Ministry of
JUSTICE

# Applying Neural Networks and other statistical models to the classification of serious offenders and the prediction of recidivism

**Min Yang, Yuanyuan Liu and Jeremy Coid**

Ministry of Justice Research Series 6/10
March 2010

# Applying Neural Networks and other statistical models to the classification of serious offenders and the prediction of recidivism

**Min Yang, Yuanyuan Liu and Jeremy Coid**

**This information is also available on the Ministry of Justice website:**

**www.justice.gov.uk/publications/research.htm**

*The Research and Development Team in the Partnerships and Health Strategy Unit exists to improve policy making, decision-taking and practice in support of the Ministry of Justice's purpose and aims to provide the public and Parliament with information necessary for informed debate and to publish information for future use.*

## Disclaimer

The views expressed are those of the authors and are not necessarily shared by the Ministry of Justice (nor do they represent Government policy).

First Published 2010
ISBN: 978-1-84099-383-7

# Contents

# List of tables

# Summary and implications in brief

Risk assessment for violent behaviour/recidivism has been widely implemented in practice and management within the criminal justice and prison systems, as in forensic psychiatric hospitals. Risk assessment has shifted away from clinical judgement by individual clinicians to the use of actuarial instruments, following the rapid development of numerous risk assessment tools in the last 20 years. Traditional logistic regression is the dominant model used to develop these actuarial instruments, typically on a set of chosen predictors combined with a predefined violent outcome among male prisoners. The growing research literature, which includes systematic reviews, has consistently reported a 'glass ceiling effect' for most major tools of around 70% accuracy, in prediction of future violence. The performance of any actuarial risk assessment is known to be determined by several elements: specificity of predictors; well-defined criteria for outcome category; homogeneity of targeted population; and use of an adequate statistical model. Previous research to improve efficacy of risk assessment instruments gave much attention to predictors and criteria for outcomes. Studies of CT models for predicting violent risk reported inconsistent performance, and this was also the case with the literature on applying NNs in this field, which has started to appear in recent years. The present study is among the first to compare predictive efficacy and performance of the CT, NNs and LR models in linking with items in four major risk assessment tools (HCR-20, PCL-R, VRAG and RM2000V), and to test their validity in different target populations (females, young adults and prisoners with any personality disorder) and in terms of different criteria for defining violent recidivism.

## The main findings of the project are as follows

- The overall predictive accuracy of all three types of model was comparable, regardless of the predictor set in any instrument. However, the LR was the most robust type of model, and performed consistently for both training and test samples; the simple CT and NN models were often shown to perform well in some training samples and poorly in small test samples.
- Both LR and NNs performed significantly better for female prisoners than for males, using PCL-R predictors.
- All models demonstrated considerably improved predictive accuracy in using HCR-20 predictors to predict violence in male prisoners with no PDs, compared to those with PDs (prisoners are particularly prone to Anti-Social Personality Disorder (ASPD).
- Using institutional and community variables together with the HCR-20 items or with the RM2000V items as predictors, only the NNs model demonstrated a moderately improved predictive accuracy for both instruments.
- In predicting a narrowed outcome category of violent recidivism against non-recidivism only (i.e. excluding other non-violent recidivism), all models demonstrated clearly and significantly improved accuracy.

The main findings lead to some important implications for researchers as well as decision makers in this field.

1. Much research is required for identifying specific predictors to develop risk assessment instruments for violent recidivism by female prisoners. The better predictive efficacy of the PCL-R for female prisoners requires further exploration.

2. Much research is also needed to identify specific risk predictors for violence prediction among high risk prisoners, such as those with ASPD or DSPD (Dangerous and Severe Personality Disorder), and perhaps to develop assessment tools specifically for them.

3. There is a need to carry out further typological study of the reoffending behavior of prisoners in order to identify more specific definitions of outcome, for example a potential new definition specific to violence, that could effectively maximise differences between categories (violent or not) within that outcome.

4. Statistical models alone do not provide a silver bullet for improving the predictive accuracy of risk assessment tools. They are all data-driven and can only deliver as much as the data can provide, but with different strengths and limitations. One may still opt for the LR model for its robustness and conservative character, as well as being easy to use in practice. Optimally, one may consider the use of the NNs for a substantial sample size combined with a large number of predictors with small effects. One may also consider the use of combined models at different stages of the risk assessment process, for different purposes. Further research is required in this direction.

Based on the knowledge and lessons learnt from this project, the research team has now successfully obtained support from the NIHR (National Institute for Health Research) for a five-year Grant Programme (2009 to 2014), *Improving Risk Management in Mental Health Service* (RP-PG-0407-10500). The research areas or methodological issues reported here will be further studied and addressed by various projects within the Grant Programme.

# 1. Context

Risk assessment for future violence is of major importance in the aftercare of offenders and public protection by professionals in criminal justice, probation and mental health. There is a need to improve risk assessment and to increase our overall knowledge in this area. Many different risk assessment and violence prediction tools have already been developed. There is still a need to find the best combination of instruments and related methodologies to improve our ability to assess risk and predict the required types of outcomes.

The main components of empirical study of risk assessment are the predictors, outcomes, different populations of interest, and the statistical methods used to integrate the predictors (Steadman & Monahan, 1994). Much previous research has been devoted to improving the predictors and outcomes. There is relatively little literature, firstly on identification of the best statistical models in testing and developing new instruments, and secondly on the homogeneity of target populations when testing instruments.

Comparisons between multiple risk assessment instruments have demonstrated a 'glass-ceiling' effect in terms of accuracy, whereby Area Under the Curve (AUC) values, the ratio between the true positive and false positive prediction from the receiver operating characteristic curve (ROC) analysis, demonstrate a maximum at 0.75 (Kroner & Mills, 2001; Coid *et al.*, 2007). Conventional Logistic Regression models used to construct these instruments assume independence among predictors, which is unlikely, and this may explain the finite level of performance.

More sophisticated approaches, such as the use of models involving data mining techniques, including decision tree and artificial Neural Networks models have also been suggested. Steadman *et al.* (2000) proposed the use of Classification Tree models, the simplest decision tree type of model, to construct risk assessment instruments. Stalans *et al.* (2004) have also used CT models to classify violent offenders. Meanwhile, the application of artificial Neural Networks (NNs) models for prediction and classification has attracted growing interest (Price *et al.*, 2000; Starzomska, 2003). However, research in applying CT models including Classification and Regression Tree (C&RT), Chi-square Automatic Interaction Detector (CHAID) and Iterative Classification Tree (ICT) in predicting violent risk, or violence or overall recidivism among psychiatric patients or criminal offenders in comparison with traditional regression models, has produced inconsistent findings over the performance of these models (Gardner *et al.*, 1996; Monahan *et al.*, 2000, 2005, 2006; Steadman *et al.*, 2000; Silver, Smith & Banks, 2000; Silver & Chow-Martin, 2002; Stalans, Yarnold, Seng, Olson, & Repp, 2004; Thomas *et al.*, 2005; Rosenfeld & Lewis, 2005). A few studies compared NNs with traditional regression models (Brodzinski, Crable, & Scherer, 1994; Caulkins, Cohen, Gorr & Wei, 1996; Palocsay, Wang, & Brookshire, 2000; Grann & Langstrom, 2007) in predicting risk of recidivism, and also reported inconsistent findings. There have been no reports of methodological study of statistical models in the assessment of violent recidivism among offenders. Among professionals in forensic mental health and criminal justice services, the applicability of the NNs and CT models, and their comparability with the traditional LR models, remain unclear.

The aims of the study were as follows.

(i) Test the applicability of NNs and CT models in predicting violent recidivism.

(ii) Compare the predictive power of NNs and CT with conventional LR and Discriminant Analysis (DA) models using the multiple risk assessment instruments previously employed in the Prisoner Cohort Study (Coid, Yang *et al.*, 2007).

(iii) The study also aimed to explore the predictive power of changeable (so called dynamic) variables including institutional behaviour and post-release factors such as social network factors, attitudes to crime, employment and life style/events in relation to violent offences and calibrate them in both NNs and LR model analysis.

(iv) Finally the study aimed to examine possible interaction between statistical models and the different elements of the target population in risk assessment. So the study further compared the differential predictive accuracy of separate models for young adults, women, and those with personality disorders. The ultimate objective of this methodological study was to provide guidelines for future research. If possible, it was even intended to develop new instruments, with specialised software, suitable for implementation in forensic and correctional services.

For aims (i) and (ii), the study explored major technical aspects of NNs and CT models by means of StatSoft software, and compared the performance of these models with LR or DA in predicting violent outcomes, using several sets of variables in existing risk assessment instruments. The study findings provide evidence showing the general comparability of NNs, CT and LR/DA, together with further details on strengths and limitations of different models to guide future work. For aim (iii), the study examined the predictive power of those variables by testing their effects in addition to static or criminal behaviour variables that were fixed for all models under comparison. The findings for model performance with additional variables, as discussed below in the Implications section, suggest potentially greater flexibility for NNs in detecting small effects when large amounts of data (variables and cases) are available. It also emerged that dynamic or changeable variables collected in the study cohort had only relatively moderate effects in predicting violence reoffending in general. Findings for aim (iv) draw conclusions on how model performance could be improved in assessing risk for specific populations such as prisoners without personality disorders and with different outcome specifications such as treating non- offenders differently from other non-violent offenders in clinical practice. Pooling all the evidence together, the study concludes that four methodological challenges are involved in risk assessment: identifying causal or powerful predictors that are specific to a specific outcome; identifying a target population with suitable homogeneity; defining specific outcome criteria for specific outcomes and finally maximising the strength of statistical models. These challenges all need to be tackled together in order to improve predictive accuracy further. This is an under-explored terrain, possibly due to technical complexity. In aiming to meet this challenge, the present study proposes a multi-stage multi-tool and multi-model approach, and tests the applicability of different models in a multi-stage process. Initial findings and recommendations for future research are presented in the report. This report should be mostly of interest to researchers in the area of methodology for developing risk assessment tools.

