)
Department of Health Statistics, School of Public Health, Sichuan University, 610041 Chengdu, China
e-mail: lixiaosong1101@126.com

M. Yang (✉)
Division of Psychiatry, School for Community Health Sciences, University of Nottingham,
9 Triumph Road, Nottingham NG7 2TU, UK
e-mail: Min.Yang@nottingham.ac.uk

M. Ramsay
Partnerships and Health Strategy Unit, Ministry of Justice, 102 Petty France (9th floor),
London SW1H 9AJ, UK

J. W. Coid
Forensic Psychiatry Research Unit, Queen Mary University of London,
William Harvey House, 61 St. Bartholomew's Close, London EC1A 7BE, UK

Published online: 01 April 2011

Introduction

Future violence is of major importance in the aftercare of offenders and for public protection, but its management is dependent on an acceptable level of accuracy in estimating future risks. Structured risk assessment instruments are increasingly used to achieve this goal as they are considered more accurate in the prediction of future violent and sexual behavior than subjective clinical judgment (Dawes et al. 1989; Grove and Meehl 1996). However, recent studies have identified limitations to these instruments, often due to their developmental origins, and have noted their loss of accuracy when applied to different offender populations, particularly those with different characteristics from the population on which the instrument was originally standardized (Gendreau et al. 2002; Glover et al. 2002; Kroner and Mills 2001). Comparison studies of different instruments suggest there may be a "glass-ceiling" effect, with most instruments demonstrating only moderate accuracy beyond which further improvement is seldom achieved. Furthermore, instruments designed to predict violence lack outcome specificity for violence and their predictive items include measures predicting general criminality (Coid et al. 2011; Glover et al. 2002; Hemphill et al. 1998). These limitations pose the question whether new statistical methodologies can be employed to improve performance.

It has been recognised that good performance in predicting re-offending can be determined by many factors. Among these, "high specificity of predictors for good predictive power", "well defined criteria of outcome for distinguishable categories", "high specificity of target population for homogeneity", and "adequate statistical methods to achieve the best discriminate effects based on the data" are key elements. Steadman and Monahan (1994) highlighted the major methodological challenges in predicting re-offending in terms of *limited range of predictors, weak criterion (outcome) variables, constricted validation samples, and unsynchronized research*. Whilst considerable research has been carried out into improvement of the first good performance factors, and especially into the specificity of predictor variables (Hanson 2005), less attention has been given to the identification of the best statistical models to use when predicting re-offending.

Previous Studies of the Three Classification Models in Violence Prediction

Logistic Regression

In both risk assessment practice and research, a dichotomous classification for an individual's risk behavior, such as recidivism (general or violent) or not, has been the ultimate outcome. For the prediction of such an outcome, Logistic Regression (LR) has emerged as the conventional statistical technique of choice in the development of new models and also in the testing of existing instruments (Hosmer and Lemeshow 1989). Many of its applications can be found in the fields of psychiatry and psychology (Hartvig et al. 2006; Lin et al. 2007; Thomas et al. 2005). LR applies maximum likelihood estimation after transforming the dependent into a logit variable (the natural log of the odds of the dependent occurring or not). In this way, LR estimates the probability of a certain event occurring. Unlike ordinary least squares regression, LR neither assumes linearity of relationship between the independent variables and the dependent, nor does it require normally distributed variables or assume homoscedasticity. Hence it generally has less stringent requirements (Harper 2005). However, it was criticized for ignoring the possibility that different variables might predict violence for different subgroups of individuals (Steadman

et al. 2000). Explanations based on LR are therefore sometimes seen as undesirable for clinicians and other offender managers.

Classification Tree Models

Certain previous authors have argued that violence risk assessment instruments should reflect actual clinical thinking processes and that Classification Tree (CT) models may be a better representation of how clinicians typically make their risk judgments (Steadman et al. 2000). The CT model was developed by Breiman et al. (1984) and is a questions-decision-tree model. A hierarchy of questions is asked and the final decision is made depending on the answers to all the previous questions. There is a family of CT models including Classification and Regression Tree (CART, also C&R or CRT) (Breiman et al. 1984), Chi-squared and Interactive Decision (CHAID) (Kass 1980), QUEST (Quick, Unbiased, Efficient Statistical Trees) (Loh and Shih 1997), C4.5 (the former version is ID3, and the latter version is C5.0/See5) (Quinlan 1993, 1996), Decision Tree Forests (also Random Forest) (Breiman 2001), and Boosting Trees (Friedman 1999a, b). CT models are widely used in applied fields such as developing diagnostic tests in medicine (Colombet et al. 2000) and decision theory study in psychology (Gardner et al. 1996; Steadman et al. 2000). In recent years, studies have emerged on predicting violent risk levels, or violence or overall recidivism among psychiatric patients or criminal offenders (Banks et al. 2004; Gardner et al. 1996; Monahan et al. 2000; Monahan et al. 2005, 2006; Rosenfeld and Lewis 2005; Silver and Chow-Martin 2002; Silver et al. 2000; Stalens et al. 2004; Steadman et al. 2000; Thomas et al. 2005).

The Iterative Classification Tree (ICT) approach was developed by Steadman et al. (Banks et al. 2004; Monahan et al. 2000, 2005, 2006; Silver and Chow-Martin 2002; Silver et al. 2000; Steadman et al. 2000). Being a new actuarial method for violence risk assessment, this method is enhanced by combining several standard CT models (e.g. CHAID or CART) in an iterative or repetitive fashion. Subjects not classified into designated groups (either high risk or low risk) in the first iteration of CHAID were pooled together and re-analyzed in a second iteration of CHAID. This iterative process continued until it was not possible to classify any additional groups of subjects as either high or low risk (with no groups allowed to contain fewer than 50 cases) (Monahan et al. 2000). However, it was argued that the iterative process generated a complex classification system no simpler than a traditional regression approach (Rosenfeld and Lewis 2005). From the ICT model introduced by Steadman et al. (2000) to the "Classification of Violence Risk (COVR)" software based on the ICT model (Monahan et al. 2006), there have been studies applying the model mainly to the data from the MacArthur Violence Risk Assessment Study (Steadman et al. 1998) and the New Jersey data (National Institute of Justice 1992; Smith 1996; Smith and Smith 1998). The aim of these studies was to classify subjects into high-risk, low-risk, and average (or unclassified) risk groups. The authors used two decision thresholds (or cut-offs) to classify three categories of subjects: high, low, and average (or unclassified) risk. One threshold was twice the base prevalence rate of violence in the full study sample; this was used to assign cases with predicted probability of violence greater than this threshold to the high-risk group. Another threshold was half the base prevalence rate of violence in the full study sample, which was used to assign cases with predicted probability of violence less than this threshold to the low-risk group. Those cases with a predicted probability of violence in between the two thresholds were classified as average (or unclassified) risk group. So far, few validation studies incorporating the gold standard of three such classification groups have been published. The only report of the

validation performance of the ICT model was in AUC values and for a classification involving two categories (violent recidivism or not), instead of three classification groups (high-risk group, low-risk group, and unclassified-risk group).

Gardner et al. (1996) compared performance between the CART and Negative Binominal Regression (NBR) among 784 psychiatric patients in a prospective longitudinal study (Lidz et al. 1993). The included predictors were demographics (e.g. age), clinical information (e.g. major diagnoses), research covariates during the follow-up in the community (e.g. compliance with psychiatric medications), and the Brief Symptom Inventory (Derogatis and Melisaratos 1983). The dependent variable in this study comprised incidents of community violence (count or continuous variable), and the CART used was the regression tree instead of the classification tree. At a specific cut-off point, the sensitivities for CART and NBR were reported as 7.7 and 9.3%, respectively, and the specificities for CART and NBR as 99.2 and 99.1%, respectively. The bootstrapped validation did not show differentiated performance between the two models. Therefore, the authors concluded that predictions based on the regression trees of the CART algorithm were as accurate as predictions based on the conventional regression model.

Stalens et al. (2004) compared classification tree analysis (CTA) via optimal discriminant analysis (Yarnold 1996) with LR to predict violent recidivism among a sample of 1,344 violent offenders on probation during the 4 weeks from October 30, 2000 to November 30, 2000 in Illinois in USA. Predictors in their models were: subsets of violent offenders, type of prior violence, frequency of prior arrests, demographics, offense characteristics, substance use, mental health, probation conditions and dynamic predictors. Results showed that AUC values were 0.67 for CTA and 0.71 for LR, and sensitivities 35.2% for CTA and 9.8% for LR, and specificities 88.4% for CTA and 98.7% for LR, and overall accuracies 78.6% for CTA and 81.8% for LR. Bootstrapped validation was conducted for CTA model only. The authors concluded that CTA was superior to LR for balancing the number of false positives and false negatives. Such a conclusion was arguable because the imbalanced sensitivity and specificity[1] could be avoided by adjusting the cut-off probability according to the base rate of violent recidivism.

Rosenfeld and Lewis (2005) applied both CART and LR models to predict violent reconviction in a sample of 204 stalking offenders referred to forensic psychiatry clinic in New York City between January 1, 1994 and December 31, 1998 (Rosenfeld and Harmon 2002). The predictors included a number of demographic, clinical, and offense-related variables. Three nested CART models were constructed by increasing the number of predictors from 5 (Model 1) to 9 (Model 2) and to 13 (Model 3) in each successive model. Similarly, three LR models were built using the same predictors as Model 1–Model 3 of CART. Results showed no statistical difference between the two models in predictive accuracy for the construction sample, with AUC values for CART from 0.79 to 0.85, as compared to 0.78 to 0.80 for the LR. However, their jack-knifed cross-validation resulted in somewhat lower accuracy for the CART with AUC values from 0.64 to 0.66 as compared to LR with AUC values from 0.71 to 0.74. This study demonstrated similar performance of the two models in terms of the construction sample and considerable

[1] $$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}}$$

$$\text{specificity} = \frac{\text{number of true negatives}}{\text{number of true negatives} + \text{number of false positives}}$$

shrinkage in the CART model among validation samples. LR models were much more resilient to cross-validation, with relatively modest loss in predictive power.

