BREAKING THE ‘GLASS CEILING’ OF RISK PREDICTION IN RECIDIVISM: AN APPLICATION OF CONNECTIONIST MODELLING TO OFFENDER DATA

Dominic A.S. Pearson

B.A., MSc., C.Psychol.

Thesis submitted for the qualification of

Doctor of Philosophy (Ph.D)

University of York

Department of Psychology

December 2011

Abstract

The present thesis explored the capability of connectionist models to break through the ‘glass ceiling’ of accuracy currently in operation in recidivism prediction (e.g., Yang, Wong, & Coid, 2010). Regardless of the inclusion of dynamic items, all risk measures rarely exceed .75 in terms of the area under the receiver operating characteristic curve (AUC) (Hanley & McNeil, 1982). This may reflect the emphasis of multiple regression equations on main effects of a few key variables tapping long-term anti-social potential. Connectionist models, not used in criminal justice, represent a promising alternative means of combining predictors given their ability to model interactions automatically. To promote learning from other fields a systematic review of the literature on the application of connectionist models to operational data is presented. Lessons were then taken forward in the development of a connectionist model suitable for the present data which comprised fields from the Offender Assessment System (OASys) (Home Office, 2002) relating to 4,048 offenders subject to probation supervision. Included in the items for modelling was the Offender Group Reconviction Scale (OGRS) (Copas & Marshall, 1998; Taylor, 1999). Combining static and dynamic items using conventional statistical methods showed a maximum cross-validated AUC of .82. Using the connectionist model however a substantial increase in accuracy was observed, AUC=.98, and this largely maintained when variations in time to recidivism were controlled. Variation to model parameters suggested that performance linked to the resources in the middle layer, responsible for modelling rare patterns and interactions between items. Model pruning confirmed that while the connectionist model exploited a wide range of variables in its classification decisions, the linear model was affected mainly by OGRS and a limited number of other variables. Results are discussed in terms of the theoretical and practical benefits of developing the use of connectionist models for better incorporating individuals’ dynamic risk and protective factors in recidivism assessments, and reducing the costs associated with false classifications.

List of Tables

List of Figures

List of Appendices

Acknowledgements

As this thesis was completed part-time, the support required has been extensive and extended. The project was originally commissioned by the board of County Durham probation area in 2005 following its interest in developing research and innovation. I put the case for the research with the support of Professor Cynthia McDougall and I would therefore like to thank Cynthia and the leadership of County Durham probation (now Durham Tees Valley probation trust) for promoting and backing the research idea. Thanks are due to the senior management of the probation trust, led by Russell Bruce, for providing the research time and computer resources.

I would particularly like to thank all members of my research committee, Dr Tom Hartley, Professor Cynthia McDougall, and Professor Padraic Monahan. Due to various changes in employment each of these experts has served time as my primary supervisor and it has been a pleasure and a privilege to work with them. Although I made many journeys over the A66 and down the A1 I am indebted to SKYPE for providing a means of supervision without the road or rail costs. Each supervisor contributed to my learning in a unique way and I would like to thank Padraic for his methodological and technical expertise, Cynthia for her sage practical advice, and Tom for his ideas and enthusiasm.

I am also grateful to Durham probation's information and technology team for their assistance in the extraction and preparation of the data. In supporting conflict between my operational work requirements and thesis completion I would like to thank Leila Sedgewick, Cynthia McDougall, and Hazel Willoughby. In providing high quality proof-reading services I would like to thank Robert McDougall.

Last but not least I would like to thank my family and a few friends for their constant moral support. These people believe in me when I have moments of doubt. My parents have taught by example the importance of remaining positive and working hard one day at a time. I hope I can and do re-pay their love and encouragement.

Declaration

This thesis comprises the candidate's own original work and has not, either in the same or different form, been submitted to this or any other University for a degree. The candidate designed all experiments, and was responsible for all testing and analyses.

Although pilot work was presented at the conference referenced below, subsequent substantive work has been withheld from publication prior to its examination as a Ph.D due to the advent of a competitive environment within the UK National Offender Management Service.

Pearson, D. (2007, July). Exploring the potential value of connectionist modeling in the prediction of further offending of sex offenders. In J. Wood (Chair), *Establishing the role of forensic psychology and economics with multi-agency public protection cases and prolific offenders*. Symposium conducted at the Division of Forensic Psychology Conference, York, UK.

CHAPTER 1

1. Overview

1.1 Background

Accurate prediction of offender recidivism, where a subject commits a further offence within a short period of sanctioning for the original offending behaviour, is at the heart of an effective and efficient criminal justice system. Current prediction measures can be divided according to the use of structured clinical judgement or statistical weighting to combine risk factors into a prediction. Despite the complexity of the task, involving the assessment and integration of a range of risk factors, reviews of the leading measures have indicated that statistical methods give no improvement in performance with all measures demonstrating equivalent 'moderate' accuracy (Campbell, French, & Gendreau, 2009; Coid et al., 2009; Kroner & Mills, 2001; Yang, Wong, & Coid, 2010). These reviews suggest that statistical measures that include changeable dynamic factors perform at the same level or below those that are limited to static factors, thus limiting criminal justice agencies' ability to link interventions to reductions in risk. Lack of benefit by incorporation of dynamic factors may reflect the reliance of existing risk measures on multiple regression equations, which do not respond well to data that are inter-correlated or that contain measurement error (Gottfredson & Moriarty, 2006). This thesis therefore presents an application of a different method, connectionist modelling, which has only rarely been applied in criminology and with inconsistent results. Connectionist modelling is a pattern recognition technology robust to non normal and noisy data, and capable of automatically discovering interactions between variables (Bishop, 1995). As a result of these properties, connectionist models have been used extensively in other fields.

1.2 Purpose and Scope of the Thesis

This thesis aims to implement a connectionist model and test its ability to predict offender recidivism. The study will employ operational data gathered routinely by one probation area of the UK National Offender Management Service (NOMS), using the dominant risk assessment framework. The sample cases for the study were at the start of their community risk, and subject to probation supervision either following the imposition of a court order or after release from prison on licence. To benchmark the performance of the connectionist model results from modelling the same data using conventional statistical models will also be considered. It is proposed that connectionist models may offer an advantage due to their ability to model interactions despite noise on the predictor and criterion variables, as characteristic of risk factors and outcomes in criminal justice (Gottfredson & Moriarty, 2006).

1.3 Outline of the Thesis Structure

The thesis will start with a review of the criminological literature relating to the prediction of offender recidivism, including the key risk factors, and the predictive validity of the existing methods of combining these into a risk estimate (Chapter 2). This will then be followed in Chapter 3 by an introduction to connectionist modelling and a systematic review of the literature in which these models have been applied to relevant problems. The precise methods used in the present study will be detailed in Chapter 4 and then in Chapter 5 they are piloted using a sub-sample of the data. The research results relating to the total data sample will first consider the general cross-validated accuracy of the models (Chapter 6), before examining the ability to classify these offenders according to their time to recidivism (Chapter 7). Finally the origin of the models' performances will be studied by examining the impact of pruning their constituent variables (Chapter 8). These results will be discussed as to their theoretical implications for our understanding of risk prediction methodologies, and practical implications for criminal justice agencies concerned in particular with managing and reducing risk of further offending by a broad range of offenders.

CHAPTER 2

2. Review of Recidivism Risk Assessment

2.1 Aims and Objectives

This chapter sets out to review those factors and concerns that are of key importance in developing predictive accuracy in the assessment of an offender's risk of recidivism. The current study uses the variables captured by probation officers using the Offender Assessment System (OASys: Home Office, 2002) and therefore the review material in this chapter will focus on related variables and risk measures accessible to a criminal justice agency. The chapter aims to introduce the features, including variables and data characteristics, to which a new model should respond in order to enhance its accuracy.

To achieve this the present chapter includes a review of the research literature regarding those variables required to be included in an assessment of criminal risk (e.g., using OASys), and also on the predictive validity of existing risk instruments. The present state of the literature will inform an assessment of the current situation in offender risk assessment. Advances in forensic risk assessment methodology that have been attempted will be considered to identify any learning points that may also inform development of a new model.

2.2 Background to Risk Assessment

2.2.1 Predictive accuracy issues.

This section will briefly review the range of issues on which the utility of a prediction depends. As such, technical and methodological aspects will be covered with the intention that this will inform the discussion throughout the remainder of the chapter.

A predictive decision requires, as a minimum for each case, information about a predictor which is associated with an outcome criterion of interest. The relationship between the predictor and the criterion is illustrated with the classic decision outcome matrix, shown in Figure 2.1 below.

	Actual Yes	Actual No
Predict Yes	True Positive :a (“hit”)	b: False Positive (“false alarm”)
Predict No	False Negative :c (“miss”)	d: True Negative (“correct rejection”)

Figure 2.1. Risk prediction terminology

A good predictor will produce a high true positive figure, i.e., the presence (or absence) of the predictor will correctly predict the outcome in most cases. This hypothetical predictor will also produce a high number of true negative decisions: the absence (or presence) of the predictor for the true positives. Whether the predictor is present in cases with or without criterion events reflects whether it is a risk factor or a protective factor. On the other hand, a false positive is found when the risk factor arises but the criterion is not present, and therefore the predictor incorrectly raised the alarm. A false negative is when the criterion outcome occurs in the absence of the risk factor, as it can be said to have ‘missed’ the correct classification. A desirable predictor is one that maximises the number of True outcomes (a and d) and minimises the number of False outcomes (b and c) in Figure 2.1.

Although false positives and false negatives are equally undesirable from a scientific viewpoint, in forensic settings they do not have equal social consequences. A high false positive rate means that many individuals who were deemed to be at risk may be detained when they were in fact safe for release. This implicates issues concerning public finances and civil liberties. A high false negative rate is of concern from a public safety standpoint since the use of the predictor leads to the release of unsafe individuals. Priority attached to reducing either error therefore reflects social value judgements about whose interests are best served. To respond to this the selection ratio, or the proportion selected as positive by the predictor, can be changed to alter the false positive or false negative rate. For example, to reduce the number of false positives the cut-off point can be raised; this has the inevitable consequence however of also reducing the rate of true positives.

Prediction studies therefore seek to identify from a set of predictor variables those that maximise the accuracy of predicting the criterion and minimise false positive and false negative outcomes. Accuracy, however, depends on reliability both of the predictors and of the criterion. Predictors such as ‘age at first conviction’ may be unreliable due to poor quality records. Similarly the reliability of the criterion may be questionable. In criminological research, reliance on official arrest or reconviction data may underestimate the actual rates of recidivism (e.g., Hall, 1987; Lloyd, Mair, & Hough, 1994). In addition, as discussed by Blackburn (1993), the presence or absence of the criterion may reflect chance or situational factors. Since a prediction is based on the presence or absence of a disposition or tendency, the conversion of this into an offensive act may reflect factors in the individual’s environment including any external controls or inhibitory treatments. Researchers have observed however, that self-reported and official offending both identify the same individuals as most serious (Farrington, 1995). This may indicate that environmental factors are secondary to individual determinants of risk.

A further key factor in the utility of a predictor is the frequency of the outcome criterion in the population of interest, or the base-rate. Further to the correlation between the predictor and the criterion, the base-rate determines the extent to which the predictor can correctly classify cases beyond chance. Chance refers to the performance if all cases were assigned to the dominant outcome class, and this implicates the base-rate as illustrated in Table 2.1 below.

Table 2.1

Performance of Measure with 80% Accuracy on Two Populations with Different Base-Rates (based on Blackburn, 1993)

Population	Base Rate	TP	FP	TN	FN
A	50 %	80% of 50	20% of 50	80% of 50	20% of 50
(n=100)	(n=50)	(n=40)	(n=10)	(n=40)	(n=10)
B	10 %	80% of 10	20% of 90	80% of 90	20% of 10
(n=100)	(n=10)	(n=8)	(n=18)	(n=72)	(n=2)

Note. TP = True Positive; FP = False Positive; TN = True Negative; FN = False Negative

Blackburn (1993) shows how in population A with a base-rate of 50%, a predictor that is accurate with 80% of recidivists and non recidivists is relatively efficient in that it correctly identifies 30% more of the population than would be identified by chance alone, i.e., the base-rate of 50%. If the base rate in a population B was 10% however, the overall accuracy is the same (80%) but this now is beneath the chance level, i.e., that achieved by predicting all the sample would be non recidivists (90%). The lower base-rate has led to a higher number of false positives thus reducing the positive predictive power. For a predictor to be efficient therefore, the effect of a low or a high base-rate is to require a very strong association ($r > .5$) with the criterion, something that is rare among the variables studied in the criminological literature (see below). Since for most criminal behaviour the base-rates fall in the 20 to 80% range, it is not acceptable to adopt a blanket strategy of classifying all cases according to the most frequent outcome, and therefore predictive accuracy is critically important to develop.

2.2.2 Evaluating predictive accuracy.

Evaluations of predictive accuracy can be done in a number of ways, the most traditional of which is to compare studies using a standardised index of performance. Due to the problems in differences in base-rate between studies, it is problematic to compare the predictive accuracy of two risk measures when both have been used on subjects released under different base-rate conditions. To avoid such biases associated with certain outcomes it has become widely accepted (e.g., Mossman, 1994; Rice & Harris, 1995), that a good measure of predictive accuracy relating risk scores to an outcome is the receiver operating characteristic (ROC) curve (Hanley & McNeil, 1982; Swets, 1988; Swets, Dawes, & Monahan, 2000). The ROC curve serves as a measure of detection that is unaffected by the decision threshold at which the receiver (e.g., criminal justice authority) operates. As shown in Figure 2.2 below, the ROC curve plots the probability of ‘hits’ or true positives against the probability of ‘false alarms’ or false positives. It gives a pictorial description of performance across all possible decision thresholds or cut-points on the risk measure.

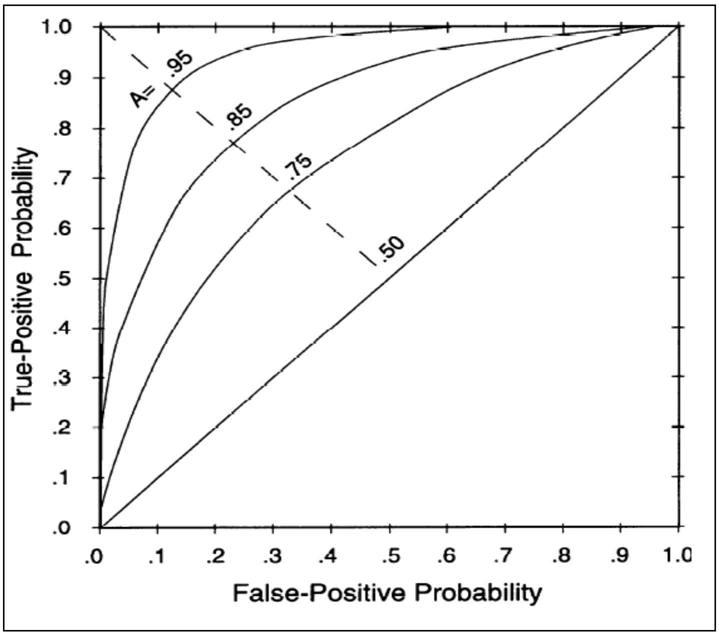


Figure 2.2. ROC curves for four levels of accuracy, each labelled by its area measure, A (taken from Swets et al., 2000).

Figure 2.2 shows ROC curves for four hypothetical risk measures. For all four, both probabilities are near 1 in the upper right portion of the figure as they would be for a very lenient decision threshold, under which the classifier almost always makes a positive decision. This strategy identifies many true positives but at a cost of misclassifying many of the negative cases (e.g., non re-offenders). Similarly for all measures the lower left portion of the figure shows a very strict decision threshold in which both probabilities are near 0 because the classifier rarely makes a positive prediction. The shape of the curve reflects the balance of errors and correct judgements within each measure's accuracy. The accuracy is greater when the curve is higher because the probability of a true positive is higher for each false positive probability. For example on the diagonal line the area measure (A) equals .50, or chance performance, because the true positive probability is no higher than the false positive probability at any point on the line. In the lowest curve, where A=.75, a false positive probability of .30 is associated with a true positive rate of approximately .65. This is better than chance but it is inferior to the other curves at the same .30 tendency to false positives.

The area measure A in Figure 2.2, known as the area under the ROC curve (AUC), therefore quantifies the extent to which a risk measure's increased sensitivity is achieved at a cost of increased false positives. The AUC is equal to the likelihood that a measure will rank a randomly chosen actual re-offender higher than a randomly chosen actual non re-offender (Hanley & McNeil, 1982). It has been proposed that AUC values of below .6 represent low accuracy, and those above .9 represent high accuracy, while those in the intervals of .7, .8, .9, represent marginal, modest, and moderate accuracy, respectively (Sjostedt & Grann, 2002). As mentioned the AUC is unaffected by changes in sample size or row and column totals (e.g., prevalence or base-rate).

A reviewer could therefore survey those primary studies using a set of instruments, either within the same study or across different studies, and compare their AUC values. Primary studies are however nevertheless difficult to compare due to differences in study design, e.g., prospective versus retrospective, differences in length of follow-up, index of recidivism employed. Sample characteristics can also vary widely by age, gender, size of sample, whether they are prisoners or forensic patients, and country or jurisdiction.

For the purpose of comparison, data from different studies are therefore often pooled using 'meta-analysis' to produce a standardised measure of effect size (Rosenthal, 1984). Meta-analysis involves the systematic empirical derivation of a common quantitative estimate of the degree of association between two variables. It has an advantage over the traditional narrative review due to its ability to quantify the size of the effect across studies (Glass, 1976). Many meta-analyses use 'fixed effects models' in which study design features are free to influence effect sizes. Effect sizes are unlikely to be the same across different study designs (e.g., observational versus randomised experimental designs). Some reviews have therefore used the '*Q* statistic' (Rosenthal, 1991) which is a measure of study heterogeneity giving an estimate of the extent to which the variation between studies is greater than would be expected by chance. In addition many reviews weight the effect size by the size of the sample, this being a known mediator of effect size differences. These issues have to be borne in mind when considering patterns of findings on predictive accuracy.

To respond to such variation between primary studies, some researchers (e.g., Farrington, Joliffe, & Johnstone, 2008; Yang, Wong, et al., 2010) have used 'random effects' models which give more equal weight to all studies and do not allow study design features such as sample size to influence the effect size estimate. These models allow a stratified comparison of weighted effect size, based on adjustment for impact of study features on the differences in effect size hypothesised to be associated with each instrument. A couple of limitations are associated with this: not all moderators can be included in the stratification and consequently it is not possible to know the interacting effect of other moderators. Second, this approach requires sub-sample data which necessarily invokes less statistical power to detect differences in predictive accuracy. These drawbacks could obscure differences between two instruments if such differences are only moderate, due to unreliable point estimates (wide or overlapping confidence intervals).

As a result of these difficulties it is important either to examine those studies in which more than one measure has been applied to the same population, under a single study design, or to refer to meta-analytic findings with consideration of a range of models and assessing the validity of those models in the face of study differences. Primary studies are always liable to issues of propriety, and are in need of replication. Given that such replication in this area is not commonplace, and that it is difficult to compare these studies with others involving different designs (as described above), where possible strong references will be made to the results that have emerged from meta-analytic reviews.

2.3 Measurement Requirements: Variables to Include

The purpose of this section of the review is to identify those variables that are important to include by a criminal justice agency in assessment and management of an adult offender's risk. In this context the selection of variables is constrained by what is accessible to probation workers. Thus, for example, early childhood risk factors were not included. Item selection also depends on the offending outcome selected for the prediction task. Individuals with specific offence histories may have different risk factors than those whose offending patterns are more general. Consequently, the evidence for factors important to general recidivism is also briefly discussed as to the extent of differences in the prediction of serious violent recidivism (section 2.3.3).

A screening of the literature relating to recidivism risk assessment revealed twelve consistently recurring areas: age, gender, criminal history, ethnicity, peers, substance misuse, learning ability, employment, motivation to change, personality disorder, mental health, and community factors. For the organisation of this section, these twelve areas were further sub-divided into five broad domains, namely: demographic, historical, dispositional, clinical, and situational factors.

The selection of variables for assessment may also depend on the time period over which the prediction is expected to be accurate. Unless stated otherwise, the following sections consider prediction over a two-year time-frame since this is the most frequent follow-up period in criminological research; furthermore it is an important time window for service delivery on court orders and post-release licences.

2.3.1 Static risk factors.

Static factors are aspects of the offender's past that are predictive of recidivism but cannot be changed. These include demographic features such as the offender's age, gender or ethnicity, but also historical events that might indicate risk such as the existence of previous convictions.

2.3.1.1 *Demographic characteristics*.

The relationship between age and crime, in which reductions in individual recidivism are associated with increases in age post adolescence, is perhaps one of the most robust findings in the field of criminology. Aggregate trends in crime suggest that crime rates peak before age 20 and then show a rapid decline that continues throughout the adult years (Blumstein, Cohen, & Farrington, 1988; Farrington, 1992; Hirschi & Gottfredson, 1983; Laub & Sampson, 2003). The average criminal career among repeat offenders may be around 10 years (Farrington, Lambert, & West, 1998). These aggregate rates do not necessarily describe individual crime rates, as they may mask wide variance in termination of careers (Piquero, Brame, & Lynam, 2004). Current age is therefore related to participation in crime but not to the frequency of offending among active offenders. Thus some older offenders still participating in offending do not show a concomitant reduction in frequency with age (Smith, Visher, & Jarjoura, 1991). Nevertheless, a general 'ageing out' of crime appears to hold for males and females, all main crime types, most Western nations, and across different centuries (Hirschi & Gottfredson, 1983). The persistence in offending among a sizeable minority may reflect a stable disposition of criminality (Gottfredson & Hirschi, 1990), which may or may not be mediated by transitions of 'state' e.g., via employment or positive relationships (Sampson & Laub, 1995). In any case the extent of variation makes it difficult to predict, at any particular age, which offender is headed for a short career and which offender is headed for a lengthy and/or prolific career (Gottfredson & Hirschi, 1990). Although there has been strong support for the proposal that, save for a chronic sub-group of offenders, most young offenders will not persist into adulthood (Moffitt, 1993; Piquero, 2008), research has also shown the existence of sub-groups including 'late-comers' sharing many of the characteristics of early onset persisters, and 'recoveries' or previously chronic offenders that unexpectedly desist from offending (Laub & Sampson, 2003). Studies of recidivism in the UK have confirmed the general premise that the likelihood of re-offending decreases with increasing age after age 20-25 (e.g., Bowles & Florackis, 2007; Lloyd et al., 1994; May, Sharma, & Stewart, 2008), however prediction for individuals clearly also hinges on other factors.

Gender differences are also highly apparent in recorded crime, with recent UK statistics for one year indicating that the number of men found guilty at court was more than three times greater than for women (Ministry of Justice, 2010b). Despite these initial gender differences in offending, which may reflect reporting or justice system processing bias (Blackburn, 1993), the recidivism rate for female offenders (54%) is not that different to that for males (58%) (Bowles & Florackis, 2007). A meta-analysis by Simourd and Andrews (1994) set out to determine whether there were differences between the sexes in factors predictive of recidivism. No differences were found, although the analyses did not include criminal history variables. These may well be important; in Bowles and Florackis (2007) which used a sample of 34,126, male offenders had a higher hazard of reconviction overall, although this difference attenuated in older groups of males and females and with additional previous convictions. Thus as the number of previous convictions in male and female offenders increased they became more similar in terms of their reconviction hazards. These results suggested that the gender differences may be moderated by the same factors increasing risk among males e.g., early criminal history. However, some authors maintain that women's risk factors are gender-specific, with women experiencing greater levels of victimisation, economic marginalisation, parenting obligations, and substance abuse (Reisig, Holtfreter, & Morash, 2006). Reisig et al. have therefore challenged the appropriateness of gender-neutral risk assessment instruments. This was refuted in a study of female offenders by Rettinger and

Andrews (2010) using a general risk/needs scale (the Level of Service / Case Management Inventory [LS/CMI]: Andrews, Bonta, & Wormith, 2004). None of the main effects or interactions involving the demographic variables of age, poverty, and ethnicity was statistically significant after general risk/need was considered. However the mean score for the LS/CMI did vary directly and positively with financial problems suggesting that lower risk women with financial problems may be under-classified by the measure. The overall evidence currently, including a recent meta-analysis, suggests that recidivism risks are not gender specific where the offending behaviour is supported by a delinquent history and other risks reviewed below (Collins, 2010; Makarios, Steiner, & Travis, 2010). Where these dysfunctional supports are absent, specific problems regarding finances and personal misfortune may be crime promoting in females. Existing risk measures seem unable to identify rare interactions such as that proposed between low-risk females and financial management problems.

A third demographic variable is offender 'ethnicity'. Statistics on race and the criminal justice system in England and Wales for 2009 showed that non-white offenders were not disproportionately represented in prison (27%) or probation (18%) compared to the general UK population (25%) (Ministry of Justice, 2010a). The proportion of offenders in each ethnic group starting court orders in the community has remained relatively stable since 2005, with the greatest difference of 0.5% being associated with the category 'mixed' ethnicity (Ministry of Justice, 2010b). The extent to which ethnicity predicts recidivism in the literature has been unclear due to the inter-relations between minority ethnicity and disadvantaged families, neighbourhoods, and economic opportunities. Indeed racial differences in offending have been shown to disappear after controlling for family social status (Cottle, Lee, & Heilbrun, 2001; Ouston, 1984). Family social status, or 'social class of origin' has also however been found to be a very weak predictor of juvenile delinquency (Simourd & Andrews, 1994; Tittle & Meier, 1990, 1991). The issue was examined in a meta-analysis of the predictors of adult offender recidivism by Gendreau, Little, and Goggin (1996) in which non-white ethnicity was a significant predictor of recidivism. Together with age and gender, ethnicity produced higher correlations with recidivism than did family factors or social class of origin. Gendreau et al.'s (1996) results were not reported by country however. It is possible that there might be an interaction between minority ethnicity and neighbourhood disadvantage, given that these conditions may be more associated with minority ethnicity. This was investigated by Wehrman (2010) using a sample of 1,917 randomly selected probation/parole cases from one county in Michigan, USA. Wehrman found that ethnicity was a statistically significant predictor of recidivism and this was unaffected not only by the interaction term for neighbourhood disadvantage, but also by a range of controls including education, age, gender, and prior convictions. Since most of the communities in the study sample were disadvantaged this limits certainty in the conclusions, however.

When data are analysed using ‘split population’ models, which control both for the probability of recidivism and its timing, as done by Schmidt and Witte (1989) ethnicity and gender are shown to affect the probability of recidivism but not its timing. This may fit with the theoretical position of Andrews and Bonta (2006) which holds that age, gender, ethnicity and social class of origin contribute indirectly to criminal behaviour and have only minimal effects after dynamic cognitive social learning variables are considered. Thus the importance of ethnicity may relate to its dependence on a range of dynamic variables, which would explain its repeated association with recidivism (Gendreau et al., 1996; Huebner & Berg, 2011; Minor, Wells, & Sims, 2003; Schmidt & Witte, 1989; Wehrman, 2010).

2.3.1.2 *Historical factors*.

The importance of current age reflected a propensity to outgrow crime among a large proportion of offenders. Longitudinal studies have shown that the offenders least likely to outgrow crime are those that showed signs of criminal activity from an early age (Blumstein, Farrington, & Moitra, 1985; Farrington & Hawkins, 1991; Francis, Soothill, & Piquero, 2007; Loeber, 1982). Blumstein et al. found that the most chronic offenders at the age of 25 had all attracted their first criminal conviction by the age of 15, and an early age at first conviction was the best predictor of which offenders would eventually become chronic. These authors also reported a positive relationship between the seriousness of the first conviction and the number of subsequent convictions (Blumstein et al., 1985). Francis et al. (2007) sought to examine whether the information collected at the first court conviction was associated with the length of criminal career. On the basis of six different birth cohorts, spanning a 25 year period and 58,407 offenders, they showed that the chances of desisting from offending were highest after the first conviction and if the offender did not then desist, the hazard remained constant for the remaining 20-25 year period. Among covariates for birth cohort, gender, offence type, disposal type (community/custody), presence of co-convictions, and age at first conviction, the latter was the most significant in the model. The probability of desisting from offending was 50% greater in those starting their criminal career aged 15-17 compared to those aged younger than 15, and those starting as an adult had twice the chance of stopping after the first conviction. The significance of this covariate changed little in separate analyses of each cohort, suggesting that the importance of age of onset in predicting persistence is not generation-specific.

The number of prior offences, or extent of 'criminal history', has also emerged as an important predictor of recidivism (e.g., Barnett, Blumstein, & Farrington, 1987; Bonta, Law, & Hanson, 1998; Gendreau et al., 1996; Huebner & Berg, 2011). In Gendreau et al., which reported mean *r* values for family, intellectual, demographic, and social achievement predictors of recidivism, adult criminal history ranked highest of any individual risk factor variable. Criminal history, which included pre-adult anti-social behaviour, more often had higher correlations with recidivism than all other factors with the exception of 'criminogenic needs' (e.g., pro-criminal attitudes). Criminal history variables were also the best predictors of recidivism in the meta-analysis by Bonta et al. (1998) focussing on mentally disordered offenders. The salience of criminal history holds also in studies of recidivism among non-disordered juvenile offenders, despite the scope for differences being limited by subjects' young age (Benda & Tollett, 1999; Cottle et al., 2001). These findings are however consistent with a social psychological explanation of crime in which behavioural habits and reinforcement histories of offenders will be key predictors of future behaviour (e.g., Andrews & Bonta, 2006). If criminal history is a marker for an underlying dimension of 'criminal propensity', it may also relate to the timing of recidivism since earlier recidivists would have higher levels of the variable than later recidivists. Huebner and Berg (2011) recently found that criminal history differentiated early and mid-failure groups from the desister group, but the relationship was not significant for the late failure-desistance contrast.

A third predictor within the domain of criminal history is 'index offence type', i.e., the offence for which the offender is currently serving a criminal sentence. If specialisation in offending exists, then knowledge of recent offending would be helpful in predicting similar future offending. Studies of juvenile offenders find weak evidence of specialisation, in terms of the relationship between successive crime types, with the possible exception of property theft and truancy type 'status' offences (Armstrong & Britt, 2004; Rojek & Erickson, 1982; Stattin & Magnusson, 1991). According to Armstrong and Britt's (2004) analysis of 2,294 adolescents followed up for 3 years, property theft and burglary had the highest probability of recurring at the first arrest, and were also the highest frequency offence pairings in subsequent arrests. The probabilities were low however and when the authors controlled for individual characteristics, including ethnicity, previous violence, age at first arrest, age at arrest and substance misuse, the probability of repeating a given offence type reduced to zero for all offence categories, except for burglary. This probability only represented a 20% chance of repeat burglary in the 'average' offender, thus not supporting a general premise of specialisation in offending in early stages of offending careers.