## Parameters of the study

- The population for this project comprised a sample of 1,353 adult men and 304 adult women prisoners, as interviewed prior to release for the Prisoner Cohort Study (Coid *et al.*, 2007). The findings on dynamic variables are based on a subsample of men (n=664) from the Phase II community survey following their release. The mean age at interview was 30.7 years for men and 28.2 years for women.

- Violent recidivism as an outcome measure for the Prisoner Cohort Study was based on the Police National Computer database. By 13 October 2005 (for men) and 9 February 2007 (for women), a mean follow-up time for the male sample after release was 1.98 years with a base rate of violent recidivism of 13.2%, and for women 2.08 years with a base rate of 8.2%.

- Predictors to test the models were individual items in four psychometric instruments: HCR-20, PCL-R, VRAG and RM2000V.

## 2. Approach

This is a methodological study focused on the value and comparability of different statistical models in the construction of risk assessment instruments. It draws on the dataset from the Prisoner Cohort Study (Coid *et al.*, 2007).

The Prisoner Cohort Study involved interviews with both men (N=1,363) and women (N=321) released between 14 November 2002 and 7 October 2005 and followed up first for the usual two years and then for four years. Detailed description of the sampling procedure, the interview, characteristics of the sample, and of the risk assessment measures can be found in Coid *et al.*, (2007).

Outcome data were derived from reconvictions recorded in the Police National Computer, an operational police database containing criminal histories of all offenders in England, Wales and Scotland up to the date 13 October 2005 for men and up to 9 February 2007 for women; 1,353 men and 304 women have outcome data. Missing data for ten men and 17 women were due to them not having been released or no PNC record being available. The mean age for men is 30.7 years (SD=11.4, min – max: 18-75), and for women 28.2 years (SD=8.8, min – max: 18 – 60). The mean follow-up time to the initial PNC search was 1.98 years (SD = 0.54, min-max: 6 days – 2.91 years) for men, and 2.08 years (SD = 0.88, min – max: 23 days – 4.14 years) for women.

Following common practice, the target population was violent reoffenders who reoffended through homicide, or major violence, or minor violence, and weapons offences. The contrasting category of ‘others’ includes those with no recorded reconvictions, and other, non-violent reoffenders during the follow-up period.

To examine the applicability and predictive accuracy of different models, the same predictors are required for each model. Items in existing risk assessment instruments are considered reliable measures. However, an instrument with different predictors may have different validity, and the number of predictors in an instrument may have an interactive effect with model performance. The following four instruments are therefore used in this study as four sets of predictors, for model comparison:

- Violence Risk Appraisal Guide (VRAG), 12 predictors;
- Psychopathy Checklist Revised (PCL-R), 20 predictors;
- Historical-Clinical-Risk Management -20 (HCR-20), 20 predictors;
- Risk Matrix 2000 – Violence (RM2000V), 3 predictors.

The study examined the performance of four models in predicting violent outcome or classifying violent reoffenders: (i) logistic regression; (ii) discriminant analysis; (iii) classification tree; and (iv) Multi-layer Perceptron Neural Networks. The LR and DA are traditional statistical models derived from probability theory. When classifying between two categories of outcome, LR and DA yield similar results, and are therefore exchangeable. The CT and NNs are data-mining tools derived from classification theory and pattern identification techniques respectively. Traditional models have been routinely used to develop risk assessment scales or instruments, and to test the validity of existing instruments (Hosmer & Lemeshow, 1989). Data-mining models are rarely used in forensic risk assessment practice. More detailed description of the model fitting aspects can be found in the Methodological notes of the report.

For each model, the same two-thirds of subjects are used to construct the model as the training sample and the remaining one-third for external testing as the test sample. A significantly greater predictive accuracy for the training sample compared to that of the test sample would suggest over-fitting of the model. The weighting or misclassification rate is fixed proportionally to the true ratio between the violent reoffender group and the other so that the wrong prediction in the two groups would be reasonably balanced. Two basic principles are applied to decide a potentially 'best' model: the highest accuracy for the overall sample or the test sample, and the least difference in accuracy between the training and test samples. These principles are used as a safeguard to avoid choosing an over-fitted model.

When fitting NNs models, many training nets are required in order to find the 'best' model that meets the above criteria. Too few runs may miss out the best models. Too many runs are considered unnecessary. The software StatSoft sets out 20 runs as default. This study has generally performed 50 runs for each instrument and asked for a return of ten best models from the software. The final 'best' model is then selected from the ten returned models, based on the two principles described above. In some exercises, when necessary, all runs are retained to observe trends and patterns of the model behaviour.

Accuracy measures include (1) sensitivity or proportion of violent reoffenders correctly predicted by the model; (2) specificity or proportion of non-violent reoffenders or non-reoffenders correctly predicted by the model; and (3) overall accuracy, a combination of sensitivity and specificity. (This is comparable to the Area Under Curve value in ROC analysis widely used in risk assessment research.)

For traditional DA and LR models, SPSS v16 was used to construct and test the models, and StatSoft and DTREG were used to fit decision tree and NNs models.

# 3. Results

## Recidivism rates during the follow-up period

By the time of the first PNC search, 1,353 men and 304 women had outcome data. There were 13.2% male and 8.2% female prisoners who reoffended violently in the two years following release. More men than women were reconvicted of robbery, acquisitive, and any crimes. Two years of follow-up is too short a time to observe adequate sexual reoffending among both men and women. Table 3.1, showing recidivism rates for different durations of follow-up, indicates that there were only 74 (5.5%) men and 35 (11.5%) women with less than 12-months follow-up. The reoffending rate for any crime during the first year after release was generally low. The likelihood of violent and acquisitive re-conviction increased after one year. A longer follow-up time is required to observe an equivalent reoffending pattern for robbery and sex offences.

***Table 3.1 Recidivism rates by duration of follow-up***

| | | Reconvictions by offence type, | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | No. of prisoners | Violence | | Robbery | | Acquisitive | | Sex | | Any reconviction | |
| | | N | % | N | % | N | % | N | % | N | % |
| **Male** | | | | | | | | | | | |
| < 12 months | 74 | 3 | 4.1 | 1 | 1.4 | 3 | 4.1 | 0 | 0.0 | 11 | 14.9 |
| 1 – 2 years | 572 | 59 | 10.3 | 24 | 4.2 | 115 | 20.1 | 4 | 0.7 | 229 | 40.0 |
| 2 – 3 years | 707 | 116 | 16.4 | 39 | 5.5 | 183 | 25.9 | 3 | 0.4 | 369 | 52.2 |
| Total | 1,353 | 178 | 13.2 | 64 | 4.7 | 301 | 22.2 | 7 | 0.5 | 609 | 45.0 |
| **Female** | | | | | | | | | | | |
| < 12 months | 35 | 0 | 0.0 | 0 | 0.0 | 1 | 2.9 | 0 | 0.0 | 2 | 5.7 |
| 1 - 2 years | 103 | 9 | 8.7 | 1 | 1.0 | 11 | 10.7 | 2 | 1.9 | 23 | 22.3 |
| 2 - 3 years | 115 | 10 | 8.7 | 4 | 3.5 | 20 | 17.0 | 2 | 1.7 | 40 | 34.8 |
| 3 – 4 years | 51 | 6 | 11.8 | 2 | 3.9 | 15 | 29.4 | 1 | 2.0 | 23 | 45.1 |
| Total | 304 | 25 | 8.2 | 7 | 2.3 | 47 | 15.5 | 5 | 1.6 | 88 | 28.9 |

Violence: homicide, major violence, minor violence, use of weapons.
Robbery: robbery, aggravated burglary.
Acquisitive: forgery, burglary and thefts.
Sex: rape, sex assaults and other sex offences.

## Overall comparison of predictive accuracy among models

Results in Table 3.2 demonstrate that for men, the performance of DA, CT and NNs models demonstrated no significant difference. There are overlapping confidence intervals in the level of their overall accuracy for both HCR-20 and PCL-R items. The same patterns are observed for the VRAG and RM2000V items. There is no evidence that one model performed better than another for male prisoners, no matter which instrument is used.

The female sample is much smaller than the male sample. The model performance is therefore less stable than that observed for males. Table 3.2 demonstrates low sensitivity in particular in the test sample. The CT training model fitted by StatSoft is clearly over-fitted. The alternative Classification & Regression Tree (C&RT) model was fitted using DTREG software. As there were overlapping confidence intervals of accuracy in the test sample among models, no statistical differences can be concluded in performance between all models.