Thomas et al. (2005) also conducted a comparison between LR and CART in predicting violence in patients with psychosis. The data on 708 psychotic patients was from the UK700 study, a large randomized controlled trial in four inner-city mental health services in the United Kingdom (UK700 Group 1999). Predictors included sociodemographic risk factors, historical risk factors, and clinical/diagnostic factors. The full LR model yielded 19% sensitivity, 96% specificity, and 78% accuracy in the construction sample, and 19% sensitivity, 94% specificity, and 77% accuracy after tenfold cross-validation. Measures for the pruned CART were 29, 98, and 82%, respectively, in the construction sample, and 14, 93, and 75%, respectively, after the tenfold cross-validation. The authors concluded that although classification trees could be suitable for routine clinical practice (based on the simplicity of their decision-making processes), their robustness and therefore clinical utility were questionable. Further research was required to compare such models in large prospective epidemiologic studies of other psychiatric populations.

To conclude, two studies out of eleven reported similar accuracy between CT and conventional regression models in construction samples and greater shrinkage of the former in validation samples; one study showed CT was slightly worse; one study found no difference between the two; and seven reported better performance of ICT models without adequate validation supports. In addition, CT methods remain less frequent in research within forensic settings compared to other clinical decision-making settings (Rosenfeld and Lewis 2005; Thomas et al. 2005). The efficacy of CT models needs to be further validated (Colombet et al. 2000; Dillard et al. 2007; Thomas et al. 2005; Trujillano et al. 2008).

Neural Networks Models

Neural Networks (NNs) models emerged from research in artificial intelligence, mostly inspired by attempts to mimic the fault-tolerance and "capacity to learn" of biological neural systems by modeling the low-level structure of the brain (Patterson 1996). As a data mining technique, they are mostly used in computationally intensive applications to identify complex patterns and relationships between multiple inputs that are not recognizable by the human brain (Bigi et al. 2005). Given the complexity in predicting human behaviour, with a large amount of clinical data and evidence relating to risk assessment research and practice, this model is potentially a useful tool with which to complete a risk assessment task. The application of NNs for prediction or classification in psychiatry has also attracted growing interest (Florio et al. 1994; Starzomska 2003; Price et al. 2000). However, the literature on predicting criminal recidivism using NNs is sparse and the findings are inconsistent.

Brodzinski et al. (1994) compared NNs to traditional discriminant analysis in developing a predictive model for a sample of 778 US juvenile probation cases between 1985 and 1986. Their eight predictors included age at first adjudication, school functioning, documented substance abuse, criminal activity by family members, peer group relations, prior adjudications for unruliness, delinquency, and probation violations. The authors randomly selected 390 cases as the construction sample and 388 as the validation sample. They demonstrated 99.48% classification accuracy on the validation data using NNs, but only 65.68% accuracy in the construction and validation combined samples using discriminant analysis. However, because the accuracy for the construction sample by NNs was not reported in the paper, it was hard to judge if the high validation accuracy was due to over-fitting by the NNs in the testing sample.

Caulkins et al. (1996) used both multilayer neural networks (MNN) and conventional statistical methods (including multiple regression, association analysis, and predictive attribute analysis) to model risk of recidivism among 3,389 offenders released from prisons in the United States from 1970 to 1972. Their models included 29 predictive variables that reflected index offense and criminal history as well as social history and institutional adjustment. Two-thirds of the data was used for the construction sample and the remaining one-third as the validation sample. To compare the performance of different models, the mean cost rating (MCR[2]) was mainly used. Their study showed that the MCR value for the MNN model was 0.460 and ranged from 0.338 to 0.440 for other models in the construction sample. For the validation sample, the MCR was 0.416 for MNN and 0.328–0.436 for other models. This study additionally compared MCR and percentage of total correct predictions (TCP) between the MNN and multiple regression based on eight predictors, including longest time free, crime group, prison punishment, living arrangement, age at first arrest, known use of synthetic and/or natural opiates, type of admission, and commitment offense. The MCR and TCP values from the construction sample were reported, respectively, as 0.412 and 0.670 for multiple regression, 0.421 and 0.677, respectively, for MNN. From the validation sample, the MCR and TCP values were 0.412 and 0.678, respectively, for multiple regression, 0.405 and 0.686, respectively, for MNN. The authors concluded that MNN did not demonstrate any advantage in predictive accuracy over the other models, and that currently available predictors had limited information for discriminating recidivists regardless of the models used.

Palocsay et al. (2000) reported an investigation of the performance of three-layer back-propagation neural networks and multivariate logistic regressions in predicting overall recidivism based on a sample of more than 10,000 offenders released from prisons in North Carolina in the United States in 1978 and in 1980. Nine predictor variables were used in both models for the classification, including: African-American ethnicity, previous serious alcohol problem, history of using hard drugs, sentenced for a felony or misdemeanor, sentenced for a crime against property or not, gender, number of previous incarcerations, age at release, and sentence served in months. In the 1978 sample, 1,357 participants were used as training sample, 183 as monitoring sample, and 3,078 as test sample. In the 1980 sample, the sample sizes for training, monitoring, and test were 1,263, 172 and 4,304, respectively. The overall rate of correct classification varied between 60 and 69% for both methods, with NNs models moderately but nevertheless significantly outperforming the LR model. The authors concluded that although performance of NNs depends heavily on network topology, NNs might be competitive with, and could offer certain advantages over, traditional statistical models in this domain.

Most recently, Grann and Langstrom (2007) took a different approach, examining weighting procedures in constructing a risk assessment scale to predict violent reconviction using a conventional approach including Nuffield's weighting, weighting generated by bivariate LR and multivariate LR, and by NNs models, with 404 forensic patients in Sweden. The 10 Historical items of the HCR-20 (a risk assessment instrument also widely used for research purposes—for further details see below) were used as predictors in this study. It demonstrated that simple weighting techniques such as Nuffield's weighting do not improve predictive accuracy over that of a non-weighted model reference. Furthermore, a complex procedure such as NNs is subject to a statistical shrinkage effect due to over-fitting problems. The authors hypothesized that including causal risk factors rather

[2] MCR is a statistic for measuring the accuracy of instruments to predict recidivism, by means of complex computation. For more details about MCR see: Greene et al. (1994).

than applying weighting procedures in any type of model could be an important factor which improves predictive accuracy.

To conclude, in the four studies above comparing NNs with traditional methods in the prediction of criminal recidivism, two studies (Brodzinski et al. 1994; Palocsay et al. 2000) showed that NNs were better while the other two reported that NNs were not superior to (Caulkins et al. 1996) or worse than (Grann and Langstrom 2007) the traditional methods.

Previous studies comparing predictive performance of these classification models reported inconsistent findings between traditional models, CT models, and NNs models. Specificity of predictors, homogeneity of samples, definition criteria of target outcomes, retrospective or prospective collection of outcomes, computational approach for model validation, and handling of model over-fitting could all contribute to inconsistency in model performance. At present there has been no study that compares traditional methods with CT and NNs models head to head in the prediction of violent reconviction.

Aims of Study

Previous studies presented some problems, such as: sample size too small, inflated predictive accuracy of models due to inadequate validation, misleading validity due to lack of technical knowledge in handling the low base rate, and lack of understanding of strength or limits of different models. As these methodological issues need to be clarified or addressed in the application of actuarial risk assessment, this study aimed to fill some of the gaps in this field. Based on a prospective cohort designed to collect violent outcomes from a high risk male prisoner population in England and Wales, and with sufficient sample size, it compared predictive performance of all three types of model: LR, CT, and NNs. More vigorous methods to control for over-fitting, to handle the low base rate, and to validate as well as test models, were applied.

Methods

Sample and Measures

This was a prospective study of a cohort of male prisoners in England and Wales released between 14 November 2002 and 7 October 2005. The offender cohort sample was generated from the Prison Service Inmate Information System, or Central System Database, if they met the following criteria: (1) serving a prison sentence of two years or more for a sexual or violent principal offence (excluding life sentence prisoners), (2) aged 17 years and over, and (3) having one year left to serve, at time of selection. Information was provided on previous criminal history using the Home Office Offenders Index on all prisoners in England and Wales meeting these criteria. Participants were interviewed during a 6–12 month period before their expected date of release by trained interviewers using a battery of clinical and risk assessment measures for violent and other criminal behavior. The violent offences comprised homicide, major violence, minor violence, weapons offences, aggravated burglary, and robbery.

A total of 1,363 prisoners were interviewed by twelve trained research assistants, who were psychology graduates. They started by establishing the criminal history and nature of the index offence. The Structured Clinical Interview for DSM-IV Axis II disorder was administered to establish diagnoses of personality disorder (PD), and modules from the Structured Clinical Interview for DSM-IV AXIS I disorders, were administered to measure

the presence of current or lifetime schizophrenia or delusional disorder, depressive disorder, drug and alcohol use disorders or dependence, followed by the risk assessment instruments. Outcome data were derived from reconvictions recorded in the Police National Computer (PNC), an operational police database containing criminal histories of all offenders in England, Wales and Scotland up to the date 09 February 2007. This source has a lower failure rate than the previous Home Office Offenders Index for non-identification and is updated more regularly (Howard and Kershaw 2000).

Out of the original sample of 1363 male prisoners, we deleted: (1) 131 cases that had missing values on predictors, and (2) a further 7 cases whose follow-up times were less than 1 year. The final sample in the present study was 1,225. The mean follow-up time was 3.31 years, with a range of 1.34–4.24 years. There were 22 cases whose follow-up time less than 2 years, 274 cases with [2, 3) follow-up years, 836 cases with [3, 4) follow-up years, and 93 cases more than 4 follow-up years. (See Table 1).

The target variable was violent re-reconviction, i.e. the predicted outcome for each prisoner in either a violent or non violent re-offending category. The violent re-offending rate was 28.0% in this cohort, namely, 343 prisoners were reconvicted for violent offences, whilst 882 prisoners had at most non violent re-offending (including 499 prisoners with no further convictions or 'no re-offending'). The choice of the target outcome category in this study was mainly for the purpose of comparison as most previous studies in this field categorized the target variable in this way. A three category outcome (violent re-offending\ other re-offending\ no re-offending) would complicate the analysis. Detailed investigation of three category outcome has been published elsewhere (Yang et al. 2010).

Predictors Used in Study: HCR-20

Performance of any statistical model is ultimately determined by the discriminate power and specificity of predictors included in the model for a given outcome. For this study, we chose the 20 items in the Historical, Clinical, Risk Management-20 (HCR-20; Webster et al. 1997) as model predictors for violent behavior. Although the instrument is recommended for structured clinical guidance and not for scoring individuals as an actuarial measure of risk, nevertheless, it is the most widely researched, empirically-based guide to risk assessment, and it has dynamic risk factors that are known to change through intervention (Dolan and Khawaja 2004).