The trend appearing in adolescence may become more established in adulthood, when sub-groups of offenders may specialise, e.g., for violence or property offences (Brennan, Mednick, & Richard, 1989; Britt, 1996; Kempf, 1987). Serious acquisitive offences of robbery and burglary appear to be the highest risk category in terms of long-term adult recidivism (Francis et al., 2007; Howard, 2011; Piquero, Sullivan, & Farrington, 2010). Piquero et al., following south London boys from age 8 to age 40, identified a group of 'long-term low rate' offenders that could be characterised by a slightly higher level of involvement in violence and robberies, compared to the adolescence-limited offenders who committed more theft type offences. This was supported by Howard's (2011) assessment of the hazards of re-offending for different re-offence categories among cases on UK probation community supervision. Theft and handling and violence re-offences had similar hazards in the first three months, but the theft and handling risk dropped away while the violent re-offending risk was persistent. Thus theft and handling was voluminous and showed a short time to re-offence, but the risk attenuated (see also Schmidt & Witte, 1989). Howard (2011) found that the re-offences with highest hazards in the fifth quarter were for drink driving and for drugs offences and those with the lowest hazards were for theft and handling. It is possible that the justice system is responsible for reducing the career longevity of theft offenders (Piquero et al., 2010), however a number of these offenders in Howard (2011) were identified as 'versatile offenders' representing nearly one-third of all offenders with a considerable early hazard of non-violent re-offending and a moderate hazard of violent re-offending. It therefore appears that, across an offender population, re-offending may relate more to offender historical and dynamic dispositional characteristics (e.g., criminal thinking) than to offence-type.

2.3.2 Dynamic risk factors.

Dynamic variables are those that are open to change and are therefore amenable to treatment (Andrews & Bonta, 2010). They can include present circumstances, current perceived problems, attitudes, and skills. Such variables may be stable dynamic, which means they change only slowly (e.g., educational deficits), or acute, in which case they change very rapidly (e.g., mood). Those variables on which change is associated systematically with variation in re-offending risk are termed 'criminogenic needs'. The distinction between criminogenic and non-criminogenic needs is important because changes on non-criminogenic needs are not necessarily associated with the probability of recidivism.

2.3.2.1 *Dispositional variables*.

Dispositional variables cover cognitive, emotional and social tendencies or traits. Deficient cognitive functioning includes lower intellectual levels and neuro-psychological deficits, but also problem-solving deficits and anti-social beliefs (Blackburn, 1993). Research has consistently indicated a small but significant negative correlation between intellectual level and delinquency, independent of social class (e.g., Hirschi & Hindelang, 1977; Moffitt, Gabrielli, Mednick, & Schulsinger, 1981), and the interpretation of this relationship is complex, with poor school achievement often implicated via its effect on cognitive development amid association with deviant peers (Loeber & Dishion, 1983). Offenders thus may have developed poor social problem-solving skills via a cognitive-behavioural process of modelling and reinforcement of anti-social contingencies (Andrews, 1980; Andrews & Bonta, 2010). Zamble and Quinsey (1997) proposed a coping-relapse model of criminal recidivism with a 'precipitating environmental trigger' (e.g., argument or job loss) as its starting point, followed by an 'acute cognitive or emotional appraisal' of the situation. The result of the appraisal gives rise to feelings of hostility, anger, or fear. A failed attempt to deal with the situation comes next, followed by a worsening cycle of negative emotions and maladaptive cognitions, culminating in offending behaviour. Importantly, whether or not a person will experience an environmental trigger or appraise the situation as threatening is supposedly mediated through static criminal history factors such as those reviewed above, but also stable dynamic response mechanisms such as coping ability, criminal attitudes, and criminal associates. Coping and problem-solving skills deficits such as those proposed by D'Zurilla and Goldfried (1971) and attitudes / peers supportive of offending, may therefore be critical in the onset and promotion of recidivism. Evidence for the validity of these dispositional factors in predicting recidivism is reviewed below (section 2.5).

The predictive relationship between criminal attitudes, maladaptive thinking, and recidivism was explored by Healy (2010). Using a sample of 73 high reconviction risk males, Healy compared current offenders who self-reported at least one offence in the previous month, with 'primary desisters' as defined by no self-reported offending in the previous month. Results showed that, although both groups had a similar level of previous contact with the criminal justice system, current offenders were significantly younger than primary desisters, and had begun offending at a younger age. There were also statistically significant differences in self-reported criminal attitudes, as measured by CRIME-PICS II (Frude, Honess, & Maguire, 1994), and endorsements on the Psychological Inventory of Criminal Thinking Styles version 4.0 (PICTS: Walters, 1995). PICTS is a measure of thinking styles that support a criminal lifestyle, such as entitlement. The relative influence of these factors on primary desistance was explored in a logistic regression model, in which only PICTS Current Criminal Thinking emerged as significant. This factor was then entered into a final model alongside current age, and age at onset. The final model indicated that all three variables were statistically significant, suggesting that current criminal thinking is important to desistance from offending (at least in the early stages of supervision) even after age and age at onset are controlled. Although Healy's (2010) study was vulnerable to normal problems with self-report in which uncontrolled factors may make offenders more or less likely to disclose, it supported Walters (2009) regarding the incremental validity of thinking skills relative to static factors in the prediction of official recidivism. Healy (2010) also supports Zamble and Quinsey's

(1997) proposal that recidivists and non recidivists experience similar problems but respond to them differently (see also Brown, St. Amand, & Zamble, 2009).

Criminal thinking may be closely related to offender motivation to change in therapy in that offenders with concrete, impulsive thinking styles may see little point in engaging with or completing therapy. The relationship between attrition on criminal justice system treatment programmes and official recidivism was investigated in a meta-analytic review of 114 studies by Olver, Stockdale, and Wormith (2011). Treatment attrition was associated with increased recidivism regardless of programme type, with non-completers' recidivism rates 10-23% higher compared to completers. Examination of the predictors of attrition showed that a treatment non-completer was most likely to be young, male, from a minority ethnic background, single, unemployed, with limited formal education, a history of previous offences and prison sentences, and actuarially higher risk (Olver et al., 2011). Clearly these factors overlap substantially with recidivism risk factors already reviewed. Such high-risk, high-need clients, precisely those who most need the full treatment, also present with responsivity factors that make engagement in treatment most difficult (e.g., low motivation, poor engagement, disruptive behaviour). These responsivity factors were strong predictors of attrition: denial of offending and low motivation for treatment for instance showed roughly 19% and 13% difference in attrition rates respectively compared to offenders admitting their offending and those that are motivated for treatment. Since non-completers tended to be higher risk than completers even before starting treatment it is not possible to conclude that their higher recidivism rates relate purely to the former group's failure to complete treatment, rather that they were already higher risk and their low motivation for treatment and other responsivity factors represented barriers to reducing this (Andrews & Bonta, 2010).

The meta-analysis of the predictors of offender recidivism by Gendreau et al. (1996) introduced earlier, included predictor categories for criminogenic needs, family factors, intellectual functioning, and personal distress, in addition to the demographic and criminal history factors already discussed. Criminogenic needs were coded as including anti-social attitudes supportive of an anti-social lifestyle and negative behaviour relating to education and employment. The largest correlations between any individual predictors and recidivism were found for adult criminal history, antisocial personality (discussed under 'Clinical variables' below), companions, and criminogenic needs.

Multiple comparison tests between all of the individual predictors revealed that adult criminal history and criminogenic needs produced the greatest frequency of significant differences in effect size, being statistically superior to family structure, intellectual functioning, socio-economic status, and personal distress (e.g., anxiety, depression, psychiatric symptomatology). The relatively weak predictive validity of personal distress variables in Gendreau et al. (1996) accords with the view that mood related variables, including anger, do not predict long-term recidivism but may operate in an acute manner more proximally to the offending behaviour (Mills & Kroner, 2003; Zamble & Quinsey, 1997).

Gendreau et al. also classified the predictors into static and dynamic factors, the latter category including criminogenic need factors, personal distress and social achievement. Comparison between the recidivism correlations of each showed a significant difference, with the dynamic factors predicting recidivism better than the static factors. The two major static and dynamic categories, criminal history and criminogenic needs, were almost identical however with criminogenic needs producing higher correlations with recidivism only marginally more than one-half of the time (54%). Nevertheless Gendreau et al.'s research synthesis attests to the importance of anti-social attitudes, which when considered together with criminal history produces good prediction estimates (Brown et al., 2009; Gendreau et al., 1996). Criminal history and deviant lifestyle variables, coded as a history of drug/alcohol abuse, were also associated with recidivism in a study of intellectually disabled offenders (Fitzgerald, Gray, Taylor, & Snowden, 2011). However, from the sample of 145 only 14 were reconvicted for a general offence in the two year follow-up. The fact that differences were observed in such a small criterion group both indicates the salience of the criminal history and deviant lifestyle variables, and also indicates the need for replication in a larger sample of intellectually disabled offenders.

2.3.2.2 *Clinical variables*.

Clinical factors include substance abuse, mental disorder, and personality disorder. Anti-social personality disorder was categorised as a dynamic factor in Gendreau et al. (1996) although the extent to which this is mutable by treatment is debatable. Psychiatric disorders including substance use disorders are modifiable with psychopharmacological treatment. The relationship between these clinical factors and recidivism is reviewed below after an explanation of the link between previously discussed risk factors and anti-social personality.

The relationship between age and crime, in which most individuals that offend do not continue into adulthood, was reviewed earlier. One of the factors associated with persistence in offending was age of onset. Longitudinal studies of the life progression of young offenders have shown that there are some important developmental differences between 'adolescence-limited' and 'life-course persistent offenders' (Farrington, 1995; Moffitt, 1993). Moffitt proposed that the two groups represent different pathways to offending, with the adolescence-limited group's offending representing a normal expression of the search for autonomy, and the life-course persistent 5% of the male population, being indicative of psychopathology. Moffitt's review cited evidence that this latter group, characterised by the earlier onset of offending and more serious and persistent offending, has underlying neuropsychological deficits. Such deficits manifest themselves as low scores on tests of language, self-control, and the inattentive, overactive and impulsive symptoms of attention-deficit hyperactivity disorder, which are linked with the early emergence of childhood anti-social behaviour and with its subsequent persistence (Moffitt, 1993). This is consistent with findings from the Cambridge Study in Delinquent development, which followed boys from age 8 to 32 (Farrington, 1995). Farrington found that the most important predictors of later delinquency fell into six categories including: anti-social childhood behaviour; hyperactivity/impulsivity; poor school attainment/low intelligence; family criminality; family poverty; and poor parental child-rearing behaviour. As these predictors included family environmental factors external to the child, Farrington constructed a measure of 'antisocial personality' including conviction, self-reported delinquency, self-reported violence, anti-social group behaviour, taking a prohibited drug, heavy smoking, drunk driving, irresponsible sex, heavy gambling, unstable job record, anti-establishment attitudes, being tattooed, and self-reported impulsivity. This measure of antisocial personality significantly correlated with itself at ages 10, 14, 18, and 32 despite the significant environmental changes between those ages. Of those cases still anti-social at age 32, sixty percent had been the most anti-social males at age 18 compared with only fourteen percent from the other group who weren’t anti-social until later (Farrington, 1995).

Not surprisingly therefore, anti-social personality disorder (APD) is the psychiatric disorder most commonly found among offenders. It is conceptualised diagnostically as “an enduring pattern of inner experience and behaviour that deviates markedly from the expectations of the individual’s culture” (American Psychiatric Association [APA], 2000, p.689). It is characterised under the *Diagnostic and Statistical Manual of Mental Disorders* (4th ed.; APA, 2000) by a persistent pattern of disregarding the rights of others manifested in for example repetitive involvement in criminal behaviour, deceitfulness, impulsiveness, aggressive irritability, irresponsibility and lack of remorse. People with APD are not necessarily psychopaths; this latter group forms a sub-group of APD and includes more emphasis on personality features such as ‘superficiality’ and ‘grandiosity’ after Cleckley (1976). Nevertheless, meta-analytic evidence suggests that the behavioural features associated with APD are better associated with recidivism than are the personality features specific to psychopathy (Gendreau, Goggin, & Smith, 2002; Hemphill & Hare, 1998; Walters, 2003).

Clinical syndromes of mental disorder, also known as Axis I or ‘major mental disorders’, include schizophrenia, major depression, bi-polar disorder, delusional disorder, and atypical psychoses. In most cases major mental disorders have their onset in late adolescence or early adulthood. Surveys of offenders suggest prevalence rates of approximately 6-10% among male prisoners and closer to 15% among female prisoners (Singleton, Meltzer, Gatward, Coid, & Deasy, 1998; Steadman, Fabisiak, Dvoskin, & Holohean, 1989). Birth cohort studies, following up life outcomes of all citizens, have suggested that individuals developing major mental disorders may become criminals to a greater extent than non disordered individuals (Hodgins, 1995). Hodgins’ data confirmed evidence from self-report studies that mentally disordered offenders appear to fall into two distinct offending groups: early- and late-starters. Early starters are characterised by a stable pattern of anti-social behaviour in childhood and resemble the anti-social children who grow up to become persistent offenders discussed above (e.g., Farrington, 1995). The late starters however progress through childhood and adolescence normally and begin offending in adulthood at about the time the symptoms of their mental disorder become apparent. Hodgins (1995) proposed that while the offending of the early starters may be more to do with APD, the offending of the late starters may result directly from the symptoms of their disorder and might therefore be treatable with pharmacotherapy and community supervision.

The predictors of recidivism by mentally disordered offenders have been subjected to meta-analysis by Bonta et al. (1998). Bonta et al. classed the predictors into four sets: demographic, criminal history, anti-social lifestyle, and clinical, with the latter class including psychiatric diagnosis. On the basis of 68 independent studies and a total sample size of 15,245 they found that the most accurate predictors were demographic and criminal history variables, similar to the pattern found for non mentally disordered offending populations. The weakest predictors of recidivism were the clinical variables; psychosis in fact negatively correlated with future recidivism, and depression was unrelated. A diagnosis of APD was a significantly better predictor than any other clinical disorder, supporting the finding regarding the importance of criminal history in the wider criminological literature.

A synergistic effect in the inter-relation among variables is however an important possibility not accounted for by Bonta et al. (1998). A large number of studies have found an increased risk of recidivism when a diagnosis of major mental disorder is conjoined with substance abuse (Castillo & Fiftal Alarid, 2011; Hartwell, 2004; Philipse, Koeter, van der Staak, van den Brink, 2006; Monahan et al., 2001; Rice & Harris, 1995; Swanson et al., 2006; Walter, Wiesbeck, Dittmann, & Graf, 2011). Philipse et al.'s (2006) final model had a four factor solution including a factor of co-morbidity of personality disorder and substance misuse disorder together with three other static factors. This achieved an AUC of .79 in the prediction of recidivism after 5 years. Walter et al. (2011) followed up 379 forensic psychiatric patients for 8 years, and found an overall recidivism rate of 41% that differed markedly with the presence or absence of substance misuse. Offenders with psychiatric Axis I disorders were the control group and 25% of these recidivated, compared to 33% of offenders with personality disorder alone, 45% of offenders with a single diagnosis of substance misuse disorder, and 69% of offenders with co-occurring substance use disorder and personality disorder. The groups also differed significantly in their time to recidivism.

The MacArthur Risk Study (Monahan et al., 2001) also merits brief discussion due to its methodological sophistication in the examination of mental health and community violence. 1,136 patients with mental disorders were monitored every ten weeks post discharge and these results were compared to a comparison group of 519 non offender controls randomly sampled from the same census tracts as the discharged patient group. Some of the findings that emerged were familiar, implicating prior criminality and substance abuse in increasing the likelihood and frequency of recidivism. Once again there was no evidence that schizophrenia was a risk factor, nor were command hallucinations or paranoid delusions predictive. The study did however find, using an interactional model described later (section 2.5), an impact of neighbourhood in which concentrated poverty including high unemployment and low income levels, contributed to increased risk over and above the effects of individual characteristics. Community level variables are discussed in the next section.

Taken together the evidence suggests that substance misuse is likely to be higher in higher risk groups and these may often contain offenders with mental disorder. These higher risk offenders have a longer and earlier starting criminal history and these features characterise APD. Other than APD however, mental disorder has not been shown to be independently predictive of recidivism although it may sometimes interact with community stressors to increase risk.

2.3.2.3 *Situational variables.*

The contingency of costs and benefits in a decision to offend may depend on the presence of social and environmental reinforcers (positive or negative), such as access to legitimate income, as much as on the personal characteristics such as anti-social tendency and impulsivity discussed above. Petersilia (2003) has argued that prisoners face significant barriers to reintegration into communities after their release from prison, making successful re-entry both difficult and unlikely. Due to implicit problems relating to their criminal record, she argues that offenders are prevented from finding housing or obtaining employment. This places pressure on offenders' families, and aggravates pre-existing problems related to anti-social peers, substance abuse and low income. Personal characteristics and situational factors may interact through a coping-relapse model as discussed earlier (Zamble & Quinsey, 1997). Within this model the offender's response to a trigger may be maladaptive, may increase negative thinking, and culminate in offending behaviour. The more stressors or triggers in the offender's environment, potentially the greater the likelihood that pre-existing risk factors such as those discussed above, will be activated.

Huebner and Berg (2011) investigated the effect of individual factors, criminal history, and community characteristics on the long-term probability of recidivism, in an eight year follow-up study of 3,786 offenders released from prison. The logistic regression and the survival models were consistent in suggesting that prisoners with more extensive criminal histories and those serving time for a property crime were least likely to desist and failed more quickly. This is consistent with the above discussion of the predictive importance of criminal history and antisocial attitudes (e.g., Gendreau et al., 1996). Turning to the measures of release setting however, Huebner and Berg found a strong role for supportive relationships, with the factor 'sustained marriage' significantly delaying failure and more than doubling the odds of long-term desistance. A similar result regarding the protective effects of sustained marriage, controlling for selection effects, was previously found by Theobold and Farrington (2010). Huebner and Berg (2011) also found that prisoners released to transitional housing, such as approved premises, were likely to fail quickly and not desist. Transitional housing was significant across all recidivism/desistance contrasts and may reflect these prisoners' weak networks of support and social capital, although some were housed in this way purely due to community risk. The majority of the sample instead returned home to live with family (e.g., mother, aunt, or sibling) however this factor was not shown to assuage risk and was associated with increased risk in later stages of follow-up. The authors suggest that compared to offenders that can afford to live alone or with a partner, those returning to families may have the least individual economic and social resources, and the tax put upon families may gradually become exhausted driving offenders to reconnect with deviant peer groups. Relationships may therefore be an important dynamic situational variable whose protective influence may depend on specific aspects of the relations.

Employment is also an important situational factor proposed to be associated with desistance (Laub & Sampson, 2003). Social achievement including employment history, income, and number of address changes, was a strong dynamic risk factor in Gendreau et al. (1996), showing higher correlations with recidivism more frequently than all other factors with the exception of criminal history and criminogenic needs. This was recently supported by Makarios et al. (2010) in a representative sample of prison releases, which found that the number of residence changes and unemployment were consistent predictors of recidivism increasing the predictive power of the regression model by 68% compared to that based solely on the static control variables. As seen with relationships, employment may relate to the timing of recidivism and may not be a key discriminating factor on the overall odds of recidivism across longer follow-up intervals (Huebner & Berg, 2011; Tripodi, Kim, & Bender, 2010). In Tripodi et al. (2010) recidivists that remained unemployed abstained from offending for 17 months, on average, which was statistically different from the 31 months crime-free among those that obtained employment.

It is possible that the measures for predicting recidivism already discussed are not valid outside of Westernised nations and cultures. Since the majority of the research discussed above has focussed on American, Canadian, British, Dutch, German, and Swedish samples, different cultural environments may interact with predictors differently, and certain predictors may lose their validity. The effect of impulsivity for example may be reduced in an environment characterised by higher levels of social control. This was investigated by Baumer (1997) with adults in Malta, and by Ang and Huan (2008) with adolescents in Singapore. These represent variations to the environments already covered; Malta is unindustrialised, and Singapore draws upon an Asian sample. Results were broadly consistent with the wider literature in finding similar levels of recidivism and similar risk factors in operation. Consistent with Farrington (1995) in the UK, Ang and Huan (2008) found that family criminality, early aggressive behaviour, and an early age of first criminal offence were significant risk factors for adolescent recidivism in Singapore. With adults Baumer (1997) found that age, male gender, number of previous convictions, and property as opposed to violent offending, were the risk factors associated with later reconviction and re-imprisonment. Thus the findings were consistent with research in other more industrialised nations in supporting the validity of static demographic and historical factors in predicting reconviction among adult offenders (Bonta et al., 1998; Francis et al., 2007; Gendreau et al., 1996).

2.3.3 Variables relating to risk of serious harm.

The prediction of serious harm resulting from violence suffers from problems of the low base-rate of violence, but also from problems in differing definitions of violence

(Monahan, 1981; Mulvey & Lidz, 1984). Hall (1987) noted that arrest rates for violence reflected just one in five actual occurrences. Similarly Monaghan and colleagues (2001) found that the base-rate altered dramatically, from 4.5 to 27.5%, depending on whether official violent recidivism was relied upon or whether the study used multiple indices (self-report, collateral report, agency records). Violence may include damage to property or to animals, or it may include dangerous behaviour such as driving while intoxicated or other irresponsible acts. Restricting the definition down to inter-personal violence further reduces the base-rate and hence makes the identification of the specific form of violent behaviour more challenging, complicated by the potential overlap among criterion and non-criterion cases. The discussion below considers what has been learned regarding the risk factors for inter-personal violence, including sexual and domestic abuse. Definitional problems with these offences, implicated for instance by the standard of evidence required to secure a conviction must also be borne in mind (e.g., criminal damage or burglary can be easier to prove than the physical or emotional abuse that may have also occurred). Quinsey, Harris, Rice, and Cormier (1998) for instance found that the recidivism of many sexual offenders, while charged as non sexual crimes, actually contained a sexual component or motivation.

Meta-analyses of studies comparing violent/non-violent outcomes among general criminal populations have generally implicated the same factors as those discriminating the any/no recidivism dichotomy (Bonta et al., 1998; Hanson & Bussière, 1998; Hanson & Morton-Bourgon, 2005; Hanson & Wallace-Capretta, 2004). Among mentally disordered offenders (MDOs), the most predictive factors included criminal history, young age, minority ethnicity, deviant lifestyle, and anti-social personality disorder (Bonta et al., 1998). Bonta and colleagues found that a violent history was a better predictor than a violent index offence; the latter was not predictive of future violence. Similarly mental disorder alone was not predictive of violence: when compared to non disordered offenders, MDOs were less likely to recidivate violently. Monahan and colleagues' (2000; Steadman et al., 2000) iterative classification tree model (described later) also found major mental disorder in the absence of substance misuse to be associated with low violence risk. In fact high violence risk was actually associated with low schizophrenia in patients with hostile symptoms who were unemployed and involuntarily admitted.

Highest violence risk was linked to the interaction between ‘seriousness of prior arrests since age 15’ and recent violent fantasies.

Factors associated with future violence among sexual offenders were studied in the meta-analysis by Hanson and Bussière (1998) and in its extension by Hanson and Morton-Bourgon (2005). In their analysis of 61 studies and 28,972 sexual offenders, Hanson and Bussière (1998) found that a criminal history was a moderate predictor of sexual recidivism (*r*=.13), as was APD (*r*=.14). The predictive estimate for criminal history increased slightly (*r*=.19) when sexual criminal history was measured. Similar to the findings with general offenders and ‘any recidivism’ outcomes, failure to complete treatment was a moderate predictor of sexual offence recidivism (*r*=.17). Although many variables specific to sexual offending were considered, e.g., clinical presentation and seriousness of the index offence, none emerged as a significant predictor of violence. The strongest of all predictors of sexual violence was dynamic and related to phallometrically assessed sexual interest in children (*r*=.32). Measures of sexual deviance, including sexual offence history, were the strongest predictors of sexual violence, and criminal lifestyle variables, including general criminal history, were the strongest predictors of non sexual violence and any recidivism (Hanson & Bussière, 1998). Hanson and Morton-Bourgon (2005) increased the number of studies considered from 61 in Hanson and Bussière to 89, and provided longer follow-up data on the overlapping studies. Results replicated those found previously in finding the strongest predictors of sexual recidivism to be the indicators of sexual deviance, and the strongest predictors of non sexual violence to be those tapping anti-social orientation (e.g., general criminal history and lifestyle instability). Anti-social orientation was strongly predictive also of sexual recidivism and was the strongest predictor of any violence, sexual or non-sexual. Hanson and Morton-Bourgon (2005) concluded that professionals concerned with the assessment and management of sexual offenders could profit from the substantial literature on the assessment and treatment of general offenders (p.1159).

Risk factors for violent recidivism among domestic abuse perpetrators were examined by Hanson and Wallace-Capretta (2004) after questions regarding the appropriateness of traditional approaches to risk assessment for abusive men. Predictor variables included in the study were therefore those commonly used with general offenders (e.g., criminal history, lifestyle instability), as well as those factors thought to be partner abuse recidivism risk factors (e.g., attitudes tolerant of abuse, marital distress). Participants were 320 male abuse perpetrators attending community treatment and followed up for, on average, 58 months. Predictive validity was assessed by correlating intake and post-treatment measures with subsequent recidivism. Results showed that the measures associated with violent recidivism, which included charges, were the same factors associated with general criminal recidivism (Gendreau et al., 1996). Violent recidivists were younger, with more criminal history and with higher scores on lifestyle instability as indexed by work/school, finances, and accommodation. The strongest correlation with violent recidivism, *r*=.32, was the criminal history sub-scale of the Level of Service Inventory-Revised (LSI-R: Andrews & Bonta, 1995). With the exception of pro-abuse attitudes which showed small but statistically significant correlations, measures of domestic abuse were unrelated to recidivism. The LSI-R total score comprises criminal history, lifestyle instability, anti-social peers and anti-social attitudes and this predicted violent and general recidivism with AUCs of .73 and .76 respectively. Interestingly the LSI-R item 'negative attitude toward helpers' correlated almost as highly with general and violent recidivism as did the total score, regardless of whether then men completed treatment. This adds to the evidence already reviewed on the validity of negative attitudes to treatment, in predicting subsequent recidivism (Brown et al., 2009; Healy, 2010; Olver et al., 2011).

This section has so far considered the correlates of violent recidivism among violent and mentally disordered offenders. Consistent findings between the risk factors for general and violent recidivism have also been identified in cohort population studies. The Cambridge Study in Delinquent Development, a longitudinal study introduced earlier, for instance found that "the causes of aggression and violence were essentially the same as the causes of persistent and extreme anti-social, delinquent and criminal behaviour" (Farrington, 1995, p.945). The best independent predictors up to age 18 of self-reported violence at age 32, included anti-sociality, no money saved, anti-social group membership, and a hostile attitude to the police (Farrington, 1989). A review for the UK Home Office, regarding the risk factors for serious harm, also suggested that the key factors among general offender samples were previous offending, being male, unemployment, family relationships, substance misuse, and APD (Powis, 2002). Howard (2011) also recently found that 'violent specialists' were less likely to re-offend violently than were ‘versatile offenders’, reinforcing the conclusion about the links between persistent general anti-social behaviour and violence.

Although pro-abuse attitudes have only shown small independent relationships to violent recidivism, the role of violent threats is clinically important and rarely evaluated. Analysis of studies of stalking has suggested that threats of harm to victims are a risk factor for violence, albeit only weakly related to the criterion with a very high false positive error (Meloy, 1996). Threats, or general negative attitudes, may however interact with other factors to discriminate violent recidivists (Rosenfeld & Lewis, 2005). This echoes Monahan’s (1981) recommendations that specific aspects of the situation such as availability of weapons or victims should also be considered and that the future of violence prediction may depend on how these variables are combined (Monahan et al., 2001). Another reason for the poor showing of variables such as anger and victim access in the prediction of violence is that they may operate more proximally to the offending behaviour and may be mediated by less labile factors such as negative attitudes and deviant lifestyle (Hanson & Harris, 2001; Zamble & Quinsey, 1997).

2.3.4 Summary of measurement requirements.

The literature on the measures of recidivism appears to be consistent across time and culture in finding that the most frequent correlates of persistent offending behaviour are the static demographic and historical factors, including:

- Young age;
- Early age of first conviction;
- Male gender; and
- Longer and more serious criminal history;

These same factors have also been linked to APD (e.g., Farrington, 1995) which may identify the sub-group of persistent offenders that fail to outgrow crime (Piquero et al., 2004). Although there is a greater likelihood of male involvement in offending, recidivism risks may depend on factors other than gender. Likewise, minority ethnicity may predict recidivism but may depend on the influence of more dynamic cognitive social learning variables such as anti-social attitudes (Gendreau et al., 1996). The following are therefore the dynamic factors underpinning recidivism risk:

- Poor cognitive skills;
- Anti-social associates;
- Anti-social attitudes supporting deviant lifestyle and behaviour in the following ‘criminogenic need’ areas:
    - Substance misuse
    - Employment / education
    - Relationships
    - Finances
    - Accommodation

Poor cognitive skills associated with poor school achievement and low IQ have been implicated in the onset and promotion of offending behaviour (Andrews & Bonta, 2010; Zamble & Quinsey, 1997). These deficits mean that recidivists and non recidivists encounter similar life problems but deal with them differently. They also explain why higher risk offenders are most likely to fail to complete offending behaviour treatment programmes (Olver et al., 2011). Motivation for treatment may therefore be a key determinant of whether the risks associated with offenders’ cognitive deficits can be mitigated. A negative attitude to treatment, education and employment is associated with the above criminogenic needs which provide the rationale for engaging in anti-social behaviour (Gendreau et al., 1996).

In the prediction of violence, the same factors responsible for offending persistence, e.g., deviant lifestyle, are implicated in offence seriousness. A violent index offence may discriminate violent recidivists less than a history of such acts (e.g., Hanson & Bussière, 1998; Monahan et al., 2000). Negative attitude to treatment/support may also be predictive of violence outcomes (Hanson & Wallace-Capretta, 2004), and may reflect lack of protection from community stressors when they occur.

Overall, although they are more vulnerable to fluctuation, there is reason to believe that dynamic variables can contribute to recidivism risk estimates. This follows meta-analytic evidence of statistically significant differences in effect size between static and dynamic predictor domains (Gendreau et al., 1996), as well as an increasing body of newer evidence supporting the incremental validity of criminogenic needs and interactions between static and dynamic factors. Discussion now turns to the existing methods of combining the risk factors into an assessment.

2.4 Methods of Assessing Risk

There are many ways of determining an individual's recidivism risk level and the following section will review those with most applicability to the present study and that have received the most research attention. Since the present study focuses on an adult criminal justice sample, this section will focus on the key methods used with offenders in that setting.