### Table 3.2 Predictive accuracy of risk assessment tools by models

| Model | Subsample | N | Sensitivity | Specificity | Accuracy | (95% CI) |
|---|---|---|---|---|---|---|
| **Male** | | | | | | |
| **HCR-20** | | | | | | |
| DA | Training | 827 | 0.75 | 0.63 | 0.65 | (0.61-0.68) |
| | Test | 426 | 0.64 | 0.58 | 0.58 | (0.54-0.63) |
| CT | Training | 827 | 0.84 | 0.67 | 0.69 | (0.66-0.73) |
| | Test | 426 | 0.61 | 0.60 | 0.60 | (0.55-0.65) |
| NNs | Training | 827 | 0.61 | 0.68 | 0.67 | (0.64-0.70) |
| | Test | 426 | 0.60 | 0.66 | 0.66 | (0.62-0.71) |
| **PCL-R** | | | | | | |
| DA | Training | 627 | 0.71 | 0.61 | 0.63 | (0.59-0.66) |
| | Test | 324 | 0.76 | 0.55 | 0.58 | (0.52-0.63) |
| CT | Training | 627 | 0.89 | 0.69 | 0.71 | (0.68-0.75) |
| | Test | 324 | 0.42 | 0.63 | 0.60 | (0.55-0.66) |
| NNs | Training | 627 | 0.73 | 0.61 | 0.63 | (0.59-0.67) |
| | Test | 324 | 0.68 | 0.60 | 0.61 | (0.56-0.66) |
| **Female** | | | | | | |
| **HCR-20** | | | | | | |
| DA | Training | 211 | 0.81 | 0.71 | 0.72 | (0.66-0.78) |
| | Test | 90 | 0.44 | 0.67 | 0.63 | (0.55-0.75) |
| CT | Training | 211 | 1.00 | 0.88 | 0.89 | (0.85-0.93) |
| | Test | 90 | 0.33 | 0.80 | 0.75 | (0.66-0.84) |
| C&RT[a] | Training | 211 | 0.86 | 0.80 | 0.81 | (0.76-0.86) |
| | Test | 90 | 0.44 | 0.75 | 0.72 | (0.62-0.82) |
| NNs | Training | 211 | 0.95 | 0.71 | 0.73 | (0.67-0.79) |
| | Test | 90 | 0.67 | 0.72 | 0.72 | (0.62-0.82) |
| **PCL-R** | | | | | | |
| DA | Training | 164 | 0.84 | 0.71 | 0.73 | (0.66-0.80) |
| | Test | 70 | 0.50 | 0.73 | 0.70 | (0.59-0.81) |
| CT | Training | 164 | 1.00 | 0.88 | 0.89 | (0.84-0.94) |
| | Test | 70 | 0.25 | 0.89 | 0.81 | (0.72-0.90) |
| C&RT[a] | Training | 164 | 0.78 | 0.86 | 0.76 | (0.69-0.83) |
| | Test | 70 | 0.50 | 0.77 | 0.74 | (0.64-0.84) |
| NNs | Training | 164 | 0.95 | 0.77 | 0.79 | (0.73-0.85) |
| | Test | 70 | 0.63 | 0.76 | 0.74 | (0.64-0.84) |

[a] C&RT: Classification and Regression Tree model fitted by the package DTREG.

Nevertheless, the accuracy in general of the two instruments among female prisoners is notably higher than that of males, with no confidence interval overlap in the male/female training samples for DA and NNs with the PCL-R, and only a marginal difference in the test sample for the C&RT model. (This finding of a difference in model performance between genders is further demonstrated and discussed in a later section on gender difference).

## Risk assessment for specific populations

### Gender difference

The predictive accuracy of a risk instrument is derived from its ability to separate individuals between the outcome categories, i.e. violent reoffenders and others in this study. The larger the difference in the mean score of an instrument between the two outcome categories, the greater its ability to separate, and hence a higher predictive accuracy. The mean difference (Table 3.3) in PCL-R score between the two outcome category groups was 5.9 among women compared to 3.7 among men, and for the HCR-20, 5.3 among women compared to 4.4 among men. Using the AUC value as a measure of accuracy, Table 3.3 demonstrates that the HCR-20, and PCL-R in particular, demonstrate higher accuracy to predict violence among women than among men with a significant level $p<0.05$ for PCL-R. However, the RM2000V has higher accuracy among men than women without reaching a significant level. This further explains why all three types of models show better predictive accuracy among female prisoners than males when using the PCL-R (see Table 3.2). However, this finding raises an important question. Is this a sampling issue or a gender difference in terms of the association between psychopathy and recidivistic behaviour? Further research on PCL-R items and its factor constructs is required to understand why this instrument works better for female prisoners than for males.

### *Table 3.3 Risk level for violent reoffending and predictive effects of risk assessment instruments*

| | No. of prison-ers | HCR-20 | | PCL-R | | VRAG | | RM2000V | |
|---|---|---|---|---|---|---|---|---|---|
| | | Mean | (SD) | Mean | (SD) | Mean | (SD) | Mean | (SD) |
| **Men** | | | | | | | | | |
| Non-violent & non-reoffenders | 1,166 | 18.5 | (7.8) | 17.6 | (7.7) | 4.2 | (1.9) | 10.7 | (11.0) |
| Violent reoffenders | 177 | 22.9 | (6.4) | 21.3 | (6.3) | 5.4 | (1.0) | 18.2 | (7.3) |
| Difference (p value)[a] | | 4.4 | (0.000) | 3.7 | (0.000) | 1.2 | (0.000) | 7.5 | (0.000) |
| AUC (95% CI)[b] | | .67 | (.63-.71) | .63 | (.59-.67) | .68 | (.65-.72) | .70 | (.66-.73) |
| **Women** | | | | | | | | | |
| Non-violent & non reoffenders | 279 | 19.7 | (7.5) | 16.0 | (7.3) | 4.1 | (1.5) | 9.5 | (10.0) |
| Violent reoffenders | 25 | 25.0 | (6.1) | 21.9 | (6.4) | 4.9 | (1.2) | 14.8 | (6.6) |
| Difference (p value)[a] | | 5.3 | (0.000) | 5.9 | (0.000) | 0.8 | (0.000) | 5.3 | (0.000) |
| AUC (95% CI)[b] | | .70 | (.60-.81) | .73 | (.63-.83) | .66 | (.55-.76) | .64 | (.55-.74) |
| Gender difference in AUC (p value)[b] | | -0.03 | (0.299) | -0.10 | (0.034) | 0.02 | (0.350) | 0.06 | (0.124) |

[a] t-test to compare with the non-violent & non-reoffenders group.
[b] t-test to compare difference in the AUC value between men and women.

## Young adults

Young offenders aged under 22 may be at the peak of their criminal career, and are therefore the subgroup of greatest concern for risk of reoffending. Examining the performance of various models on this subgroup and using predictors from established risk assessment instruments therefore has important implications for clinical practice.

There were 308 (22.3%) young adults aged under 22 in the male sample. Of these, 62 (20.1%) reoffended violently within a two-year follow-up, contrasting with 116 (11.1%) of 1,045 men aged 22 years or older. Only HCR-20 predictors were examined. Among the young offender group, C&RT model demonstrated the lowest accuracy (Table 3.4). The LR and NNs did not show notable differences in their predictive accuracy, which could be due to an insufficient statistical power of these models as fitted to this small sample. Among the older age group, NNs demonstrated higher accuracy than the other two models. However, this was only just above chance and may have been the result of a trade-off with higher specificity (true negative) predicted by NNs compared to the other two models. A comparison of predictive accuracy between the two age groups in the performance of LR and NNs demonstrated no significant differences. Better performance of C&RT for the older age group was observed due to poor fitting of the model in the younger age group which has a smaller sample size. More vigorous comparison of model performance, in particular for C&RT between the two age groups, as based on a larger number of people in the younger age group than the current sample, is required to confirm the pattern found in this study.

***Table 3.4 Predictive accuracy of three models for violent recidivism by age group (HCR-20)***

| Model | Subsample[a] | N | Sensitivity | Specificity | Accuracy | (95% CI) |
|---|---|---|---|---|---|---|
| **Male** | | | | | | |
| **≤21 years** | | | | | | |
| LG | Training | 173 | 0.73 | 0.63 | 0.65 | (0.58-0.72) |
| | Test | 100 | 0.40 | 0.72 | 0.64 | (0.55-0.73) |
| C&RT | Training | 173 | 0.94 | 0.29 | 0.42 | (0.35-0.49) |
| | Test | 100 | 0.75 | 0.27 | 0.40 | (0.30-0.50) |
| NNs | Training | 173 | 0.85 | 0.73 | 0.75 | (0.69-0.81) |
| | Test | 100 | 0.40 | 0.75 | 0.66 | (0.57-0.75) |
| **>21 years** | | | | | | |
| LG | Training | 654 | 0.79 | 0.63 | 0.65 | (0.61-0.69) |
| | Test | 326 | 0.67 | 0.57 | 0.58 | (0.53-0.63) |
| C&RT | Training | 654 | 0.81 | 0.59 | 0.61 | (0.57-0.65) |
| | Test | 326 | 0.69 | 0.59 | 0.60 | (0.55-0.65) |
| NNs | Training | 654 | 0.62 | 0.72 | 0.71 | (0.68-0.74) |
| | Test | 326 | 0.53 | 0.72 | 0.70 | (0.65-0.75) |
| **Women** | | | | | | |
| **≤21 years** | | | | | | |
| DA[b] | Training | 82 | 0.73 | 0.74 | 0.73 | (0.63-0.83) |
| | Test | 82 | 0.36 | 0.66 | 0.62 | (0.51-0.73) |
| C&RT[c] | Training | 82 | 0.91 | 0.66 | 0.70 | (0.60-0.80) |
| | Test | 82 | 0.91 | 0.59 | 0.63 | (0.52-0.74) |
| NNs[c] | Training | 82 | 0.82 | 1.00 | 0.98 | (0.95-1.00) |
| | Test | 82 | 0.18 | 0.81 | 0.72 | (0.62-0.82)[d] |
| **>21 years** | | | | | | |
| LG[b] | Training | 234 | 0.79 | 0.73 | 0.74 | (0.68-0.80) |
| | Test | 234 | 0.37 | 0.68 | 0.65 | (0.59-0.71) |
| C&RT[c] | Training | 234 | 0.95 | 0.73 | 0.75 | (0.69-0.81) |
| | Test | 234 | 0.90 | 0.48 | 0.51 | (0.44-0.58) |
| NNs[c] | Training | 234 | 0.90 | 0.97 | 0.97 | (0.95-0.99) |
| | Test | 234 | 0.16 | 0.91 | 0.85 | (0.80-0.90)[d] |

[a] Due to some missing data on some items of HCR-20 scale, the total number for men does not add up to 1,353.
[b] Leave one out for testing in SPSS 16.0.
[c] ten-fold cross validation for testing.
[d] Possible over-fitting.