The predictors were the 20 items of the HCR-20, which consisted of: 10 items related to historical factors (e.g. employment problems), 5 items related to current clinical presentation (e.g. lack of insight) and 5 items related to future risk factors (e.g. lack of personal support) (Webster et al. 2002). Each item was scored as 0 (not present), 1 (partially or possibly present) or 2 (present), leading to a maximum total score of 40.

Table 1 Numbers and rates of any re-offending and violent re-offending by follow-up years

Follow-up years	Prisoners	Any re-offenders (rate %)	Violent re-offenders (rate %)
[1, 2)	22	7 (31.8)	2 (9.1)
[2, 3)	274	142 (51.8)	51 (18.6)
[3, 4)	836	507 (60.6)	254 (30.4)
≥4	93	70 (75.3)	36 (38.7)
Total	1,225	726 (59.3)	343 (28.0)

As the HCR-20 was developed in North America, much of the evidence base for its efficacy is in that population. Although the HCR-20 is increasingly adopted into clinical practice in European forensic settings including Germany, Sweden, and recently the Netherlands (Dahle 2006; Grann and Langstrom 2007; Vogel et al. 2004), there is relatively little UK research that uses the HCR-20 as a core component of routine clinical practice. Gray et al. (2003) found the HCR-20 was an excellent predictor of inpatient violence in mentally disordered offenders over a short period. Grevatt et al. (2004), also looking at short-term in-patient violence, agreed on the efficacy of the Clinical scale of the HCR-20, but did not find any efficacy for the Historical scale. Doyle and Dolan (2006) showed that both the total score and the Historical scale of the HCR-20 were good predictors over a short interval of 24 weeks for the prediction of violence outside the institution. They also noted that the addition of dynamic variables (such as the Clinical and Risk scales of the HCR-20) would improve upon purely historical baseline measures. Gray et al. (2008) concluded that the HCR-20 has good validity in predicting both violent and non-violent offending following release from medium secure units for male forensic psychiatric patients in the UK. Examining each item of the HCR-20 for its efficacy of predicting violent reconviction and violent reconviction leading to imprisonment among 190 prisoners in a Scottish prison, and using survival analysis, the study (Cooke et al. 2001) indicated that many of them were significantly related to the outcome variables. The study used items from the HCR-20 as predictors primarily for the purpose of model comparison.

Statistical Models

Logistic Regression

LR was used to classify individuals in the target categories based on the logistic function. It is related to the probability of the chosen outcome event, for instance, in this study, the probability of a prisoner incurring a violent reconviction. Unlike the standard multiple regression function, the logistic function forces estimated probabilities to lie within the range 0–1, which is more sensible than linear regression as a means of modelling probabilities (Manly 2005). Equation 1 represents the definition of the general logistic function (Hosmer and Lemeshow 1989).

$$\log[p/(1-p)] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i \quad (1)$$

where p is the estimated probability of the outcome of interest, β_0 is called the intercept term, β_1, …, β_i are called the regression coefficients associated with predictors x_1, …, x_i accordingly, and i indicates the ith predictor. Based on the estimated probability, a classification cut-off can be defined to determine two target groups and all cases are classified into the defined category groups, violent or non-violent in this study.

By default, LR sets cut-off probability at 0.50 for classification of two categories, namely equal *misclassification cost* for the two categories, which in turn results in a classification in favour of the category with greater weight in terms of the overall proportion or *prior probability*. For example in this study, the proportion of violent reoffending category is 28%, which is lower than that of the contrasting category (72%). Equal misclassification cost means no constraint to the prior distribution in proportion. The final classification weighting is 0.14 (0.50×0.28) for the former category and 0.36 (0.50×0.72) for the latter. As a result, the prediction or classification is weighted in

favour of the majority category (the non-violent category), with high specificity and low false negative prediction in contrast to low sensitivity and high false positive prediction for the violent category. Such a consequence in reality is less desirable than the other way round, if no balanced result can be achieved. In order to achieve a better balance for the prediction we changed the classification cut-off from the equal weight 0.50/0.50 to 0.28/0.72, to assign a higher misclassification cost to the smaller category, so as to achieve equal weighting for both categories. LR models were built using SPSS 16.0 software package (SPSS Inc. 2008).

Classification and Regression Tree (CART)

We chose CART from the range of CT models. The CART is one of the most widely used CT models and has been successfully used in many medical decision-making areas (Harper 2005). There is increasing recognition that CART might be both clinically feasible and useful for predicting violent outcomes (Gigerenzer et al. 1999; Silver et al. 2000; Thomas and Leese 2003), especially for large samples with a relatively low base rate of violence (Thomas et al. 2005). CART was developed by Breiman et al. (1984) and improved further by Ripley (1996). This model is a non-parametric technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numerical, respectively. Similar to many other tree models, CART uses a sequential process to identify the predictor variables that best differentiate groups along the outcome variable of interest. The sample was then split into two "branches" based on this predictor.[3] Subsequent steps identified the best predictor within each of these branches and this process was repeated until no more variance could be explained with the remaining variables or until another criterion, such as a minimum group size, had been reached. The end points of these branches, referred to as "nodes," represent subgroups of the original sample that differ in terms of the probability of the outcome variable.

The splitting procedure is very important in CART. Gini index is one of the most popular split selection methods, and it is often the default option for the impurity measure and goodness-of-fit measure in classification problems. The basic gini index is defined as Eq. 2 (Breiman et al. 1984):

$$\text{impurity}(t) = \sum_{i \neq j} p(i|t)p(j|t) \tag{2}$$

where $p(j|t)$ is the estimated probability that an observation belonging to group j given that it is in node t, $p(j|t) = p(j, t)/p(t)$; and $p(j, t)$ is the estimated probability that an observation being in group j and at node t, $p(j, t) = [\pi(j)\, N_j(t)]/N_j$; $p(t)$ is the estimated probability that an observation is at node t, $p(t) = N(t)/N$; $\pi(j)$ is the prior probability for group j; $N_j(t)$ is the number of group j members at node t, and N_j is the size of group j.

As the prior probabilities play a role in every gini index computation at every node, there will be higher misclassification rates in under-represented groups (the violent group in our case) when the prior probabilities are estimated from the data. Therefore, both *Prior probability* and *Misclassification cost* in this study were specified to be equal for each category of the target variable. This has a similar effect to adjusting the classification

[3] CART is designed to fit binary classification trees, while CHAID and some other CT models perform multi-level splits rather than binary splits when computing classification trees. It should be noted that there is no inherent advantage of multi-level splits, because any multi-level split can be represented as a series of binary splits. Please see http://www.statsoft.com/textbook/classification-trees/.

cut-off probability to balance misclassification cost in favor of the smaller category in LR to reduce false positive prediction of violence. The SPSS 16.0 software package (SPSS Inc. 2008) was used to build CART models.

Multi-Layer Perceptron Neural Network (MLPNN)

The MLPNN is among the most commonly used network architectures (Rumelhart and McClelland 1986). As shown in Fig. 1, the MLPNN has a feed forward structure: inputs layer, hidden layer(s), output layer, and *bias* term, with the weights as the adjustable parameters of the model. Such a structure has stable behavior and fault tolerance, and it can model functions of almost arbitrary complexity, with the number of layers and the number of units in each layer determining the complexity of function. Like most NNs analysis, MLPNN consists of training and testing processes. The prediction is made by learning from training data. The training (or learning) and testing process of a MLPNN is shown in Fig. 1 (Price et al. 2000).

Algorithmically, "training" is carried out using the following sequence of steps (Price et al. 2000; StatSoft 2008).

(1) Present the network with an input-target pair. When the 'signals' are sent from input layer to hidden layer,

$$v_{pj} = \sum_{i=1}^{N} w_{ij} x_{pi} \tag{3}$$

where x_{pi} is the input units, and p is the number of observations, i is the number of predictors; w_{ij} is the weights between the ith input unit and the jth hidden unit, supposing there are j hidden units; v_{pj} is the middle set of 'signals' in this process.

$$x_{pj} = f(v_{pj}) + B_{inpj} \tag{4}$$

where $f(v)$ is the activation function (or transfer function) in the hidden layer, which is often set as logistic function or tanh function for the classification problem; B_{inpj} is the input bias unit between input and hidden layers, corresponding to the intercept term of regression models (Bishop 1995); x_{pj} is the last of the 'signals' in this process.

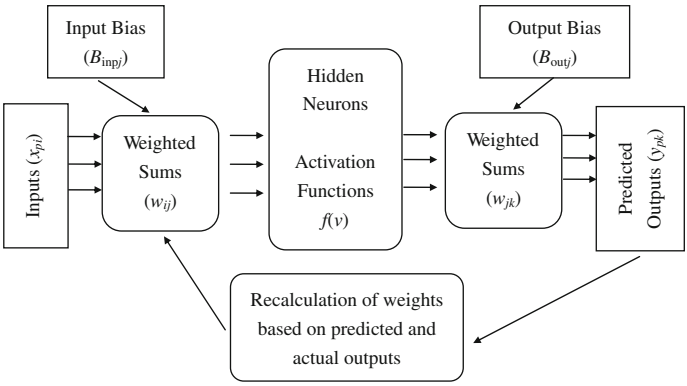

Fig. 1 A flow diagram of the training process in a multi-layer perceptron neural network

(2) Compute the predictions of the network for the targets. When the 'signals' are sent from hidden layer to output layer,

$$v_{pk} = \sum_{j=1}^{N} w_{jk} x_{pj} \tag{5}$$

where w_{jk} is the weights between the jth hidden unit and the kth output unit, supposing there are k output units; and v_{pk} is the middle set of 'signals' in this process.

$$y_{pk} = f(v_{pk}) + B_{outj} \tag{6}$$

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

(3) Use the error function to calculate the difference between the predictions (y_{pk}) of the network and the actual target values. Sum squared error and cross entropy are the two most popular kinds of error functions. The latter is preferred for use in classification problems, and the maximum likelihood estimation method is used in cross entropy to estimate the error rate of classification prediction.

(4) Use the training algorithm to adjust the weights (w_{ij} and w_{jk}) of the networks. Back propagation is the most famous training algorithm in the past. More recently, the improved 'second generation' training algorithms, including gradient descent, conjugate descent, and BFGS (Broyden–Fletcher–Goldfarb-Shanno, or Quasi-Newton), have been developed (Bishop 1995; Shepherd 1997).

(5) Repeat steps (1)–(4) again for a number of training cycles or iterations until the network starts producing sufficiently accurate outputs.