As described by Bonta (1996), over the past thirty years methods of predicting risk have moved from first generation unstructured clinical assessment, to checklists of warning signs, and then structured clinical judgement. During this time 'second generation' risk assessments have also emerged, empirically based and using actuarial methods for prediction. The empirical drive has led to 'third' and 'fourth generation' measures, frequently actuarial but supplemented with a wider sampling of theoretically informed dynamic risk items, sometimes at multiple time points. The enduring distinction is between clinical assessment and mechanical prediction, with the former relating to the use of clinical skills by the evaluator to develop a formulation based on observation of the offender's behaviour and the collection of background information. Mechanical prediction meanwhile sums information using statistical algorithms, rather than clinical skills, in order to generate risk scores from specific items. These scores represent an estimated likelihood of occurrence of the event of interest, or alternatively the score can be allocated into bands such as 'low', 'medium' or 'high' risk. The algorithms utilised within mechanical predictions are usually derived from large cohort longitudinal studies in which a range of risk factors have been collected and tested for their association with the behaviour in question. This allows mechanical methods to assign different predictive weights to the individual risk predictors and in new cases enables consideration of the relative contribution of high or low levels of a certain variable. Statistical methods therefore ensure that each individual is judged using the same criteria and that risk levels are comparable. Ensuring parity between subjects is inherently more problematic with clinical methods since clinical skills and assessment criteria are less fixed and can vary between evaluators. Thus, reliability and validity are more easily empirically tested with mechanical methods than they are with clinical methods, and therefore mechanical prediction methods more readily offer empirical evidence of any predictive relationship to the outcome in question.

While clinical methods may be limited in consistency of their application they appear to have the advantage that their application to new cases relies less on the similarity of the case to the cases in the test's construction sample. Some argue that this presents problems for actuarial methods that classify cases based on the statistical groups in which they fit (Dingwall, 1989; Hart, Michie, & Cooke, 2007). Arguably clinical methods can suffer from the same limitation since practitioners may focus on a few conspicuous variables based on a selection of their experience (Grove & Meehl, 1996). While actuarial methods have an explicit construction sample, clinical methods are based on the assessor's training or experience which varies with the individual and may be subject to human cognitive biases and short-cuts (Tversky & Kahneman, 1974). In clinical practice this can include ignoring base-rates, assuming two variables are correlated, and seeking out confirmatory rather than disconfirmatory evidence (Faust, 1986). Proponents of the clinical method maintain however that statistical models are limited by their reliance on normative information and therefore overlook valuable ideographic insights relevant to individual cases (Litwack, 2001; Pollack, 1990).

In the absence of advice as to the preferred method of risk assessment, practitioners are liable to adopt a consensus estimation of risk by taking the most typical estimate from a range of estimates (Doren, 2002). This practice can lead to problems (Mills & Kroner, 2006; Vrieze & Grove, 2010). Mills and Kroner (2006) used four recognised risk assessments to predict post-release violence and general recidivism. For most offenders there was agreement between the measures, but for some cases predictive accuracy was seriously affected where there was a marked disparity in standardised risk scores. Given the requirement to select the risk assessment with the very best accuracy this highlights the need for research that identifies the most appropriate measure for a given offender population, forensic setting, and purpose of assessment. Lack of clear guidance on the most appropriate methodology for a given setting and client group promotes a drift away from evidence-based practice and towards measures that were designed instead for treatment screening (Boothby & Clements, 2000). The next section therefore reviews the major risk prediction methodologies: clinical, behavioural, and mechanical (psychometric and actuarial). In section 2.5 the competing approaches are then evaluated as to their relative predictive accuracy.

2.4.1 The clinical method.

2.4.1.1 *Unstructured clinical judgement*.

The ‘first generation’ of risk assessments was clinical assessment, characterised by informal, intuitive, non-observable criteria for making decisions. Psychometric tests, reviewed below, may be incorporated but these are not consistently applied and which ones selected may vary between cases. Files may be reviewed but what is attended to in these files is at the discretion of the evaluator: no *a priori* theory is in place to prioritise the importance of the data obtained.

Unstructured clinical judgement is often a preferred method on account of the freedom and flexibility afforded to the assessor. This is seen as allowing the assessor full reign to apply his/her training and experience to the unique characteristics of the individual case. Objections to statistical criteria for prediction decisions have also been levelled on the basis that they oversimplify and sometimes confuse factors involved in an offender’s trajectory to offending (e.g., Grubin & Wingate, 1996). Grubin and Wingate (1996) describe how although prior criminal history is seen as the best predictor of future offending it may do little more than to distinguish a group that has demonstrated they are prepared to continue with certain behaviours regardless of sanctioning. Alternatively it may reflect that this group is more known to the police and therefore more likely to be apprehended for misdemeanours than offenders with no criminal history.

A counter-argument to the benefits of unguided individualised evaluation is that assessment should be made on the basis of a transparent procedure with explicit criteria open to scrutiny in court. Furthermore, assessment should only be made on the basis of factors that have been demonstrated to be statistically associated with a criterion (e.g., violence). Given the extensive research literature linking various dynamic factors to reconviction, a research-guided method was advanced to lend transparency and consistency to clinical judgement.

2.4.1.2 *Research guided clinical judgement*.

In research guided clinical judgement, also known as 'structured professional judgement', lists of risk and protective factors drawn from empirical results are used to steer risk assessment. Like unstructured clinical assessment the evaluator is free to assign their own weighting to factors based on their judgement as to the level of evidence for the presence or the seriousness of the factor. In theory, this approach can use any list of risk or protective factors gleaned from the empirical results pertaining to a single topic. This means that different devices employ different risk factor lists, due to different sub-sets of the literature, or different emphases in the risk assessment (e.g., recidivism by psychiatric inpatients compared to recidivism by those on general criminal sentences). Empirical tests of the efficacy of this method have involved comparing the rating of subjects' risk and protective factors to the specified outcome criterion.

The popularity of research guided clinical approaches may stem from the flexibility given to evaluators to give different weights to the different risk considerations based on the case dynamics, while also ensuring that the same empirically informed risk factors are reviewed across cases. The meaning of the total score is left to the evaluator as it may be based on different subjectively derived weightings.

The best researched of the research guided clinical approaches is called the HCR-20, a label that stands for its three sub-scales (Historical, Clinical, and Risk Management) and its 20 items. The HCR-20 (Webster, Douglas, Eaves, & Hart, 1997) was developed to provide structure to the assessment of violence risk and the items were chosen due to their connection with violence in the research literature. The assessor is asked to score the individual on each item, including a justification on the match with the factors specified in the scheme (Table 2.2). On the basis of the presence of the risk variables and clinical experience the HCR-20 asks the assessor to make a judgement on the individual's risk level. As shown in Table 2.2 one of the items within the Historical scale of the HCR-20 concerns the presence of psychopathy as measured by the psychometric test PCL-R (Hare, 1991, see below).

A key benefit of the HCR-20 is that, in encouraging a formulation of the risk issues and associated future scenarios, the assessment is structured to assist with risk management. A change in scores in the risk management section can be used as evidence to moderate immediate risk levels. In addition, a judgement can be given for the risk of harm relating to different events, thereby allowing distinction between the risk for different forms of violence such as violence to others, verbal aggression, or violence to self. The HCR-20's accuracy in predicting violence has been validated in forensic psychiatric (Strand, Belfrage, Fransson, & Levander, 1999), civil psychiatric (Douglas, Ogloff, Nicholls, & Grant, 1999), and criminal justice samples (Belfrage, Fransson, & Strand, 2000; Douglas & Webster, 1999).

Table 2.2

Components of the HCR-20 (Webster et al., 1997)

Historical Items		**Clinical and Risk Management Items**	
H1.	Previous violence	C1.	Lack of insight
H2.	Young age at first violent incident	C2.	Negative attitudes
H3.	Relationship instability	C3.	Active symptoms of major mental illness
H4.	Employment problems	C4.	Impulsivity
H5.	Substance use problems	C5.	Unresponsive to treatment
H6.	Major mental illness	R1.	Plans lack feasibility
H7.	Psychopathy	R2.	Exposure to de-stabilisers
H8.	Early maladjustment	R3.	Lack of personal support
H9.	Personality disorder	R4.	Noncompliance with remediation attempts
H10.	Prior supervision failure	R5.	Stress

2.4.2 Behavioural assessment.

Observation of behaviour represents a long standing tradition of assessing risk and an alternative means to clinical judgement based on interview data. Forensically, the transfer of behaviour between settings may be indicated by a lack of willingness or ability on behalf of the offender to control anti-social behaviour (Zamble & Porporino, 1988). Zamble and Porporino observed that it was those prisoners that coped poorly in prison, by showing little control over their impulses or behaviour that coped in a similar way in the community. That low self-control persistently manifests itself in behaviour is consistent with the firmly-held criminological finding, reviewed earlier, that one of the major predictors of recidivism is the type and frequency of previous convictions (Bonta et al., 1998; Farrington, 1995; Gendreau et al., 1996; Piquero et al., 2010).

A number of institutional studies have used observed behaviour as the basis for a prediction of community recidivism risk, with results suggesting that formal adjudications are a strong predictor (Clark, Fisher, & McDougall, 1993; Heil, Harrison, English, & Ahlmeyer, 2009; Hill, 1985). The review by Hill (1985) replicated earlier research on young adults suggesting that official records of institutional misconduct were sufficient to outperform other measures in the prediction of reoffending post-release (e.g., Mannheim & Wilkins, 1955). Hill's review however called for further research into the extent to which institutional events, such as disciplinary infractions, add to known predictors of recidivism such as prior criminal history. This was first investigated empirically by McDougall and colleagues (McDougall & Clark, 1991). Life sentence prisoners such as murderers, frequently present with no history of offending behaviour as required by actuarial risk assessments (see below). Index offence and subsequent prison behaviours were independently identified, and then referred to psychologists in a separate institution to rate the similarity between the two sets of behaviours. Statistical analysis showed that the behaviours were similar 60% of the time, while random pairs of behaviours were similar for only 20%, with the difference between the conditions being statistically significant (Clark et al., 1993). The Clark et al. study led to the development of the Wakefield Risk Assessment model (McDougall, Clark, & Woodward, 1995), a system of behaviour monitoring in use for a number of years in Her Majesty's Prison Service as a means of life sentence prisoner risk assessment. Behaviour monitoring in this way could therefore be used to detect risk that is not apparent in criminal history records, or supplement existing indications.

In a forensic hospital setting behavioural assessment has subsequently developed along similar lines, using the concept of 'offence paralleling behaviour' (Jones, 2004) to identify persistent pathological patterns of behaviour that may be related to an ongoing risk of offending. Consistent with Clark et al. (1993) the behaviour is expected to be functionally similar across environments, thereby offering an opportunity for intervention to alter the offender's pattern of responding to situations. The notion of behavioural consistency has however been criticised, with the observation that consistency may be dependent on the presence of certain trigger stimuli (Mischel, 1968). In forensic institutions there is a commonly held view among prison administrators that the environment is qualitatively distinct and therefore provokes unusual behaviours, e.g., as a result of sexual deprivation. Mischel and Shoda (1995) however have argued that behaviour could be consistent between altogether different environments if similar psychological features were activated, and this promotes the importance of cognitive factors in the perception and interpretation of situational cues. The extent to which these psychological features are activated should therefore correlate with the observed behaviour, promoting a link between risk and behavioural frequency (particularly if the behaviours are still observed in a relatively controlled custodial environment). 'Act frequency' may be a reasonable measure of cross-situational consistency (Buss & Craik, 1989; McAdams, 1997). Thus offenders regularly committing offences in prison, regardless of whether they had relevant convictions prior to custody may be reconvicted more quickly than prisoners not offending in custody (Heil et al., 2009). Heil et al. (2009) showed that sexual arrests at one- and five-year follow-up were disproportionately associated with the prisoners that offended sexually in custody. The objective and reliable nature of behavioural information may give it an advantage relative to other measures in recidivism prediction.

2.4.3 Mechanical approaches.

2.4.3.1 *Psychometric evaluation*.

Psychometric measures are distinguished by the fact that they are commonly used in the assessment of psychological traits that may be related to psychopathology, whereas actuarial measures estimate risk based on aspects of behaviour. Both have in common, however, the use of statistical rather than clinical indices for prediction. Examples of psychopathology targeted by psychometric measures include i) tendency to high-risk affective states, ii) dangerous personality traits, and iii) characteristics of sexual deviancy.

As a result of the importance of psychopathology to the treatment of offenders, psychiatric and correctional settings routinely use psychometric personality inventories. One such is the structured clinical assessment known as the Psychopathy Checklist – Revised (PCL-R; Hare, 1991) The PCL-R is designed to assess the presence of psychopathy traits and, underpinned by the work of Cleckley (1976), was designed by Hare after an empirically based assessment of the clinical factors that comprise psychopathy (Hare, 1980). The latest version comprises twenty items considered central to psychopathy, of which eight are interpersonal (factor 1) and the majority of the remainder are behavioural (factor 2). Factor 2, overlapping with APD, is shared by a large number of offenders (Cooke & Mitchie, 2001; Hare, 2003), while factor 1 personality characteristics are distinct to a small sub-group of individuals in the offender population. This explains why the base-rate for psychopathy in criminal justice and forensic psychiatric populations, between 15-23%, is much lower than the base-rate of 50-80% for APD (Hare, 2003).

The PCL-R rating scale uses information from a semi-structured interview, case-history information, and specific scoring criteria to rate each of the 20 items on a three point scale according to the extent to which it applies to a given individual. Although the PCL-R was not originally designed to assess risk for violent recidivism (Hemphill & Hare, 2004), it has come to dominate the literature of violence prediction because the traits associated with psychopathy are correlated with violence outcomes (Hare, Clark, Grann, & Thornton, 2000; Hemphill, Hare, & Wong, 1998). As a result, a number of violence prediction schemes call for an assessment of psychopathy to be considered within the assessment (e.g., HCR-20: Webster, et al., 1997).

A number of psychometric tests are in routine use in the UK and other criminal justice systems (e.g., Beech, 1998; Craig, Franklin, & Andrews, 1984; Frude et al., 1994; Hathaway & McInley, 1967; Walters, 1995). In a study of the performance of a range of psychometric tests in predicting recidivism, Walters (2006) found in favour of the comparison category of structured ‘risk appraisals’. When the analysis was confined to the comparison of *crime-relevant* psychometrics and risk appraisals however, the difference became statistically non-significant. Furthermore, integration of these content-relevant self-report measures with the risk appraisal, added to the validity of structured assessments in more than one-half of comparisons. This supports the use of crime relevant psychometrics in augmenting structured risk assessments.

2.4.3.2 ***Actuarial risk assessment*.**

Bonta (1996) described the earliest actuarial risk assessments (ARAs) as the ‘second generation’ of risk assessment. ARA is the most mechanical means of prediction, empirically based but atheoretical in terms of items included. A regression equation is estimated on the basis of the characteristics of a development sample, including their recidivism outcome as collected after a prescribed interval. The predictor variables are statistically weighted as to the level of their association with the outcome, and during the prediction this weights matrix is then used to apply to new cases whose outcome is unknown. Items that are found statistically to be unimportant to the prediction equation are dropped, leaving a formal set of items to be rated by an assessor performing a new prediction.

The ARA procedure therefore ensures that practitioners remain focussed on the key variables, rather than becoming distracted by seemingly salient features of an unusual case. Since the weighting of each item is determined empirically, ARA may have an advantage in adversarial contexts promoting transparency and consistency in decision-making. In many risk assessment contexts, the question is not whether or not the subject will fail, but whether the subject's risk of failure is beyond a specified legal threshold (e.g., more likely than not), and ARAs producing risk percentages may be more suited to this task than are clinical judgement approaches producing yes/no estimates (Doren, 2002).

The consistency afforded by ARAs is not always viewed as an advantage since this prevents the evaluator from adding weight to certain items that may be seen as important in terms of imminence or frequency of the high risk behaviour. Since ARAs to date have generally focussed on historical indicators that are static and unchanging, it is argued that these do not take account of important changes in the case that have a strong bearing on imminence, for example indications that the offender has attempted to access victims (Hanson & Harris, 2000). Whether overall such adjustments increase predictive accuracy is an empirical question and related evidence is reviewed below.

Another common criticism of ARAs is that they are highly context specific and therefore may have problems in generalisation beyond the cases on which they were developed (Grann, Belfrage, & Tengstrom, 2000). If a new case does not closely resemble the cases in the construction sample, it can be argued that the ARA is not valid for the specific case. This is particularly problematic when the prediction outcome occurs at a low base-rate, such as with sexual and violent recidivism. Grubin and Wingate (1996) suggest that actuarials are good at determining those cases at low risk for recidivism but are poor-moderate in identifying true positive cases. This has led to the suggestion that

ARAs should be limited to programme screening to sort those individuals for prioritisation (Campbell, 2003; Sjostedt & Langstrom, 2001).

One of the best known ARAs is the Violence Risk Appraisal Guide (VRAG: Quinsey et al., 1998). The VRAG was constructed by taking variables known to predict violent behaviour among criminal offenders as well as among men with mental disorders who have records of violent behaviour, and then summing those variables into one scheme. No clinical training is required to use the scheme, except that required for the assessment of psychopathy, the scheme's most heavily weighted item. In addition to the PCL-R score the VRAG items include: elementary school maladjustment, non-violent offence history, never married, DSM-III diagnosis of personality disorder, victim injury, alcohol abuse and female victim in index offence. The 12 risk factors are weighted according to how the presence of each affected the base-rate of violent failure in a sample of 618 mentally disordered offenders (Harris, Rice, & Cormier, 2002; Harris, Rice, & Quinsey, 1993). A simple weighting scheme was used in which a weight of 1 was assigned for each full deviation of 5% in the base-rate associated with the presence of the item. The total score is reached by summing the scores on each of the 12 weighted items and the participant being placed accordingly into one of nine risk categories assumed to reflect the probability of recidivism (Quinsey et al., 1998). The VRAG authors (Harris et al., 1993) report probability estimates for violent recidivism within seven years of upwards of .76 for high VRAG scores (over +21).

Other renowned ARAs include the Salient Factor Score (Hoffmann, 1983, 1994), the General Statistical Information on Recidivism (GSIR: Bonta, Harman, Hann, & Cormier, 1996; Nuffield, 1982), the Wisconsin Classification System (Baird, 1981), the Static-99 (Hanson & Thornton, 2000), the Offender Group Reconviction Scale (OGRS: Copas & Marshall, 1998; Taylor, 1999), and Risk Matrix 2000 (RM2000: Thornton et al., 2003). The latter two are used extensively in the UK National Offender Management Service (NOMS) to determine resource allocation pre- and post-sentence. The OGRS is an ARA based solely on history of offending and certain demographic variables. It estimates the probability that an offender will be reconvicted of any offence within two years of release. The nine variables included in the revised version, OGRS-II, are shown in Table 2.3 below.

As apparent from Table 2.3, OGRS does not use self-report or clinical judgement and there is no assessment or weighting of mental health variables. All ratings are computer generated, thereby eliminating rater reliability issues and ensuring the ease and practicality associated with its popularity. A score cannot be calculated for persons without previous convictions. The OGRS measure has been evaluated on a wide range of offender populations, including criminal justice (Coid et al., 2009; Lloyd et al., 1994; Wakeling, Howard, & Barnett, 2011), and forensic mental health (Gray et al., 2004; Snowden, Gray, Taylor, & MacCulloch, 2007).

Table 2.3

Variables used within OGRS-II (Taylor, 1999)

1.	Offender age at commencement of risk
2.	Gender
3.	Number of custodial sentences as a youth
4.	Current offence type
5.	Age at current conviction
6.	Age at first conviction
7.	The Copas rate variable (the rate at which the offender has been convicted)
8.	History of burglary offences
9.	History of breach (of community orders)

The RM2000 (Thornton et al., 2003) is intended for use with men aged 21 and older to assess risk of violence including sexual violence. The instrument was developed as a simple, cost-effective actuarial predictor on the basic premise that most criminal behaviour is predictable from a simple combination of age and some indicators of reoffending of the type being predicted (Friendship, Thornton, Erikson, & Beech, 2001). The RM2000(V) includes only three items, while the RM2000(S) includes six items. A combined (C) score for risk of sexual and violent recidivism is then produced, although in practice the RM2000(S) and RM2000(V) are each permitted to be used as a stand-alone with their own recidivism estimates spanning 5 to 15 years. The three scales have been independently validated with UK sexual offenders, showing moderate-high accuracy (Barnett, Wakeling, & Howard, 2010; Grubin, 2008).

ARA is not limited to the review of static indicators, and the incorporation of dynamic variables has proceeded with the Level of Service Inventory-Revised (LSI-R: Andrews and Bonta, 1995). The LSI-R is a 'third generation' risk assessment (Bonta, 1996), given that it is theoretically informed and incorporates dynamic risk factors useful in the allocation of offenders to treatment. It therefore provides an assessment of risk of re-offending as well as information relating to the treatment needs of the offender. Designed originally using Canadian data (Andrews, 1982) it has now been used extensively with a variety of offender samples within Europe and North America. Scores are produced in relation to one static component, criminal history, and nine dynamic sub-components: education/employment; finances; family/marital; accommodation; leisure/recreation; companions; alcohol/drug problems; emotional/personal; and attitude/orientations. To produce a composite score the domain scores are weighted according to how efficiently previous research has shown them to relate to reconviction, with the strongest weighting awarded to the criminal history domain. As such, the LSI-R is an actuarial risk-needs tool and allows minimal room for clinical over-ride based on anecdotal evidence.

An extension of the LSI-R to include case management plans saw the development of the Level of Service/Case Management Inventory (LS/CMI: Andrews et al., 2004). The facility to update risk assessment and case management plans with new information arising from treatment, characterises 'fourth generation' instruments (Andrews et al., 2006). The LS/CMI potentially enhances the LSI-R by addition of an 'anti-social personality pattern' sub-component focussing on early and diverse problems. Girard and Wormith (2004) have provided predictive criterion validity of the LS/CMI with diverse samples including male sex offenders, domestic abusers, and offenders under psychiatric care.

Another example of an ARA in regular use is the Offender Assessment System (OASys: Home Office, 2002) for use with all offenders under NOMS in England and Wales. Similar to the LSI-R, in OASys ARA information using static items akin to those in OGRS, is given the strongest weighting but actuarially adjusted on the basis of the clinically determined extent to which dynamic risk factors are also present in the case. The dynamic factors are more comprehensively reviewed than under the LSI-R, with the time required to complete the OASys assessment spanning two and a half hours, compared to the ten minutes required by the LSI-R (Raynor, 2007). Dynamic factors reviewed are shown in Table 2.4 (items 3-12). Each dynamic factor is clinically assessed on separate items, and the resulting score on each factor is given a pre-determined weighting according to the strength of its prior relationship with recidivism (Howard, Clark, & Garnham, 2006). This means that the scope for individual differences in how professionals adjust the ARA is constrained thereby assisting in the measure's reliability. Following the assessment of offending-related needs OASys includes a sentence-planning section to follow supervision through to case closure and allow linkage between intake assessment, service delivery, re-assessment, and medium- or long-term outcomes. As such OASys would be characterised by Andrews, Bonta and Wormith (2006) as a 'fourth generation' risk assessment, similar to the LS/CMI.

Table 2.4

OASys Risk of Reconviction and Offending-Related Factors (Home Office, 2002)

1.	Offending Information
2.	Analysis of offences
3.	Accommodation
4.	Education, training, and employability
5.	Financial management and income
6.	Relationships
7.	Lifestyle and associates
8.	Drug misuse
9.	Alcohol misuse
10.	Emotional well-being
11.	Thinking and behaviour skills
12.	Pro-criminal attitudes

2.4.4 The clinically adjusted actuarial procedure.

The idea behind clinically adjusted ARA is that actuarial methods are used as the foundation for the risk assessment and then the evaluator is permitted to subjectively adjust or override the instrument in at least some circumstances. The rationale for adding in clinical considerations to actuarial findings is that existing actuarials focus their strongest weighting on unchanging historical indices and therefore take insufficient account of the current state of important dynamic factors. Such dynamic factors might signal the timing at which the predicted event may occur, or they may indicate time-limited protection from the event. Clinical adjustment therefore has good face validity; however, this may not translate into external validity. The danger inherent in clinically adjusting ARAs is that the adjustment may itself be associated with clinical biases and therefore could serve to lessen rather than to increase the predictive accuracy compared to the performance of unadjusted ARA instruments (Hart, Laws, & Kropp, 2003; Quinsey et al., 1998).

The illegitimate face of clinical adjustment is when it is merely a guise for clinical judgement. For example, a risk assessment report in which an ARA and its results are reported, but then overlooked in the remainder of the report including the conclusion regarding the final risk judgement. In this situation the assessment has substituted ARA for clinical judgement, rather than adjusted the former with the latter. According to proponents of actuarials, if the ARA is overlooked or automatically discounted then the assessment cannot be said to be actuarially grounded, and is considered to be 'irrational, unscientific, unethical, and unprofessional' (Zinger, 2004, p.607).

Adjustments may be made legitimately in a few distinct scenarios: i) where the case characteristics are clearly different to those in the ARA's construction sample (e.g., female offenders); ii) where the outcome being predicted in the assessment or the follow-up time period covered is different to that which the ARA was designed to predict; iii) where the case shows particularly 'rare' characteristics for which there is also a supposedly clear link to risk or protection, even if this has never been researched. This area is the most clearly anecdotal cause for clinical adjustment. A final category of appropriate adjustment relates to instances where research has demonstrated the information to add incrementally to the ARA's predictive accuracy.

Research evidence exists to show that completion of a relevant cognitive-behavioural treatment programme targeting the offender's criminogenic needs, can protect against ARA identified recidivism risk (Andrews et al., 1990; Beech, Erikson, Friendship, & Ditchfield, 2001; McGrath, Cumming, Livingstone, & Hoke, 2003; Thornton, 2002). The ARAs that incorporate dynamic factors into their statistical algorithm, such as LSI-R and OASys introduced above, provide for change based on treatment or other intervening factors, however many of the items included in the scoring are highly stable and are not open to great variation over the short term, e.g., due to treatment completion. Clinical factors such as motivation for treatment and ongoing self-risk management are embedded in the risk factor scores and are not given the emphasis deemed necessary by practitioners.

The Structured Risk Assessment (Thornton, 2002) classifies sexual offenders on the basis of a static risk measure, and then considers initial deviance, evaluation of treatment progress, and risk management based on acute risk factors. Step 1 assesses actuarial risk, using Static-99 (Hanson & Thornton, 2000); step 2 is the 'initial deviancy assessment' measuring sexual interests, distorted attitudes, socio-affective functioning, and self-management; and step 3 reconsiders risk potential following treatment intervention. Therefore step 2 and step 3 are based on clinical interview and psychometric deviancy assessment (Beech, 1998), with criteria for adjustment not related to statistical validity in terms of an empirical link with recidivism. Clinical adjustment is permissible because treatment change is an empirically supported factor thus practitioners feel justified in making subjective clinical adjustment (Doren, 2002). Evidence exists to suggest that adjustment can safely be made if this is on the basis of psychometric deviancy information (Beech et al., 2001; Craig, Thornton, Beech, & Brown, 2007; Thornton, 2002). For example, Thornton (2002) found that none of the offenders in the high static risk category on Static-99 that were also 'low deviancy' reconvicted after 3 years, while two-thirds of the men classified as 'high deviancy' within the same static risk band reconvicted. There are however no tables linking the summary scores to recidivism rates, and therefore the measure does not have an empirical basis for adjusting the ARA.

The Sex Offender Need Assessment Rating (SONAR) (Hanson & Harris, 2001) performs a similar function in adjusting actuarial risk, again using Static-99, by combining the static risk score with one based on consideration of stable and acute dynamic factors. Stable factors purportedly alter the probability of recidivism but only change under purposeful activity (e.g., treatment). Stable factors include the offender's significant social influences, intimacy deficits, sexual self-regulation, general self-regulation, attitudes, and cooperation with supervision. Acute factors fluctuate on a daily or weekly basis and may relate more to the timing of recidivism. Hanson and Harris included the following seven items: access to victims; emotional collapse; collapse of social supports; hostility; substance abuse; sexual preoccupations; and rejection of supervision. Hanson and Harris (2001) originally proposed rules on how the ARA should be adjusted with the dynamic information, and this was subsequently updated with empirically informed rules (Hanson, Harris, Scott, & Helmus, 2007). Hanson et al. found the combined static/stable categories were more accurate than either the static or stable variables individually. The AUC value for the Static-99 assessment of any recidivism by sexual offenders over three years was .69, while for the static/stable combination it was .70. In addition, Hanson et al. showed that the acute factor information added to the short-term prediction of recidivism. While structured dynamic information added (marginally) to risk prediction based on Static-99, times when the clinical override judgement was used were associated with a decrease in predictive accuracy compared to the unadulterated ARA.

2.5 Success / Validity of Methods

This section covers the empirical evidence regarding the predictive validity of the methods described above. To facilitate understanding of the relative value of each approach, the material is organised first to compare unstructured prediction against structured/mechanical approaches, and then to compare structured judgement approaches against the statistically derived ARAs.

2.5.1 Unstructured clinical judgement versus structured prediction.

Concerns over the accuracy of unstructured clinical assessments stemmed from the findings of a pair of ‘natural experiments’ exploited by researchers in the 1970s (Steadman & Cocozza, 1974; Thornberry & Jacoby, 1979). These experiments involved the comparison of outcomes for patients that had been detained under civil commitment laws and then all released after a landmark legal ruling. Since all of the patients had originally been detained on account of the interaction between their mental health and their violence risk, their simultaneous release provided an opportunity to review the accuracy of these assessments. In the first study 98 patients were tracked for two and a half years following their release into the community. Few of these individuals had further contact with the law: 20% were arrested and just 11% were reconvicted (Steadman & Cocozza, 1974). Only two cases were involved in violent acts, the reason for their civil commitment; indicating that the criterion had been over-predicted. The Steadman and Cocozza results were reinforced just a few years later in a parallel case

(Thornberry & Jacoby, 1979). Although these cases illustrate alarmingly high levels of false positive predictions, this may relate to a number of present factors: a low base-rate of the criterion; problems in reliability of the criterion measure; attention to the wrong predictive factors; or problems in the reliability of the predictor variables. This latter factor may relate to unreliable case records or it may relate to inconsistent assessment of those materials. Alternatively, it may relate to erroneous assessment in terms of the variables selected for measurement and the means by which they were combined.

Given these confounding factors in establishing the validity of clinical prediction it is instructive to compare it with an alternative prediction approach using the same criterion measure. Grove and colleagues (Grove & Meehl, 1996; Grove, Zald, Lebow, Snitz, & Nelson, 2000) have examined the relative predictive effectiveness in a meta-analysis, focussing on studies comparing clinical judgements with statistical procedures. Grove et al. (2000) is of particular interest since it encoded into the analyses covariates for study design variables, e.g., the experience of the assessor. Based on analysis of 136 studies the clinical method was superior to mechanical prediction in only 8 studies (6%), while in 65 (48%) it was outperformed by the mechanical method. Grove et al. (2000) concluded that mechanical procedures were therefore equal or superior to clinical prediction in a wide range of circumstances. The only design variable that significantly influenced the relative accuracy of the two methods was whether the clinical method was based on a clinical interview, with presence of this factor increasing the margin of difference in favour of mechanical prediction. When the comparisons by setting were examined there was a trend for superiority of mechanical prediction across all settings, in particular medical and forensic.