In the female sample, only 79 were aged under 22, and 222 were aged 22 or more. The rate of violent reoffending was 13.4% and 8.5 % for the younger and older group respectively. Due to the small sample size, the model has to be trained using all the data, and different cross-validation methods for different models are used with different software for the purpose of testing. Table 3.4 shows that there was a large discrepancy in accuracy between the training and test results for each model in the female sample due to the small sample size. The higher accuracy of NNs in the training sample for both age groups was accompanied by an extremely low sensitivity in the test sample, indicating marked shrinkage in accuracy of the HCR-20. Overall, for women, no difference in performance of any model was observed between the younger and older age groups. In other words, the predictors of the HCR-20 do not have specificity among young offenders. The type of model made no difference to the predictive accuracy of the HCR-20 within any age bands.

## Male prisoners with PDs

Personality disorder, in particular ASPD, is considered a high risk factor for serious recidivism among prisoners and has been used as a criterion for the Home Office/Ministry of Justice/ Department of Health DSPD Programme. There are fundamental differences between the PD and non-PD prisoners based on our data. Compared to the non-PD group, the PD group was approximately seven years younger when interviewed, 3.7 years younger at their first court appearance for violence, more were single, more were of White ethnicity, more were long-term unemployed before imprisonment, and more were diagnosed as drug-dependent and with alcohol misuse disorder. They had a higher prevalence of major psychosis and depressive disorder. The PD group was convicted of more previous violent, robbery and acquisitive offences, and was also more likely to be reconvicted of the same type of offences following release.

In the male sample, a total of 967 (73.0%) had one or more Axis II PD diagnoses. Among this PD group, 88.9% (N=860) had ASPD. The base rate for violent reoffending over the two-year follow-up was 15.8% for the PD group and 6.2% for the non-PD. The CT and NNs model do not perform stably for data with low base rates, from our experience. For this exercise, the second set of PNC recidivism outcomes, for the period up until 9 February 2007, was used to extend the follow-up period, with a mean duration of 3.3 years for both PD and non-PD groups. This increased the base rate for violent reoffending to 26.1% for the PD and 12.1% for the non- PD group. Of the original sample of 1,363 male prisoners, 136 were omitted; this was due to missing PD diagnosis (39 cases), less than one-year follow-up period (7 cases), and missing data on some items from the HCR-20 (90 cases). The final sample of 1,227 in this analysis consisted of a total of 908 (74%) with one or more Axis II PD diagnoses and 319 (26%) without any PD. Among the PD group, 94% (N=854) had Cluster B PD diagnosis, of which 95% (N=814) included ASPD.

The entire sample was split into three subsamples, 4/6 for model training, 1/6 for test and 1/6 for external validation.

### *Table 3.5 Predictive accuracy of three models by PD diagnosis (HCR-20, men, N=1,227)*

| Model | Subset | N | Sensitivity | Specificity | Accuracy | (95%CI) |
|---|---|---|---|---|---|---|
| **PD (N=908)** | | | | | | |
| LR | Train | 595 | 0.69 | 0.61 | 0.63 | (0.59, 0.67) |
| | Test | 163 | 0.57 | 0.50 | 0.52 | (0.44, 0.60) |
| | Validate | 150 | 0.61 | 0.57 | 0.59 | (0.51, 0.67) |
| CT | Train | 595 | 0.70 | 0.66 | 0.67 | (0.63, 0.71) |
| | Test | 163 | 0.50 | 0.54 | 0.53 | (0.45, 0.61) |
| | Validate | 150 | 0.59 | 0.60 | 0.60 | (0.52, 0.68) |
| NNs | Train | 595 | 0.69 | 0.58 | 0.61 | (0.57, 0.65) |
| | Test | 163 | 0.58 | 0.57 | 0.58 | (0.50, 0.66) |
| | Validate | 150 | 0.67 | 0.56 | 0.60 | (0.52, 0.68) |
| **Non-PD (N=319)** | | | | | | |
| LR | Train | 203 | 0.81 | 0.72 | 0.73 | (0.67, 0.79) |
| | Test | 55 | 0.86 | 0.73 | 0.75 | (0.64, 0.86) |
| | Validate | 61 | 0.63 | 0.77 | 0.75 | (0.64, 0.86) |
| CT | Train | 203 | 0.81 | 0.71 | 0.72 | (0.66, 0.78) |
| | Test | 55 | 0.86 | 0.77 | 0.78 | (0.67, 0.89) |
| | Validate | 61 | 0.75 | 0.70 | 0.70 | (0.58, 0.82) |
| NNs | Train | 203 | 1.00 | 0.80 | 0.83 | (0.78, 0.88) |
| | Test | 55 | 0.71 | 0.79 | 0.78 | (0.67, 0.89) |
| | Validate | 61 | 0.75 | 0.89 | 0.87 | (0.79, 0.95) |

It can be seen from Table 3.5 that, for the PD group, all three models performed at the same low level of accuracy, around 0.6 or lower ($\chi^2_{CMH}$ = 4.10, d.f. =3, p = 0.129). For the non-PD group, models showed significant difference in their performance ($\chi^2_{CMH}$ = 9.08, d.f.=3, p = 0.011), with the mean accuracy at 0.83 for NNs, 0.74 for LR and 0.73 for CT. The NNs showed the most improvement in accuracy from a mean of 0.60 for the PD group to 0.83 for the non-PD group ($\chi^2_{CMH}$= 40.82, d.f. = 2, p < 0.001), followed by LR from 0.58 to 0.74 ($\chi^2_{CMH}$ = 11.12, d.f. =2, p = 0.009) and finally CT from 0.60 to 0.73 ($\chi^2_{CMH}$ = 4.16, d.f. = 2, p = 0.04).

The improved accuracy of all models for the non-PD male prisoners was not limited to the HCR-20 alone. The simple comparison of the AUC value between the PD and non-PD groups for each of the four instruments in Table 3.6 further indicates that the same finding is repeated for the PCL-R with an AUC difference of 0.10, and for the VRAG with a difference of 0.17. For the RM2000V the difference was 0.10.

**Table 3.6 Risk level for violent reoffending and predictive effects of risk assessment instruments by PD diagnosis (men, N=1,227)**

| | Prisoners[a] | | HCR-20 | | PCL-R | | VRAG | | RM2000V | |
|---|---|---|---|---|---|---|---|---|---|---|
| | N | % | Mean | SD | Mean | SD | Mean | SD | Mean | SD |
| **PD** | | | | | | | | | | |
| Violent reoffenders | 255 | 26.1 | 23.7 | 6.0 | 22.2 | 6.0 | 5.4 | 1.0 | 19.4 | 6.4 |
| Non-violent & non-reoffenders | 720 | 73.9 | 21.2 | 6.4 | 19.8 | 6.5 | 4.6 | 1.6 | 14.7 | 8.0 |
| Group difference (p value) | | | 2.5 | <0.001 | 2.4 | <0.001 | 0.8 | <0.001 | 4.7 | <0.001 |
| AUC (95% CI) | | | .61 | .57-.65 | .59 | .55-.63 | .63 | .60-.67 | .67 | .63-.70 |
| **Non-PD** | | | | | | | | | | |
| Violent reoffenders | 42 | 12.1 | 15.9 | 6.5 | 15.6 | 7.0 | 5.1 | 1.1 | 8.2 | 6.7 |
| Non-violent & non-reoffenders | 305 | 87.9 | 11.0 | 5.8 | 11.1 | 6.6 | 3.0 | 2.0 | -0.48 | 9.8 |
| Group difference (p value) | | | 4.9 | <0.001 | 4.5 | <0.001 | 2.1 | <0.001 | 8.68 | <0.001 |
| AUC (95% CI) | | | .70 | .61-.79 | .69 | .61-.78 | .80 | .74-.85 | .77 | .70-.83 |

[a] Excluded cases with missing items in different instruments, and on the PD diagnosis.

Investigation of the predictive accuracy of HCR-20 items revealed a marked discrepancy between the two groups of participants on a number of static items, including H1 (Previous violence), H2 (Young age at first violent incident), H3 (Relationship instability), H4 (Employment problems) and H6 (Major mental illness). These items showed higher predictive power among the non-PD group compared to the PD. Once these items were removed from HCR-20, the scale with the remaining items showed the same predictive accuracy for violent recidivism between the two groups. These static items are known to be effective predictors of future violent behaviour and have a strong association with other forms of anti-social and criminal behaviour. In this context, it could be hypothesised that these static items may largely be constituted by anti-social behaviour traits or Factor 2 of psychopath traits which contribute a large amount of the efficacy of any risk assessment instrument for distinguishing between criminals and non-criminals within a mixed population. When applying such an instrument to a high risk population, among which ASPD traits are highly prevalent, as in this analysis, the predictive efficacy contributed by the anti-social component within the instrument may be limited or even reduced. Given the fact that in the PD group, 94.1% were Cluster B PD, and with 95% having ASPD, the findings support this hypothesis to explain why the HCR-20 underperformed for the PD group.