In this study, the networks were trained by the BFGS algorithm together with the cross entropy error function as built into the STATISTICA8.0 software package (StatSoft Inc. 2008). The input layer consisted of 20 predictors and the output layer contained two units for the two categories in the dependent variable. As the hidden units are different in different training samples, the hidden layer was composed of 18–26 hidden units in this study. Through repeated trainings, we finally selected the "best" models using two criteria: (1) the highest level of area under curve (AUC) value from the ROC analysis and overall accuracy from the testing and validation samples to ensure that the selected model was not over-fitted, and (2) the least difference in accuracy between the violent and non-violent output categories, by both training and testing samples for balanced results.

Evaluation Indicators

The predictive accuracy of a risk assessment tool can be expressed in various ways. Traditional evaluation indices (e.g. sensitivity, specificity, accuracy; positive and negative predictive values) are popular and useful (Hart et al. 1993), but have been criticized for instability with varying predictor base rates (Rice and Harris 1995a). The area under the curve (AUC) of the receiver operating characteristic (ROC) is regarded as an effective method of quantifying performance of a risk assessment tool that is relatively immune to changes in base rate (Green and Swets 1966; Henderson 1993; Mossman 1994; Rice and Harris 1995a), and has been used in many previous studies of risk prediction efficacy (Dolan and Khawaja 2004; Grann and Langstrom 2007; Gray et al. 2008). By plotting the hit rate (the rate of true positives) against the false-alarm rate (the rate of false positives)

for all observed predictor values, the ROC curve graphically depicts the tradeoff in specificity that occurs as sensitivity is increased with lower cutoff scores and vice versa. AUC value is the effect size estimate derived from the ROC analysis and ranges from 0.0 (perfect negative prediction) to 1.0 (perfect positive prediction). There is still no consensus as to the proper interpretation of AUC estimates for predictive validity, although different interpretations have been suggested by different researchers (Rice and Harris 1995b; Sjöstedt and Grann 2002). For a full-scale comparison, we chose AUC value and 95% CI of AUC value, sensitivity, specificity, accuracy and 95% CI of accuracy as indicators of predictive validity in the present study.

Multi-Validation

Cross-validation is simply an empirical approach to the problem of attempting to obtain an unbiased estimate of predictive accuracy (Gottfredson and Moriarty 2006). However, cross-validation based on one sample is usually not sufficient, especially if the sample size is small (Cohen 1990), or the ratio between sample size and number of predictor variables is low (Grann and Langstrom 2007). A ratio of at least 5:1 is usually recommended (Cicchetti 1992). In our sample, the ratio is more than 6:1 and the sample size ($N = 1{,}225$) is sufficient. We further divided this sample into four subsets randomly, and combined these subsets into four different validation samples. This multi-validation method is expected to yield a more reliably true result than single-sample validation, as the latter process may coincidently result in high model fit values (Grann and Langstrom 2007).

The four subsets generated randomly in the SPSS 16.0 software package (SPSS Inc. 2008) were: Subset A, $N = 306$, base rate = 27%; Subset B, $N = 306$, base rate = 28%; Subset C, $N = 307$, base rate = 27%; Subset D, $N = 306$, base rate = 29%. These subsets then formed four "different" full samples, which have a different training sample, testing sample, and validating sample (see Tables 4, 5). The proportion of training sample, testing sample, and validating sample in each full sample was 50, 25, and 25%, respectively. As the aim of this study was to compare the predictive performance of different models rather than to develop a risk instrument, the proportions above were assigned to guarantee a sufficient size of sample in each of the training, testing, and validating samples. This is necessary to avoid negative effects potentially caused by small sample size in applying the three models in this study, especially for CART. Finally, it is worth noting that the same training, testing, and validating samples were then used for each of the three types of model, for comparison purposes in each full sample.

Results

Description of the Sample

Participants in the cohort had a mean age of 30.7 years (SD = 11.3 years, range = 17–75 years) at the time of interview. 966 (78.9%) were white British, 186 (15.2%) black Caribbean or black African or black other, 35 (2.9%) Asian, and 38 (3.1%) other ethnic origin. Most (74.0%) participants ($N = 906$) had a diagnosis of personality disorder, with 813 (66.4%) antisocial personality disorder (ASPD), 134 (10.9%) had a lifetime history of schizophrenia/schizophreniform disorder, and 278 (22.7%) delusional disorder.

In the total sample, there were 726 (59.3%) prisoners subsequently convicted of a criminal offence: 343 (47.2%) were convicted for violence, 367 (50.6%) for acquisitive

offences, and 16 (2.2%) for sex offences. Table 1 shows that the rate of any re-offending was 31.8, 51.8, 60.6, and 75.3%, respectively, among prisoners whose follow up time was <2 years, [2, 3) years, [3, 4) years, and ≥4 years. The violent re-offending rate was 9.1, 18.6, 30.4, and 38.7%, respectively, among prisoners whose follow up time was <2 years, [2, 3) years, [3, 4) years, and ≥4 years.

Table 2 shows that the violent re-offender group scored significantly higher on the HCR-20 total than the non-violent re-offender group, and that the Historical subscale demonstrated the highest differential effect amongst the three subscales. Figure 2 further demonstrates the predictive efficacy of the HCR-20 total score: the higher the HCR-20 score, the higher the reconviction rate for violent offences. Table 3 shows that most items (except H3, H6, C3, R3, and R5) had a significant correlation with violent reconviction, the coefficients ranging from 0.09 to 0.24.

Comparison of Predictive Accuracy Among Different Models

To compare the three models, traditional ROC analysis was initially conducted as a point of reference, based on the total scores for the HCR-20. The overall AUC value of the HCR-

Table 2 The descriptive results for the HCR-20

HCR-20 scale	Violent re-offender (N = 343) Mean ± SD	Non violent re-offender (N = 882) Mean ± SD	Cohen effect size (p value)
Historical (10 items)	13.07 ± 3.38	10.37 ± 4.69	0.62 ($p < 0.001$)
Clinical (5 items)	4.08 ± 2.04	3.15 ± 2.11	0.45 ($p < 0.001$)
Risk (5 items)	5.06 ± 2.58	4.36 ± 2.48	0.28 ($p < 0.001$)
Total (20 items)	22.21 ± 6.55	17.88 ± 7.84	0.58 ($p < 0.001$)

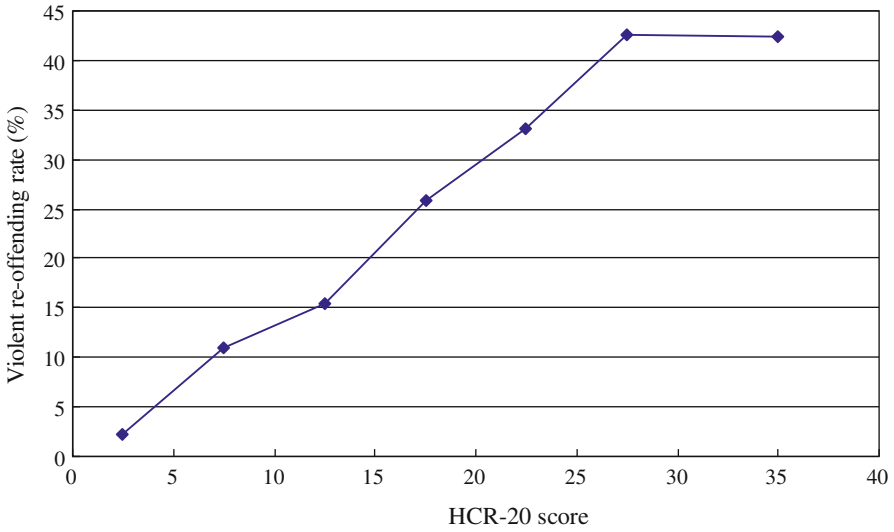

Fig. 2 The relationship between HCR-20 total score and the violent re-offending rate

Table 3 Correlation coefficient of individual HCR-20 predictors with violent reconviction

HCR-20 predictors	Coef.[a] (*p value*)
H1 Previous violence	0.10 ($p < 0.001$)
H2 Young age at first violent incident	0.24 ($p < 0.001$)
H3 Relationship instability	0.05 ($p = 0.108$)
H4 Employment problems	0.16 ($p < 0.001$)
H5 Substance use problems	0.17 ($p < 0.001$)
H6 Major mental illness	0.01 ($p = 0.772$)
H7 Psychopathy (21–29 = 1, 30–40 = 2)	0.16 ($p < 0.001$)
H8 Early maladjustment	0.20 ($p < 0.001$)
H9 Personality disorder	0.20 ($p < 0.001$)
H10 Prior supervision failure	0.21 ($p < 0.001$)
C1 Lack of insight	0.14 ($p < 0.001$)
C2 Negative attitudes	0.17 ($p < 0.001$)
C3 Active symptoms of major mental illness	0.02 ($p = 0.551$)
C4 Impulsivity	0.17 ($p < 0.001$)
C5 Unresponsive to treatment	0.09 ($p = 0.003$)
R1 Plans lack feasibility	0.09 ($p = 0.001$)
R2 Exposure to destabilizers	0.10 ($p < 0.001$)
R3 Lack of personal support	0.05 ($p = 0.105$)
R4 Noncompliance with remediation attempts	0.13 ($p < 0.001$)
R5 Stress	0.05 ($p = 0.063$)

[a] Spearman's rho at two-tailed sig. level

20 was 0.66 (95% *CI*: 0.63–0.70), with consistent results from all four sub-sets of samples. This remained robust using both testing and validation samples (Table 4). Based on individual predictors, the LR model demonstrated AUC values from 0.69 to 0.70 among the four sub-sets of samples, with a slightly better performance from the training samples (range: 0.72–0.75) than the testing (0.63–0.68) and validation (0.64–0.66) samples. The same pattern was found for the CART model, with AUC values ranging as 0.70–0.71 for the training sample, 0.60–0.66 for the testing, and 0.58–0.66 for the validation samples. The MLPNN yielded 0.71–0.78 for the training, 0.65–0.70 for the testing, and 0.64–0.70 for the validation samples. Combining the performance of training, testing, and validation samples for each sub-set of data, the overall performance for the LR model was 0.69–0.70, 0.65–0.67 for the CART model, and 0.70–0.72 for the MLPNN. All models performed with a modest level of accuracy but good robustness. Although the performance of the CART models appeared poorer than the other two, no significant differences were detected given that their 95% confidence intervals overlapped.