Since Grove et al. (2000) included medical as well as psychological diagnosis it is appropriate to review the success of clinical methods in meta-analyses on specifically forensic samples (Bonta et al., 1998; Gendreau, et al., 1996; Hanson & Bussière, 1998; Hanson & Morton-Bourgon, 2009; Mossman, 1994). Mossman (1994) provided a comparison of 17 studies using clinical methods and 13 studies using ‘behaviour-based predictions’ to predict violent behaviour among psychiatric patients. The behavioural prediction strategy yielded a higher AUC value of .78 compared to .67 using the clinical method. Statistical risk assessment procedures predicted sexual recidivism better than did clinical procedures (*r* = .42 vs. *r* = .11) across 61 studies in Hanson and Bussière (1998).

This supported Bonta et al. (1998) and Hanson and Morton-Bourgon (2009); both meta-analyses found that objective risk measures predicted general and violent recidivism better than did professionals' judgements of risk. Summarising across all of the above mentioned meta-analyses, Andrews et al. (2006) estimated that the mean predictive validity for the prediction of general recidivism was .10 for unstructured clinical judgement, compared to .42 for general risk scales. A very similar result was found for the prediction of violence in the meta-analyses where the mean for unstructured clinical judgement (.13) was outstripped by that for actuarial risk scales (.39) (Andrews et al., 2006). The clarity of the findings, together with the volume and variety of studies included leads one to conclude that unguided clinical judgement offers little in terms of relative efficacy when systematically compared to other available approaches.

2.5.2 Research guided clinical judgement versus actuarial prediction.

Attention therefore turns to the success of measures of research guided clinical judgement compared to ARA procedures. The Hare PCL-R is a psychometric measure and somewhat bridges the divide between a research guided approach and an ARA. It is incorporated in the research-guided HCR-20, but also features in the VRAG actuarial device. Comparisons between devices where one depends on this measure must be cognisant therefore of the role of the PCL-R and indeed whether this is responsible for any difference. The PCL-R has consistently been found to be a significant predictor of recidivism, although the average correlations with offending have been modest (Hemphill et al., 1998) or inconsistent (Salekin, Rogers, & Sewell, 1996).

In the prediction of violence the research guided HCR-20 has frequently been found to add incremental validity to the PCL screening version (Belfrage et al., 2000; Douglas et al., 1999; Strand et al., 1999), and therefore indicates the value of the total HCR-20 measure as a supplement to one of its own items. Evidence exists however to suggest that removal of the PCL item from the HCR-20 also removes its predictive advantage over the PCL-R alone (de Vogel, de Ruiter, de Hildebrand, Bos, & van den Ven, 2004). This may indicate the value of the PCL-R as a mediator of violence risk rather than a determinant in itself, consistent with the advice from its developers (Hare, 2003).

Direct comparisons of ARAs and research guided approaches in the prediction of violence appear to be limited to a small number of primary studies. On the face of it there appears to be a lack of consensus, with some finding in favour of ARAs (Bonta & Yessine, 2005; Coid et al., 2009; Gray et al., 2004; Kroner & Mills, 2001; Loza & Green, 2003; Mills & Kroner, 2006; Sjostedt & Langstrom, 2002), and some finding in favour of research guided approaches (Cooke, Mitchie, & Ryan, 2001; Dahle, 2006; Doyle & Dolan, 2006; Grann, Belfrage, & Tengstrom, 2000). These studies vary considerably in sample size, type of population being tested, and length of follow-up, such that it is difficult to draw conclusions using a qualitative review methodology.

Comparisons of research guided instruments to ARAs under meta-analysis are limited to the prediction of violent recidivism, which follows from the focus of most research guided approaches on this outcome. A handful of meta-analyses have been conducted in recent years, each including the HCR-20, the LSI-R, the PCL-R, and static ARAs such as the GSIR or OGRS (Campbell et al., 2009; Farrington et al., 2008; Hanson & Morton-Bourgon, 2009; Yang, Wong, et al., 2010). The study by Campbell et al. (2009) covered published and unpublished data from 1980 to 2006, taking in 88 studies and 185 effect size estimates for violent recidivism, the majority of which were based on general offenders rather than forensic psychiatric samples. Results, weighted for sample size, showed that the VRAG had the strongest predictive relationship with violent recidivism (.27) followed by the HCR-20 (.25), the LSI-R (.25) and the GSIR (.22). However, using the *Q* statistic (Rosenthal, 1991) it was apparent that in all but the HCR-20, the effect size estimates were more variable that would be expected by chance. This suggested that there are other important moderators within the data which were not controlled. Since the confidence intervals all overlapped, the authors concluded that all instruments were likely to be sampling from the same population parameter. Hence Campbell et al. (2009) could not find clear evidence for a difference between ARAs and research guided methods, with all measures seeming to pick up on the same variation within the data.

Farrington et al. (2008) expanded on the study by Campbell and colleagues by the inclusion of data from a large prisoner cohort study by Coid et al. (2009). Farrington et al. also included 'random effects' models giving more equal weight to all studies regardless of sample size. Particular moderator features were then analysed separately. This is important in those instances, such as seen in Campbell et al. (2009) where study heterogeneity is at issue. Farrington et al. found that all of the effect size estimates had statistically significant *Q* values with the exception of those for the LSI-R and the GSIR.

Therefore, examining the random effects models only, the HCR-20 performed best with an AUC value of .70 (‘modest’ accuracy: Sjostedt & Grann, 2002). The GSIR and the OGRS measures did marginally better than the HCR-20, with AUCs of .73 and .71 respectively, but the estimates were only based on four effect sizes in each case and so were deemed in need of further evaluation and replication. Andrews et al. (2006) previously made a similar point about the promising OGRS measure.

Similar to Campbell et al. (2009), Farrington and colleagues found that in almost all cases, the confidence intervals for the average effect sizes overlapped, the only significant difference being between the GSIR and the LSI-R (the best and the worst performer). This suggested that, without distinguishing the purpose of the assessment, the type of sample etc., all measures were essentially interchangeable. When the analysis by moderator variables was performed, there was an interesting difference according to the type of sample. Consistent with their origins, the VRAG and the PCL-R both performed better on psychiatric samples compared to general criminal justice samples and the HCR-20 showed a similar trend, the difference just falling short of statistical significance. The AUC on psychiatric samples reached approximately .73 with each of the three measures. Farrington et al. (2008) concluded that, based on the available evidence, a static measure such as the GSIR or OGRS was likely to have the best predictive validity in unselected offender samples, although the HCR-20 should be used for the assessment of dynamic change.

Both of the foregoing meta-analyses found that studies could not be easily compared due to lack of homogeneity across variables such as length of follow-up, size of sample, type of sample, and other study design features. In addition there are likely to be differences between studies on features that are not measured. Random effects models would allow these to influence the results but would not be able to stratify the post-hoc analysis to account for these moderators. Yang, Wong, et al. (2010) countered this by using a within-group design including only independent studies that compared the predictive validity of more than one tool on the same individual. The study also used random effects models to compare weighted effect sizes, and then to examine and adjust for the impact of study features on the differences in effect sizes arising. Search criteria identified 28 studies between 1999 and 2008 and 174 effect size estimates. Follow-up time was for an average of 43.8 months. All of the recognised risk assessment tools featured in the study, with the PCL-R featuring most regularly. For this reason the PCL-R formed the reference category against which all other instruments were benchmarked. Eight other risk assessment devices were reviewed, including the VRAG, the HCR-20, the LSI-R, the GSIR, and OGRS measures discussed above.

Similar to previous meta-analyses in this area, Yang, Wong, et al. (2010) found that the random effects model showed considerably improved goodness-of-fit over the fixed effect regression model, due to study heterogeneity. The HCR-20 was the only instrument showing a statistically larger effect size that the PCL-R under the random effects model. This difference remained after controlling for instrument differences and study/sample characteristics. After taking into account the data structure, the country of study, participant sex, mean age, follow-up time, and prospective/retrospective nature of study however, the predictive accuracy of instruments all fell within an AUC range of .56 and .71, with the majority falling within a narrow range of .65-.69.

The pattern of results showed that for all instruments larger effect sizes were evident depending on certain study features: prospective rather than retrospective data collection; longer rather than shorter follow-up time; and studies on women or mixed samples rather than studies on men only. Consequently Yang, Wong, et al. went on to explore the existence of interactive effects, including whether there were any gender-related interactions with instrument type. Specifying these interactions significantly improved the goodness-of-fit of their model. While no significant sex differences were found for the other instruments, the accuracy of OGRS with men was significantly larger than that of the PCL-R with men, with estimated effect sizes of .94 and .63 respectively. For women the effect size of OGRS was considerably reduced at .14 compared to .74 for the PCL-R. In the presence of these gender-related interactions the previously observed significant difference between prospective and retrospective designs disappeared.

Overall Yang, Wong, et al. (2010) showed that predictive accuracy between the tools was essentially very similar, regardless of whether the instruments include static indicators, a combination of static and dynamic factors, or whether the summing of instrument scores are empirically weighted or left to human judgement. The variation in effect sizes between studies was mainly due to factors other than instrument differences, with participant age, follow-up time, outcome criteria, gender and gender-interactions with instrument and country accounting for 85% of the variation. This is supported by the meta-analyses of studies with general offenders (Campbell et al., 2009; Farrington et al., 2008), with adult sexual offenders (Hanson & Morton-Bourgon, 2009), and with juvenile offenders (Schwalbe, 2007). Schwalbe (2007) estimated that instrument differences contributed a mere 17% of the variance in effect sizes, while methodological moderators such as existence of cross-validation and sample type contributed 42% of the variance. Thus effect sizes were on average eight percent larger if the instruments had not been cross-validated, and a similar increase was present if the instruments were developed on a heterogeneous probation sample, rather than on a more uniform (uniquely high risk) prison sample.

These studies therefore highlight the importance of sampling in the development of risk assessment measures. It seems likely that all existing measures are drawing from the same pool of variance; one that can be captured equally and to the same (limited) extent by historical indices and current dynamic indicators. Kroner et al. (2005) illustrated, using a hybrid model that performed as well as the developed models, that the existing risk factors in the tools may be essentially interchangeable (see also Coid et al., 2011). The consistently 'modest' performance in predictive accuracy suggests that the method of combining or summing the variables may have reached a ceiling, and that alternative methods of doing this may add value in achieving the required improvements (Yang, Wong, et al., 2010).

2.6 Innovations in Forensic Risk Assessment

This section focuses on new methods being developed in forensic psychology and psychiatry for combining information in risk assessment, following recommendations that these should be explored (Borum, 1996; Yang, Wong, et al., 2010). On account of the complexity of ARAs, Borum (1996) advocated further exploration of classification and regression trees (Brieman, Friedman, Olshen, & Stone, 1984). In pursuit of improved predictive accuracy Yang, Wong, et al. (2010) similarly recommended the investigation of tree modelling, but also development of neural networks. The potential benefits of artificial neural networks (also known as connectionist models) with offender data are considered in detail in Chapter 3.

Classification tree (CT) modelling in the prediction of recidivism has proceeded both to increase clinical applicability of statistical risk assessment (Monahan et al., 2000), and to improve predictive accuracy (Steadman et al., 2000). CT modelling uses recursive partitioning which is a nonparametric form of discriminant analysis often operationalised via a standard package such as CHAID (chi-squared automatic interaction detector, SPSS Inc., 1993). CHAID partitions a sample of cases into smaller and smaller subgroups developed based on contingent associations between select risk factors and an outcome. The advantages of the CT model include that it allows many different combinations of risk factors to classify a person as low- or high-risk. A first question is applied to all cases subject to assessment and contingent on each case's answer to that question one or another second question is posed, and so on, until each subject is classified into a high- or low-risk category. The goal is to sort cases into subgroups that consist entirely of individuals who are later found to recidivate or not recidivate (Cook & Goldman, 1984). This contrasts with a linear regression approach in which a common set of questions is asked of everyone being assessed and every answer is weighted and summed to produce a score used for categorisation. Thus while a main effects regression model implies a single solution fits for all persons being assessed, CTs use an “interactive and contingent” model (Steadman et al., 2000, p.84). A second advantage proposed by Steadman and colleagues is that the CT model can acknowledge the practical difficulty of adequately classifying all cases into a high- or low- recidivism risk group. Therefore rather than relying on a single threshold for distinguishing among cases, their CT approach employed two thresholds: one for identifying high-risk cases and another for identifying low-risk cases, leaving an intermediate group unclassified. Unclassified cases were considered indistinguishable from the base-rate of the sample as a whole.

Using data from the MacArthur Violence Risk Assessment study (Steadman et al., 1998), high-risk cases were those reconvicting at at least twice the base-rate (>37% over 12 months), while low risk cases were those re-offending at half the base-rate or below (<9% over 12 months). With these cut points, 43% of the sample were unclassified using a regression analysis approach compared with 49% using the simple CT approach. This indicated that both the traditional ARA and the CT approach failed to distinguish nearly one-half of cases from the base-rate.

Monahan et al. (2000; Steadman et al., 2000) went on to look at the use of repeated iterations using a CT approach; repeated (iterative) analyses were undertaken on the group that had not been distinguished from the population base rate (*n*=462). A second iteration allocated 119 of these individuals to either high- or low-risk groups, and a third and fourth iteration allocated an additional 63 and 60 subjects respectively. Using this form of recursive partitioning, 77% of the sample could be allocated to the high- or low-risk groups, representing a significant improvement. The AUC for the main effects regression was .81, while that for the simple CT and the iterative CT were .79 and .82 respectively. Although the accuracy rates were not that different overall, the precision of the iterative CT model was much better, classifying an additional 20% of cases into high- or low-risk groups (Steadman et al., 2000). The iterative procedure may be beneficial in increasing the classification of cases into more precise outcome classes because the re-analysis in the second iteration is based only on a subset of cases from the original sample. This new focus results in a different distribution of risk factor characteristics with which the statistical procedure has to work. Thus potent relationships with the outcome measure could be uncovered that did not exist in the total sample. For instance, while the presence of high psychopathy combined with a history of childhood victimisation and substance abuse history classified approximately one-half of the high-risk offenders, the remainder were not all classified until iteration four (Steadman et al., 2000).

The iterative (I)CT technique was subsequently replicated using a large criminal justice sample of released prisoners (Silver, Smith, & Banks, 2000). Silver et al. (2000) went on to cross-validate the model since it had not been subjected to this previously. They found a greater degree of shrinkage, or reduction in performance when the device was applied to a new sample, using the CT and the ICT models than they found for their logistic regression model whether this was iterated or not. This was unexpected because the process of iteration might have increased the risk of overfitting by making the model more reliant on the relationships in the data. Instead it seemed that the CT process of successive partitioning itself may disproportionately capitalise on chance relationships in the data.

To improve the accuracy of ICT models Silver et al. (2000) suggested that a 'multiple models' approach might be beneficial. Consistent with meta-analytic theory in which combining results from multiple studies is held to produce a more reliable estimate of the true predictor-criterion relationship, cases scoring as low- or high-risk across several models could be classified confidently into the relevant prediction category. It would also allow each model to emphasise a different underlying causal process (e.g., a prooffending model emphasising anti-social personality disorder versus a desistance model emphasising supportive relationships). This had been indicated by the differences between the clinically feasible (Monahan et al., 2000) and the empirically optimal (Steadman et al., 2000) ICT models. These models produced comparable AUC values but the predictions correlated with one-another only modestly (.52). Thus each model appeared to tap into an important, but different, interactive process relating to violence (Banks et al., 2004). Combining discordant scales into one super-model with scale scores as items in the super-model is a solution that has also been suggested by Vrieze and Grove (2010).

Multiple models research on the MacArthur data found that combining the clinically feasible ICT and the empirically optimal ICT into one model produced an overall AUC of .83, i.e., higher than either model individually. This was extended by using only the clinically feasible variables and combining this model with nine additional models each forced to use a different key predictor as the starting point for recursive partitioning. This combined model produced an overall AUC of .89 (Banks et al., 2004). Results held up under cross-validation in which risk classification scores for a combined model based on approximately 1,175 cases was applied to nine additional unseen samples each of similar size (Silver & Chow-Martin, 2002). With the exception of the model predicting imprisonment within one year, which had the lowest base-rate (6%), a 'modest' degree of shrinkage was associated with each of the other three outcome measures. The MacArthur risk studies have demonstrated that models designed to detect interacting variables can be developed to predict recidivism reliably. Predictive accuracy for these emerging models (AUC > .80) appears better than seen in mainstream ARAs (AUC < .80).

2.7 Conclusion and Recommendation

This chapter reviewed the key factors associated with recidivism risk in psychological research, how these have contributed to risk assessment practice, and the predictive accuracy of the extant risk scales. The review of risk factor variables indicated the importance of static factors in the prediction of recidivism but suggested that, although reliably measured, these may be insensitive measures masking wide variation. Research has identified the potency of dynamic factors, particularly anti-social attitudes, which relates to poor problem-solving and low motivation for treatment. Although these attitudes are hard to measure, partly due to offender dissimulation, they may be reflected in 'deviant lifestyle' variables which may discriminate time-limited from more persistent offenders. Deviant lifestyle may be measured by a combination of the presence of high criminal history, substance abuse, negative associates, and low employability, including interactions between these variables. Indeed among some offenders, being employed or having positive relationships may delay or protect against recidivism. Among certain groups such as females, these factors may be more important than general criminal history. Such findings attest to the potential for recidivists and non recidivists to present overlapping but subtly different data patterns.

The importance of dynamic factors in research and clinical intuition has led to their inclusion in risk assessment practice. Unstructured clinical judgement is typically inferior to structured assessment, in which specific risk factors are included based on theory and empirical evidence linking them to the criterion being predicted. These structured clinical assessments have been found to perform as well as the leading risk assessment measures which are actuarial and based mostly on static criminal history variables. The review of the predictive validity of the measures showed that all measures achieve a moderate level of accuracy, predicting at or below an AUC of .75 (Coid et al., 2009; Kroner & Mills, 2001; Yang, Wong, et al., 2010). Despite the different construction samples and conceptual bases of the different measures, the risk measures are essentially interchangeable (Coid et al., 2011; Kroner et al., 2005). Thus neither the inclusion of dynamic variables in clinical assessment nor their inclusion within actuarial assessment has advanced the field further than the level of accuracy seen with the purely static measures.

The apparently limited influence of dynamic factors in risk prediction may relate to existing methodology for statistically combining the variables which has been strongly focused on simple weighting schemes (e.g., Nuffield, 1982) and linear regression analysis (e.g., Brown, 1978). These conventional models assume independence among the predictor variables, which is unlikely. Alternatively it may be because recidivism risk is typically measured over long follow-up periods, in which time some dynamic variables may have changed several times and are therefore more weakly associated with the criterion than static factors (Coid et al., 2011). Accurate risk prediction therefore needs to occur within a context of i) high measurement error within the predictors and on the criterion measure, as well as ii) inter-dependencies among the predictor measures. These conditions are known to impair the accuracy of conventional statistical methods when applied to a validation sample due to their assumption of equal between group variance in the predictor variables (Gottfredson & Moriarty, 2006).

Yang, Wong, and colleagues (2010) suggested that areas of development should include studies of CT and connectionist models, capable of detecting potential interactions. CT models were reviewed in the present chapter and were shown to perform similarly in the prediction of recidivism to actuarial models. While repeatedly iterating the CT model over the unclassified cases improved the precision of the final estimates, it did not substantially raise the overall AUC accuracy under cross-validation. Combining two different CT models did raise accuracy however, and extending the number of combined CT models to nine increased accuracy slightly further (Banks et al., 2004). This work suggests that offender data contains many sub-models and that this process may be necessary for advancing accuracy beyond the plateau experienced using conventional statistical methods.

Connectionist modelling may be a fruitful approach given the potential for multiple contingent relationships within offenders' data and related methodological problems including unreliable predictor and criterion measures. Its potential will therefore be explored in the remainder of this thesis. The connectionist approach will be introduced in Chapter 3 where its performance will then be reviewed across a range of fields with data that is characterised by the problems discussed above. The contribution of any prior applications of the approach to offender data will also be evaluated and discussed.

CHAPTER 3

3. Systematic Review of the Application and Development of Connectionist Models in Risk Management Systems

3.1 Introduction

This chapter aims to provide a systematic review of the real-world application of connectionist modelling (introduced below). A 'systematic' review methodology has been adopted in order to apply "scientific strategies, in ways that limit bias, to the assembly, critical appraisal, and synthesis of all studies that address a specific clinical question" (Cook, Mylrow & Haynes, 1997, p.376). Due to the sheer breadth of possible applications it was felt that a consistent research strategy should be used to locate studies of potential relevance. The final set of materials to be included in the review was therefore selected on the basis of predefined inclusion and exclusion criteria. The premise was that this method is transparent and therefore understandable and replicable.

This chapter begins by giving a brief description of connectionist modelling and the rationale for conducting a systematic review in this topic area. A detailed breakdown of the review's methodology is then provided, followed by a narrative summary of the results. The discussion brings all the conclusions together and focuses on drawing out recommendations for ways to address the concerns of relevance to operational offender data.

3.1.1 Description of connectionist modelling.

A connectionist model is a computational, mathematical model for information processing based on excitatory and inhibitory connections (McCulloch & Pitts, 1943). The first use of the concept as a learning associator is attributed to Frank Rosenblatt (1958), who used contemporary knowledge of neurophysiology to create a mathematical model to simulate information processing in the human brain. This early design was called a 'perceptron' and today's more advanced models are variously referred to as 'multilayer perceptrons', 'artificial neural networks', or 'connectionist models'.

Connectionist models are characterised by a multi-layer structure of interconnected basic computational processing elements (nodes) performing simple mathematical tasks. A typical model will have at least three layers (see Figure 3.1). The first layer consists of input nodes. Each node in this layer can be thought of as one variable (a covariate in the terminology of regression). The second layer consists of hidden nodes. Hidden nodes are internal representations within the model that mediate between the input layer and the final layer. These nodes act as 'feature detectors' by performing transformations in ways that emphasise certain distinctions within the input pattern. The third layer consists of the output node(s). The output node represents the desired classification (e.g., likelihood of recidivism).

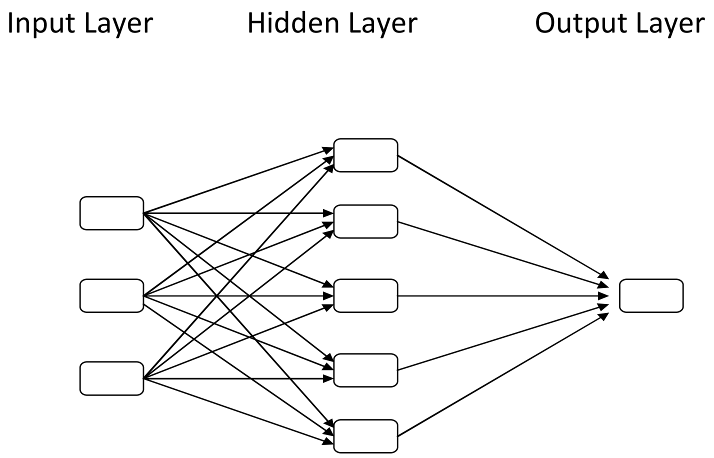


Figure 3 1. The design of a simple connectionist model with forward flow of activation from the input nodes to the output node.

The activation of units in the input layer reflects the characteristics of the case being processed (e.g., items on a risk measure). The summed activation value is then fed forward along weighted connections to the output layer via the hidden nodes. At each stage an activation function converts the net (weighted) input to each node into its activation value which is propagated forward through the outgoing connections. The classification decision is made on the basis of the activation of the output node. A network without hidden nodes, with inputs connected directly to a single output, would be identical to a multiple logistic regression model with main effects and no interactions.

In a connectionist model, all interactions are discovered in a data driven way, including nonlinear association rules.

The model is trained in a 'supervised' process involving repeated exposure to known input-output patterns. The difference between the activation signal and the target output is used to adjust the weights on the network's interconnections in a process of error backpropagation (Rumelhart, Hinton, & Williams, 1986). Weights are recalculated after each observation until all observations have been processed and training stops when the error between the target and the actual output values has reached its minimum. The learning of the network is therefore stored in the weights of the connections of the final trained model that produced an output value for each case as close as possible to the target output. To validate the model the trained network is tested on a number of separate cases not used during training (e.g., a split-half sample). The independence of this sample from the training data ensures an unbiased estimate of the performance of the connectionist model.

Connectionist modelling is new to the field of forensic psychology. In other fields however, such as engineering and some areas of medicine, the method has a long-standing history (Dayhoff & Deleo, 2001). Progress has since been made in a range of different behavioural applications and these are reviewed using a systematic strategy in the results section below. There is a growing need in criminology and forensic psychology for defensibility given the limitations of clinical assessment (see Chapter 2). One aim of the review therefore was to learn lessons from related fields that may assist in applying and developing the method on offender data.

3.1.2 Objectives of the systematic review.

The present review focused upon published work related to methods of operational risk prediction. Specifically, it aimed to review the application of connectionist modelling in social / human settings. This overall aim was broken down into its component objectives. First, the emphasis on 'social' rather than 'agricultural' or 'biochemical' settings for example, was important to ensure that the material would yield lessons relevant to the offender data. Given the large, complex, incomplete, and subjective nature of this data, the review was particularly interested in solutions to design issues related to tackling these problems. Second, the application of a new method is only meaningful in comparison to the simultaneous use of a pre-existing method. A further objective of the review was therefore to appraise the relative effectiveness of connectionist modelling compared to standard statistical methods and / or clinical (non-statistical) predictions.

3.2 Method

3.2.1 Criteria for inclusion and exclusion of studies in the review.

The aim and objectives of the review discussed above implied a number of criteria for inclusion / exclusion. First of all, to be included in the final set of studies for review the material had to evaluate the application of connectionist modelling. Therefore studies that referred to connectionist modelling without actually employing the method were excluded. Second, following the review's objectives, material needed to be relevant to operational offender data, addressing data problems experienced in criminal justice settings including numerous cases or variables, low base rate for the event, and data that is subjective, incomplete, and time-varying. Studies of connectionist modelling that did not clearly apply to an operational setting or were not being developed to address these data problems were excluded. Third, all studies that did not relate to social or human behavioural problems were excluded from the review. This was necessary in order to focus attention on scientific research that was relevant to the prediction of human behaviour. Fourth, studies of connectionist modelling that did not report a comparison with a prediction from either a clinical or a statistical methodology were excluded. Although this restricted the number of studies in the final set, this latter criterion focussed attention of the review on those studies that showed interest in the *relative* performance of their connectionist model.

The final two criteria related to wider needs for the research. The need to focus on contemporary research meant that studies were excluded if they had a publication date earlier than 1985. Similarly the need for the investigator to fully understand the research meant that the study could only be included if it was written in English (translations were included). This was a pre-requisite for inclusion in the review and as such was considered first. The sequential process of study selection is summarised in Figure 3.2.

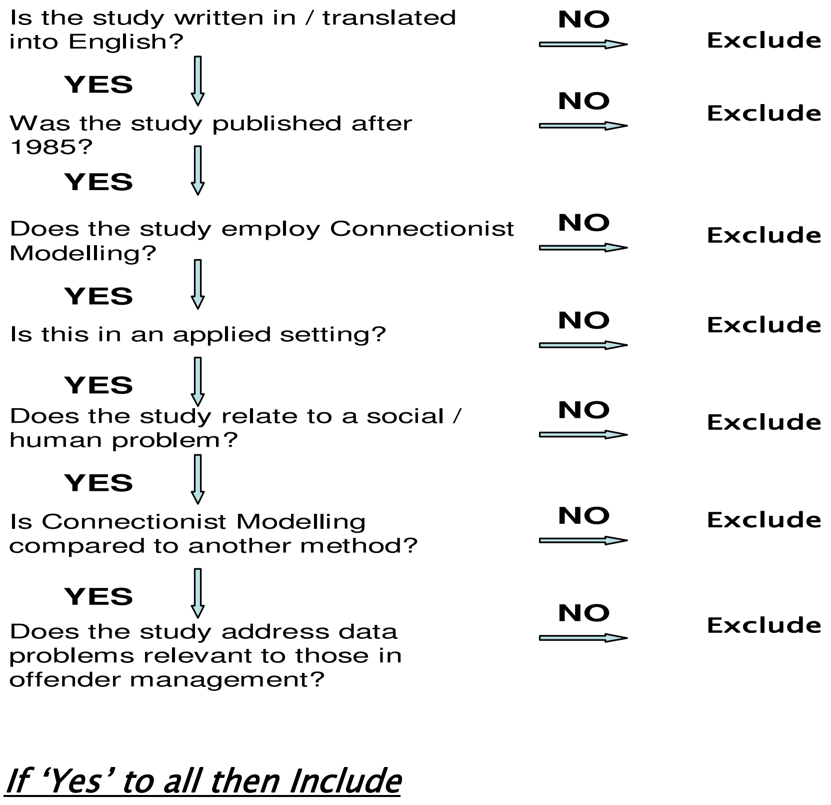

Is the study written in / translated into English?
NO
Exclude
YES
Was the study published after 1985?
NO
Exclude
YES
Does the study employ Connectionist Modelling?
NO
Exclude
YES
Is this in an applied setting?
NO
Exclude
YES
Does the study relate to a social / human problem?
NO
Exclude
YES
Is Connectionist Modelling compared to another method?
NO
Exclude
YES
Does the study address data problems relevant to those in offender management?
NO
Exclude
If 'Yes' to all then Include

Figure 3.2. Criteria specified for study selection

3.2.2 Search strategy for identification of relevant studies.

A systematic search strategy was selected, not unlike that espoused by the Centre for Reviews and Dissemination (CRD, 1996) for systematic reviews of the effectiveness of interventions. Since the effectiveness of a statistical method rather than a therapeutic intervention was the subject of the review, the search strategy was customised.

First, the objectives of the review were specified along with the inclusion criteria (see above). Then the databases to be searched and search terms were specified (see below). After the electronic search yielded the titles of the qualifying studies these were reviewed individually for relevance according to the required selection criteria. Finally each study remaining in the review was screened in more detail by reading the article's abstract. The review of titles and abstracts was done by the reviewer without additional support. Study selection was therefore not done in duplicate by two independent raters. Study quality was not emphasised in the selection of material due to the need for the review to examine the performance of connectionist models using imperfect operational data.

3.2.2.1 *Search terms*.

Two different searches were conducted to capture the relevant material from across fields. The parameters used for searches A and B are set out in Table 3.1 below.

Table 3.1

Search Terms and Parameters

	SEARCH A	SEARCH B
Search Term #1	(artificial neural network*)	Connectionis*
Search Term #2	And appl*	And appl*
Search Parameter #1	Find words in title, or find Term #1 in key words	Find words in title, or find Term #1 in key words
Search Parameter #2	Publication year from 1985-2011	Publication year from 1985-2011

3.2.2.2 *Resources searched*.

The following electronic databases were selected for interrogation: PsychINFO, IBSS (International Bibliography of the Social Sciences), MEDLINE, ISI Web of Knowledge, SCOPUS, ASSIA, Criminal Justice Abstracts, Science Direct, EDINA BIOSIS. The Ministry of Justice website and the Home Office's 'Research, Development and Statistics' archives were also separately searched in an attempt to locate other relevant material.

3.2.2.3 *Description of comparison statistical methods used*.