The superior performance of NNs over the LR and CT models in the non-PD group could partly be due to items within the HCR-20 that were more valid for prisoners without a Cluster B PD than for those with a Cluster B PD when predicting violence. It could be partly due to this specific characteristic of the sample together with the chosen predictors suiting NNs better than the other models. There could be some interactive effects between model and predictors and the homogeneity of population.

## Model performance with additional predictors

Additional variables for institutional behaviour (difficult and disruptive behaviour whilst in prison), motivational factors for the Index Offence, and dynamic measures for the male sample following release were examined individually for their association with violent re-conviction at two years, based on simple descriptive statistics. Those with statistical significance in their association with violent reconviction were selected to form three sets of new variables.

- The set of ten variables for institutional behaviour comprised: being placed in a stripped cell; assaulted others, used a weapon; assaulted other prisoners, assaulted prison staff; self-injured whilst fighting; other person injured whilst fighting; hostage taking; damaging property or own cell; and making weapons.
- The set of 11 crime motivation variables consisted of: expressive aggression; sexual gratification; paraphilia or sexual deviation; hyperirritability; compulsive urge to harm/kill; blow to self-esteem; under-controlled aggression; revenge, financial gain; gang/group activity; and intoxication.
- The set of nine of dynamic or changeable factors included no address to go to on release; currently living with family members friends; bored at all times; frequency of seeing friends; thoughts of hurting others; social network score; psychosis score (PSQ), drinking (Audit Score); and any drug dependence.

Most are dichotomous variables. They all have demonstrated small predictive effects on violent recidivism, individually and collectively. Each of the three sets was retained in one model but used as individual variables in the modelling exercise. Predictive accuracy of HCR-20 and RM2000V, with and without those variables, using all models, is presented in Tables 3.7 and 3.8.

***Table 3.7 Predictive accuracy of HCR-20 or RM2000V alone, with institutional behaviour and motivation factors***

| Model | HCR-20 | | | RM2000V | | |
|---|---|---|---|---|---|---|
| | Sample | Acc. % | (95%CI) | Sample | Acc. % | (95%CI) |
| LR: Original scale | Train: 827 | 0.64 | (0.61, 0.67) | Train: 815 | 0.66 | (0.63, 0.69) |
| | Test: 426 | 0.57 | (0.52, 0.62) | Test: 412 | 0.70 | (0.66, 0.74) |
| LR: Plus 10 institutional behave. variables | Train: 812 | 0.67 | (0.64, 0.70) | Train: 800 | 0.66 | (0.63, 0.69) |
| | Test: 416 | 0.59 | (0.54, 0.64) | Test: 413 | 0.66 | (0.61, 0.71) |
| LR: Plus 11 motivation variables | Train: 821 | 0.67 | (0.64, 0.70) | Train: 809 | 0.64 | (0.61, 0.67) |
| | Test: 423 | 0.61 | (0.56, 0.66) | Test: 408 | 0.66 | (0.61, 0.71) |
| CT: Original scale | Train: 827 | 0.64 | (0.61, 0.67) | Train: 815 | 0.64 | (0.61, 0.67) |
| | Test: 426 | 0.59 | (0.54, 0.64) | Test: 412 | 0.67 | (0.62, 0.72) |
| CT: Plus institutional behave. variables | Train: 812 | **0.70** | **(0.67, 0.73)** | Train: 800 | **0.72** | **(0.69, 0.75)** |
| | Test: 416 | 0.63 | (0.58, 0.68) | Test: 413 | 0.67 | (0.62, 0.72) |
| CT: Plus 11 motivation variables | Train: 821 | **0.69** | **(0.66, 0.72)** | Train: 809 | 0.64 | (0.61, 0.67) |
| | Test: 423 | 0.63 | (0.58, 0.68) | Test: 408 | 0.68 | (0.63, 0.73) |
| NNs: HCR-20 | Train: 827 | 0.65 | (0.62, 0.68) | Train: 815 | 0.64 | (0.61, 0.67) |
| | Test: 426 | 0.62 | (0.57, 0.67) | Test: 412 | 0.67 | (0.62, 0.72) |
| NNs: Plus 10 institutional behave. variables | Train: 812 | **0.74** | **(0.71, 0.77)** | Train: 800 | **0.73** | **(0.70, 0.76)** |
| | Test: 416 | **0.71** | **(0.67, 0.75)** | Test: 413 | 0.68 | (0.63, 0.73) |
| NNs: Plus 11 motivation variables | Train: 821 | **0.71** | **(0.68, 0.74)** | Train: 809 | **0.73** | **(0.70, 0.76)** |
| | Test: 423 | **0.72** | **(0.68, 0.76)** | Test: 408 | 0.67 | (0.62, 0.72) |

***Table 3.8 Predictive accuracy of HCR-20 or RM2000V alone, with dynamic or changeable variables (Phase II community male sample)***

| Model | HCR-20 | | | RM2000V | | |
|---|---|---|---|---|---|---|
| | Sample | Acc. % | (95%CI) | Sample | Acc. % | (95%CI) |
| LR: Original scale | Train: 399 | 0.69 | (0.64, 0.74) | Train: 396 | 0.66 | (0.63, 0.72) |
| | Test: 205 | 0.63 | (0.56, 0.70) | Test: 194 | 0.67 | (0.60, 0.73) |
| LR: Plus 9 dynamic variables | Train: 383 | 0.73 | (0.69, 0.77) | Train: 378 | 0.67 | (0.62, 0.72) |
| | Test: 195 | 0.65 | (0.58, 0.72) | Test: 186 | 0.68 | (0.62, 0.75) |
| CT: Original scale | Train: 399 | 0.68 | (0.63, 0.72) | Train: 396 | 0.68 | (0.63, 0.72) |
| | Test: 205 | 0.61 | (0.54, 0.68) | Test: 194 | 0.67 | (0.60, 0.73) |
| CT: Plus 9 dynamic variables | Train: 383 | 0.69 | (0.64, 0.73) | Train: 378 | **0.82** | **(0.77, 0.85)** |
| | Test: 195 | 0.64 | (0.57, 0.70) | Test: 186 | 0.70 | (0.63, 0.77) |
| NNs: HCR-20 | Train: 399 | 0.82 | (0.78, 0.85) | Train: 396 | 0.73 | (0.69, 0.78) |
| | Test: 205 | 0.76 | (0.70, 0.82) | Test: 194 | 0.73 | (0.67, 0.79) |
| NNs: Plus 9 dynamic variables | Train: 383 | 0.80 | (0.76, 0.84) | Train: 378 | **0.82** | **(0.78, 0.86)** |
| | Test: 195 | 0.80 | (0.74, 0.85) | Test: 186 | 0.76 | (0.70, 0.83) |
| NNs: Plus 11 motivation variables | Train: 821 | 0.71 | (0.68, 0.74) | Train: 809 | 0.73 | (0.70, 0.76) |
| | Test: 423 | 0.72 | (0.68, 0.76) | Test: 408 | 0.67 | (0.62, 0.72) |

The overall accuracy of the HCR-20 and RM2000V scales does not change after introducing each of the three sets of variables into the LR model. The CT training model showed some moderately increased effects for both HCR-20 and RM2000V after including the institutional behaviour variables, and there were some increased effects for the HCR-20 after including the motivation factors. There were some increased effects for the RM2000V after including the dynamic variables. The NNs training and test models all showed some increased effects for the HCR-20 after introducing either institutional behaviour variables or crime motivation factors, and with increased effects for the RM2000V after introducing any of the three sets into the training models. These findings seem to suggest that the LR model is too conservative to detect small effects and that new variables could be added on to existing established risk assessment instruments. Data-mining models, the NNs in particular, are more sensitive and flexible than the LR, and are better able to reflect additional effects (even from new variables which have a relatively limited impact). However, this advantage of the CT or NNs model could also be to their disadvantage due to over-fitting, given small sample for test/validation of dynamic variables, which results in high accuracy in the training sample but poor fitting on test data. The findings should be treated as tentative and further research based on a larger sample is required.

## Model performance with different outcome specifications

The usual, relatively broad definition of the outcome for violence subsumes two categories, that is, violent recidivism is compared to other non-violent recidivism, including non-reoffending. However, the non reoffenders and non-violent reoffenders are arguably different categories. Non-violent reoffenders or general reoffenders share more similarities with the violent group than the non-reoffenders. Most risk assessment instruments predict both general and violent reoffending. Mixing the general reoffenders with non-reoffenders makes the category less homogenous or less specific. It is assumed that such heterogeneity or lack of specification in the outcome category may limit the predictive accuracy of instruments, and hence is reflected in a lower performance of any statistical model. By excluding the general reoffenders from the 'other' category, the performance of risk assessment should improve using any model.

The study demonstrated (see Table 3.9) that, when defined only by non-reoffenders in contrast to violent reoffenders, the new two-category outcome is more homogenous within each category or represents a more clear-cut contrast. All types of model demonstrated a significantly increased predictive accuracy in the HCR-20 from 12% to 17% compared to their performance on the outcome category as more usually defined.