Table 5 presents the predictive validity of LR, CART, and MLPNN models using sensitivity, specificity, accuracy, and 95% *CI* of accuracy as the evaluation indicators. With regard to accuracy and 95% *CI* of accuracy, based on individual predictors, the LR model generated an accuracy value from 0.62 to 0.64 among the four sub-sets of samples, with slightly better performance from the training samples (range: 0.65–0.68) than the testing (0.58–0.63) and validation (0.59–0.64) samples. A similar pattern was found for the CART model, with 0.61–0.70 for the training, 0.57–0.65 for the testing, and 0.57–0.65 for the validation sample. For the MLPNN, the accuracy value ranged from 0.62 to 0.66 for the training, 0.59–0.65 for the testing, and 0.60–0.67 for the validation samples.

Table 4 The comparison among four methods by AUC and 95%*CI* of AUC

Sub-sets combination[a]	Sub-samples[b]	ROC	LR	CART	MLPNN
AB/C/D	Train (443:169)	0.65 (0.60, 0.70)	0.72 (0.67, 0.76)	0.71 (0.66, 0.75)	0.71 (0.66, 0.75)
	Test (223:84)	0.65 (0.58, 0.71)	0.68 (0.62, 0.75)	0.65 (0.58, 0.71)	0.70 (0.64, 0.76)
	Validate (216:90)	0.70 (0.64, 0.76)	0.66 (0.60, 0.73)	0.63 (0.56, 0.70)	0.70 (0.64, 0.76)
	Total (882: 343)	0.66 (0.63, 0.70)	0.69 (0.66, 0.73)	0.67 (0.64, 0.70)	0.70 (0.67, 0.73)
BC/D/A	Train (443:170)	0.63 (0.59, 0.68)	0.74 (0.70, 0.78)	0.70 (0.65, 0.74)	0.71 (0.66, 0.75)
	Test (216:90)	0.70 (0.64, 0.76)	0.64 (0.58, 0.71)	0.64 (0.57, 0.70)	0.69 (0.63, 0.76)
	Validate (223:83)	0.68 (0.62, 0.75)	0.64 (0.57, 0.71)	0.66 (0.60, 0.73)	0.67 (0.61, 0.74)
	Total (882: 343)	0.66 (0.63, 0.70)	0.69 (0.66, 0.72)	0.67 (0.64, 0.71)	0.70 (0.67, 0.73)
CD/A/B	Train (439:174)	0.68 (0.63, 0.72)	0.75 (0.71, 0.79)	0.70 (0.66, 0.74)	0.78 (0.74, 0.81)
	Test (223:83)	0.68 (0.62, 0.75)	0.67 (0.60, 0.73)	0.66 (0.59, 0.72)	0.69 (0.63, 0.76)
	Validate (220:86)	0.62 (0.56, 0.69)	0.65 (0.59, 0.72)	0.63 (0.56, 0.69)	0.64 (0.57, 0.70)
	Total (882: 343)	0.66 (0.63, 0.70)	0.70 (0.67, 0.73)	0.67 (0.64, 0.70)	0.72 (0.69, 0.75)
DA/B/C	Train (439:173)	0.69 (0.65, 0.73)	0.75 (0.71, 0.79)	0.70 (0.65, 0.74)	0.74 (0.70, 0.78)
	Test (220:86)	0.62 (0.56, 0.69)	0.63 (0.57, 0.70)	0.60 (0.53, 0.67)	0.65 (0.58, 0.71)
	Validate (223:84)	0.65 (0.58, 0.71)	0.66 (0.59, 0.72)	0.58 (0.51, 0.66)	0.67 (0.60, 0.73)
	Total (882: 343)	0.66 (0.63, 0.70)	0.70 (0.66, 0.73)	0.65 (0.61, 0.68)	0.70 (0.67, 0.73)

[a] The letters on the left side represent the training sample (about 50% of the whole sample), the letter between two slashes stands for the test sample (about 25% of the whole sample), and the letter in the right side represents the validation sample (about 25% of the whole sample)

[b] The number inside the brackets on the left side is the number of offenders without any subsequent violent offence, while the number on the right side inside the brackets is the number of offenders with a subsequent violent offence

Table 5 The comparison among LR, CART, and MLPNN by sensitivity, specificity, accuracy, and 95% *CI* of accuracy

Sub-sets combination[a]	Sub-samples[b]	LR			CART			MLPNN		
		Sen.	Spe.	Acc. (95%*CI*)	Sen.	Spe.	Acc. (95%*CI*)	Sen.	Spe.	Acc. (95%*CI*)
AB/C/D	Train (443:169)	0.67	0.64	0.65 (0.61, 0.69)	0.70	0.61	0.63 (0.59, 0.67)	0.62	0.62	0.62 (0.58, 0.66)
	Test (223:84)	0.61	0.63	0.63 (0.58, 0.68)	0.56	0.57	0.57 (0.51, 0.63)	0.63	0.64	0.64 (0.59, 0.69)
	Validate (216:90)	0.64	0.64	0.64 (0.59, 0.69)	0.58	0.61	0.60 (0.55, 0.65)	0.66	0.68	0.67 (0.62, 0.72)
	Total (882: 343)	0.65	0.64	0.64 (0.61, 0.67)	0.64	0.60	0.61 (0.58, 0.64)	0.63	0.64	0.64 (0.61, 0.67)
BC/D/A	Train (443:170)	0.73	0.62	0.65 (0.61, 0.69)	0.68	0.65	0.66 (0.63, 0.69)	0.64	0.63	0.63 (0.59, 0.67)
	Test (216:90)	0.57	0.60	0.59 (0.53, 0.65)	0.62	0.64	0.63 (0.58, 0.68)	0.66	0.65	0.65 (0.60, 0.70)
	Validate (223:83)	0.66	0.59	0.61 (0.56, 0.66)	0.68	0.64	0.65 (0.60, 0.70)	0.69	0.57	0.60 (0.55, 0.65)
	Total (882: 343)	0.67	0.61	0.62 (0.59, 0.65)	0.67	0.64	0.65 (0.62, 0.68)	0.65	0.62	0.63 (0.60, 0.66)
CD/A/B	Train (439:174)	0.72	0.62	0.65 (0.61, 0.69)	0.81	0.53	0.61 (0.57, 0.65)	0.83	0.57	0.65 (0.61, 0.69)
	Test (223:83)	0.70	0.56	0.60 (0.55, 0.65)	0.76	0.51	0.58 (0.52, 0.64)	0.76	0.56	0.61 (0.56, 0.66)
	Validate (220:86)	0.69	0.56	0.59 (0.53, 0.65)	0.72	0.51	0.57 (0.51, 0.63)	0.73	0.55	0.60 (0.55, 0.65)
	Total (882: 343)	0.71	0.59	0.62 (0.59, 0.65)	0.78	0.52	0.59 (0.56, 0.62)	0.79	0.56	0.63 (0.60, 0.66)
DA/B/C	Train (439:173)	0.73	0.65	0.68 (0.65, 0.71)	0.49	0.78	0.70 (0.67, 0.73)	0.74	0.63	0.66 (0.63, 0.69)
	Test (220:86)	0.58	0.57	0.58 (0.52, 0.64)	0.40	0.75	0.65 (0.60, 0.70)	0.59	0.59	0.59 (0.53, 0.65)
	Validate (223:84)	0.66	0.63	0.64 (0.59, 0.69)	0.33	0.76	0.64 (0.59, 0.69)	0.63	0.62	0.62 (0.57, 0.67)
	Total (882: 343)	0.67	0.63	0.64 (0.61, 0.67)	0.43	0.77	0.67 (0.64, 0.70)	0.68	0.62	0.64 (0.61, 0.67)

[a] The letters on the left side represent the training sample (about 50% of the whole sample), the letter between two slashes stands for the test sample (about 25% of the whole sample), and the letter in the right side represents the validation sample (about 25% of the whole sample)

[b] The number inside the brackets on the left side is the number of offenders without any subsequent violent offence, while the number on the right side inside the brackets is the number of offenders with a subsequent violent offence

Sensitivity was given more emphasis than specificity in deciding a cut-off point, given that both in managing and studying a violent offender population there was a pressure to enhance the level of 'true positives' (re-offenders correctly identified as such). Here the LR model produced a slightly better sensitivity value for the training (0.67–0.73) than testing (0.57–0.70) and validation (0.64–0.69) samples. The same pattern was found for the CART model, with a range of 0.49–0.81, 0.40–0.76, and 0.33–0.72 for the training, testing, and validation samples, respectively. For the MLPNN, the sensitivity value ranged from 0.62 to 0.83 for the training, 0.59–0.76 for the testing, and 0.63–0.73 for the validation samples.

As with the AUC and 95% *CI* of AUC comparisons, both Tables 4 and 5 suggest that, in general, the three models demonstrated similar levels of accuracy in predicting violent reconviction. The MLPNN showed a small improvement, but this was not significant. However, while the performance of CART was not very stable, LR was robust.

Discussion

This study is the first to simultaneously compare the validity and predictive accuracy of LR, CART, and MLPNN models in the prediction of violent reconviction. We found that, in general, the three models demonstrated similar levels of accuracy in predicting violent reconviction. Whereas MLPNN moderately outperformed the other two, this did not reach significance. The efficacy of the HCR-20 for violent reconviction among UK high risk male prisoners ranged from 0.58 to 0.70 (mainly 0.62–0.68) in four different validation samples in terms of AUC values and 0.57–0.67 in terms of overall accuracy. This finding is largely consistent with the previous reports for the HCR-20 in terms of AUC values (0.56–0.82) (Farrington et al. 2008). However, compared with the better results (AUC value: 0.75–0.82) that are reported in some studies (Douglas et al. 1999; Doyle and Dolan 2006; Nicholls et al. 2004; Vogel et al. 2004), the performance of the HCR-20 in our study is lower. This instrument was originally developed for forensic psychiatric patients (Webster et al. 1997). However, our sample was drawn from a high risk population of serious offenders in UK prisons, which is different from the developmental origin of the HCR-20 and may help to explain the poor performance of the HCR-20 for this sample. Furthermore, there were 813 subjects (66.4%) diagnosed as having antisocial personality disorder in our sample. The heterogeneousness of the sample may be another reason for the poor performance. We carried out a further study to check the effect of heterogeneousness on the predictive validity, which will be reported in the future.