The review sought to compare connectionist modelling with existing measures of applied prediction. The main methods of prediction for comparison with connectionist modelling in the component studies were multiple regression analysis (MRA), discriminant function analysis (DFA) and logistic regression (LR). The aim of MRA is to derive an equation relating a criterion variable and several predictor or explanatory variables. The basic idea is that the mean of a criterion variable (e.g., vocabulary size) lies on a straight line when plotted against values of an explanatory variable (e.g., age). It is then possible to use the derived equation relating average vocabulary size score to age, to predict vocabulary size for different ages. Regression coefficients give the amount of change in the criterion variable associated with a unit change in the corresponding explanatory variable (conditional on the other explanatory variables in the model remaining unchanged). The end goal is to arrive at a set of values for the regression coefficients which make the values of the criterion variable predicted from the model as close as possible to the target criterion values. A measure of the amount of variance explained by the model is also given, along with a statistical test of the level of improvement resulting from fitting the model, relative to the inaccuracy that still exists within it.

MRA comes with certain assumptions (see Field, 2005 for more on the assumptions). The most prominent assumption is that the criterion variable is normally distributed with a mean whose relationship with the explanatory variables is linear (i.e., follows a straight line). This also implies that the criterion variable should be continuous rather than dichotomous. In addition the criterion variable should not be totally dependent on the values of the explanatory variables, nor should the explanatory variables be too highly inter-correlated because this limits the ability of the individual predictors to explain variance on the criterion variable.

DFA and LR analysis are both linear statistical procedures and therefore are similar in many respects to MRA. The important difference is that these techniques can be used to predict category membership rather than a continuous score. LR achieves this by computing the log-odds of the predicted values to give a probability score for the dichotomous outcome. Thus LR gives the probability that a case will belong to a particular outcome category. DFA on the other hand calculates the variates associated with each outcome category using a formula that maximises the differences between the groups. DFA assumes that the predictors are normally distributed, while LR makes no such assumption. Like MRA both DFA and LR seek to explain the criterion variable on the basis of a single equation or function separating the predictors and therefore produce weak or unstable solutions when the predictors are highly inter-correlated. Given that many outcomes of interest are categorical these techniques are both employed relatively frequently in the studies included in the review.

3.3 Results

The quantitative results of the search strategy outlined at 3.2.2 above are summarised in Figure 3.3 below. This shows that a total of 11,064 articles were identified by the search of electronic databases. Data were downloaded into an EndNote software library system. No relevant articles were found in the Home Office 'Research, Development and Statistics' archives, and only one was found on the Ministry of Justice website.

Screening of titles for relevance reduced the number to 532 articles. This was further reduced to 172 following screening of abstracts and removal of duplicates. Upon reading the full article it was apparent that a further 13 articles were not primary research, and 76 did not address the data problems relevant to the present study. Thus 83 separate studies remained following elimination (0.75% of the initial total). An additional 4 closely related studies, found after a hand search of reference lists, were added to this to give a final total of 87 studies for review.

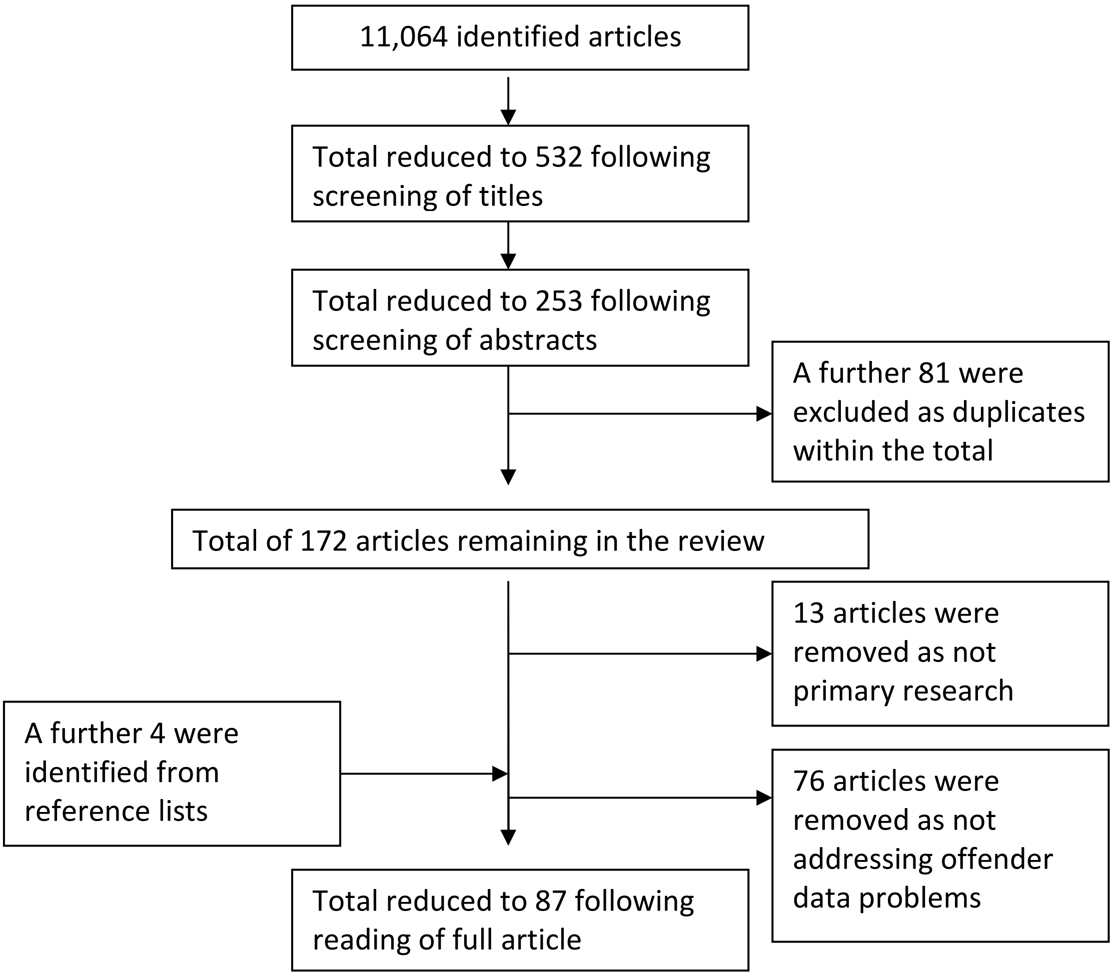

11,064 identified articles
Total reduced to 532 following screening of titles
Total reduced to 253 following screening of abstracts
A further 81 were excluded as duplicates within the total
Total of 172 articles remaining in the review
13 articles were removed as not primary research
A further 4 were identified from reference lists
76 articles were removed as not addressing offender data problems
Total reduced to 87 following reading of full article

Figure 3.3. Flow chart showing the process of elimination

3.3.1 Summary of included studies.

The 87 included studies related to fields ranging from predicting school drop-outs to predicting the behaviour of drivers on the motorway. Medicine was the dominant field, comprising over one-half of included studies. Approximately ten different fields are represented within the review.

As mentioned under 'objectives of the review' (section 3.1.2 above) the data problems against which applications of connectionist modelling were selected were those which are relevant to offender data. The categories of data problems were:

- Large data-sets
- Numerous predictor variables
- Missing / incomplete data
- Low base-rates of occurrence
- Time to event modelling
- Narrative / subjective data

Below each of these categories in turn is a focus for discussion of developments made in the identified studies. In summarising the results a narrative description, rather than a quantitative meta-analysis, is provided due to the lack of consistent outcomes between studies for measuring accuracy.

3.3.2 Studies involving large data sets.

In statistical modelling it is generally considered good practice to use large, representative samples. Overfitting is a particular danger with smaller data sets. This is the tendency for models to 'over-learn' or to memorise the precise characteristics of the training data and thereby reduce the ability to generalise to new samples. Large well-prepared data sets are more likely to show up complex interactions and non-linear relationships, if present, while high levels of measurement error or noise may obscure the benefit of statistical weighting schemes upon cross-validation (Dawes & Corrigan, 1974). Although the effect of noise will be specifically investigated in a later section of the chapter, it is considered important to review the success of connectionist models relative to linear models in real-world applications involving large data sets. Such practical applications are likely to reflect the operational reality of offender data. A total of 12 studies are included here for review under the heading ‘large data sets’. Each of these had an initial study population of at least 1,000 cases.

Various methodological steps have been taken to avoid overfitting. The general approach taken has been to assign the sample randomly to training and testing (e.g., Finne et al., 2000; Ravdin et al. 1992; Song, Mitniski, MacKnight, & Rockwood, 2004). Testing is done on a reserved part of the original sample to evaluate the performance of the model and monitor error levels. When error stops reducing, training is halted. In this way experimenters select the best model for later evaluation on an independent testing set. Palocsay, Wang, and Brookshire (2000) employed ‘Neuroshell 2’ (Ward Systems Group, 1996). To avoid requiring the researcher to decide when to stop training, Neuroshell 2 contains an option which automatically reserves a portion of the data for monitoring and uses this during training to compute the optimum point to save the network based on its performance on the monitoring data.

Some researchers (Finne et al., 2000; Flaherty & Patterson, 2003) trained their model using Bayesian regularisation, which involves adding a penalty term to the error function. This helps avoid overtraining the model by limiting the size of the connection weights. In Finne et al. (2000), models were trained on all cases bar one and tested sequentially on the one case that was totally unseen in the training phase. This procedure was repeated until every case in turn had been the unseen case. This ‘leave-one-out’ cross-validation method, has the advantage of maximising the amount of training data and should ensure that the model has learned the features of the cohort in the screening set.

A similar approach, 10-fold cross-validation (Stone, 1974), was used in two studies involving large data-sets (Alonso-Betanzos, Mosqueira-Rey, Moret-Bonillo, & del Rio, 1999; Ciampi & Zhang, 2002). In this approach, the cases in the data-set are randomly divided into 10 mutually exclusive test partitions of approximately equal size. Upon selecting a test partition, the remaining cases are independently used for training, and the resulting model is tested on the test partition. Following this procedure the first test partition is returned to the data-set and a new test partition is selected. This is repeated until each of the 10 ‘folds’ have been tested. The average correct classification rate is then taken as the performance measure. This method is claimed to be more effective than split-sample in controlling overfitting (Tourassi & Floyd, 1997). Tourassi and Floyd (1997) showed that, as the size of the training set increased so did the predictive accuracy. Since the estimate was based on fewer and fewer test cases this resulted in the need for a greater number of randomisations of the data as the size of the test set decreased in size (to reduce the variance). 10-fold arguably strikes an appropriate balance between perturbing the sample too much thereby compromising training (by taking out too many cases e.g., split-half testing), and perturbing the sample too little, making the cross-validation suspect (by testing on too few cases).

A number of other model parameters are also open to manipulation. These include the learning rate, the momentum, and the model's structure. Some studies sought to 'tune' these features experimentally (Betechuoh, Marwala & Manana, 2008; Caulkins, Cohen, Gorr, & Wei, 1996; Palocsay et al., 2000; Song et al., 2004). Betechuoh and colleagues used a 'genetic algorithm' to find the optimum number of hidden units for their connectionist model. In a genetic algorithm model parameters are coded as 'chromosomes' which are evaluated and reinforced according to their average predictive accuracy within the training data.

The study by Caulkins et al. (1996) was one of the few to consider the merits of connectionist modelling in predicting criminal recidivism. These authors developed their model on a previously studied large data-set (Gottfredson & Gottfredson, 1980, 1985). In tuning the parameters of their connectionist model, Caulkins et al. used a grid system in which models with different combinations of the available parameters corresponded to grid points in the space of the system parameters. Each grid point represented a separate model with a different configuration of aspects such as number of hidden units, learning rate, and type of training algorithm. These models were each trialled and a final model was developed based on performance on a tuning data set. This 'tuning' data was later returned to the training data to enable the final model estimation procedure on the full training data. Caulkins and co-workers found an effect of the learning rate and the number of training iterations, but no benefit from varying the number of hidden units.

Palocsay et al. (2000) did however find better results depending on the number of hidden units employed, suggesting the existence of non-linearity and/or complex interactions in the data that may be suitable for analysis with a connectionist model. The number of hidden units was varied from 5-50 and the corresponding training and testing results were examined. Two separate configurations produced equal best results so the researchers adopted the least complex variant. Since the initial values on network weights are normally set to ‘random’ Palocsay et al. then trained their selected model using 50 different random number starting weights. The average performance was taken for the results. The Palocsay et al. study is of interest alongside the Caulkins et al. study since both used offender data to predict recidivism (see also Yang, Liu, et al., 2010).

3.3.2.1 *Summary of results.*

Table 3.2 and Figure 3.4 summarise the results of all of the studies included in this section of the review. Model accuracy relates to predictions made on the testing data where cross-validation on separate data was performed. This applies to all studies bar one (Betechuoh et al., 2008). The results indicate a benefit of sample size for model accuracy: Figure 3.4 indicates a positive relationship between sample size and the performance of connectionist models. There also appears to be a relationship between the method of sampling and performance. The outstanding result is that by Betechuoh et al. (2008). This study however did not externally validate its model and therefore it is not clear whether the model would be able to generalise beyond the training cases. If the model had memorised the training cases rather than the principles of the data, this would preclude accurate generalisation.

Table 3.2

Predictive Accuracy of Studies with Large Data-Sets

Author(s)	Field	Size of training sample [% of whole]	Method	Accuracy of model	Accuracy of comparator
Alonso-Betanzos et al. (1999)	Medicine	167[1] [94%]	10-fold	.64	.59 (‡)
Betechuoh et al. (2008)	Medicine	16,150 [100%]	Genetic Algorithm	.92	.80 (‡)
Buzatu et al. (2001)	Medicine	1,600 [85%]	Split sample	.78	.77 (‡)
Chun et al. (2007)	Medicine	1188 [23%]	Split sample	.67	.71 (‡)
Ciampi & Zhang (2002)	Medicine	• 936[2] [64%] • 1,739 [66%]	10-fold	Error rate of model improved on LR (‡) by • .17 • .16	
Finne et al. (2000)	Medicine	656 [98%]	Leave one out	.56[3]	.52 (‡)
Song et al. (2004)	Medicine	4,200 [49%]	Split sample	.86 (AUC)	.62 (AUC)
Stephan et al. (2007)	Medicine	656 [52%]	Split sample	.83 (AUC)	.85 (‡) (AUC)
Marshall & English (2000)	Child protection	9,084 [70%]	Split sample	.79	.87 (‡)
Flaherty & Patterson (2003)	Child protection	492 [23%]	Split sample	.63	.61 (‡)
Caulkins et al. (1996)	Criminal Justice	2,385 [70%]	Split sample	.68	.69 (‡)
Palocsay et al. (2000)	Criminal Justice	1,357 [24%]	Split sample	.66	.64 (‡)
Yang, Liu, et al. (2010)	Criminal Justice	827[4] (66%)	Split sample	.66 (AUC)	.58 (‡) (AUC)

Note. AUC = Area Under the receiver operating characteristic Curve; ‡ =linear statistical; ∞ = clinical

[1] This sample contained information from 3,209 cases

[2] Ten data-sets were studied in Ciampi & Zhang (2002): only two samples are classed as 'large data-sets'

[3] Accuracy level at 90% sensitivity

[4] Sample trained on variables from the HCR-20 risk measure as this was the largest sample.

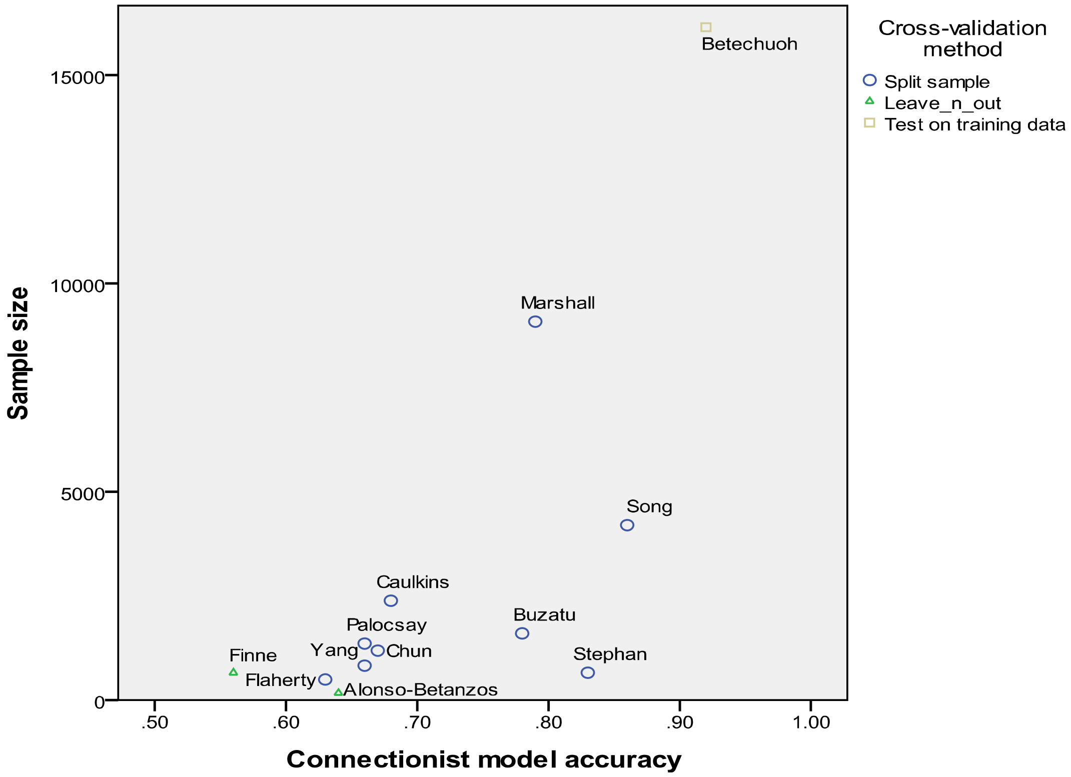


Figure 3.4. Relationship between sample size and model accuracy

Studies using the split sample design seem to have produced better rates of predictive accuracy than have studies employing leave *n*-out cross-validation. This may be an artefact of the number of training cases however, since the latter method is normally selected in order to maximise the size of the training set. Figure 3.5 splits the performance of connectionist models by the sampling method to examine this issue further. Within each main sampling design, the connectionist model's performance is compared to that of the comparison method. Both studies using leave *n* out showed improved performances with the connectionist model over the comparison models (LR and DFA). Differences were less consistent in the split-sample design studies. Leave-one-out ensures that the results reflect performance on every case in the population. Although this means that the sampling bias will be low, it also means that the variance will be high since the model will reflect the construction sample. This may result in poor generalisation to an external sample. Only one study in the review applied a model optimised by leave-one-out to a new sample (Stephan et al., 2007).

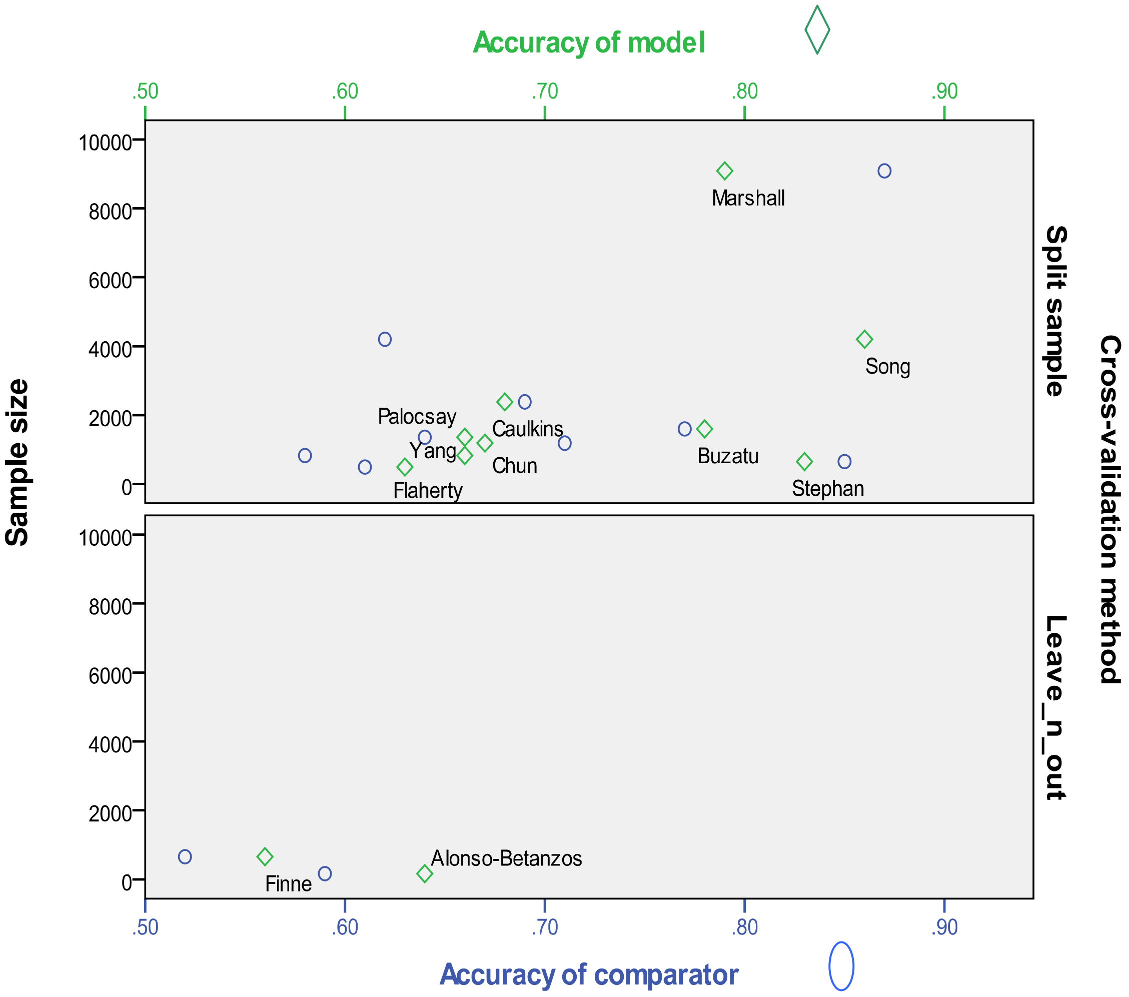


Figure 3.5. Comparison between models within sampling design

In Stephan et al.'s (2007) study the original model by Finne et al. (2000) was applied to a cohort in an altogether separate screening setting and compared to LR and an alternative connectionist model developed with a split sample design (590:66). Results revealed non-significant differences between the connectionist methods, but a significantly lower AUC value compared to the LR approach. The authors suggest that the moderate sample size may have impeded connectionist performance, although there were also differences in the criterion base-rate between the training and testing settings (see section 3.3.5 for methods to address this). Another external validation study found similarly in favour of the regression-based approach (Chun et al., 2007), although here the connectionist model was developed abroad while the regression-based method was developed on historic cases assessed at the test setting, giving it an advantage.

One study developed an approach combining the strengths of both connectionism and MRA (Ciampi & Zhang, 2002). Ciampi and Zhang were able to show a significant improvement in reducing the error rate produced by MRA by initialising the connectionist model with the coefficients corresponding to the MRA model. As the authors point out, this does not constitute a fair comparison between the two models but does indicate a basis for using one method to improve the other.

The clearest difference between the connectionist model performance and that of the comparison measure, were in the studies by Song et al. (2004) and by Marshall and English (2000). Song et al. developed a model to compare with the existing measure, an un-weighted index in which the presence of risk factors increased the risk score in equal units. The superiority of the connectionist model was consistently evident at the optimal AUC value over ten simulations. Marshall and English developed their model using three internal testing sets and then applied the fitted model to a separate sample. They compared their results against a model developed using LR analysis. The impetus for the research was accurate identification of criterion occurrences against a rare outcome. Although total accuracy favoured the linear statistical approach, the results showed the connectionist model to be a better performer due to its greater sensitivity in identifying true positives, with only slight deterioration on true negatives. This study will be referred to in section 3.3.5 (low base-rates).

Finally, the results pertaining to Caulkins et al. (1996), Palocsay et al. (2000), and Yang, Liu, et al. (2010) are worthy of discussion given that each examined connectionist modelling using offender data. Caulkins et al. found consistency in levels of accuracy achieved between training and testing data. All models performed to a similar standard and the authors concluded that there was as yet no apparent advantage to the connectionist approach. This was attributed to the lack of discrimination ability amongst the offender patterns: many offender patterns in the data had nearly identical numbers of recidivists and non-recidivists. They called for greater attention to theory building to develop improved measures of predictive factors.

Palocsay et al. (2000) studied prisoner release data used previously by Schmidt and Witte (1989). On internal validation, using a portion of offenders from the same year, and on external validation using an offender cohort from a later year, Palocsay et al. found that the connectionist model predicted significantly more outcomes successfully.

This also held when recidivists alone were considered. For non-recidivists the LR model was better on internal validation, but equivalent to the connectionist model under external validation. Although Yang, Liu, et al. (2010) found a non-significant overall difference between approaches, contrary to Palocsay et al. (2000) they similarly found increased ‘shrinkage’ between training and testing performance under regression analysis. This particularly occurred among the larger sub-group, male offenders. A reversed effect was found however when the data were re-trained on different age groups. Here it was evident that, although the connectionist model remained the strongest in predictive accuracy on the test set, its shrinkage was greater, particularly among smaller sub-groups. An inadequate sample size for training the models for each age group seemed to lead to low true positive rates in the test sample (Yang, Liu, et al., 2010). Yang and colleagues suggest that the flexibility of connectionist models may make them better suited to the detection of small effects when large amounts of data are available.

In conclusion the ability to retain a sufficient sample size for analysis has appeared to be an issue in the studies so far reviewed. Although each study had an initial sample of at least 1,000 cases, not many retained this number of cases for analysis. This appears to have related mainly to the problems of missing data and noise; these will be specifically reviewed later in the chapter (see section 3.3.4). Notwithstanding, a number of large sample studies were available for review and a trend for improved accuracy with increasing size of sample was apparent. Thus the importance of a sufficiently large sample in the development of a connectionist model is a key factor to keep in mind when considering the performance of connectionist models in later sections of this chapter. However, in line with a review of studies on medical data sets by Sargent (2001), the present review does not relate the performance of connectionist models solely to the size of the sample. Other important considerations are parameter selection, data sampling during validation, and differentiation of patterns between groups. Without a sufficient number of separable patterns on each class of the output variable it will be very hard for any statistical method to differentiate successfully between cases. Since many types of criminal recidivism occur at a low rate (or go undetected), responses to this specific issue will be considered in a separate section (section 3.3.5).

3.3.3 Studies involving numerous predictor variables.

Overfitting the training data is likely to be a particular danger when the number of connections in the model is much greater than the number of variables. The way in which the number of connections varies with an increase in the number of predictor variables can be illustrated by showing the number of paired interaction terms in a 20 variable model and then comparing this with a 40 variable model (Table 3.3).

Table 3.3

Comparison of Paired Interaction Terms Depending on the Number of Variables in a Model

No. of paired interactions = $n_{\text{variables}}$ x ($n_{\text{variables}}$-1) / 2

20 variable model	40 variable model
$n_{\text{variables}}$ x ($n_{\text{variables}}$-1) / 2 20 x 19 / 2	$n_{\text{variables}}$ x ($n_{\text{variables}}$-1) / 2 40 x 39 / 2
190	780

Table 3.3 shows that doubling the number of variables, brings about a disproportionate increase in the number of paired interaction terms needed to specify a function. The number of paired interactions for each variable is one less than the total number of variables. Multiplying this by the number of variables and then dividing this in half gives the number of separate terms. In the example of a model with 40 variables, the number of interaction terms is four times greater than that in a model with half the number of variables.

Although this may be somewhat problematic for the generalisability of a connectionist model, it is a major difficulty for traditional statistical models. Here multiple interaction terms must be specified and entered into a model before running the analysis. The nature of regression equations is to select the smallest number of variables that independently contribute to outcome. Thus variables which are not independently associated with the criterion may therefore end up getting forced out of the equation by the selection process. This may be a problem clinically since important dynamic factors may end up being overlooked.

Studies were identified as relevant to this section of the review if the ratio of predictor variables to training cases exceeded 1:10 (after Concato, Feinstein & Holford, 1993). Consequently a total of 18 studies were identified for review under the heading 'numerous predictor variables'. A further study which does not meet this standard (Frize, Ennett, Stevenson, & Trigg, 2001), but which is of methodological note, is also discussed.

The studies included in this section of the review have all sought, in one way or another, to optimise the structure of their input layer. This may not be surprising given the recognised propensity for overfitting when the number of predictor variables is high and the number of cases is proportionately low. The way in which this has been approached has varied from using standard statistical procedures (e.g., Edwards, Hollingsworth, Zazulia, & Diringer, 1999; Moreno et al., 1995) to using connectionist approaches with and without a genetic algorithm (e.g., Cevenini et al., 2007; Dybowski, Weller, Chang, & Gant, 1996; Ladstätter, Garrosa, Badea, & Moreno, 2010; Risser et al., 2008; Wu, Huang, & Meng, 2008). Other studies have found ways to open the 'black box' by seeking to understand the contribution of each variable within the connectionist model (e.g., Peng et al., 2007; Peng & Peng, 2008; Santori, Fontana, & Valente, 2007).

Connectionist approaches are frequently criticised for not being transparent in a similar way to regression analysis. Since the pattern of activation is distributed in connectionist models, it has not proved possible to isolate the contribution of each variable. However there have been some recent applications of a procedure known as 'sensitivity analysis' (Peng & Peng, 2008; Santori et al., 2007). This method helps evaluate the importance of each of the predictors within the model, by measuring the impact on the predictive error of the model when each variable in turn is no longer available. Santori and colleagues reduced their input layer from 20 units to 8 units, by retaining only the most important variables according to the sensitivity analysis. This improved network performance on the training set from an AUC of .83 to .94.

Cevenini et al. (2007) also found an optimal set of predictors though a connectionist stepwise procedure (see also Hayashi, Hsieh, & Setiono, 2010). To do this they used leave-one-out iterative pruning and then tested the AUC value, for every set of predictors available. For each set of predictors they used 1,000 different random samples generated by the ‘bootstrap’ re-sampling method. In this procedure a sample with which to train the connectionist model is taken from the overall dataset. The bootstrap sample is then returned to the overall dataset, whereupon the model is tested. This procedure is then repeated several (e.g., 1,000) times, to reduce the variance, and the average performance is taken. Cevenini and colleagues defined the optimal number of predictors as the minimum number where no significant difference in AUC pertained compared to the baseline model with the original set of predictors.

One of the most methodologically interesting studies in this category was reported by Dybowski et al. (1996). This study involved 168 training patterns and 157 possible predictor variables. Aware that using all of these clinical fields might diminish model accuracy, Dybowski et al. decided to filter out those predictors least likely to influence outcome. They did this using a decision tree technique, namely Classification and Regression Trees (CART: Breiman et al., 1984). Running CART over the training set allowed Dybowski et al. to determine those predictors that differed greatly in their output class membership, which could then be used to generate a decision tree. Retaining those predictors most clearly associated with the outcome classes reduced the number of inputs from 157 to 11. A linear stepwise method selected 9 variables, only two of which were in the sub-set identified by the decision tree. The 18 unique variables were then entered as inputs into the connectionist models.