## Table 3.9 Predictive accuracy of HCR-20 for violent outcome of two definitions (male sample)

| Model | Sample | HCR-20 | | | | |
|---|---|---|---|---|---|---|
| | | Sensitivity | Specificity | Accuracy | (95%CI) | Improved acc. of (1) over (2), % |
| LR[a] | Train: 563 | 0.78 | 0.70 | 0.72 | (0.68, 0.76) | 12.5 |
| | Test: 282 | 0.67 | 0.65 | 0.65 | (0.59, 0.71) | 14.0 |
| LR[b] | Train: 827 | 0.77 | 0.62 | 0.64 | (0.61, 0.67) | |
| | Test: 426 | 0.62 | 0.56 | 0.57 | (0.52, 0.62) | |
| CT[a] | Train: 563 | 0.61 | 0.75 | 0.72 | (0.68, 0.76) | 12.5 |
| | Test: 282 | 0.59 | 0.70 | 0.68 | (0.63, 0.73) | 15.3 |
| CT[b] | Train: 827 | 0.75 | 0.63 | 0.64 | (0.61, 0.67) | |
| | Test: 426 | 0.64 | 0.58 | 0.59 | (0.54, 0.64) | |
| NNs[a] | Train: 563 | 0.76 | 0.75 | 0.76 | (0.72, 0.80) | 16.9 |
| | Test: 282 | 0.63 | 0.71 | 0.70 | (0.65, 0.75) | 12.9 |
| NNs[b] | Train: 827 | 0.69 | 0.64 | 0.65 | (0.62, 0.68) | |
| | Test: 426 | 0.65 | 0.61 | 0.62 | (0.57, 0.67) | |

[a] Sample of narrowly defined outcome: violent reconviction versus non-reoffenders (N=845).
[b] Sample of broadly defined outcome: violent reconviction versus non-reoffender/other non-violent reoffender (N=1,253).

As shown in Table 3.10, further examination of effects, using the AUC value, for other instruments including PCL-R, VRAG and RM2000V for the two definitions of outcome categories over a 3.3-year follow-up also demonstrated a significant increase in accuracy due to the more clear-cut or better-differentiated categorisation for all instruments. This was strong evidence to confirm that such a marked increase of effects for all instruments will also be reflected in an improved performance by all models.

## Table 3.10 Predictive accuracy (AUC values) of risk instruments by outcome criterion (Phase II male sample, 3.3 years follow-up in average)

| | Violent reoffender vs. any others N=804 | Violent reoffender vs. non reoffender N=535 | t (p value)[a] |
|---|---|---|---|
| Base rate % | 21.0 | 31.6 | |
| Instrument | AUC (95%CI) | AUC (95%CI) | |
| VRAG | 0.73 (0.69-0.79) | 0.82 (0.78-0.85) | 4.50 (0.000) |
| RM2000V | 0.71 (0.66-0.75) | 0.79 (0.75-0.83) | 3.89 (0.000) |
| PCL-R | 0.67 (0.63-0.71) | 0.73 (0.69-0.78) | 2.73 (0.006) |
| HCR-20 (total) | 0.70 (0.65-0.74) | 0.76 (0.71-0.80) | 2.73 (0.006) |

[a] Hanley & McNeil's method (1983).

This confirmed that the definition of outcome category had a strong impact on the predictive ability of the models. Statistical models can only reveal a pattern if such a pattern truly exists in the data. In practice, accuracy of risk assessment instruments can vary considerably depending on different populations. It is, ideally, important to use an instrument developed **specifically** for assessing a **specific** type of risk and for a **specific** population.

## Prediction with three categories of outcome

We have already noted that the common, broadly-defined outcome did not separate non-offenders from other, non-violent reoffenders. This kind of definition lacked specificity, with a consequent marked impact on the predictive accuracy of the risk assessment instruments as reflected in all models. The obvious option is therefore to consider three categories of outcome in future risk prediction: (i) violent reoffenders; (ii) non-violent reoffenders; and (iii) non-reoffenders. These three groups of individuals should have different risk factors. However, in reality, there is no single risk assessment instrument that has been constructed or designed to predict for these three categories. Using any single existing instrument to predict for individuals in the three categories would yield a low level of accuracy. As anticipated, using the HCR-20 to predict the three categories in one step yielded very poor accuracy for the NNs at around 0.45-0.54 (Table 3.11), and 0.63 for the LR, with only 17.3% accuracy for the most important violent category.

However, as shown in Table 3.11, if an instrument was applied in more than one stage, for example to predict any reoffending compared to non reoffending at stage one, and then, in stage two, violent reoffending and non-violent reoffending among the ‘any reoffenders’ group, it is possible to observe improved accuracy. Using HCR-20 predictors once again, the LR and NNs models fitted in two steps showed an improved performance compared to the one-step approach.

The case for this new classification strategy is strengthened by two practical considerations. Firstly, throughout the modelling experiment, the low base rate for the violent category resulted in unbalanced proportions between the two categories of outcome, and a lack of power for the smaller category of violent recidivism, posed a difficulty in optimising model performance. This was particularly the case for CT and NNs. Although proper handling of the cut-off probability, or specifying prior probability and misclassification costs, or case weighting, did often balance the predictive results in the two categories, this problem still resulted in poor performance of the CT and NN models in many test samples. This was due to unbalanced outcome combined with the small sample size. The new approach involved more balanced categories at each stage, 53.8% of non-reoffenders versus 46.2% of any reoffenders in stage one; and, in stage two, 29.5% of the convicted groups being classified as violent, and 70.5% as ‘other’. Secondly, it appears a natural and logical process of risk screening first to assess which prisoners are most likely to be re-convicted of any crime, and then in the second stage to assess what type of crime they committed. Clearly, the second-stage assessment is conditional on the first-stage outcome. The input of information (or predictors) in the two stages should not be the same, however, due to different outcome categories in the two stages.

### Table 3.11 Percentage of correct classification by outcome category, by model and by classification strategy (HCR-20 predictors, training sample N=827 and testing sample N=426, men)

| Model | Subset | Non- offender | Non- violent offender | Violent offender | All |
|---|---|---|---|---|---|
| **Two stage approach** | | | | | |
| LR | Train | 67.3 | 35.6 | 50.9 | 55.0 |
| wt (1:1: 3.3) | Test | 67.0 | 60.4 | 42.6 | 61.3 |
| NNs | Train | 70.9 | 48.9 | 66.4 | 63.2 |
| wt (1:1:3) | Test | 66.5 | 41.7 | 47.5 | 55.4 |
| NNs | Train | 57.6 | 66.7 | 86.4 | 64.3 |
| wt (1:2:6) | Test | 48.0 | 59.0 | 62.3 | 53.8 |
| **Three categories in one step** | | | | | |
| LR | Train | 79.9 | 47.0 | 10.0 | 60.1 |
| wt (1:1: 3.3) | Test | 75.1 | 43.1 | 6.6 | 54.5 |
| NNs | Train | 72.2 | 18.2 | 63.6 | 53.8 |
| wt (1:1:3) | Test | 67.4 | 11.8 | 55.7 | 46.9 |
| NNs | Train | 41.9 | 53.4 | 47.3 | 46.3 |
| wt (1:2:6) | Test | 44.3 | 46.5 | 44.3 | 45.1 |

Note: Multinomial logistic regression model was used for fitting three categories in one step.

The two-stage approach could not be realised using CT models due to certain technical difficulties. Both LR and NNs models were feasible. When using HCR-20 predictors, compared to the one-stage prediction method, the two-stage approach to the outcome of two categories (violent versus all others), did not markedly improve overall accuracy for either LR (multinomial LR for three categories) and NNs. Further improvement of their performance might be anticipated in the two-stage procedure if the predictors in the first stage were specific to general reoffenders, and those in the second stage specific to violence reoffenders. Conducting such a study, for example to identify predictors or risk assessment instruments for general offending, to be used in stage one and those for violence in stage two, might be an important step in improving the methodology.

# 4. Implications

## Strengths and limitations of different models

- Traditional LR and DA models are the most robust and least subject to over-fitting. This implies that a risk assessment instrument appropriately developed using a reasonably large sample and using LR or DA models can generally maintain predictive accuracy when subsequently applied to an external sample with similar characteristics to the construction sample. Number of predictors and specificity of predictors in the models will not affect their robustness. However, whilst predicting outcome satisfactorily at an average level, the models are less sensitive when predicting extreme cases. Hence they have limited flexibility and efficiency in clinical practice.

- For the CT model different computational algorithms are available which can significantly improve the predictive accuracy when developing or 'training' a model (a process whereby only part of the population is involved). However, this improvement in most cases is at the cost of a much reduced accuracy in the testing sample (involving the remainder of the population). This means that a considerable shrinkage in accuracy will be observed when the developed model, or scale, is applied to any external sample, even with very similar characteristics. This is particularly true when the size of the test sample is small. However, after controlling for such shrinkage, the decision tree models show comparable accuracy when predicting violent recidivism to that of LR or DA.

- The most common NNs (Multi-layer Perceptron) are highly parameterised models with the greatest flexibility to cope with large and complex data. High flexibility also means they are highly subject to over-fitting. To develop or train a model, one can use methods such as increasing the sample size, increasing the number of neutrons (hidden units facilitating mathematic transformation between input variables and the defined outcome category), or manipulating misclassification errors to achieve high accuracy. However, the improvement of accuracy in the training sample is mostly accompanied by poor accuracy or shrinkage in the external testing sample. After controlling for such problems during the model training process, NN models show comparable accuracy for predicting violent recidivism to that of LR or DA.