This study has additional strengths. Firstly, it used a prospective forensic cohort sample followed up for 3.31 years on average after release, whilst most previous studies involved retrospective designs (Brodzinski et al. 1994; Caulkins et al. 1996; Grann and Langstrom 2007; Palocsay et al. 2000; Rosenfeld and Lewis 2005; Silver and Chow-Martin 2002; Silver et al. 2000). Secondly, the sample, selected from male prisoners jailed for violent, sexual, and robbery convictions, was relatively large for testing the constructive and predictive validity of the three models. Thirdly, the outcome was restricted to violent convictions, and sexual offences were not included. It was not necessary to increase the base rate of offending by combining additional categories as in previous studies (Brodzinski et al. 1994; Caulkins et al. 1996; Palocsay et al. 2000; Silver and Chow-Martin 2002; Silver et al. 2000). Fourthly, the validation method used to test the reliability of different prediction models is comprehensive. In contrast to many previous studies only using construction and validation samples (Brodzinski et al. 1994; Caulkins et al. 1996; Rosenfeld and Lewis 2005; Thomas et al. 2005), this study employed a multi-validation method.

Is NN a Better Model for Risk Prediction than LR or CART?

Since the first article on neural network modelling (McCulloch and Pitts 1943), NNs have attracted increasing attention from different application areas. One important difference between NNs and traditional methods is that in developing an application, NNs are not programmed but are trained to solve a particular type of problem. This "learning" ability to solve a problem makes NNs well able to tackle a wide variety of problems, some of which have proved intractable using traditional computing approaches (Florio et al. 1994). Another main advantage of NNs is their capability to model extremely complex functions and data relationships using a parallel computing procedure. NNs therefore have a remarkable ability to derive and extract meaning, rules, and trends from complicated, noisy, and imprecise data (StatSoft 2008). Furthermore, the amenability of this technique to any ongoing reclassification of risk could have important implications in the area of risk prediction and assessment. In another study (Yang et al. 2010), NNs was found to some extent to be more sensitive than the other two types of model, after adding the dynamic variables. This implied that NNs might be more responsive to changing behavior and responsiveness. *NN models can include dynamic information on offenders, for example as they move through a probation or correctional setting.* Further research is required to explore this feature of NNs in assessing risk by incorporating dynamic information in an on-going process system.

However, over-fitting and generalization of the NNs are two critical issues. NNs can suffer from either under-fitting or over-fitting. A network that is too complex may fit the noise, not just the signal, leading to over-fitting. Over-fitting can then lead to predictions that are far beyond the range of the training data for many of the common types of NNs (Comp.ai.neural-nets FAQ), hence resulting in poor generalization. In practice, over-fitting can be expected in a model where one may observe high accuracy in the training data but a much lower accuracy in the testing data.

To avoid over-fitting, we set aside a test sample that was not used in training but to test the model after training, and chose a model with better, or at least similar accuracy, in the test sample compared to the training sample. For model generalization, we additionally used a third independent set of data as a validating sample to further confirm the performance of the built model in generalization (Bishop 1995). If the performance of the network was found to be consistently good on both the test and validation samples, then it was reasonable to assume that the network would generalize well on unseen data (StatSoft 2008). We therefore took advantage of our sample size and dealt with possible over-fitting of the NNs and information on generalization using the recommended procedures for reliable and robust models.

The issue of model transparency is another problem for NNs. Transparency of the model implies that it is possible to go back and reconstruct what went wrong in cases where the model failed. It is important for models for which usefulness in the context of actual clinical and legal decision-making may be required (Grann and Langstrom 2007). It is well known that NNs have a "black-box" nature (Guerriere and Detsky 1991; Ning et al. 2006). Bigi et al. (2005) indicate that NNs do not allow the easy determination of which variables contribute mainly to a particular output. This may therefore be interesting when searching not for a single but a complex set of predictors. Grann and Langstrom (2007) also pointed out that it was possible to inspect the contributing factors for each of the input variables and thereby got a picture of the relative importance of each factor in NNs. However, the contributing factors can only be used to compare input variables within the same neural network and do not generalize even across the seeds of a study (Tam and Kiang 1992). The MLPNN model built in this study has the same problem. Compared with LR and CART,

the transparency of MLPNN is much lower. Therefore, the interpretability of the results from MLPNN may not in practice be desirable for clinicians or offender managers.

Findings in the present study generally support the argument of Price et al. (2000) that NNs are only as good as the data analyzed. It is not surprising to see that the three models demonstrated similar levels of accuracy in predicting violent reconviction, and that while the MLPNN moderately outperformed LR and CART, this did not reach significance. More studies on the performance of NNs models for different outcomes with different predictors among different populations would be useful to provide more information to help offender managers and clinicians to decide when and how to apply NNs to specific data.

In addition, we found that the performance of the CART was no better than the other two models in this study, and there was some instability in the validation set DA/B/C according to the AUC value. Rosenfeld and Lewis (2005) and Thomas et al. (2005) reported reduced performance (or over-fitting) of the CART, which was also reported in our early study when the model was tested on small samples by less vigorous validation procedures (Yang et al. 2010). The tendency to over-fitting was reported as a major problem of CT models (Thomas et al. 2005). These models may fit the training sample well because they can capture complex combinations of factors associated with violent reconviction in the particular sample, but do not perform so well on new data sets. This problem can be detected by validation measures. When the sample size is large enough, the multi-validation approach used in this study can help to select a more robust CT model according to the validated performance. If the sample is not very large, other validation measures (e.g. cross-validation, bootstrapping etc.) are necessary to account for large variation. To minimize over-fitting problems, pruning and stopping rule are two main approaches suitable for tree models (Loh and Shih 1997). Through careful pruning, CT model can be simplified by chopping off the smaller branches to get a smaller and more robust tree, while stopping rule (or growth limit) can decide the complexity of CT models by setting its parameters, which include maximum tree depth and minimum numbers of cases in parent and child nodes. For a more robust performance, both measures were applied in the present study. Therefore, the less marked instability of the CART model in this study could be due to the fact that we used the multi-validation approach and pruning procedures to minimize the problem of over-fitting. The large-enough sample sizes for each validation and test sample could be another reason. As the application of this methodology is relatively new in violence risk assessment, there is still a requirement for more validation studies of different populations using different types of CT models (e.g. CHAID, C5.0/See5 et al.) with different predictors.

It is worth noting that our study found that LR, the classic traditional regression model, is robust in terms of any evaluation indicator (AUC values or sensitivity or specificity or accuracy). This finding is consistent with that of previous studies in forensic settings (Gardner et al. 1996; Hartvig et al. 2006; Rosenfeld and Lewis 2005; Stalens et al. 2004; Thomas et al. 2005) and in the broader medical field (Harper 2005; Tu 1996). There should be a benefit from the robustness properties of the LR model estimated by the maximum likelihood estimation procedure (Maria-Pia Victoria-Feser 2000). However, conventional regression models (including LR) are mostly used to determine the average effect of independent variables on the dependent variable (Lemon et al. 2003), which tends to lean in the direction of the average members of the population, without consideration of the special needs of subgroups (Forthofer and Bryant 2000).

To conclude, all three methods compared in this study have their own merits and disadvantages. LR is robust, but it is criticized for ignoring the possibility that different variables might predict violence for different subgroups of individuals (Steadman et al. 2000) and its

explanations are not ideal for clinicians and offender managers. The CT is a better representation of how these professionals typically make their risk judgments and it can be used to identify different risk groups easily through its modeling principles. However, its performance may not be very stable and it is prone to over fitting if no preventive measures are taken. Therefore, the efficacy of CT models still needs to be validated (Colombet et al. 2000; Dillard et al. 2007; Thomas et al. 2005; Trujillano et al. 2008). NNs may be particularly useful when the primary goal is outcome prediction and important interactions or complex non-linearities exist in a data set (Tu 1996). When new data are collected, it is very convenient to complete the outcome prediction based on the built NNs model. However, as with any other data mining technique, care should be taken to avoid over-fitting. In addition, the "black-box" nature of NNs reduces the interpretability of the results, which is not attractive to clinicians and offender managers (Grann and Langstrom 2007; Guerriere and Detsky 1991; Ning et al. 2006). It is unlikely that any one model will be the technique of choice in all circumstances. Researchers should therefore choose the most suitable model according to their goals and the characteristics of their data sets.

Although the performance of risk assessment could be improved in some degree by alternative statistical models, it is not sufficient to rely simply on the development of models for the future improvement of recidivism prediction. We hold similar opinions to Caulkins et al. (1996) and Grann and Langstrom (2007) that *theory building to delineate behavioral mechanisms and contextual influences involved in criminal recidivism should be prioritized over development of complex statistical prediction models*. In other words, high accuracy of risk prediction is determined by four key factors described above: (1) high specificity of predictors for good predictive power, (2) well-defined criteria of outcome for distinguishable categories, (3) high specificity of target population for homogeneity, and (4) adequate statistical methods to achieve the best discriminate effects based on the data. Because all four factors influence predictive performance synthetically, deficiencies in any one would lead to a poor outcome. In the current study, we focused mainly on the classification models and did not explore the other factors (e.g. predictors, population, and outcome). The HCR-20 items used in this study may not be the best predictors to enhance model performance. Correlation among these items could add noise in the model that affects predictive accuracy, in particular when strong correlation across items becomes a collinearity issue in classical models. However, the test of multicollinearity of the items was not considered in this study, as the HCR-20 is internationally a well validated psychometric instrument, and correlation between items within certain constructs or domains should be expected in line with theory to ensure consistency across items. In addition, as we used the same predictors for all models, the results should be comparable no matter multicollinearity exists or not. The HCR-20 items were chosen for the convenience of model comparison and, while they have been used elsewhere for this kind of purpose, they have not previously been employed on violent male prisoner samples in England and Wales. Future studies should consider investigating whether the predictive performance could be improved by further analysis of predictor issues (e.g. multicollinearity) and the other factors. With regard to the population factor, we have found that model performance and predictive power of predictors were significantly different between prisoners with and without an ASPD diagnosis in the same cohort sample.