Other studies in the review have also used linear statistical procedures to reduce the number of input variables (Edwards et al., 1999; Ladstätter et al., 2010; Moreno et al., 1995). Moreno and colleagues (1995) used principal components analysis to transform their set of correlated variables into a smaller set of uncorrelated ones. This reduced eighty predictors down to fifteen components which nevertheless accounted for 87% of the original variance. Ladstätter et al. (2010) used a self-organising feature map (SOM) clustering technique, to find groups of neurons in the input layer associated with the output nodes. Thus high dimensional data in this study, with 246 input nodes, were mapped into a lower dimensional space containing 100 nodes. The SOM achieves this by isolating ‘winning’ input units which best represent the input-output mapping and adjusting network weights around these to approximate the data distribution.

The review also identified a study by Frize et al. (2001) which aimed to test out the impact of several novel technical approaches. These were i) an additional penalty term to reduce the weights of the least important variables so that their influence is removed from the model; ii) using this weight elimination approach to then assign cases' scores on a variable to one of a pair of 'high' and 'low' nodes, depending on the value of the parameter, to facilitate interpretation of higher- or lower- than normal input values in predicting outcomes; iii) reducing the number of inputs from 51 to 6 variables based on those with the highest weights; iv) variation to the criterion base rate. This latter manipulation involved an examination of model performance with changes to the size of the dominant class on the outcome variable. The impact of all of these techniques is summarised below, with the exception of (iv) as this is returned to in a later section.

3.3.3.1 ***Summary of results*.**

Table 3.4 below summarises the results of the studies in this section of the review. The column showing the 'number of predictor variables' contains the number *after* the procedures described above to reduce the number of predictor variables. Inspection of the table reveals that the majority of studies did manage to reduce the proportion of predictor variables to cases to within the recommended 1:10 ratio. However, where the number of input variables has remained high this has not appeared to harm performance of the connectionist models (e.g., Edwards et al., 1999; Ladstätter et al., 2010; Meccoci et al., 2002; Zou et al., 1996). Accuracy figures in Table 3.4 are generally higher than those reviewed earlier.

Sensitivity analysis for identifying the key predictor variables seemed to be successful in appropriately reducing redundancy. Santori et al. (2007) found an improved rate of accuracy with 8 key variables than achieved with 20 predictors. In particular the positive predictive power (PPP) increased, i.e., the rate of identification of criterion occurrences among cases scoring above the cut-off. The performance of both structures of connectionist model however surpassed that of the LR model in percent correct as well as in the identification of true positives. Of the methods to refine the number of predictor variables in the input layer, prior accuracy rates were not always assessed / reported. Where they were, difference in performance has been positive (Andriulli et al., 2003; Dybowski et al., 1996).

Table 3.4

Model Accuracy with Numerous Predictor Variables

Author(s)	Field	Number of predictor variables [no. of training cases]	Elimination Method	Accuracy	
				Model	Comparator
Andriulli et al. (2003)	Medicine	31 [144]	Genetic algorithm	.89	.53 (‡)
Cevenini et al. (2007)	Medicine	13 [545]	Iterative pruning	.78	.78 (‡)
Dybowski et al. (1996)	Medicine	18 [168]	CART and linear statistical	.86 (AUC)	.75 (AUC)
Edwards et al. (1999)	Medicine	14 [67]	Linear statistical	1.00 .98 (AUC)	.85 (‡) .92 (AUC)
Fukushima et al. (2004)	Medicine	33 [130]	Clinical selection	.97 (AUC)	.96 (AUC) (∞)
Kennedy et al. (1997)	Medicine	53 [90]	None	.92	.81 (‡)
Mecocci et al. (2002)	Medicine	37 [34]	None	.93	.82 (‡)
Moreno et al. (1995)	Medicine	15 [86]	Principal components analysis	.81	.75 (‡)
Peng & Peng (2008)	Medicine	31 [478]	Sensitivity analysis	.85 (AUC)	.79 (AUC) (‡)
Santori et al. (2007)	Medicine	8 [107]	Sensitivity analysis	.87	.79 (‡)
Ladstätter et al. (2010)	Nursing	100 [462]	SOM / clustering	R^2=.45	R^2=.39
Gioftsos & Grieve (1996)	Physio-therapy	242 [36]	None	.86	∞ = "significant difference" (lower)
Hayashi et al. (2010)	Marketing	55 [534]	Iterative pruning	.75	.74 (‡)
Wu et al. (2008)	Education	4 [142]	Genetic algorithm	.82	.82 (‡)
Zou et al. (1996)	Psychiatry	396 [60]	None	.96	.91 (‡)

Note. AUC = Area Under the receiver operating characteristic Curve; **‡** =linear statistical; ∞ = clinical

Depending on the data it may be advantageous to retain variables and the information they contribute to a classification function. Edwards et al. (1999) attributed their accuracy figures to the contribution of a number of inter-correlated variables that were not able to remain in the stepwise LR model. Where there are indications of nonlinearity in the data accuracy should be increased by retaining all variables (Meccoci et al., 2002; Zou et al., 1996). Zou and colleagues for example found their connectionist model significantly outperformed a two-layer model when applying 396 inputs to sixty cases and attributed this to nonlinearity.

Removing variables actually reduced performance in two studies, not considered in this section thus far due to the ratio of predictor variables to training cases not exceeding 1:10 (Price et al., 2000; Song et al., 2004). Song et al. found that the performance error of their model increased when each variable was removed from the input vector. Similarly, Price et al. found that the regression and connectionist models performed at the same level (AUC=.81) using 8 predictors selected by backwards elimination, but after inclusion of all 23 variables only the connectionist model improved (AUC=.96). Since both studies had large sample sizes, the input layer was not disproportionate to the number of cases even prior to elimination of predictor variables. A small sample may lead to overfitting using connectionist models despite refinement to the input layer (Wu et al., 2008).

One method to eliminate unimportant variables in the context of a smaller sample was the 'weight elimination cost function' (Frize et al., 2001). Frize et al. found that accuracy was better than that of the same model without this function. By using only those 6 variables whose weights did not go to zero in the weight-elimination experiment, Frize et al. found that they could improve accuracy from 89% to 91% after only a short cycle of training (130 epochs). This was unexpected as they had thought that reducing the number of input variables from 51 to 6 would also eliminate unknown interactions. They concluded that reducing the complexity of their model increased its generalisation ability by requiring minimal training. The authors also conducted an experiment to evaluate the impact of assigning cases to a high / low input node. This produced worse performance than regular data presentation which the authors suggest may be due to unnecessary additional resources in the model (it required more hidden nodes).

In conclusion connectionist researchers have recognised the value, both practically and scientifically, of minimising the number of predictor variables. This is particularly important where limited sample sizes are being considered. For these reasons a variety of input elimination procedures have been developed. Sensitivity analysis, iterative pruning, weight elimination, and genetic algorithms are all methods which are able to take into account and retain those variables which are important in combination but unimportant in independent relevance to the outcome variable. In small sample contexts this might improve both connectionist and conventional model performance, relative to retaining all variables. Reducing model complexity in this way may encourage a faster rate at which sub-models are discovered (i.e., separate models for certain groups of patterns in the data). When a high number of predictor variables must be considered connectionist models appear to have an advantage over traditional statistical methods, particularly where the number of predictor variables exceeds the number of cases.

3.3.4 Studies with incomplete / noisy data.

Real-world clinical data is likely to contain missing fields of data for each case. Incomplete case data can be problematic for statistical models. The parallel nature of processing performed by connectionist models enables them to accept a certain amount of inaccurate data without a serious effect on predictive accuracy, known as fault tolerance (O'Reilly & Munakata, 2000). This occurs due to greater flexibility because of the multiple paths from the input units to the output unit. The model can therefore learn something from the (compromised) pattern as the information from the input vector is distributed and the hidden representation is less dependent on one piece of information. However, variables missing at training will be of little value when validating the model. Where input data is partially missing, it is possible that this uncertainty would compromise predictive accuracy during validation.

A related problem is that of random feature noise. This is characteristic of operational data in that there will be a degree of variability and error according to the assessment of each variable. This is problematic because errors in the description of cases may confuse the classification rule used for generalisation. Overfitting the data arises where a model has learned this noise during its training with a negative impact on classification results for new cases. The ability to learn decision rules in the presence of random noise as well as incomplete data is therefore a necessity for an operational learning method. This section includes studies wherein noisy or incomplete data was a key concern. A total of 6 studies were identified as relevant within this standard. A further study, excluded late in the search strategy due to the absence of a comparison model, developed an interesting approach to the problem of missing data (Buscema, Mazzetti di Pietralata, Salvemini, Intraligi, & Indrimi, 1998) and will also therefore be discussed.

A number of studies decided simply to exclude all of those cases without complete case data (e.g., Marshall & English, 2000). One study omitted those variables associated with incomplete data to create a model with a smaller input layer (Nguyen, Malley, Inkelis, & Kuppermann, 2002). Another study retained incomplete cases in the connectionist model and compared this with a LR model using the same variables but with fewer cases due to the missing data exclusions (Gonzales & DesJardins, 2002). Gonzales and DesJardins were therefore able to evaluate the benefit of a connectionist approach brought by the ability to retain the additional / degraded cases. This was also examined by Collins and Clark (1993) in the area of white collar crime. These workers studied the effect of incomplete data experimentally by purposely degrading data in one of four predictor variables. Two manipulations were introduced: in the first experiment 33% of values in this variable were replaced with a missing value indicator; in the second study 100% of values in this variable were set to missing. This was done to test the existence of unique functional capability of connectionist modelling over and above traditional methods. Specifically Collins and Clark wanted to test fault tolerance.

Some studies within the review specifically investigated the effect of accepting problematic case examples into their models on the basis that the connectionist performance may be less impeded by this (Fallah-Tafti, 2001; McMillen & Henley, 2001; Nolan, 2002). Nolan (2002) systematically introduced random changes to the values of the predictor variables. He then compared performance using 10-fold cross validation to the performance of DFA. McMillen and Henley (2001) did three analyses, the first with an error-free connectionist model, the second with the data for two of the predictor variables rendered statistically redundant, and the third with a further two noisy variables added. They tested their models on a separate sample from within the same population and compared predictive accuracy to that obtained by LR analysis.

Other studies used connectionist modelling techniques to solve the problem of incomplete data (Abe et al., 2004; Buscema et al., 1998). Abe and colleagues prepared and trained 32 connectionist models each with a different combination of the clinical parameters available. For example, three cases had 9 clinical variables in a single combination: this made up one connectionist model. Abe et al. then applied their 'team' of connectionist models to 96 naturally occurring individual cases. By contrast, Buscema et al. (1998) sought to reconstruct their missing data. They did this by training a connectionist model on a small set of cases with complete data. Using the memorised weights matrix, based on the identity of the complete variables they then asked their model to predict the values of the incomplete variables. After 500 epochs of training the model reconstructed the data relating to the 13 missing variables on the case records. A counter-check against the original complete data-set with the data relating to the same 13 variables set to zero, revealed a mean accuracy of .84.

3.3.4.1 ***Summary of results*.**

Table 3.5 suggests that the inclusion of cases with missing values does not impede cross-validation accuracy (Gonzales & DesJardins, 2002; Nguyen et al., 2002). In Gonzales and DesJardins (2002) the connectionist model achieved a higher percentage correct on the test set despite having been trained on incomplete data containing a lower base-rate of occurrence on the target variable. Although the true positive rate was little different, the false positive rate was improved relative to the LR model. Gonzales and DesJardins (2002) concluded that one of the key advantages of connectionist methods is their ability to determine the relative importance of each of the input values whether they are legitimate or missing.

In Collins and Clark (1993) the connectionist models outperformed DFA especially in larger training samples. With one-third missing data the connectionist model achieved 85% correct, while the DFA model achieved 82%. The one-third degraded connectionist model even outperformed the DFA method that had no missing data. This supports Gonzales and DesJardins (2002) regarding the fault tolerance of connectionist models. Interestingly with 100% missing data in the variable the connectionist model nevertheless correctly classified over three-quarters of cases. The shrinkage estimate in this connectionist model was smaller even than that in the one-third degraded DFA model.

Table 3.5

Model Accuracy with Incomplete / Noisy Data

Author(s)	Field	Method	Accuracy	
			Model	Comparator
Abe et al. (2004)	Medicine	32 Models trained with each of multiple combinations of the data available	.91 (AUC)	.81(AUC) (∞)
Nguyen et al. (2002)	Medicine	Elimination of degraded variables and use of smaller model	.82 (AUC)	.82 (AUC) (‡)
Buscema, di Pietralata et al. (1998)	Psychiatric	Model deprived of key information	.87	- -
Collins & Clark (1993)	Personnel	Variation to extent of missing data	.85	.82 (‡ DA)
Fallah-Tafti (2001)	Transport	'Kalman filter' to remove the effects of noise	R=0.92	R=0.62 (‡)
Gonzales & DesJardins (2002)	Education	Inclusion of cases excluded by LR model due to missing data	.76	.72 (‡ LR)
McMillen & Henley (2001)	Criminal Justice	Introduction of 'noisy' and redundant predictors	.78	.78 (‡ LR)
Nolan (2002)	Computer Science	Introduction of 'noisy' data (20% noise)	.91	.91 (‡ LR)

Note. AUC = Area Under the receiver operating characteristic Curve; **‡** =linear statistical; ∞ = clinical

McMillen and Henley (2001) and Nolan (2002) investigated the effect of noise on performance. Although the performance of the connectionist model and the LR model seemed to be equivalent, McMillen and Henley found that as noise level increased the connectionist models were "substantially more accurate in the prediction of high (drink driving) risk" (p. 15) than the regression models. Connectionist models were particularly adept at predicting which heavy drinkers were a high risk for driving – the key category in terms of risk management. With lower levels of noise in the data, regression analysis provided better levels of accuracy. In support of McMillen and Henley, Nolan (2002) found that his connectionist model significantly outperformed LR and a decision tree as the level of noise increased to 90% in ten percent increments, although similar accuracy was found when the level of noise was 20% as shown in Table 3.5. Across levels of noise the connectionist models ranked higher with an average of .79 compared to .77 using LR. Nolan postulated that the finer granularity of the connectionist model enabled it to converge on logical concepts with greater accuracy than regression analysis (Buscema et al., 1998; Smith & DeCoster, 1998).

In conclusion, studies of the use of connectionist models on incomplete or noisy data suggest that such models may perform similarly to regression analysis under the traditional approach of excluding either the cases or the variables containing the degraded data (Nguyen et al., 2002). The ability to include these cases or variables within large databases is an advantage of connectionist modelling and has been shown to raise the relative performance of these models (Gonzales & DesJardins, 2002; Collins & Clark, 1993). Under conditions of insufficient data, work-around solutions have included using connectionist models to predict the values of the missing variables based on complete case training (e.g., Buscema et al. 1998), and employing a team of connectionist model 'judges' each trained with various data characteristics and applied simultaneously to a problem (e.g., Abe et al., 2004). Good performance in the absence of key data reflects the fault tolerance of connectionist models.

3.3.5 Studies with low base-rates.

There is reason to suppose that the criterion base-rate is likely to be low among some offender samples or when using short reconviction follow-up outcomes as the criterion (see Chapter 2). In their review of risk assessment instruments in child protection, Lyons, Doueck and Wodarski (1996) point out that the assumptions of normality and equal covariance are likely to have been violated in these low base-rate circumstances. This would pose particular problems for linear models but may be less of a difficulty for connectionist models since they do not make assumptions about the distribution of the data. Nevertheless, training a connectionist model with a disproportionate representation of cases across the target variable may cause the network to learn to predict all cases as belonging to the larger group instead of learning factors to discriminate between the groups (SPSS Inc., 1997). Mobley, Schechter, Moore, McKee and Eichner (2000) suggest that the number of cases at each category of the output variable should follow the ‘rule of ten’ (Concato et al., 1993; see section 3.3.3). Hence responders on the target variable arguably should not outnumber non-responders by more than ten to one. Given that base-rates at shorter time intervals or for violence are often in the region of 10%, methods developed to optimise prediction accuracy are likely to be important. This section considers 14 studies wherein classifying rare outcomes was a key concern.

Although criminal recidivism is common in some offenders its occurrence in large sample populations is typically less than .50 within a two year follow-up (e.g., Farrington et al., 2006). The applications of connectionist modelling by Palocsay et al. (2000) and by Grann and Langstrom (2007) focussed on the model parameters and particularly on the number of units assigned to the hidden layer. In Palocsay et al. the number of hidden units was tested experimentally with training done for each of 50 different random starting weights. Grann and Langstrom assigned hidden units using a heuristic suggested by Ward Systems Group (1996). This suggested taking half the sum of the input and output units plus the square root of the number of cases used for training. They separated their data into five sub-sets with a different configuration of train/test sub-sets. Grann and Langstrom set their starting connection weights to 0.30.

Given the importance of accurate identification of cases at risk for the rare outcome, it is not surprising that this has also been a key concern in two studies involving child protection services’ data (Flaherty & Patterson, 2003; Marshall & English, 2000). In order to avoid models learning to classify by default to the larger group, Flaherty and Patterson (2003) opted to equalise the ratio of outcomes on the target variable during the training phase (see also Wichard, Cammann, Stephan, & Tolxdorff, 2008). To maintain the ecological validity of test conditions, they then tested out their model on cases with the actual proportion of target cases. Marshall and English (2000) made no such alteration to the proportion of ‘high risk’ outcomes within their training data (21%). Having realised that the identification of these cases was superior in the connectionist model to the comparison regression model, they sought to pinpoint the origin of this improvement. To explore whether the connectionist model was using the prior probabilities of group membership more effectively, they ran a DFA with the same inputs as the connectionist model and using prior probabilities of group membership. Table 3.6 below shows the results of this procedure.

Other researchers have attempted to influence statistically the training of their connectionist models (Das et al., 2003; Mobley et al., 2000; Price et al., 2000). This has involved attempts to determine the optimal type of network architecture (Das et al., 2003), the optimal time to stop training (Mobley et al., 2000), and the benefit of pattern weighting in which the error function during training is adjusted to give weighting to individual patterns (Price et al., 2000). Price et al. set pattern weights as the inverse of their base-rate (.079).

Since different connectionist models will perform differently due to different architectures, different initial weights, different learning algorithms or different training sets, a selection of studies in the review developed methods to combine different approaches within a ‘committee’ of connectionist models (da Silva, Hernandez & Rangayyan, 2008; Maqsood & Abraham, 2007; Wichard et al., 2008). In a committee machine, the individual results of each connectionist model are combined to achieve better generalisation. Da Silva et al. (2008) employed the AdaBoost algorithm (see Haykin, 1999). This technique gradually builds a committee of models via sequential learning. In each iteration a new model (expert) is added to the committee. A sample of the training set that is misclassified by one model will have its weight increased when it is considered during the training of a new expert model. The probabilistic weights of correctly classified samples are reduced at the next iteration. In the later iterations therefore, the expert models are forced to focus on the difficult samples of the training set. The results were compared with those achieved by a single connectionist model as well as by DFA.

3.3.5.1 *Summary of results*.

The figures for classification accuracy in Table 3.6 are those of the most successful model developed in each study. Where the study has sought to predict one of several levels of an outcome, the accuracy on the level of interest is reported instead. Many studies found an advantage of connectionist modelling in low base-rate samples (e.g., Baxt & Skora, 1996; Das et al., 2003; Marshall & English, 2000; Palocsay et al., 2000). Palocsay et al. (2000) using offender data observed a shrinkage effect on application of the connectionist model to independent data with the same criterion base-rate. The model nevertheless outperformed LR analysis in terms of true positives, true negatives, and overall accuracy. The improvement was statistically significant for each of these with the exception of true negatives. Yang, Liu, et al. (2010) did not find statistically significant differences between models using offender data, although based on the AUC statistic the connectionist model scored higher across all classification thresholds than the DFA model. Interestingly, when Yang et al. added additional ‘crime motivation’ variables into the models the difference in accuracy increased and became statistically significant.

Marshall and English (2000) hypothesised, after the connectionist model outperformed the regression model, that the connectionist model was making more effective use of the group membership prior probabilities information within the data. To test this proposal they implemented an ‘impoverished’ model with no risk factors as inputs. This was designed to show how the connectionist model would assign cases purely on the basis of prior group membership. The results were the worst for any of the models studied, with a TPR of .33 and total accuracy of .60. Aside from the impoverished connectionist model, the performance of the traditional statistical models was clearly inferior to the connectionist approach, including when the prior probabilities of group membership were specified (see Table 3.6). One model did however emerge with classification results equivalent to the developed connectionist model. This was the DFA model with equal prior probabilities in which all risk factors were entered and stepwise variable selection was performed. This achieved comparable accuracy to the connectionist model, albeit with a considerable loss in parsimony (27 predictors as compared to 5). Since fewer variables were required for the same result, Marshall and English concluded that the improvement in connectionist performance is likely to be a better treatment of variable interactions through the hidden nodes in the network.

Table 3.6

Model Accuracy with Low Base-Rate Data

Author(s)	Field	Method	Accuracy			
		(% = base rate)	Model		Comparator	
			TPR	FPR	TPR	FPR
Baxt & Skora (1996)	Medicine	Training = 34% Test set = 7%	.96	.04	.73 (∞)	.19 (∞)
Cazzaniga et al. (2008)	Medicine	Training = 14% Test = 35.2%	.72		.66	
Da Silva et al. (2008)	Medicine	Committee machine	.88 (AUC)		.79 (AUC) (‡ DFA)	
Das et al. (2003)	Medicine	Training = 3% Test set = 6%	.88	.03	.38 (‡ LR)	.28 (‡ LR)
Lee et al. (2007)	Medicine	Prevalence = 16.5%	Not given	FPR=.075	Not given	FPR=.23 (‡ LR)
Mobley et al. (2000)	Medicine	ROC analysis on internal validation set Training = 12% Test set = 19%	1.00	.53	1.00	.89 (‡ LR)
Yamamura et al. (2003)	Medicine	Excluding outlying data from the test set	1.00	.07	.84 (‡ LR)	.23 (‡ LR)
Wichard et al. (2008)	Medicine	Balanced outcomes	.45	.05	.46	.05
		Unbalanced	.45	.05	.48	.05
Price et al. (2000)	Medicine	Pattern weights Prevalence = 7.9%	.96 (AUC)		.81 (AUC) (‡ LR)	
Flaherty & Patterson (2003)	Child protection	Baysean network Training = 50% Test set = 6.3%	.60	.36	.62 (‡ LR)	.39 (‡ LR)
Marshall & English (2000)	Child protection	Adaptive gradient descent Prevalence = 21%	.72	.14	.50 (‡ DFA)	Not given
Grann & Langstrom (2007)	Criminal Justice	Prevalence = 21%	.64 (AUC)		.71 (AUC) (‡ LR) .72 (AUC) (∞)	
Palocsay et al. (2000)	Criminal Justice	Prevalence = 37%	.39	.18	.36	.19
Yang, Liu, et al. (2010)[5]	Criminal Justice	Prevalence = 12%	.64	.31	.54	.63

Note. TPR = True Positive Rate; FPR = False Positive Rate; AUC = Area Under the receiver operating characteristic Curve; **‡** **=**linear statistical; ∞ = clinical

[5] Data in this row are for men and women combined (they are reported separately in Yang, Liu, et al., 2010)

Yamamura et al. (2003) developed the concept of 'predictive ranges' in making difficult decisions. Upon analysis of the incorrectly predicted cases, Yamamura and colleagues realised that over one-quarter of cases had outlying values in at least one of the input parameters. Deeming that the training data did not support accurate testing of variables with outlying values, Yamamura et al. excluded the relevant cases from the testing set. This raised the correct classification rate from .53 to .73. The equivalent improvement for MRA was not reported, although the connectionist model had already significantly outperformed the comparison method despite the existence of unpredictable test cases.

The recent implementation of committee machines also represents a promising approach to the problem of classifying rare outcomes. As described above, da Silva et al. (2008) used a committee machine with a continuous learning process in which the weights from the previous expert are taken forward to the new expert in the committee. Although no difference was found compared to the traditional method in which random initial weights are used at each step, the committee machine did outperform the single connectionist model, as well as the logistic perceptron and linear DFA. After the committee machine the single connectionist model was the most powerful choice for accurate classification. The difference in performance was significant for some sets of models but not for others, with the probability that a committee machine will have an AUC value larger than that of a single connectionist model being .88. This is different to Tsai and Wu (2008),[6] who found that a single connectionist model classifier outperformed a committee machine of connectionist networks. The benefit of committee machines may be clarified further in the following two sections where committee machines have been applied to subjective and time-varying data.

In conclusion it is clear that the generalisability of statistical models can depend on the prevalence of cases on the target variable. A number of approaches have been taken to address the difficulties caused, including balancing the training cases, screening the characteristics of testing cases for predictability, varying the stopping time for training, and manipulating the type or configuration of models used. Given sufficient data on the minority class including cases and variables, connectionist models may be capable of achieving good sensitivity values on rare patterns (Marshall & English, 2000; Price et al.,

[6] Not included in the review due to lack of a comparison statistical model

2000; Yang, Liu, et al., 2010). The use of committees of connectionist models is a promising approach, able to amalgamate a number of variations implied by the particular predictor variables or starting weights selected.

3.3.6 Studies with time to event outcomes.

Modelling individual case trajectories may be important for a variety of reasons. Population statistics may be problematic for accurate case predictions since individuals may have event times that differ widely from the mean. Related to this, the prognostic impact of a predictor variable may vary over time. Hence the presence of a factor may have more significance for an individual after one year than after five years, for example. Different offence types have different lengths of interval before re-offending (e.g., Howard, 2011). Service providers may consider that cases that recidivate within Year 1 are more important than cases who do not relapse until Year 5, for example. Connectionist approaches may be of benefit because they carry no assumptions about the relationship between predictor and criterion variables. Thus time dependencies and missing data can be more easily tolerated in connectionist models than in conventional statistical analysis. In addition connectionist models can provide probability estimates for recurrence at a specific follow-up period by presenting the entire time step to the model as input at each training cycle (Elman, 1990). Ten studies were selected for review under the heading of survival modelling. Each of these studies followed cases for a minimum average of 16 months.

To deal with time dependencies, some studies developed a separate connectionist model for each time point and presented cases to the model once for each time point (Alon, Qi, & Sadowski, 2001; Lundin et al., 1999; Naguib, Robinson, Neal, & Hamdy, 1998; Palmer, Montano, & Franconetti, 2008; Ravdin et al., 1992). If a case had incomplete outcome data within a later period they were removed from that point forward, but retained in the model for the earlier time point. Connectionist models were generally benchmarked against MRA, but also traditional time-series approaches (Alon et al., 2001; Palmer et al., 2008). In addition to comparing results between methods, Alon et al. (2001) compared results for ‘one-step’ and ‘multi-step’ models based on results for two differing seasonal periods.

3.3.6.1 *Summary of results*.

Studies in Table 3.7 are generally limited by small sample size and a short follow-up period; these factors work to limit the power of both connectionist and linear models. Despite limitations in these areas, some accuracy rates indicate the potential of connectionist models in modelling event probabilities for discrete time intervals (Naguib et al., 1998; Poulakis et al., 2004). Generally equivalent performances have been reported in comparisons between connectionist modelling and regression analysis (Bryce, Dewhirst, Floyd, Hars, & Brizel, 1998; Lundin et al., 1999; Poulakis et al., 2004; Ravdin et al., 1992). Although models performed similarly on parsimonious models, when more predictor variables were included the AUC for regression was significantly less (Bryce et al., 1998; Poulakis et al., 2004). This is consistent with results from section 3.3.3 above describing the effect on model performance of numerous predictor variables.

Bottaci et al. (1997) reported a comparison with clinical prediction in a 5 year prospective study. Although both the clinicians and the connectionist model scored well when predicting survival this result was expected given that 93% of the patients survived. The superiority of the connectionist method was apparent upon inspection of the value for the positive predictive power (PPP): this gives the probability that a subject scoring above the cut-off actually failed. The PPP for connectionist modelling was almost twice that for the clinicians' assessment (36% vs. 16%). Although Bottaci et al. did not employ traditional linear statistics on their data, they did report the result for a two layer / logistic perceptron. Accuracy rates of the logistic model were the same as those for the clinicians.

Table 3.7

Model Accuracy in Studies with Time Varying Outcomes

Author(s)	Field	Method	Accuracy	
			Model	Comparator
Bottaci et al. (1997)	Medicine	• 100 test cases • 2 year follow up	.90	.79 (∞)
Bryce et al. (1998)	Medicine	• 95 cases • 2 year follow up	.78 (AUC)	.67 (‡) (AUC)
Lundin et al. (1999)	Medicine	• 300 test cases • 17 year follow up (median) • separate model for 5, 10, 15 years	.88 (AUC)	.86 (‡) (AUC)
Naguib et al. (1998)	Medicine	• 41 cases • 56 months follow up (median)	.80	.75 (‡)
Poulakis et al. (2004)	Medicine	• 40 test cases • 61 months follow up (median)	.77 (AUC)	.74 (‡) (AUC)
Ravdin et al. (1992)	Medicine	• 960 test cases • 16 month follow up (median) • Relapse predictions for each year of follow up	χ^2=48.59	χ^2=42.06 (‡)
Song et al. (2004)	Medicine	• 2,850 test cases • 72 months follow up	.86 (AUC)	.62 (∞) (AUC)
Alon et al. (2001)	Retail	• One-step vs. multi-step forecasts • 1 year time series	Error = 1.50%	Error = 2.75% (‡)
Chang (2005)	Transport	• 492 test cases • 1 year follow up	.614	.608 (‡)
Palmer et al. (2008)	Transport	• 10,000 test cases • 1 year follow up	*M.*Error = 4.13	*M.*Error = 5.48 (‡)

Note. AUC = Area Under the receiver operating characteristic Curve; **‡** =linear statistical; ∞ = clinical

In Alon et al. (2001) contrary to expectations the connectionist model did better on the 12 month forecast (multi-step) than on the one-month forecast (one-step). Their connectionist model significantly outperformed their MRA model on the multi-step forecast. The MRA’s best performance was on the one-step forecast, but this was inferior to the connectionist model on the same forecast. Table 3.7 reports the average error across the two time periods. In the period with turbulent fluctuating conditions, the connectionist model performed better than all other methods. In the stable period error rates were generally lower, as expected, but the connectionist model did not outperform the other time series forecasting methods (but nevertheless continued to outperform MRA). The connectionist model also demonstrated an improvement on the traditional time-series approach in Palmer et al. (2008), particularly where the terms were selected by sensitivity analysis (Naguib et al., 1998; see section 3.3.3 above).

In conclusion the evidence suggests that connectionist models can be designed to model time to event data. This can be done by developing a separate model for each time period. Until the data are censored, data can be used in model training of earlier time periods. While many studies showed equivalent performances, connectionist models may offer more value than other extant approaches in modelling time-varying data where this is chaotic or fluctuating (e.g., Alon et al., 2001). This may not be surprising in view of the material reviewed in section 3.3.4 regarding the performance of connectionist approaches with noisy data. Committee machines may also add value in this regard with time to event data (see Maqsood & Abraham, 2007).