## Model performance in assessing risk for specific populations

- Both DA and NNs performed significantly better for female prisoners than for males using PCL-R predictors. This suggests that the PCL-R has better validity in predicting female violent recidivism than for men. Further study of predictors specific to violent recidivism among psychopathic female prisoners would facilitate the development of a new instrument for female prisoners.

- Using HCR-20 predictors, no model demonstrated a significant difference in predictive accuracy among prisoners aged less than or equal to 21 and over 21. The same pattern was observed among both men and women. This implies that HCR-20 items are not age-specific in the prediction of violent recidivism among UK prisoners. However, the sample size of the subgroup aged less than or equal to 21 for women was less than 100. With 20 predictors in the HCR-20, the models tested may not have had enough statistical power to show differences. This finding should, therefore, be treated as tentative. When compared to prisoners with no Axis II personality disorder, those with PD (in particular ASPD) were significantly younger at interview, younger at first violent offence, were more likely to have substance abuse disorders and also major mental illness. They were previously convicted of more violent, robbery and acquisitive offences. All the risk assessment instruments tested failed to predict violent recidivism among this population at an acceptable level of accuracy (the results ranged between 0.50 and 0.57). This finding points to a further research area: to develop risk assessment tools for high-risk groups.

- When using HCR-20 predictors to assess men with no PD, all models (NNs in particular) demonstrated considerably improved predictive accuracy compared to the total sample. This finding has two practical implications: new risk assessment tools are needed for high risk groups such as the DSPD population; risk assessments should be carried out for those with and without ASPD entirely separately when using existing assessment instruments. This would improve accuracy among the non-ASPD subgroup if not those with ASPD.

## Model performance with additional variables

- Ten variables describing institutional behaviour, 11 variables describing motivations for the original Index Offence, and nine dynamic or changeable variables were identified which were independently associated with violent recidivism among men at a moderate level of accuracy in the two-year follow-up after release. These three sets of variables were added separately to the HCR-20 predictor list and the RM2000V predictor lists to examine which model demonstrated improved accuracy of the two scales following inclusion of the new variables.

- The LR model did not detect any additional predictive effects from the combination of any set of variables over and above the accuracy already achieved by the HCR-20 or by the RM2000V predictors alone. The LR model was not sufficiently sensitive to reflect minor effects of those variables being added to the existing instruments. Non-specificity and intercorrelation of these variables could have distorted any effects and limited the power of the LR model.

- CT models demonstrated marginally improved accuracy over the original HCR-20 and RM2000V predictors, notably in the training sample, after including institutional, motivation or dynamic variables. However, the improvement was not significant in the testing sample which casts doubt on the future performance of the model when applying it to a new sample.

- NNs demonstrated considerable improvement in accuracy after adding institutional or motivation variables to the HCR-20 in both the training and test samples, and in the RM2000V to the test sample. Including dynamic or changeable variables in the RM2000V significantly improved its predictive accuracy using NNs in both the training and test samples. NNs appeared more efficient and powerful than LR and CT using data which included multiple variables with small effects. However, further research is needed into the methodology of NN models to stabilise and enhance such a feature in order to generalise their use to clinical practice.

## Model performance with different outcome specifications

- Heterogeneity or lack of specification of the outcome category will limit the performance of any statistical model. This study compared the accuracy of all models when predicting violent recidivism as between two definitions: (A) violent reoffenders versus all others (consisting of non reoffenders and other non-violent reoffenders); (B) violent reoffenders versus the narrower group of non-reoffenders. All models demonstrated significantly higher accuracy when predicting outcome B compared to A, using HCR-20 predictors. Most static variables in existing risk instruments predict both violent and general offences. Static variables do not change over time, for example, number of previous offences, gender, age at first conviction. Trying to use these instruments to specifically classify between non-violent reoffenders (mostly general offenders) and violent offenders, as in outcome A, does not work well. Hence the predictors showed reduced power.

- Predicting the outcome for three categories (violent reoffender, non-violent reoffender and non-reoffender) has not often been carried out before and is explored in this study using only HCR-20 predictors. One-stage classification for the three categories produced poor results with both multinomial logistic and NN models. Two-stage classification (any reoffenders versus non-reoffenders in stage one, and other reoffenders versus violent reoffenders among any reoffenders in stage two) yielded a level of accuracy comparable to that of HCR-20 for outcome A of two categories. The two-stage approach can be achieved easily by the LR and NNs models. Further study to develop a new risk instrument specific to the outcomes for three categories should in theory improve predictive accuracy for all models. Identifying risk scales specific to general reoffending and scales specific to violence (versus other non-violence), and then applying them accordingly at different stages of the two-stage approach may be another pathway leading to improved performance of all models.

- Predictive accuracy in risk assessment is affected by multiple elements including predictors, target population, outcome and statistical models. Among those elements, statistical models are the last to be considered as they are all data-driven and can only identify patterns of data shaped by the outcome, predictors and target populations. Improvement of risk assessment could best be achieved by addressing all elements simultaneously if this is possible. Future risk assessment should be aimed at the use of specific tools, with specific predictors, to classify outcomes with homogenous categories on specific target populations. There should be a mechanism in place to screen, or to reduce, false prediction. The model process should produce clear information, useful for risk management and intervention, in clinical practice. This may involve using multiple risk scales to predict multiple categories in multiple stages from different models for different target populations. This study has piloted and discussed the technical as well as practical aspects of this problem.

# 5. Conclusions

This project involved an extensive exploration of the applicability of the more commonly used Neural Networks (NNs) and Classification Tree (CT) models in predicting violent reoffending in a relatively large UK prisoner sample of both men and women. It has collected sufficient empirical evidence to make scientific conclusions on the original aims and objectives of the study and to make recommendations on methods and models for future research, together with clinical practice in this area.

The traditional Logistic Regression or Discriminant Analysis models and the CT or NNs of data mining models are generally comparable in terms of their predictive accuracy, with different strengths and weaknesses.

The traditional models are more robust and controllable. The LR has been used in classifying outcome to two categories, and DA or multinomial LR can classify outcomes of three or more categories. There are well-established and easy to use software packages are readily available. Given predictors, estimation methods, and construct sample, one can always anticipate a model with the same parameter estimates and the same level of accuracy, regardless of time, place, and software. DA models are not subject to over-fitting in the process of developing a model. The predictive accuracy achieved with the construct sample can always be repeated in the cross-validation sample or in the external validation sample, providing that the validation sample is large enough to have sufficient statistical power. Their output parameters are easy to interpret in practice for the purpose of risk management or intervention. However, their efficiency can be distorted or reduced by a large number of intercorrelated variables, instead of making good use of all variables.

Although the CT or CT&RT models are flexible, comparable, and not restricted to large data sets with intercorrelated variables involving small effects, they are nevertheless less plausible in risk assessment practice for certain reasons. They can be manipulated technically to achieve a rather high predictive accuracy when developing a model. However, the common consequence is either poor performance in other external samples or very low accuracy in prediction of the outcome category that is relatively small (violent recidivism in this case). While classifying individuals into different categories (within the output target groups), with group probability determined at each stage, they do not provide sufficient information on the variables for management at an individual level.

NNs are highly parameterised models with the greatest flexibility to reflect complex relationships between inputs and outputs in the data. Their application may be restricted by the following features: firstly, the MLP (Multi-layer Perceptron) NNs explored in this study are sensitive to changes of parameters such as sample size, misclassification error, number of predictors, number of training nets, number of hidden units, and hence subject to over-fitting. This has the consequence of poor performance on an external sample. A change in any parameter often causes a change in model performance in terms of predictive accuracy. Most importantly, these parameters in NNs are interrelated, or dependent on each other. Following a change in one parameter, the NNs will adjust everything else and generate initial weightings both automatically and randomly to start a model fitting process. One may fix as many parameters as one can, but one can never obtain the same model in any two runs. This feature creates difficulty in controlling the model fitting process, hence the term ‘black box’. Secondly, it relies entirely on computerised software for the assessment, with no possibility of a ‘paper and pencil’ approach for clinicians. Thirdly, the output parameters are not the original form of the variables, and less interpretable. However, the NN models performed somewhat better than the LR and CT when there were a large number of variables or in a target population with better homogeneity. They can be considered as a new type of model in risk assessment. Further research in applying the NNs to a larger sample with more variables could accumulate evidence of their applicability in this field.

Predictive accuracy in risk assessment is shown to be affected by multiple elements, predictors, target population, outcome and statistical models. Among these elements, statistical models may be the least important as they are all data-driven and can only identify patterns of data shaped by the outcome, predictors, and target populations all together. Improvement of risk assessment could be effectively realised by addressing all elements at the same time. The following research areas identified by this study may further improve the risk assessment practice.

- Identify specific predictors and develop risk assessment instruments for violent recidivism of female prisoners.
- Develop risk assessment instruments specific to high-risk prisoners, such as those with ASPD or DSPD, for their reoffending behavior, and assess them separately from other prisoners.
- Carry out typology study for the reoffending behaviour of prisoners in order to identify the most specific definition of outcome, for example a potential new definition specific to violence that could effectively maximise differences between categories (violent or not) within the outcome.

It is clear that predictive accuracy in this context is not unlimited. Achieving high accuracy of prediction should not be the sole purpose of risk assessment. A good risk assessment tool should have an acceptable accuracy level to help with clinical or professional decision making as well as being able to provide additional information for risk management and intervention. Focusing on these two aspects, a future risk assessment approach should be aimed at using specific tools with specific predictors to classify outcome using homogenous categories on a specific target population. There should be a mechanism in place to screen or to reduce false prediction, and the model process should produce clear information useful for risk management and also intervention in clinical practice. This may involve using multiple risk scales to predict multiple categories in multiple stages by using a different model in each stage for different target populations. The project team is currently undertaking further investigation and experimentation with this new approach.