Base Rates

It was necessary to consider the potential problem of base rate in this study. The base rate for any given event is defined as the relative frequency of occurrence of that event in the

population of interest and is typically expressed as a proportion or percentage (Gottfredson and Moriarty 2006). The difficulty in predicting events of interest increases as the base rate differs from 0.50 (Meehl and Rosen 1955). The more frequent or less frequent an event, the greater the likelihood of inaccurate prediction. In most studies of offenders, the frequency of violent reconviction is small (about 0.10–0.30). When developing predictive models for use in criminal justice applications (and certain other fields), it is important to consider base rates in the development process and to avoid making predictions or classifications based on criteria that produce larger errors than would occur with the simple use of the base rate (Gottfredson and Moriarty 2006; Reiss 1951). In most softwares that fit models for classification, the usual default weighting is the same for each category of outcome. Such weighting procedure produces a more balanced classification if the base rate is 50%, as it will otherwise inevitably result in a model that favors the category with the largest proportion, and underperforms with respect to the underrepresented category (Ripley 1996). As mentioned earlier, our sample had a base rate of 0.28 for violent offending and 0.72 for a combined category of other. Default weighting will produce results in favour of the non-violent category. To avoid such a consequence, in fitting the LR model, the cut-off probability was changed from the default 0.5 to the base rate of the violent category. Using the same principle, when fitting the CART model, prior probabilities and misclassification costs were specified to be equal for each category. For classification using MLPNN models, case weights were used and interpreted as measures of category importance, which is useful for imbalanced data. We considered that the cost for misclassification of a violent re-offender to a non-violent category could be larger than the other way round, so we used the inverse base rate that adjusted MLPNN case weights to increase the penalty for error in misclassifying violent outcome as non-violent, and also to achieve a propensity of classification comparable to the LR and CART models. The relationship between priors, misclassification costs, and case weights becomes quite complex in all but the simplest situations (for discussions, see Breiman et al. 1984; Ripley 1996).

Limitations

This study has not investigated issues relating to the definitions or criteria for outcome variables nor the quantity and specificity of predictors. A recent study using the same data as in this study showed that, based on a narrow definition of the outcome categories as violent re-offenders versus non re-offenders, and leaving out other non-violent re-offenders completely, several risk assessment instruments including the HCR-20, PCL-R (Psychopathy Checklist-Revised), VRAG (Violence Risk Appraisal Guide), and OGRS (Offender Group Reconviction Score) demonstrated an improved predictive accuracy when compared to the more broadly defined outcome categories as used in the present study (Coid et al. 2010). The selection of variables may not necessarily demonstrate an improved predictive accuracy using traditional regression models (Kroner et al. 2005). A larger number of variables, with more complex information on participants, may lead to better predictive accuracy in other data mining procedures, and NNs in particular, given the nature of these models which cope with non-linear and interactive relationships between predictors (Tu 1996; Yang et al. 2010). Further research in these areas is required.

Another important problem, as mentioned earlier, concerns the heterogeneity of the target population. Some indicator variables (e.g. whether the offender was diagnosed as ASPD) may divide the population into different subgroups (ASPD vs. non-ASPD) which have better homogeneity when analyzed separately, therefore also better predictive

performance. Some results of further work in testing this hypothesis already point in this direction, although the issue is sufficiently complex to merit a separate article.

Recommendations

Based on the results of this study and some related research, we put forward the following recommendations.

(1) Improving violence prediction should focus on a number of methodological areas: homogeneity of target population, specificity of predictors, definition of outcome criterion, and choice of statistical models. Within each area there are different issues of concern. This study only examined some key issues in three types of statistical models currently used in risk assessment practice. Further research in examining issues and concerns in other areas is required.

(2) All statistical models are data driven. Sample size and base rate are major issues in ensuring a goodness fit of a model. They are interlocked, and should be dealt with in an integrated manner. When there are many predictors, small sample size combined with low base rate often leads to model over-fitting, hence low validity or false predictive performance of the fitted model. Therefore efforts should be put into solving the problem of over-fitting when applying any model.

(3) Over-fitting can be detected by using an appropriate validation approach regardless of prediction models. When sample size is large enough, multiple validation samples in addition to training and testing samples are often necessary. Simulation based or Bayesian validation approach can be used for smaller sample size. There are other procedures to control for this problem, for example, appropriate pruning and stopping rules for classification tree models, and weight decay for NNs.

(4) Low base rate, also recognized as 'the class imbalance problem' in some other research areas, should always be addressed in validating a predictive model, and can be solved in several ways. One way is to adjust for the cut-off probability; other ways are to specify prior probabilities and misclassification costs, or to adjust for case weights. The ultimate aim of those procedures is to minimize prediction error and achieve a balanced classification.

(5) In reporting predictive performance of a model, various accuracy measures should be considered, e.g. AUC value, accuracy, sensitivity, and specificity. While AUC and accuracy present average performance of the prediction differently, sensitivity and specificity show how balanced the classification is between the two groups in the target population. In addition, 95% CI of AUC value (or accuracy) can report predictive performance more objectively. It is especially useful when comparison is needed.

(6) Although this study showed similarly moderate accuracy of three types of model in prediction of violent reconviction, each model has its advantages and shortcomings. While the traditional LR is robust to over-fitting and CART can reflect actual clinical thinking processes, the NNs seem more responsive to large number of predictors when sample size is large enough. It is impossible to find one model that fits all situations well. Researchers should select appropriate model(s) based on data characteristics and the aims of the study.

(7) To advance the methodological field in the future, one potential area for research is to maximize the advantage of different models in a multiple stage assessment process. For example, a two-stage assessment could be considered, in which a LR model with

a specific instrument for general re-offending could be used in the 1st stage, and a conditional NNs model with specific static as well as dynamic predictors for violent re-offending could be used in the 2nd stage. Some preliminary findings of a two-stage assessment exercise can be found in Yang et al. (2010). While it was only a trial, ways of improving and practicing this strategy might potentially be helpful in enhancing this kind of methodology.

Acknowledgments The project was funded by Ministry of Justice (England and Wales) and a grant from China Scholarship Council. Professor Min Yang and Professor Jeremy Coid were funded from the National Institute of Health Research Programme Grant (RP-PG-0407-10500). Malcolm Ramsay works for the Ministry of Justice. His contribution here is made in a personal capacity.

Conflict of interest None.

References

SPSS Inc. (2008) SPSS for windows, Rel. 16.0.1. 2008. SPSS Inc., Chicago. http://www.spss.com/

StatSoft Inc. (2008) STATISTICA (data analysis software system), version 8.0. http://www.statsoft.com

Banks S, Robbins PC, Silver E, Vesselinov R, Steadman HJ, Monahan J (2004) A multiple-models approach to violence risk assessment among people with mental disorder. Crim Justice Behav 31:324–340

Bigi R, Gregori D, Cortigiani L, Desideri A, Chiarotto FA, Toffolo GM (2005) Artificial neural networks and robust Bayesian classifiers for risk stratification following uncomplicated myocardial infarction. Int J Cardiol 101:481–487

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

Breiman L (2001) Decision tree forests. Mach Learn 45:5–32

Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth and Brooks/Cole, Monterey, CA

Brodzinski JD, Crable EA, Scherer RF (1994) Using artificial intelligence to model juvenile recidivism patterns. Comput Hum Serv 10:1–18

Caulkins J, Cohen J, Gorr W, Wei J (1996) Predicting criminal recidivism: a comparison of neural network models with statistical methods. J Crim Just 24:227–240

Cicchetti DV (1992) Neural network and diagnosis in the clinical laboratory: state of the art. Clin Chem 38:9–10

Cohen J (1990) Things I have learned (so far). Am Psychol 45:1304–1312

Coid JW, Yang M, Ullrich S, Zhang TQ, Sizmur S, Farrington DP (2010) Improving accuracy of risk prediction for violence: does changing the outcome matter? Int J Offender Ther (Submitted)

Coid JW, Yang M, Ullrich S, Zhang TQ, Sizmur S, Roberts C, Farrington DP (2011) Most items in structured risk assessment instruments do not predict violence. J Forensic Psychiatr Psychol 22. doi: 10.1080/14789949.2010.495990

Colombet I, Ruelland A, Chatellier G, Gueyffier F, Degoulet P, Jaulent MC (2000) Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. In: Proceedings/AMIA annual symposium. AMIA symposium 2000, pp 156–160

Comp.ai.neural-nets FAQ, Part 3: Generalization. http://www.faqs.org/faqs/ai-faq/neural-nets/part3/preamble.html

Cooke DJ, Michie C, Ryan J (2001) Evaluating risk for violence: a preliminary study of the HCR-20, PCL-R and VRAG in a Scottish prison sample. Scottish Prison Service occasional papers

Dahle KP (2006) Strengths and limitations of actuarial prediction of criminal reoffence in a German prison sample: a comparative study of LSI-R, HCR-20 and PCL-R. Int J Law Psychiat 29:431–442

Dawes RM, Faust D, Meehl PE (1989) Clinical versus actuarial judgement. Science 243:1668–1674

Derogatis LR, Melisaratos N (1983) The brief symptom inventory: an introductory report. Psychol Med 13:595–605

Dillard E, Luchette FA, Sears BW, Norton J, Schermer CR, Reed RL (2007) Clinician vs mathematical statistical models: which is better at predicting an abnormal chest radiograph finding in injured patients? Am J Emerg Med 25:823–830

Dolan M, Khawaja A (2004) The HCR-20 and post-discharge outcome in male patients discharged from medium security in the UK. Aggress Behav 30:469–483