3.3.7 Studies with subjective / narrative data.

Material reviewed above suggested that connectionist modelling may be of real value where data is missing or incomplete (section 3.3.4). Connectionist modelling may be most successfully applied where pattern-recognition is required in particularly vague or complex situations (Zahedi, 1993). Much of the data collected by probation officers in the process of offender assessment can be described as subjective data based on individual perception. Some medical researchers claim that traditional statistics are not sensitive, accurate or convenient enough to assess such data (Zou et al., 1996). Narrative pattern-recognition and the allocation of imprecise information into loose categories are tasks commonly performed in human decision-making. Given that human neural mechanisms were the prototype for the design of connectionist models, it may be that a connectionist approach is well suited to tasks which humans are good at solving. Fuzzy set theory (Zadek, 1965) resembles human reasoning in its use of approximate information to generate decisions. Fuzzy logic is derived from fuzzy set theory where membership values based on qualitative data are used for assignment of cases between 0-1 on a variable (e.g., For example a case could be labelled ‘tall’ rather than ‘short’: IF male IS true AND height >= 1.8 THEN is_tall IS true; is_short IS false). The application of this to connectionist modelling may be useful where the data available is not numerically defined. Nine studies identified within the review were deemed to address issues related to subjective or narrative data in an operational setting.

The majority of studies sought to limit predictor variables to those that could be used in both linear and connectionist models. Some used statistical elimination procedures ensuring that only those factors that were able to show unique variance were included in the model (Brodzinski et al., 1994; Wang, Ohno-Machado, Fraser, & Kennedy, 2001). Brodzinski et al. (1994) for example determined two criteria for the inclusion of factors; i) a canonical loading at or above 0.30 after preliminary DFA; and ii) a significant difference on the factor between recidivists and non-recidivists using ANOVA. These authors included the practitioner's own subjective risk rating as a separate factor, but this did not meet the selection criteria for inclusion of factors in the linear model. Other studies sought to extract objective features that could replace subjective inputs (Nakamura et al., 2000). Nakamura and colleagues also eliminated two further objective features based on connectionist model performance with each independent feature withheld.

Another study, albeit outside of the systematic review, sought to use narrative data from clinical records (Bassoe, 1995). Fuzzy set theory was used to extract knowledge from narratives into input variables. Bassoe (1995) developed a data-base of clinical findings for each clinical entity extracted. Each of their connectionist models was trained by between 4-14 clinical entities depending on the specific association with the clinical findings. Diagnosis involved finding the best match between the test clinical findings and the knowledge stored in the connectionist model.

3.3.7.1 *Summary of results.*

Accuracy rates reported in Table 3.8 are those attained by the best model developed in each study. This shows that studies using clinical records have produced consistently superior accuracy figures using connectionist models compared to traditional models. Where the data were not refined by linear stepwise methods connectionist models have demonstrated improved accuracy compared to the prevailing methods, whether un-weighted clinical indices (Song et al., 2004; Zou et al., 1996), or classic statistical models (Brodzinski et al., 1994; Connor, Symons, Feeney, Young, & Wiles, 2007). Zou et al. account for the difference by pointing out that the Diagnostic and Statistical Manual psychiatric system ([DSM-III-R] American Psychiatric Association, 1987) uses a logical decision tree approach whose consistency is limited with intuitive / fuzzy data. The difference was particularly pronounced, and statistically significant, when the authors studied accuracy on a sub-set of 'hard-to-diagnose' cases.

Table 3.8

Model Accuracy with Subjective / Narrative data

Author(s)	Field	Method	Accuracy	
			Model	Comparator

Bassoe (1995)	Medicine	Transform clinical narratives to input layer	TPR = .97
			FPR = .04
				--

Nakamura et al. (2000)	Medicine	Substituted subjective features with objective correlates	.85 (AUC)	.75 (∞)(AUC)
Song et al. (2004)	Medicine	Used self-reported health from interview data	.86 (AUC)	.62 (∞)(AUC)
Wang et al. (2001)	Medicine	Self-reported clinical history Statistical variable selection	.85 (AUC)	.84 (‡) (LR) (AUC)
Ladstätter et al. (2010)	Nursing	Used data from self-report surveys	R^2=.45	R^2=.39
Brodzinski et al. (1994)	Criminal Justice	Data from case records Statistical selection criteria	.995	.63 (‡) (DFA)
Connor et al. (2007)	Psychiatry	Psychological and Quality of Life data discretised using software (PowerPredictor)	.73	.42 (‡) (DFA)
Zou et al. (1996)	Psychiatry	Data from DSM-III-R and ICD-10. Psychiatric consensus on diagnosis	.96	.91 (‡) (Two layer)
Iyer & Sharda (2009)	Sports	Classified into subjective output categories determined by heuristic rules	TPR=.87	TPR=.44

Note. TPR = True Positive Rate; FPR = False Positive Rate; AUC = Area Under the receiver operating characteristic Curve; **‡** **=**linear statistical; ∞ = clinical

Connor et al. (2007) found that the connectionist model correctly predicted an additional 23% of cases that were missed by the DFA. This was achieved by drawing upon a broader range of data categories many of which had been excluded by the DFA's variable selection procedure. Similarly, in Brodzinski et al. (1994) while only five variables met the statistical inclusion criteria for the DFA method, all 22 variables could be input into the connectionist model. The results reported in Table 3.8 represent an increase in classification accuracy of 59% by using a connectionist model. Brodzinski et al. suggest that the advantage of connectionist models is in their response to the difficulties caused to traditional statistical methods by complicated, vague and subjective data.

The study by Nakamura et al. (2000) included a direct comparison of models employing subjective factors and models employing objective factors. The AUC of .85 reported in Table 3.8 is that for a model developed using objective factors. The corresponding figure for their model developed using the subjective factors was inferior (AUC = .71), although this difference was reduced with subjective variables selected by sensitivity analysis (AUC = .76). The accuracy of the clinicians was inferior to the best performing connectionist model, but similar to the best model derived from few subjective factors. This is surprising given the larger amount of data available to the clinicians. Nakamura et al. suggest that this may reflect that the clinicians do not make full use of the information available to them, simply depending on a limited number of conspicuous variables. This is quite different to a connectionist model which is comprehensively affected by all of the data available. The relative success of models based upon objective features may reflect their more reliable identification and consistent use. Studies that refined the number of subjective variables may therefore have achieved their accuracy due to isolating a sub-set that is more reliably identified and that varies in a more consistent way.

In the study of qualitative data by Bassoe (1995) the connectionist model was able to learn a wide range of clinical cues from one site and remember them when tested in a separate site. Good TPR and FPR values were achieved and the associations made were medically relevant. Bassoe concluded that the findings strongly suggested a clear discrimination between stored patterns and noise.

The material reviewed in this section has clearly suggested that connectionist models can make use of subjective data. They seem to be able to do this more successfully than either clinical or linear statistical prediction methods. Subjective self-report data, particularly in the context of offender assessment, may contain response bias including ‘fake good’ and/or central tendency. This can make self-report data problematic for inclusion in conventional statistical models where connectionist models may be less affected. This may be due to these models' better ability to discriminate patterns from noise, a finding that seems to have emerged in the course of this review. This does not mean that subjective data does not benefit from refinement to arrive at a set of more reliable outcome correlates. The evidence reviewed suggests that inclusion of objective and reliable measures (e.g., observable behaviours) may be important.

3.4 Discussion

3.4.1 Statement of principal findings.

The review was summarised using a narrative rather than a quantitative meta-analytic method due to insufficient use of a single measure of accuracy across studies. In the studies systematically identified by the review, connectionist modelling was rarely outperformed by traditional linear models or by clinical prediction methods. A variety of design methods have been used to achieve the better performance of connectionist modelling, each one specific to the data structure under consideration.

Where very large data-sets are concerned connectionist models have been applied with some success. Although data fitting is promoted by the paradigm of training and internal testing to optimise the learning function, tendency to over-fit the precise characteristics of the data was minimised by larger training samples. In moderate sized samples this raises the importance of minimising the size of the sample reserved for cross-validation.

Subject to sufficient data, including cases and variables, under each class of the outcome of interest connectionist models have evidenced good performance on low base-rate classification problems. This may also be contingent upon sufficient test data in the predictable range. The review therefore highlights the importance of the testing cases being analogous to those used for training. Although connectionist models can provide predictions for individual cases, the use of confidence intervals or an examination of the likelihood ratio as an indication of the precision of population predictions is recommended.

In contexts of incomplete or noisy data a clear superiority of connectionist models over other approaches was seen. Even where a high proportion of error was introduced into the training, connectionist models were able to reproduce valid results. At low levels of noise, traditional statistical methods may be more accurate. In view of connectionist models' competence with noisy data, it was perhaps unsurprising to see evidence of respectable accuracy on predictions based on subjective data. Nevertheless examples of improved performance by all models when more objective data were used, point to the importance of reliably identified data. This may be why a number of studies in subjective data contexts were able to improve their model's accuracy rates by refining the number of predictor variables to input into their model.

Although one of the main advantages of connectionist modelling is the capacity of a network to retain inter-correlated variables, refining the input layer has been associated with improved performance. This may relate to the consequent need for less extensive training or fewer hidden layer resources. Methods used to achieve this have included linear statistical stepwise procedures, and clinical consensus. This latter method helps user credibility in the predictor variables used by the model. Automated methods to refine the input layer have included 'genetic algorithms', 'sensitivity analyses', 'iterative pruning', and 'weight elimination cost functions'. In the development of a connectionist model, these methods are likely to be more appropriate than linear stepwise procedures as they can retain those variables which are particularly important when in combination with other variables. This may be highly relevant to criminal justice system data since they may be particularly subjective and inter-related.

It is also clear from the review that time to event modelling is possible using connectionist approaches. A separate connectionist model can be developed for each time point prediction, with cases repeatedly presented to the model for each time point. This allows the modelling of time-varying factors. Equivalent or improved accuracy rates were seen in comparison with the traditional and prevailing prediction methods.

3.4.2 Limitations of the review.

The conclusions of the systematic review should be considered together with the following limitations. First, due to the non-specific language sometimes used to refer to connectionist modelling (e.g., computer support) it is possible that not every application of the method was identified by the search strategy. It is hoped however that the majority of applications would have been identified given that the two main reference terms were used: ‘artificial neural network’ modelling and ‘connectionist’ modelling.

Second, although the Ministry of Justice website and the Home Office ‘Research, Development and Statistics’ archives were searched for relevant material, all other sources of information focussed on peer-reviewed published articles. Owing to the relative novelty of connectionist modelling, publication bias may have therefore operated in its favour. Although not every study favoured connectionist modelling and indeed many sought to defend existing practices, it would be difficult to assert that there was no possibility of publication bias in the current review.

3.4.3 Recommendations for the present thesis.

Analysis of the material identified by the review has yielded a number of recommendations as to how to improve statistical models made up of operational offender data. The aim of these recommendations is to inform and develop prediction accuracy using a connectionist methodology in the present thesis.

The recommendations, set out below, are presented in sequential order. The sequence is specified to respond first to those data concerns which, if left untreated may undermine all other work to improve classification accuracy. Recommendations are therefore made for model development work in order of priority.

First reviewed are the recommendations for tackling the problems of complexity caused by large data-sets and numerous predictor variables. The complexity of the data-structure is an unavoidable fact that must be tackled first. Following second are the recommendations for low-base rates and time to event outcomes. Both areas are central in any development work aiming to maximise model sensitivity to criterion cases. Finally the recommendations relating to missing and subjective data are summarised. Given evidence from the review of the robustness of connectionist modelling in the face of missing or noisy data, this concern assumes lesser priority.

Making best use of large data-sets.

Evidence from the review suggested that large data-sets were beneficial in model development. The way in which the data-set is prepared is an important part of responding to its complexity. The first recommendation therefore is to develop a programme to scale the data automatically. This would also facilitate the automatic streaming of new data for future analyses. Making best use of the available data is the subject of recommendations two and three. The first of these, recommendation two, surrounds the methodology for modelling reoffending outcomes. Since it is not known whether a generic model comprising all offenders and all offences is likely to be any more effective than the use of multiple differently trained models, early pilot work should test out the merits of training specific models for specific outcome predictions. If this is productive these could potentially be combined into a 'committee' of expert models.

In making full use of the available data, a number of studies in the review experimented with methods such as leave-one-out cross-validation. The advantage of these methods is that they avoid the need to sacrifice large portions of the training data for model validation. The third recommendation is therefore to investigate the value of alternatives to the traditional split sample method of cross-validating a model.

Generalisation of learning will also be dependent on the length of training, given that the fitting capacity of connectionist models makes them liable to over-learn the specifics of the training data. The fourth recommendation is therefore to investigate the timing at which training is stopped so as to achieve the required level of generalisation of learning.

Recommendation 1:

To implement a programme to prepare data automatically for modelling.

Recommendation 2:

To explore the benefit of developing specific models each differently trained according to different offending concerns.

Recommendation 3:

To pilot the use of alternative cross-validation methods to the split sample, so as to preserve more data for model training but without increasing sampling bias.

Recommendation 4:

To investigate the optimum training stopping point for maximum accuracy.

Responding to low base-rates and uneven follow-up intervals.

Operational data contain cases with differing amounts of time at risk. This is a result of numerous factors including differing points of entry into the study, and time-limited outcomes. This causes a problem in terms of the prevalence of the target and a potential problem for the model's PPP (positive predictive power). The fifth recommendation therefore involves the development of separate models for the different follow-up periods. Areas for exploration are separate environments or separate output units for each follow-up interval.

Recommendation 5:

To pilot the use of separate output units for each period of time at risk.

Responding to missing and subjective data.

Degraded data quality is known to impact on the reliability of data analyses. Evidence from the review suggests that connectionist models are tolerant of this even when it reaches the point where other methods are unable to perform analyses. As it is important to verify the robustness of the proposed methods, the sixth recommendation is for pilot work to investigate model accuracy on unadulterated operational data. Should the existing data be associated with reduced accuracy relative to classical statistics a number of recommendations from the review would follow. The first would be to take a complete data-set, set a proportion of variables to zero, and use connectionist methods to predict those data. The results can be correlated with the actual data as a measure of the reliability of the procedure. The second procedure for investigation would be to train a 'team' of connectionist models each with different constellations of the available data, and use this team of models for classification on incomplete case data.

The review has highlighted the importance of including objective data whose variance in relation to the outcome variable is reliable. This may preclude the value of the more subjective data collected by offender managers. The final recommendation is therefore to test out methods suggested by the review for the selection of predictor variables. The merits of refining the input layer are unclear given that a connectionist model is tolerant of a highly populated input vector. Methods employed by studies in the review included the use of a connectionist model to reduce the number of variables (e.g., using sensitivity analysis or iterative pruning); the use of a linear statistical stepwise reduction method; and the use of current theory to select variables. When refining the input layer it is recommended that each selection method is compared to a baseline model involving all available predictor variables. The evidence from the review is inconsistent regarding whether one of these selection methods would demonstrate improved performance over the baseline model. Should model pruning be beneficial, it is expected that a connectionist method would yield the most promising results as this is expected to avoid the inadvertent elimination of inter-related variables.

Recommendation 6:

To implement the connectionist model on un-corrected operational data, and to use connectionist methods to address any apparent problems.

Recommendation 7:

To pilot the use of a connectionist selection method to refine the number of predictor variables and to compare each method to the baseline model involving all predictor variables.

Finally, if clear differences in performance are not discernable between connectionist modelling and traditional linear techniques, the approach suggested by Ciampi and Zhang (2002) could be adopted. This involved estimating a LR model, saving the coefficient weights and applying these to the connectionist network. The optimal weights developed by the regression model are thus the starting point for training the connectionist model. The result of this is a model that classifies at least as well as the developed regression-based model. However, connectionist researchers recommend setting initial weights to small random values to allow a model to adjust its weights according to its own learning rather than getting trapped in a premature solution at the start of the training process (Patterson, 1996).

CHAPTER 4

4. General Method

4.1 Introduction

Chapter 3 indicated the importance of the methodology used to develop model performance. Consequently the present chapter focuses on describing in some detail the components of the models, including the offender assessment protocol that contributed the data fields, the architecture used to structure the models, and the procedures used for analysis and interpretation. To avoid duplication this information is not given repeatedly in introducing each experiment in subsequent chapters; instead the present chapter serves as a general methodological description and a reference source for the chapters that follow.

4.2 Participants

Participants were all offenders on the caseload of County Durham probation service commencing supervision between June 2003 and June 2005. 4,048 offender participants with complete data were included in the data-set (see Procedure section 4.4 for details on how data were selected). The offender commenced supervision either following the imposition of a court order or release from custody. 1,281 (31.6%) had been released from custody, and the remaining 2,767 (68.4%) were on community sentences. Table 4.1 below summarises the demographic characteristics of participants in the sample. Also included in the table is information on the criminal history of participants.

Table 4.1

Demographic and Historical Characteristics of the Sample

Demographic/history	Prison releases		Community sentences		Total participants	
	(*n*=1,281)		(*n*=2,767)		(*n*=4,048)	
Age, *M* (SD)	29.73	(9.95)	29.86	(10.21)	29.82	(10.13)
Gender Male, *n* (%)	1,151	(89.85)	2,346	(84.78)	3,491	(86.39)
Ethnicity White, *n* (%)	1,211	(94.54)	2,665	(96.31)	3,862	(95.41)
Number of prior convictions, *M* (SD)	10.96	(12.30)	8.32	(10.53)	9.15	(11.19)
Current serious offence, *n* (%)	411	(32.08)	479	(17.31)	890	(22.00)
OGRS score, *M* (SD)	53.42	(30.09)	48.78	(27.68)	50.25	(28.55)
OGRS risk band, *n* (%)						
Low	486	(37.94)	1,186	(42.86)	1,672	(41.30)
Medium	379	(29.59)	953	(34.44)	1,332	(32.91)
High	416	(32.47)	628	(22.70)	1,044	(25.79)
OASys static risk, *n* (%)						
Low	457	(35.67)	1,246	(45.03)	1,703	(42.07)
High	824	(64.32)	1,521	(55.00)	2,345	(57.93)

Table 4.1 shows that participants in the sample were predominantly white and male, with the proportion of males even greater among prison releases. On average participants were approximately 30 years of age, and were generally previously known to prison and probation having previously accrued approximately 9 convictions on average. Prison releases had a higher mean number of prior convictions than did participants on probation (t=-6.444, df=2179.08, p<.001), which is not unexpected due to prison traditionally being reserved for the most persistent or serious offenders. 62% of prison release cases were assessed as medium or high reconviction risk using the Offender

Group Reconviction Scale (OGRS),[7] compared to 57% of probation offenders. Table 4.1 shows that offenders are similarly divided into high and low reconviction risk using OASys static risk (OASys is described in the Materials section 4.3, below). OASys static risk measures past offending behaviour, as opposed to the OASys dynamic risk factors which measure current social and psychological problems.

Tables 4.2 and 4.3 provide information on the current dynamic characteristics of participants, as assessed by the probation officer using OASys. A chi-square analysis of table 4.2 showed a statistically significant association between sentence type and risk score (χ^2=157.26, df=2, p<.001). This is unsurprising since prison is supposed to hold the most persistent or serious offenders, as mentioned above. Inspection of table 4.2 shows that there were more offenders at high risk of reconviction among prison releases (30%) than among offenders sentenced to probation (14%).

Table 4.2

Distribution of OASys Risk in the Sample

OASys reconviction risk	Sentence type		Total
	Prison releases	Community sentences	
Low	331	1,001	1,332
(Within risk %)	(24.85)	(75.15)	(100.00)
(Within sentence %)	(25.84)	(36.18)	(32.91)
Medium	562	1,380	1,942
(Within risk %)	(28.94)	(71.06)	(100.00)
(Within sentence %)	(43.87)	(49.87)	(47.97)
High	388	386	774
(Within risk %)	(50.13)	(49.87)	(100.00)
(Within sentence %)	(30.29)	(13.95)	(19.12)
Total	1,281	2,767	4,048

[7] OGRS is an assessment of demographic and criminal history variables which together with the Offender Assessment System (OASys) is described under the Materials section. For further information see Chapter 2.

Table 4.3

OASys Dynamic Risk Areas Causing Concern

Dynamic risk area	Prison releases		Community sentences		Total	
	(*n*=1,281)		(*n*=2,767)		(*n*=4,048)	
Accommodation, *n* (%)	378	(29.51)	511	(18.47)	889	(21.96)
Education / Employability, *n* (%)	800	(62.45)	1,487	(53.74)	2,287	(56.50)
Financial Management and Income, *n* (%)	380	(29.66)	490	(17.71)	870	(21.49)
Relationships, *n* (%)	31	(2.42)	53	(1.92)	84	(2.08)
Lifestyle and Associates, *n* (%)	508	(39.67)	569	(20.56)	1,077	(26.61)
Drug Misuse, *n* (%)	404	(31.54)	447	(16.15)	851	(21.02)
Alcohol Misuse, *n* (%)	531	(41.45)	1,196	(43.22)	1,727	(42.66)
Emotional Well-being, *n* (%)	328	(25.60)	753	(27.21)	1,081	(26.70)
Thinking and Behaviour, *n* (%)	614	(47.93)	993	(35.89)	1,607	(39.70)
Attitudes, *n* (%)	295	(23.03)	373	(13.48)	668	(16.50)

The most frequently occurring dynamic risk area was ‘education, training, and employability’ which was seen in approximately 57% of cases. In line with the overall risk categorisation in table 4.2, there was a significant relationship between each need area and sentence type, with the exception of ‘relationships’ ‘alcohol misuse’, and ‘emotional well-being’. Table 4.3 shows that relationships was a low frequency need area regardless of sentence type, while alcohol misuse and emotional well-being were the only need areas that were proportionately higher in the community sentences group.

In the sample as a whole, offenders had an average of 4 dynamic risk areas, as measured by the mean and the median (not shown in Table 4.3). This differed according to whether the offender was a prison release or a community sentence, with 58% of prison releases having 4 or more dynamic risk factors occurring simultaneously compared to 43% of offenders on probation orders. Most offenders had fewer than six dynamic risk areas operating at one time (37% of prison releases, 23% of probation orders).

4.3 Materials

4.3.1 The Offender Assessment System

The Offender Assessment System (OASys) is the common tool for use across prison and community environments to assess each offender's risk of general reconviction and risk of serious harm. It was originally introduced to increase consistency, improve quality and systematise resource allocation (Aubrey & Hough, 1997).

1. Risk of reconviction and offending-related factors
    - a. Offending Information
    - b. Analysis of offences
    - c. Accommodation
    - d. Education, training, and employability
    - e. Financial management and income
    - f. Relationships
    - g. Lifestyle and associates
    - h. Drug misuse
    - i. Alcohol misuse
    - j. Emotional well-being
    - k. Thinking and behaviour
    - l. Attitudes
    - m. Health and other considerations
2. Risk of serious harm, risks to the individual and other risks
3. OASys summary sheet
4. Sentence planning
5. Self-assessment

Figure 4.1. The components of the OASys risk and need assessment tool

In practical use, OASys aims to i) assess how likely it is for an offender to be reconvicted; ii) identify and classify offending related needs including basic personality characteristics, and cognitive / behavioural / social problems; iii) assess risk of harm to self, the general public, known adults, children, staff and other prisoners; iv) assist with the management of risk of harm; v) link assessments, supervision plans and sentence plans; vi) indicate any need for further specialist assessments; and vii) gauge how an offender changes during the sentence. Measures for each offender fall into five categories as in Figure 4.1 above. See Appendix A for a full list of variables included in the statistical models.

4.3.1.1 *Risk of reconviction / offending related factors*.

During the development of OASys, it was considered important to combine clinical and actuarial models of assessment, in recognition of the omission of dynamic social and personal needs within prevailing actuarial assessments (Home Office, 1998, cited in Howard et al., 2006). The risk of reconviction and offending-related factors section of OASys comprises a series of criminogenic need areas that are considered functional in the commission of offending behaviour (Andrews & Bonta, 2010).

Each factor is described in detail in the OASys manual (Home Office, 2002). First assessed is the 'static' risk factor entitled Offending Information which examines current and previous offences. Its inclusion follows from national and international research indicating that criminal history is a strong predictor of future reconviction (Andrews, 1983; Barnett, Blumstein, & Farrington, 1987; Bonta et al., 1998; Cornish & Clarke, 1975). The current offence is further detailed in the Analysis of Offences, section 1b of OASys, which links to the later section regarding Risk of Serious Harm.

Many of the items in OASys sections 1a and 1b are based on the items used by a measure named the Offender Group Reconviction Scale (OGRS-II: Taylor, 1999), and are thus given the greatest weighting in prediction of reconviction. As described in chapter 2, OGRS is a purely actuarial instrument for measuring risk of reconviction, based on a study of over 44,000 offenders commencing community supervision in the 1990s (Copas & Marshall, 1998). Its algorithm produces a probability score for reconviction within two years, based solely on history of offending and certain demographic variables. OGRS has been shown to have a high level of predictive validity with a range of offender populations (Coid et al., 2007; Gray et al., 2004; Lloyd et al., 1994; Wakeling, Howard, & Barnett, 2011).

The remaining risk factors are more dynamic in nature and cover human social and personal needs. The offender's cognitive abilities are important in each of these sections, as they assess the offender's motivation to improve his/her life in each area. Education, training and employability for instance considers the offender's history of employment and training as well as his/her current attitudes to work and work-related training. This follows from research indicating that offenders are more likely to be unemployed, have a poor history of employment and express opposition to the work ethic (MacKenzie, 1997). Accommodation looks at whether the location encourages offending or creates a risk of harm in terms of relationships within or close to the household. Another overtly practical risk factor is 'financial management and income' which may be an indicator of general ability to cope, in turn related to re-offending (Zamble & Porporino, 1988).

'Relationships' and 'lifestyle and associates' both assess the impact of the offender's relationships on the likelihood of further offending behaviour. Supportive relationships are often considered a protective factor against re-offending (e.g., Nuttall, 1960), and the existence of criminal convictions within the close family has been associated with negative reconviction outcomes (e.g., Farrington, 1978). Family relationships are assessed within the 'relationships' section. Lifestyle and associates relates more to the offender's peer interactions: a clear link has been shown to exist between how offenders spend their time, the associates with whom they mix, and likelihood of reconviction (Andrews & Bonta, 2010; Raynor, Knych, Roberts, & Merrington, 2000; Rogers, 1981).

The offender's use of substances and the link between this and the likelihood of re-offending are assessed in the 'drug misuse' and 'alcohol misuse' sections. Drugs in particular are consistently linked with re-offending (e.g., May, 1999), indeed the majority of drug possession constitutes an offence and is thus a direct risk factor. The assessment of both of these dynamic risk factors has implications for the risk of harm assessment as well as that for reconviction risk. In the summing of items within the OASys summary sheet (see below), alcohol misuse is given less weight than drug misuse reflecting evidence of the relative outcome correlations (May, 1999). However it may be that the interaction of substance misuse with other risk factors raises the risk of serious harm (Monahan et al., 2001).

Psychological deficits are most explicitly covered during the assessment of 'emotional wellbeing', 'thinking and behaviour' and 'attitudes'. Emotional wellbeing examines the extent to which emotional problems interfere with the offender's functioning or create a risk of harm to self or others. Psychological and emotional factors have been moderately correlated with re-offending under meta-analysis (Gendreau et al., 1996), and the emotional wellbeing section therefore makes only a small contribution to the overall OASys prediction of reconviction risk. Thinking and behaviour assesses the offender's application of reasoning, especially to social problems. Research has indicated that offenders tend to cope poorly with life on account of various cognitive deficits (Ross & Fabiano, 1985), including a lack of impulse control, poor problem-solving, and rigid or inflexible thinking. Reducing these deficits through offending behaviour programmes may be effective in reducing rates of reconviction (Lipsey, Chapman, & Landenberger, 2001; Tong & Farrington, 2006). 'Attitudes' meanwhile assesses pro-criminal attitudes, including negative attitudes towards supervision, and regarding the offender's own offending. Attitudes are known to be difficult to measure objectively, however there is a body of research evidence which suggests that negative attitudes, even when measured by self-report measures, are predictive of reconviction, parole violation, and general misconduct (Simourd, 1997; Walters, 1992).

The final risk factor reviewed within the category 'risk of reconviction and offending related factors' is named 'health and other considerations'. Under this factor the existence of any practical barriers to engagement with protective interventions is assessed. Barriers explicitly covered include physical or mental disabilities, and commitments such as childcare or employment. Alongside these practical barriers the officer is also asked to assess the offender's understanding of the importance of completing programmes. This factor is not included in the risk score on the OASys summary section; however it is considered within the probation officer's overall sentence planning assessment.

4.3.1.2 *Risk of serious harm, risks to the individual and other risks.*

Risk of serious harm is defined as "a risk which is life-threatening and/or traumatic, and from which recovery, whether physical or psychological, can be expected to be difficult or impossible" (Home Office, 2002, p.128). All offenders are subject to a screening for risk of harm, and a sub-set are then given the full risk of harm assessment. These are those offenders for whom there are indications of risk, or for whom it would not be defensible to overlook a full review. This section therefore records whether the offender has one from a series of ‘serious harm’ offences (e.g., sexual offence against a child), and whether he/she has shown signs of risk behaviour (e.g., possession of a weapon).

Assessment of risk of harm draws together information from the earlier sections of OASys to allow the assessor to make an informed judgment on the risk of harm. This follows evidence from a literature review commissioned for the OASys project (Powis, 2002), which concluded that the risk factors most frequently connected with risk of harm are often the same factors as those associated with general offending. The risk of serious harm assessment therefore reviews static and dynamic information, with an emphasis on the static risk factors including whether any indicators of serious harm have been evident in the offender’s index offence or previous behaviour. The following risks are reviewed: harm to the public; harm to known adults; harm to staff; harm to prisoners; harm to children; harm to the individual (suicide, self-harm, coping in custody or in a hostel setting, vulnerability); and other risks including escape/abscond, control issues, and breach of trust.

Unlike the risk of reconviction and offending related factors assessment, the risk of serious harm assessment is not actuarial because it is not scored, and the final rating is based purely on clinical judgment. Within this risk judgment, ‘low’ denotes no identifiable current indicators of risk of harm; ‘medium’ refers to the presence of indicators of risk of harm where any potential event is deemed unlikely unless there is a change in circumstances; ‘high’ is given where there are identifiable indicators of harm and the potential event could happen at any time with serious consequences; and ‘very high’ is reserved for those cases where there is an imminent risk of serious harm, i.e., the potential harmful event is more likely than not to happen imminently. The output from this section, together with that of the summary sheet and sentence planning section, seeks to inform the employment of risk management procedures.

4.3.1.3 *OASys summary sheet*.

The summary sheet section of OASys sums the items within the risk of reconviction and offending related factors section, and provides the total weighted score or risk estimate. The profile of criminogenic needs across the section scores shows those need areas most worthy of input to reduce the probability of reconviction. The weighting of each section score was determined by logistic regression of the relationship between the predictors and reconvictions after a two year follow-up (Howard et al., 2006). Howard et al. reported that 26% of offenders actuarially assessed as 'low' risk reconvicted within this time-frame, while the rate for offenders at medium and high risk of reconviction was 58% and 80% respectively. There were no statistically significant differences in the accuracy of the prediction of reconviction between different sub-groups of offenders.