# Methodological notes

The LR model has been the conventional statistical technique of choice in the development of new models and in testing existing instruments for predicting dichotomous outcomes. By default, the LR sets cut-off probability at 0.5 for classification of two categories, namely equal misclassification cost for the two categories. An individual is classified within the category with greater probability. This setting always works in favour of the larger outcome category, namely the non-violent and non-reoffender group, and consequently yields a worse prediction for the minority category, the violent reoffenders. In reality, the best cut-off probability should be decided based on consideration of predictive accuracy as well as cost-effectiveness in managing individuals who may or may not go on to reoffend. However, to serve the purpose of this study of model comparison, one needs the predictive accuracy to be balanced by sensitivity and specificity. Therefore, this study changed the classification cut-off probability according to the real violent reoffending rate, to obtain a better balanced accuracy.

The simple CT was developed by Breiman, Friedman, Olshen, & Stone (1984) and is a questions-decision-tree model. A hierarchy of questions is asked and the final decision to split individuals into categorical groups of the outcome is made depending on the answers to all the previous questions. Decision tree methods include two types of classification: Classification Tree with an endpoint involving a qualitative response (yes or no), and Regression Tree with an endpoint involving a quantitative response (regression function). The CHAID programme typically produces CT models, and the C&RT programme produces both CT and RT models. If all predictors are categorical variables, CT method is an appropriate choice. If predictors are mixed with categorical variables and continuous scores, CT&RT is the model of choice. The process of computing Classification Trees involves four basic stages:

i) specifying the criteria for predictive accuracy;

ii) selecting splits;

iii) determining when to stop splitting;

iv) selecting the ‘right-sized’ tree (pruning).

There are many types of decision tree models such as Classification and Regression Tree model, and Chi-square Automatic Interaction Detector. This study mainly applied the simplest CT. Sometimes the C&RT model was used instead when the CT model performed poorly in some analyses.

There are many types of NNs. The most commonly used is the Multi-layer Perceptron (Rumelhart & McClelland, 1986). The NNs have a feed-forward structure: inputs layer, hidden layer(s), output layer, and a **bias** term which bears the weights as the adjustable parameters of the model. The input layer is determined by the number of predictors and the output layer by the number of outcome categories. The hidden layer(s) accommodate complex mathematical operations or 'brain process of thinking'. Such a structure has stable behaviour and fault tolerance and it can model functions of almost arbitrary complexity, with the number of hidden layers and the number of hidden units in each layer determining the complexity of function. NNs analysis consists of training and testing processes. NNs prediction is made by learning from training data. Algorithmically, 'training' is carried out using the following sequence of steps:

i) present the network with an input-target set;

ii) compute the predictions of the network for the targets;

iii) use the error function to calculate the difference between the predictions (output) of the network and the target values;

iv) use the training algorithm to adjust the weights of the networks; and

v) repeat steps i to iv again for a number of training cycles or iterations until the network starts producing sufficiently accurate outputs.

Steps i to iv form one training cycle which is called one iteration (E-manual of STATISTICA8.0).

Theoretically, several technical aspects can influence training performance of MLP NNs: sample size for training; choice of weighting and misclassification functions; and number of hidden units in the hidden layer. This study has explored these aspects. More detailed findings can be found in the full report of the project.

# References

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

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

Caulkins, J., Cohen, J., Gorr, W. & Wei, J. (1996). 'Predicting criminal recidivism: A comparison of neural network models with statistical methods.' *Journal of Criminal Justice*, *24*, pp 227-240.

Coid, J., Yang, M. & Ullrich, S., *et al.* (2007). 'Predicting and understanding risk of reoffending: the Prisoner Cohort Study.' *Research Summary*, Ministry of Justice, 6.

DTREG, *Software for Predictive Modeling and Forecasting*, http://www.dtreg.com.

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

Grann, M. & Langstrom, N. (2007). 'Actuarial Assessment of Violence Risk: To Weigh or Not to Weigh?' *Criminal Justice and Behavior*, *34*, pp 22-36.

Hanley, J. A. and McNeil, B. J. (1983). 'A method of comparing the areas under receiver operating characteristic curves derived from the same cases.' *Radiology*, *148*, pp 839-843.

Hosmer, D. W. and Lemeshow S. (1989). *Applied Logistic Regression*. New York: John Wiley & Sons.

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

Liu, Y.Y., Yang, M., Ramsay, M., Li, X. S. and Coid, J. (2009). 'A comparison of logistic regression, classification and regression tree, and neural networks models in predicting violent offending.' *Criminal Justice and Behavior* (submitted).

Monahan, J., Steadman, H. J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., Silver, E., Roth, L. H. & Grisso, T. (2000). 'Developing a clinically useful actuarial tool for assessing violence risk.' *British Journal of Psychiatry, 176,* pp 312-320.

Monahan,J., Steadman, H. J., Robbins, P.C., Appelbaum, P., Banks, S., Grisso,T., Heilbrun, K., Mulvey, E. P., Roth, L. & Silver, E. (2005). ‘An Actuarial Model of Violence Risk Assessment for Persons with Mental Disorders.’ *Psychiatric Service, 56*,pp 810-815.

Monahan, J., Steadman, H. J., Appelbaum, P. S., Grisso, T., Mulvey, E. P., Roth, L. H., Robbins, P. C., Banks, S., & Silver, E. (2006). The Classification of Violence risk. *Behavioral Sciences and the Law, 24*, 721–730.

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

Price, R. K., Spitznagel, E. L., Downey, T. J., Meyer, D. J., Risk, N. K. & El-Ghazzawy, O. G. (2000). ‘Applying artificial neural network models to clinical decision making.’ *Psychological Assessment*, *12*(1)*,* pp 40-51.

Rosenfeld, B. & Lewis, C. (2005). ‘Assessing violence risk in stalking cases: A regression tree approach.’ *Law and Human Behavior*, *29*, pp 343-357.

Rumelhart, D. E. & McClelland, J. (eds.) (1986). *Parallel Distributed Processing.* Vol 1*.* Cambridge, MA: MIT Press.

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

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

Stalans, L. J., Yarnold, P. R., Seng, M., Olson, D. E. & Repp, M. (2004). Identifying Three Types of Violent Offenders and Predicting Violent Recidivism While on Probation: A Classification Tree Analysis. *Law and Human Behavior*, 28(3), pp 253-271.

Starzomska, M. (2003). ‘Use of artificial neural networks in clinical psychology and psychiatry.’ *Psychiatria Polska*, *37*(2), pp 349-357.

Steadman, H., & Monahan, J. (1994). Toward a rejuvenation of risk assessment research. In J. Monahan & H. Steadman (Eds.), *Violence and Mental Disorder* (pp. 10−16). Chicago: University of Chicago Press.

Steadman, H. J., Silver, E., Monahan, J., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., Grisso, T., Roth, L. H. & Banks, S. (2000). ‘A Classification Tree Approach to the Development of Actuarial Violence Risk Assessment Tools.’*Law and Human Behavior*, *24*, pp 83-100.

Thomas, S., Leesea, M., Walsh, E., McCrone, P., Moran, P., Burns, T., Creed, F., Tyrer, P. & Fahy, T. (2005). 'A comparison of statistical models in predicting violence in psychotic illness.' *Comprehensive Psychiatry*, *46*, pp 296–303.

**Ministry of Justice Research Series 6/10**

**Applying Neural Networks and other statistical models to the classification of serious offenders and the prediction of recidivism**

This methodological study aims primarily to explore the applicability of data mining techniques, including Neural Networks (NNs) and Classification Tree (CT) models, in predicting the risk of violent recidivism, together with their predictive validity in comparison with conventional logistic regression (LR). Its ultimate aim is to lead to the development of better risk assessment. The study sample comprised 1,353 male prisoners and 304 female prisoners. The outcome of violent reconviction was followed up prospectively through the Police National Computer database (PNC) for a mean period of 1.98 years for men and 2.08 years for women, after their release from prison. Individual items in the Historical Clinical Risk – 20 (HCR-20), Psychopathy Check List Revised (PCL-R), Violence Risk Appraisal Guide (VRAG) and Risk Matrix 2000 for Violence (RM2000V) were used separately as predictors, and the predictive accuracy of the most commonly used Multi-layer Perceptron (MLP) of NNs and CT models was tested and compared. Since the performance of predictive models is also determined by the specificity of predictors, by well-defined criteria for any outcome category and by the homogeneity of the targeted population, in addition to the choice of statistical models, the project further examined the efficacy of all models in terms of different sets of variables such as institutional behaviours or community factors, different definitions of violent recidivism, and different populations such as women, young adults and prisoners with any personality disorder (PD). Most experiments showed similar predictive accuracy for all models, with significantly improved accuracy for all models when predicting violence in prisoners without any PD or the narrowed outcome category of violent recidivism versus non-recidivism only. Both the NNs and CT models were shown to be more prone to over-fitting or poor performance than the LR model, when the sample size was small. Finally the implications of the study findings are addressed and further research areas are recommended.

ISBN 978 1 84099 383 7

Produced by the Ministry of Justice

Alternative format versions of this report are available on request.

E-mail: research@justice.gsi.gov.uk

http://www.justice.gov.uk/publications/research.htm