Douglas KS, Ogloff JRP, Nicholls TL, Grant I (1999) Assessing risk for violence among psychiatric patients: the HCR-20 violence risk assessment scheme and the psychopathy checklist: screening version. J Consult Clin Psych 67:917–930
Doyle M, Dolan M (2006) Predicting community violence from patients discharged from mental health services. Brit J Psychiat 189:520–526
Farrington DP, Jolliffe D, Johnstone L (2008) Assessing violence risk: a framework for practice. Institute of Criminology, Cambridge University, Cambridge
Florio T, Einfeld S, Levy F (1994) Neural networks and psychiatry: candidate applications in clinical decision making. Aust NZ J Psychiat 28:651–666
Forthofer MS, Bryant CA (2000) Using audience-segmentation techniques to tailor health behavior change strategies. Am J Health Behav 24:36–43
Friedman JH (1999a) Greedy function approximation: a gradient boosting machine. IMS 1999 Reitz Lecture
Friedman JH (1999b) Stochastic gradient boosting. Stanford University, Stanford
Gardner W, Lidz CW, Mulvey EP, Shaw EC (1996) A comparison of actuarial methods for identifying repetitively violent patients with mental illnesses. Law Human Behav 20:35–48
Gendreau P, Goggin C, Smith P (2002) Is the PCL-R really the "unparalleled" measure of offender risk? A lesson in knowledge cumulation. Crim Justice Behav 29:397–426
Gigerenzer G, Todd PM, Group AR (1999) Simple heuristics that makes us smart. Oxford University Press, New York
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. Crim Justice Behav 29:235–249
Gottfredson SD, Moriarty LJ (2006) Statistical risk assessment: old problems and new applications. Crime Delinquency 52:178–200
Grann M, Langstrom N (2007) Actuarial Assessment of Violence Risk: To Weigh or Not to Weigh? Crim Justice Behav 34:22–36
Gray NS, Hill C, McGleish A, Timmons D, MacCulloch MJ, Snowden RJ (2003) Prediction of violence and self-harm in mentally disordered offenders: a prospective study of the efficacy of HCR-20, PCL-R and psychiatric symptomology. J Consult Clin Psych 71:443–451
Gray NS, Taylor J, Snowden RJ (2008) Predicting violent reconvictions using the HCR-20. Brit J Psychiat 192:384–387
Green DM, Swets JA (1966) Signal detection theory and psychophysics. Wiley, New York
Greene MA, Hoffman PB, Beck JL (1994) The mean cost rating (MCR) is Somers' D: a methodological note. J Crim Justice 22:63–69
Grevatt M, Thomas-Peter B, Hughes G (2004) Violence, mental disorder and risk assessment: can structured clinical assessments predict the short-term risk of inpatient violence? J Forensic Psychiat Psychol 15:278–292
Grove WM, Meehl PE (1996) Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. Psychol Public Pol L 2:293–323
Guerriere MR, Detsky AS (1991) Neural networks: what are they? Ann Intern Med 115:906–907
Hanson RK (2005) Twenty years of progress in violence risk assessment. J Interpers Violence 20:212–217
Harper PR (2005) A review and comparison of classification algorithms for medical decision making. Health Policy 71:315–331
Hart SD, Webster CD, Menzies RJ (1993) A note on portraying the accuracy of violence predictions. Law Human Behav 17:695–700
Hartvig P, Alfarnes S, Ostberg B, Skjønberg M, Moger TA (2006) Brief checklists for assessing violence risk among patients discharged from acute psychiatric facilities: a preliminary study. Nord J Psychiat 60:243–248
Hemphill JF, Hare RD, Wong S (1998) Psychopathy and recidivism: a review. Legal Criminol Psychol 3:139–170
Henderson AR (1993) Assessing test accuracy and its clinical consequences: a primer for receiver operating characteristic curve analysis. Ann Clin Biochem 30:521–539
Hosmer DW, Lemeshow S (1989) Applied logistic regression. Wiley, New York
Howard P, Kershaw C (2000) Using criminal career data in evaluation. British criminology conference: selected proceedings, 3. Available online at www.lboro.ac.uk/departments/ss/bsc/bccsp/vol03/howard.html
Kass GV (1980) An exploratory technique for investigating large quantities of categorical data. Appl Stat 29:119–127
Kroner DG, Mills JF (2001) The accuracy of five risk appraisal instruments in predicting institutional misconduct and new convictions. Crim Justice Behav 28:471–489

Kroner DG, Mills JF, Reddon JR (2005) A coffee can, factor analysis, and prediction of antisocial behavior: the structure of criminal risk. Int J Law Psychiat 28:360–374
Lemon SC, Roy J, Clark MA, Friedmann PD, Rakowski W (2003) Classification and regression tree analysis in public health: methodological review and comparison with logistic regression. Ann Behav Med 26:172–181
Lidz CW, Mulvey EP, Gardner W (1993) The accuracy of predictions of violence to others. J Am Med Assn 269:1007–1011
Lin CH, Chou LS, Lin CH, Hsu CY, Chen YS, Lane HY (2007) Early prediction of clinical response in schizophrenia patients receiving the atypical antipsychotic zotepine. J Clin Psychiat 68:1522–1527
Loh WY, Shih YS (1997) Split selection methods for classification trees. Stat Sinica 7:815–840
Manly BFJ (2005) Multivariate statistical methods: a primer, 3rd edn. Chapman and Hall/CRC, Boca Raton
Maria-Pia Victoria-Feser (2000) Robust logistic regression for binomial responses. University of Geneva, Geneva
McCulloch WS, Pitts W (1943) A logical calculus of ideas immanent in nervous activity. Bull Math Biophys 5:115–133
Meehl PE, Rosen A (1955) Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. Psychol Bull 52:194–216
Monahan J, Steadman HJ, Appelbaum PS, Robbins PC, Mulvey EP, Silver E (2000) Developing a clinically useful actuarial tool for assessing violence risk. Brit J Psychiat 176:312–320
Monahan J, Steadman HJ, Robbins PC, Appelbaum P, Banks S, Grisso T (2005) An actuarial model of violence risk assessment for persons with mental disorders. Psychiatr Serv 56:810–815
Monahan J, Steadman HJ, Appelbaum PS, Grisso T, Mulvey EP, Roth LH (2006) The classification of violence risk. Behav Sci Law 24:721–730
Mossman D (1994) Assessing predictions of violence: being accurate about accuracy. J Consult Clin Psych 62:783–792
National Institute of Justice (1992) Data resources of the National Institute of Justice, 5th edn. National Institute of Justice, Washington, DC
Nicholls TL, Ogloff JRP, Douglas KS (2004) Assessing risk for violence among male and female civil psychiatric patients: the HCR-20, PCL: SV, and VSC. Behav Sci Law 22:127–158
Ning GM, Su J, Li YQ, Wang XY, Li CH, Yan WM (2006) Artificial neural network based model for cardiovascular risk stratification in hypertension. Med Biol Eng Comput 44:202–208
Palocsay SW, Wang P, Brookshire RG (2000) Predicting criminal recidivism using neural networks. Socio Econ Plan Sci 34:271–284
Patterson D (1996) Artificial neural networks. Prentice Hall, Singapore
Price RK, Spitznagel EL, Downey TJ, Meyer DJ, Risk NK, El-Ghazzawy OG (2000) Applying artificial neural network models to clinical decision making. Psychol Assess 12:40–51
Quinlan JR (1993) C4.5: programs for machine learning. Morgan Kaufmann Publishers, San Mateo
Quinlan JR (1996) Improved use of continuous attributes in C4.5. J Artif Intell Res 4:77–90
Reiss AJ (1951) The accuracy, efficiency, and validity of a prediction instrument. Am J Sociol 61:552–561
Rice ME, Harris GT (1995a) Violent recidivism: assessing predictive validity. J Consult Clin Psych 63:737–748
Rice ME, Harris GT (1995b) Comparing effect sizes in follow-up studies: ROC area, Cohen's *d* and *r*. Law Human Behav 29:615–620
Ripley BD (1996) Pattern recognition and neural networks. Cambridge University Press, Cambridge
Rosenfeld B, Harmon R (2002) Factors associated with violence in stalking and obsessional harassment cases. Crim Justice Behav 29:671–691
Rosenfeld B, Lewis C (2005) Assessing violence risk in stalking cases: a regression tree approach. Law Human Behav 29:343–357
Rumelhart DE, McClelland J (1986) Parallel distributed processing, vol 1. MIT Press, Cambridge, MA
Shepherd AJ (1997) Second-order methods for neural networks. Springer, New York
Silver E, Chow-Martin L (2002) A multiple-models approach to assessing recidivism risk: implications for judicial decision making. Crim Justice Behav 29:538–568
Silver E, Smith WR, Banks S (2000) Constructing actuarial devices for predicting recidivism: a comparison of methods. Crim Justice Behav 27:733–764
Sjöstedt G, Grann M (2002) Risk assessment: what is being predicted by actuarial "prediction instruments"? Int J Forensic Ment Health 1:179–183
Smith WR (1996) The effects of base rate and cutoff point choice on commonly used measures of association and accuracy in recidivism research. J Quant Criminol 12:83–111
Smith WR, Smith DR (1998) The consequences of error: recidivism prediction and civil-libertarian ratios. J Crim Just 26:481–502

Stalens LJ, Yarnold PR, Seng M, Olson DE, Repp M (2004) Identifying three types of violent offenders and predictinv violent recidivism while on probation: A classification tree analysis. Law Human Behav 28:253–271

Starzomska M (2003) Use of artificial neural networks in clinical psychology and psychiatry. Psychiat Polska 37:349–357

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

Steadman HJ, Monahan J (1994) Toward a rejuvenation of risk assessment research. In: Monahan J, Steadman HJ (eds) Violence and mental disorder. University of Chicago Press, Chicago, pp 10–16

Steadman HJ, Mulvey E, Monahan J, Robbins P, Appelbaum P, Grisso T (1998) Violence by people discharged from acute psychiatric inpatient facilities and by others in the same neighborhoods. Arch Gen Psychiat 55:393–401

Steadman HJ, Silver E, Monahan J, Appelbaum PS, Robbins PC, Mulvey EP (2000) A classification tree approach to the development of actuarial violence risk assessment tools. Law Human Behav 24:83–100

Tam KY, Kiang MY (1992) Managerial applications of neural networks: the case of bank failure predictions. Manage Sci 20:879–888

Thomas S, Leese M (2003) A green-fingered approach can improve the clinical utility of violence risk assessment tools. Crim Behav Ment Health 13:153–158

Thomas S, Leesea M, Walsh E, McCrone P, Moran P, Burns T (2005) A comparison of statistical models in predicting violence in psychotic illness. Compr Psychiat 46:296–303

Trujillano J, Sarria-Santamera A, Esquerda A, Badia M, Palma M, March J (2008) Approach to the methodology of classification and regression trees. Gaceta Sanitaria/SESPAS 22:65–72

Tu JV (1996) Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J Clin Epidemiol 9:1225–1231

UK700 Group (1999) Predictors of quality of life in people with severe mental illness. Study methodology with baseline analysis in the UK700 trial. Brit J Psychiat 175:426–432

Vogel V, Ruiter C, Hildebrand M, Bos B, Ven P (2004) Type of discharge and risk of recidivism measured by the HCR-20: a retrospective study in a Dutch sample of treated forensic psychiatric patients. Int J Forensic Ment Health 3:149–165

Webster CD, Douglas KS, Eaves D, Hart S (1997) HCR-20: assessing risk for violence (version 2). Simon Fraser University, Vancouver, Canada

Webster CD, Muller-Isberner R, Frannson G (2002) Violence risk assessment: using structured clinical guidelines professionally. Int J Ment Health 2:185–193

Yang M, Liu YY, Coid JW (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

Yarnold PR (1996) Discriminating geriatric and nongeriatric patients using functional status information: an example of classification tree analysis via UniODA. Educ Psychol Meas 56:656–667