Howard et al. (2006) also compared section scores with reconviction outcome. In 'offending information' and 4 of the dynamic need areas, reconvicted offenders with the need significantly outnumbered those without the need. Mean OASys sub-section scores for drug misuse, accommodation and offending information (criminal history) were over twice as large in offenders reconvicted compared to offenders not reconvicted. Alcohol misuse, emotional well-being and thinking & behaviour did not predict reconviction when offenders' other needs were statistically controlled.

4.3.1.4 *Sentence planning*.

This section of OASys is designed to cover the sentence plans for each offender, incorporating those risk factors deemed important in the actuarial part of the assessment. The section also reviews behaviour since the last assessment, including changes in motivation and number of acceptable / unacceptable absences. The offender's compliance with the order or licence is also reviewed upon consideration of information from other agencies (e.g., court services). As the OASys risk assessment is periodically reviewed, any changes to the dynamic risk factors should be incorporated into the review sentence plan. Integrating supervision and sentence planning into the overall assessment process in this way is intended to help practitioners draw together and manage information systematically. In the current study only the earliest OASys review was selected for modelling to provide a consistent baseline prediction for each offender.

4.3.1.5 *Self-assessment*.

The purpose of this section is to provide a more complete picture by allowing the offender the chance to comment on how they see their life. This may raise disparities with the officer assessment, and offers the opportunity for professionals to gain an insight into the offender's view of his/her needs. There is some evidence that offenders are able to recognise some of their own problems and the level of difficulties they report may link to reconviction probability (Cookson & Clark, 1998). An evaluation of the self-assessment questionnaire completed by over 100,000 offenders suggested however that they tend to be more optimistic about their chances of remaining offence free than their OASys would predict (Moore, 2007a). The self-assessment questionnaire is not scored and was not included in the present data.

4.3.2 Re-offending outcome measure.

Re-offending can be indexed in a variety of ways, the most widespread being official reconviction. This undoubtedly under-estimates true re-offending levels with the possibility that a number of offenders routinely go undetected and then appear to resemble non re-offenders. Official reconviction data should however minimise false positives, i.e., labelling of desisters as re-offenders, which is a limitation of using a more inclusive index such as arrest data. The procedure for obtaining the official reconvictions is described below.

4.4 Procedure

4.4.1 Data collection.

Data were collected from participants and recorded electronically as part of routine practice by probation officers. The context for this is described below, followed by an explanation of how the data were extracted for analysis.

4.4.1.1 *Collection from participants*.

OASys is an extensive assessment of the offender, taking approximately two hours to complete (Home Office, 2002). Given the amount of fields within OASys and the length of time required for its completion, officer practice is to complete only those fields that are relevant to the offender. It is expected that the tool is completed electronically, to enable the timely transfer of information across the different environments of the National Offender Management Service (NOMS).[8]

The offender interview and file review required for OASys are first done before sentencing to inform directly a pre-sentence report which gives advice to the judge or magistrates on suitable sentencing options. The assessment is then repeated within 15 days of an offender's sentence, and upon his/her release from custody to ensure that the information collected at the pre-sentence stage is accurate. Accuracy of self-report information may be compromised by offender anxiety or by deception associated with interrogation and/or legal advice regarding sentencing (Gudjonnson & Bownes, 1992; Moston, Stephenson, & Williamson, 1992). After the offender has received his/her sentence, disclosure may be qualitatively different and more extensive regarding social and personal needs. If the offender is sentenced to custody OASys must be updated at the point of release to ensure the dynamic factors are up-to-date.

For all offenders in the community subject to supervision, OASys is reviewed every sixteen weeks under national standards (National Probation Directorate, 2000, 2002) for the supervision of offenders, although a review could be brought forward in response to any important change in circumstances. This means that every offender receives multiple assessments of their dynamic risk factors. As mentioned, the current study took the earliest available assessment to baseline the offender's risk at release or start of supervision.

4.4.1.2 *Extraction of data for analysis*.

Participant data regarding their characteristics and risk factors were imported from OASys into SPSS version 19.0 (SPSS Inc., 2010). Scores for all items, for all cases on

[8] NOMS is an executive agency of the UK Ministry of Justice, responsible for commissioning and delivering adult offender management services in custody and in the community.

community supervision[9] in County Durham probation area between June 2003 and June 2005 were captured. This resulted in a data-set of 16,349 lines representing different assessments. Since many of these assessments related to the same participant, and to respond to the need to have different participants on each line of the data-set, only the earliest dated assessment for each offender was selected for inclusion. This ensured that the assessment data related to the start of the supervision period, rather than during custody or at the end of community supervision. This resulted in a data-set of 4,234 cases or lines of data, of which 4,057 had an identification number on the Home Office Police National Computer (HOPNC). Since this identification number is essential for tracking official reconvictions, only these cases were selected for study.

Participant data regarding their conviction history was gathered from the HOPNC in December 2007, meaning that this was the latest point at which the sample's convictions could be observed. Specifically, December 13, 2007 was the cut-off for any further offending. This meant that the follow-up time in the sample ranged from 31 months (2 years, 7 months) to 55 months (4 years, 7 months), depending on the sentence start date in the sample. For each case the conviction date selected was the first occurring after the sentence date, and the length of time between sentencing and re-offending was calculated. This new variable was merged with the OASys data so that each case had a time to re-offending outcome, and this was then converted into a binary value representing whether or not there had been a reconviction. The final sample size was 4,048 after excluding duplicate cases and cases that were still incarcerated in December 2007. Initial analysis focused on any re-offending within the total follow-up time; subsequent analyses considered offending within different follow-up intervals.

4.5 Design and Analysis

The research was focussed on the application of connectionist models, interactive models which are designed to mimic the functioning of the human brain whereby learning is held in the strength of the connections between two neurons (see chapter 3 for further background). Neural network analysis was therefore the principal method of analysis used, alongside a comparison statistical model: discriminant function analysis. Discriminant function analysis is an example of a traditional statistical analysis. Both

[9] Includes offenders on court orders and those on licences following release from custody

methods of analysis, as well as the method selected for cross-validation, are described below.

4.5.1 Discriminant function analysis.

Given that the criterion variable is reconviction measured on a binary scale (0/1), and the predictor variables are both continuous and categorical, discriminant function analysis (DFA) was used as a comparison model. Logistic regression analysis was an alternative, but this did not provide the required option for cross-validation analysis within the software utilised. The main point of these classical statistical models is to predict group membership on the basis of a linear combination of the available predictor variables. DFA is based on modelling the predictor variables for each criterion group, and this provides probability estimates of a particular score on each variable given membership in a particular group. The discriminant function can be calculated for each case, or a mean value of the function can be calculated for each group.

4.5.2 Connectionist model.

The way in which a connectionist model learns has already been described (Chapter 3, section 3.1.1). To build the networks employed in the study a feed-forward architecture was used as illustrated in Figure 4.2. The input layer initially comprised 236 nodes/units, which were the offender characteristics from OASys described above. The input layer was fully connected to an intermediate hidden layer, which comprised a set of hidden units. The number of hidden units was determined empirically to maximise correct classification as is common in research using connectionist models. The hidden unit layer was fully connected to 1 output unit representing the occurrence or lack of occurrence of any further offending.

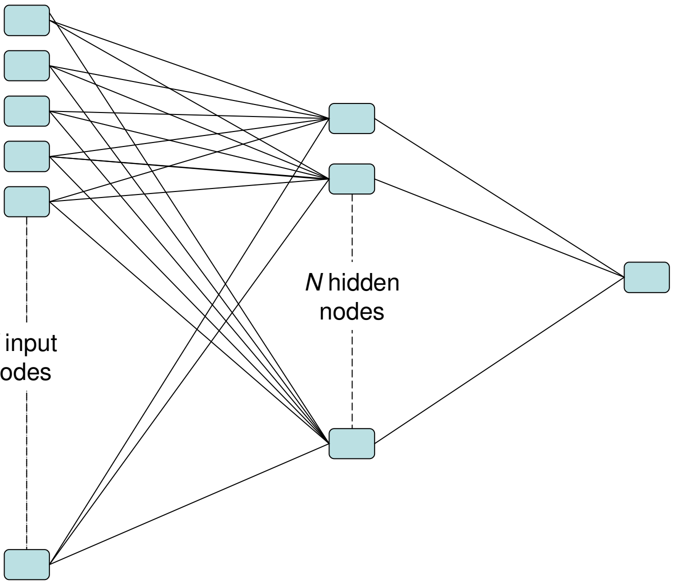


Figure 4.2. Architecture of a connectionist model

As shown in Figure 4.2, all input units were connected to each of the hidden layer units and every hidden layer unit was connected to the single output unit. The output of a unit was determined by a function of the sum of the inputs from all of the units connected to it, and each individual input was determined by the activation of the unit multiplied by the strength of the weight on the connection between the units. The activation function of units was the logistic sigmoid function which produces an output value between 0 and 1 for each case in the sample, referred to as the model's classification output.

The learning rate of the model was set at 0.25, and the momentum at 0.9. The model was implemented using the PDP++ neural network simulation software (O'Reilly, Dawson, & McClelland, 1995). The input data from OASys were scaled into a 0-1 activation range. The model adjusted weights on connections according to the backpropagation learning algorithm with gradient descent, which provides a gradual adjustment of weights that minimise the error in the model, resulting in an output activation closer to the target. Weights were adjusted after data from each individual were presented to the model.

4.5.3 Two layer model (logistic perceptron).

The results of the connectionist model were compared with those from DFA using the same data. In addition, as a direct comparison with the connectionist model, results were also compared with those from a network with the intermediate hidden layer removed. The results of a connectionist with no hidden layer and a single output (hereafter referred to as a 'two layer model') should be approximately equivalent to that of linear regression since connectionist modelling and regression both use a weighted sum of the predictor variables to produce an output. The two layer model however does not have the facility, provided by the hidden layer in the connectionist model, of automatically modelling interactions between variables and instead emphasises the main effect of individual variables in explaining the observed outcomes. Like linear regression the two layer model therefore assumes a simple additive relationship among the variables.

4.5.4 Cross-validation.

Due to the possibility of over-fitting the data (see Chapter 3) the accuracy or predictive validity of a model should be assessed on cases other than those on which it is built. This is achieved using 'cross-validation' whereby the model is trained on one set of cases and then tested on a separate set. The cases input into the model were the offender participants (*n*=4,048) described above with input nodes corresponding to factors based on offender characteristics. To retain as many of these cases as possible for training, the 'leave-one-out' procedure was adopted (Efron, 1983). In implementing this all cases bar one were put forward for training and the remaining case was reserved for testing the trained model. To minimise the impact of the specific characteristics of the case reserved for testing, this case was then returned to the sample whereupon a different case was selected and withheld pending the training of the remaining cases. This process continued until all 4,048 cases in the sample had served as the test case. Model accuracy was taken by the average of the predicted values of the test case. 'Leave-one-out' was implemented for cross-validation of each model: connectionist model, DFA, and the two-layer model. The leave-one-out procedure is likely to be of benefit when considering the utility of a model in diagnosing individual cases. If all cases from the screening setting are sampled, as in the present study, it can provide an almost unbiased estimate while profiting from exposure to the maximum amount of training data possible.

The weights given to the input nodes before training commenced are a potentially important influence on their contribution to model performance. The starting weights in all models were therefore randomly initialized to eliminate any systematic influence of these starting weights on the prominence of any of the input nodes in the developed model (connectionist and two-layer model) and multiple versions of the model were tested to ensure that initial random weights did not unduly influence the model's performance.

The order in which the cases were selected by the model for training was also randomised. This ensured that there could be no systematic order effects in terms of the exposure of the model to either re-offenders or non re-offenders.

The stopping time for training is also a variable in model performance. Each model was trained for a minimum of 1,000 passes through the training data (epochs). This stopping time was selected based on pilot testing of the data, which suggested that the model's performance reached asymptote by 1,000 epochs. The cross-validation performance was reported at each 100 epoch increment. During testing the model was judged to have accurately learned the task if activation in the output unit was greater than 0.5 for matching patterns (when the target in the output unit is 1); and less than 0.5 for non-matching patterns (when the target in the output unit is 0).

4.5.5 Measure of performance.

It was necessary to select measures of performance that were independent of pre-existing aspects of the data such as the criterion base-rate. As discussed in Chapter 2, a decision strategy based on the re-offending base-rate may appear to help accuracy but would be unrelated to the discriminating power of the models. In these conditions overall accuracy levels may be misleading as a reflection of classification performance. Thus measures of performance had to be able detect re-offenders from the level of background noise caused by overlapping non re-offender cases.

Accuracy in discriminating between offenders that reconvict and those that did not, was determined using d-prime (Green & Swets, 1966; Stanislaw & Todorov, 1999). D-prime (*d'*) addresses the signal-to-noise ratio by measuring the distance between the re-offender (signal) and non re-offender (noise) means in standard deviation units. A *d'* of 2.00 for example, indicates that the distance between the means is twice as large as the standard deviations of the two distributions. Analysis of *d'* is affected by its limits for whole numbers. This only applies in circumstances where no cases or all cases are classified correctly. A recommended way of responding to hits or false alarms with zero or perfect 100% accurate responses, is the 'loglinear' approach (Hautus, 1995). This involves adding 0.5 to both the number of hits and the number of false alarms and adding 1 to both the number of re-offender trials and the number of non re-offender trials, before calculating the hit and false alarm rates.

Due to widespread use of ROC analysis in offender risk prediction, model outputs were also used to determine points on a ROC curve (see Chapter 2, section 2.2.2). As described previously this plots the hit rate as a function of the false alarm rate for all possible decision thresholds. Like *d'* the area under the ROC curve (AUC) is a measure of sensitivity that is unaffected by response bias, e.g., bias associated with value judgements or skewed base-rates. The AUC ranges from 0.5 (signals cannot be distinguished from noise), to 1 (perfect discrimination), with each prediction result representing one point in the ROC space. The AUC can be interpreted as the proportion of times a measure would correctly identify an actual re-offender, if presented with a randomly chosen re-offender and non re-offender simultaneously (Hanley & McNeil, 1982). There is some disagreement on how to interpret the size of the effect represented by the AUC value. Sjostedt and Grann (2002) proposed a conservative interpretation in which values below 0.60 are low, and those above 0.90 are high, with scores with upper intervals of 0.70, 0.80, and 0.90 are marginal, modest, and moderate respectively. This differs from the interpretation given by Rice and Harris (2005) who suggest that values above 0.64 represent a moderate effect size and those above 0.71 represent a large effect size. There is general agreement however, that the AUC is the preferred measure of predictive or diagnostic accuracy in forensic psychology and psychiatry (Mossman, 1994; Rice & Harris, 2005; Swets et al., 2000).

CHAPTER 5

5. Pilot Testing

5.1 Introduction

Recidivism risk assessment, reviewed in Chapter 2, has advanced from unstructured clinical judgement to more structured and mechanical approaches. Validation studies of the relationship between scores on these instruments for individual offenders and subsequent general or violent recidivism have shown that these later generations of risk assessment consistently classify re-offenders and desisters with statistically significant but yet consistently only 'moderate' accuracy levels (Coid et al., 2009; Gray et al., 2004; Hanson & Morton-Bourgon, 2009; Kroner & Mills, 2001; Yang, Wong, et al., 2010). In the study by Coid et al. (2009) for example, none of the measures were able to achieve predictive accuracy significantly exceeding that produced using a simple sum of the number of each individual's previous convictions. Thus current measures may just be measuring long-term risk potential that can be captured by historical factors without benefiting from the incorporation of dynamic risk factors. Current actuarial measures all use least-squares regression techniques for the combination of risk variables, which assume independence among predictors. This may explain the 'glass ceiling' of predictive accuracy experienced by the field (Yang, Wong, et al., 2010). In response to this the present chapter presents a pilot study of the application of connectionist modelling to offender data.

Previous research, reviewed in Chapter 3, suggested that connectionist modelling may offer a number of advantages over traditional statistical analytic methods under conditions of complex (Brodzinski et al., 1994), inter-correlated (Edwards et al., 1999; Marshall & English, 2000), incomplete (Gonzales & DesJardins, 2002) and noisy data (McMillon & Henley, 2001; Nolan, 2002). A pervasive theme in this collection of work was the need to develop the architecture of a connectionist model according to the nature of the data, and Chapter 3 resulted in a number of recommendations for development. The present chapter sought to learn lessons from modelling a sample of the data, relevant to later full scale application of a connectionist model with offender data. As described in section 5.2 below, this data related to a complete and distinct class of cases from within an overall caseload. All cases had committed sexual offences and were subject to community supervision relating to a court order or post-release licence.

Results of previous applications of connectionist models indicate that it may be beneficial to include all variables during experimental piloting to avoid excluding potentially important variables (Edwards et al., 1999; Marshall & English, 2000; Meccoci et al., 2002). Marshall and English (2000) suggested that the better performance of connectionist models relative to rival statistical approaches lies in better treatment of variable interactions through the hidden nodes. This is enabled via the sigmoid activation function, which allows each hidden layer unit to produce sub-model components that are switched on or off for certain patterns in the input layer. In an attempt to model a large number of independent variables, the impact of variation to the number of hidden nodes is discussed in the present chapter. Using offender data, researchers have found different results concerning the resources required in the hidden layer. Grann and Langstrom (2007) found that the level of over-estimation to the training set based on subsequent performance on an independent test set, or 'shrinkage', was greater in the connectionist model than in traditional methods. However, their model incorporated a number of hidden units greater than the heuristic recommended by the authors (Ward Systems Group, 1996), thus raising the risk of over-fitting the training data. Palocsay et al. (2000) also using offender data did find a variation in results depending on the number of hidden units employed. In the Palocsay et al. (2000) study the improvement in the identification of re-offenders using connectionist modelling compared to using regression analysis was statistically significant. Two different hidden layer configurations, one using 39 units and the other using 26 units, had the same best results (Palocsay et al., 2000). Caulkins et al. (1996) also experimented with the number of hidden units, but found no beneficial effect. Caulkins and colleagues questioned the quality of the variables in the data upon the realisation that many offender patterns had identical numbers of re-offenders and non re-offenders. Noise on the data will often mean that few patterns are identical and this may be better tolerated by connectionist modelling (McMillen & Henley, 2001). However a learning rule to separate the two classes is not possible unless there is a systematic difference in the relationship between the variables of the cases in each outcome class. This is a potential issue with the particular sub-sample of offenders identified for pilot testing. Sexual offenders can be slower to re-offend (or to have this detected) than general samples of offenders. This means that many of the re-offenders and apparent non re-offenders may overlap in their characteristics.

Given that accuracy is very much related to the representativeness of the testing set a large number of randomisations of the train/test division is required to reduce the sampling error. This can be a prohibitive task since it can take hundreds of randomisations to ensure that the data has been sampled sufficiently. Initial exploratory testing used a split sample method and this necessarily excluded a number of cases from the training set, limiting training to those remaining cases. In problems of low-base rates, with imbalanced outcomes, this is important because it could mean that the model misses out on the opportunity to train on some of the limited number of ‘re-offender’ exemplars. In these circumstances it is important to increase the positive predictive accuracy, i.e., the number of re-offenders correctly predicted when a positive prediction is made by the model. This meant that it seemed sensible to increase the exposure of the connectionist model to re-offender cases in order to assist in the discrimination from non re-offender cases. It was therefore decided to employ the leave-one-out procedure (see Chapter 4, section 4.5.4) to maximise the number of re-offender exemplars available for model development.

5.1.1 Goal of the pilot study.

The pilot study aimed to test the use of a connectionist model on a sub-sample of cases. The purpose of this was to test the feasibility of a connectionist model, and to do this on a well-studied sub-population of particular importance to the UK criminal justice system. The sample was characterised by the important data considerations, including low base-rates and numerous predictor variables. Consistent with the discussion above about sampling error, model development and validation for the study was via the leave-one-out method.

A key objective of the pilot study was to investigate the cross-validity of the statistical model on untrained cases. Chapter 3 suggested that ‘overfitting’ was a likely problem in applied research using connectionist modelling (e.g., Mobley et al., 2000; Wu et al., 2008). Pursuit of the pilot study’s objective therefore required an examination of the extent to which a connectionist model is liable to memorise patterns in the data as opposed to the underlying principles that generalise best in matching new cases. To achieve this, a connectionist model was applied to the offender patterns using various connectionist model architectures. Since the origin of any improvement is likely to be based on the detection of additional interactions between variables important for separating re-offenders and non re-offenders, this was addressed by systematic variation to the number of hidden units available in the middle layer of the model architecture. As an additional test of capacity for overfitting, the performance of the connectionist model against randomised patterns of the same data was also examined. The methods used are described in full in the following section.

5.2 Method

5.2.1 Data sample.

As described in Chapter 4 the data were from one probation area within the NOMS executive agency of the UK Ministry of Justice. A sub-set of the total caseload was selected for pilot testing. To ensure that the selection of cases was fully representative of its class, an entire sub-population of cases was selected (*n*=154). All of these cases were serving sentences in the community for the same category of offence (sexual offending). This was thought to be of benefit in piloting a new model in terms of the clarity required for learning patterns within the data. Although many offenders commit cross-over offences, some groups of offenders are thought to be more distinct in the nature of their offending and their offending related problems (see Chapter 2). The strategy of selecting a particular group of cases that are uniform in terms of the current main offence therefore avoided any potential problems related to cohort irregularity.

Participant characteristics were the variables that were subject to the modelling in each experiment. These characteristics were all items from the probation database containing items from OASys (Home Office, 2002). For details on the structure and content of OASys, see Chapter 4, and for full information on the extant validation of OASys, see Chapter 2. In the pilot testing experiments an additional 15 variables were included related to the offender's performance on objectives during supervision.[10] The

[10] Although included in pilot testing, as they were response variables these 15 were excluded from the results of modelling described in Chapters 6, 7 and 8 (listed in Appendix A).

outcome variable was further offending, as indexed by officer information on reconviction or recall to custody.[11]

5.2.2 Model architecture.

Networks were built with the feed-forward architecture as illustrated previously in Figure 4.2. For pilot testing the input layer comprised 251 nodes, which were the offender's characteristics. Exploratory work on the sub-sample was undertaken in an attempt to understand the pattern of correlations between variables and the nature of the dimensions underlying them. All 251 variables of the data were subjected to principal components analysis (PCA). Some of the variables could not be correlated because they showed no variance in the sub-sample used in the pilot study. These were duly excluded, leaving 135 variables for analysis. This reduction in independent variables improved the relative balance of cases (154) to the number of variables (135). Although this is below the recommended 2:1 ratio (Kline, 1994), Kline points out that the more important consideration is the number of cases to eventual factors.

The remaining variables showed a good factor structure: 115 variables indicated multiple observed correlations (2+), and residual values were small across the variables suggesting the presence of underlying factors. Communality is an estimate of the common variance within a single variable based on its correlations with the other variables. Communalities indicated that the solution explained between 66% and 99% of the variance in each individual variable. 41 components explaining 84% of the variance were identified. Varimax rotation was used to identify factor loadings on the components extracted. None of the components was alone responsible for a disproportionate amount of the variance in the data. For example, the first component explained 6.6% of the variance on its own; and the second and third components explained a similar albeit reduced amount of the variance (6.0% and 4.6% respectively). This suggested that there was minimal overlap between the variables making up the components, and signalled that retaining individual variables for modelling may be fruitful in making use of numerous sub-models within the data.

[11] During the early pilot testing stage described in the present chapter official reconviction information was not available because the analyses were performed too soon after the OASys assessments.

The number of hidden units was different in each pilot testing experiment as described below. In each experiment the hidden unit layer was fully connected to 1 output unit representing the occurrence of any further offending.

The learning rate of the model was set at 0.01, and the momentum at 0.9. As recommended after Chapter 3, in preparation for piloting the data a programme was written to scale the data automatically into the 0-1 activation range. This scaling programme was applied to the data prior to reading the cases into the PDP++ software (O'Reilly et al., 1995).

5.2.3 Training and testing.

The cases input into the model were the offender patterns (*n*=154). The process of training and testing, the weighting scheme for initial inputs, and the method for case sampling during training and testing (leave one out), was as described in Chapter 4.

The stopping time for training, 200 epochs, was selected based on previous experience of modelling the data suggesting that the model's performance attenuated most rapidly after this point. During testing the model was judged to have accurately learned the task if activation in the output unit was greater than 0.5 when the target in the output unit was 1; and less than 0.5 when the target in the output unit was 0.

5.2.4 Measure of accuracy.

As described in the General Method (Chapter 4, section 4.5.5), model performance at discrimination was measured using d-prime analysis (*d*'). Near zero values of *d'* indicate poor discrimination of re-offenders from non re-offenders, higher values indicate better discrimination. Analysis of *d*' is affected by its limits for whole numbers: a perfect score of 0 or 1 is considered invalid. Therefore in the pilot study a score of 0 was replaced with 0.00000000000000000001 and a score of 1 was replaced with 0.99999999999999999999.

5.2.5 Design and analysis.

To examine the feasibility of a connectionist model an objective was to test the impact of additional resources by varying the number of nodes available in the hidden layer. Seven connectionist models were created each with the general architecture described above but with differing numbers of hidden layer units. Models with 0, 2, 5, 10, 20, 50 and 100 hidden units were constructed and applied to the offender data. Inclusion of a comparison model with 0 hidden units represented traditional regression analysis within the models. In order to examine data fitting at each level of model complexity, performance of the models was first tested against the same cases seen during training (test on training cases). Models were then re-initialised and tested on unseen cases using the 'leave-one-out' cross validation procedure.

Since the eventual performance of each connectionist model is also a function of the starting weights (randomly) assigned, this was minimised by performing the leave-one-out procedure 20 times for each of the models. As each simulation began its training with a new random initialisation this was intended to limit the noise caused by the variation in starting weights.

A more formal comparison of traditional statistical modelling and connectionism was provided by testing each case with a model derived by discriminant function analysis (DFA). The same procedure for training and testing applied to the connectionist models was applied to the DFA model (leave-one-out). Thus each test case was classified by the function derived from the remaining cases in the analysis.

To control for data fitting the data used by the connectionist model were also randomised. This was done by re-assigning the output labels (re-offender or non re-offender) in the sample, ensuring that the overall proportion of re-offenders was retained (23%). Retaining the proportion of re-offenders ensured that an alternative ratio of cases was not an additional factor in the decision-making of the random baseline model. The random assignment was repeated twenty times to create 20 different random baseline models to use for training and testing. In the control condition, the number of units in the hidden layer was set to 130. This was an arbitrary selection based on $n_{\text{inputs}}/2$, rounded upwards, but made the random baseline model roughly comparable to the 100 hidden unit model learning the actual data. Each of the 20 different random baseline models had this configuration.

5.3 Results

5.3.1 Effect of hidden unit resources.

Each of the connectionist models was trained to 200 epochs whereupon predictive accuracy was tested by re-presenting the training cases to the model as new cases for classification. Results of testing on the training cases are shown in Table 5.1 below.

Table 5.1

Accuracy of Models Tested on Training Cases

Model (*n* hidden units)	Overall accuracy	TPR	FPR	*d′*
0	.97	.87	.00	4.18
2	.97	.88	.00	10.66
5	.98	.90	.00	10.77
10	.98	.91	.00	10.83
20	.98	.91	.00	10.88
50	.98	.92	.00	10.91
100	.98	.93	.00	10.95

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); *d′*= dprime

Table 5.1 shows that, after 200 epochs of training, models were fitting the data well with a high level of accuracy in positive predictions both in terms of the rate of true positives (hits) and avoidance of false positives (false alarms). This is reflected in the *d′* values which are high in all models. A difference in data fitting is evident between the 0 unit model and the remaining models. This is also illustrated in Figure 5.1 to show more clearly the difference in data fitting between the models with a hidden layer and the model without, as apparent after 100 epochs of training.

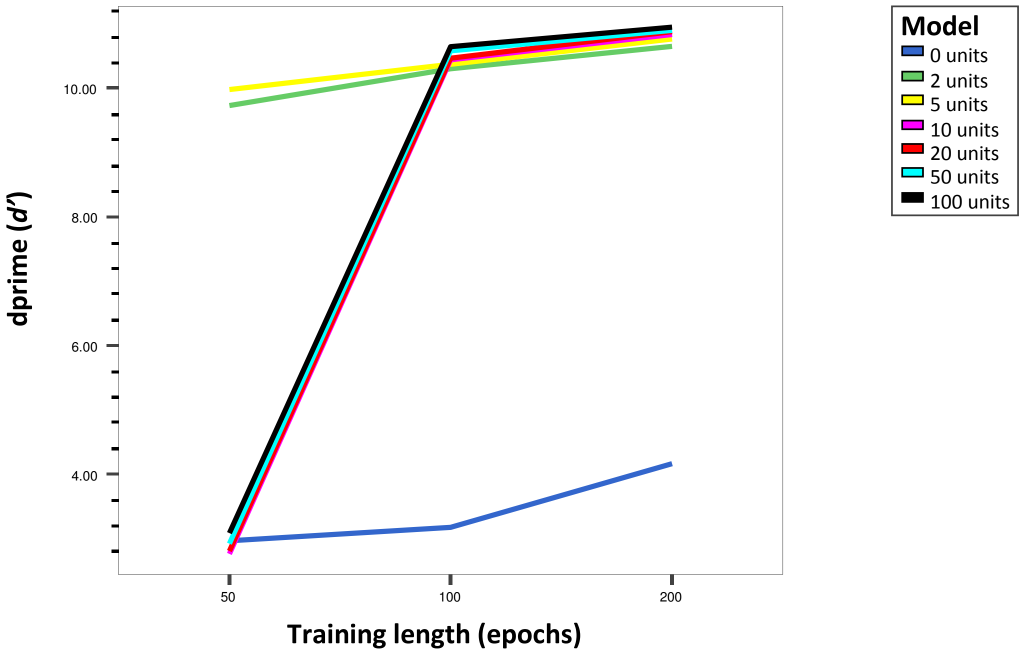


Figure 5.1. Performance of models tested on training cases

To validate the models on unseen cases, each model was re-initialised and subjected to the leave-one-out cross-validation procedure. Results in Table 5.2 below show the average level of accuracy on unseen cases. Positive predictions are notably less accurate than they were on the training cases, with the percentage of 'hits' roughly half their previous levels (Table 5.1). Contrary to the results of testing on training, on cross-validation the models with fewer resources in the hidden layer appeared to do best. This is evident from slightly higher *d'* values in the 0 and 2 unit models relative to the other models. Figure 5.2 shows how the cross-validation performance of each model varied prior to 200 epochs.

Table 5.2

Accuracy of Models Tested on Withheld Cases after 200 Epochs of Training

Model (*n* hidden units)	Overall accuracy	TPR	FPR	*d'*
0	.79	.43	.10	1.10
2	.79	.49	.12	1.15
5	.76	.37	.13	0.82
10	.71	.37	.19	0.54
20	.77	.43	.13	0.93
50	.76	.40	.13	0.85
100	.73	.40	.17	0.71

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); *d'*= dprime

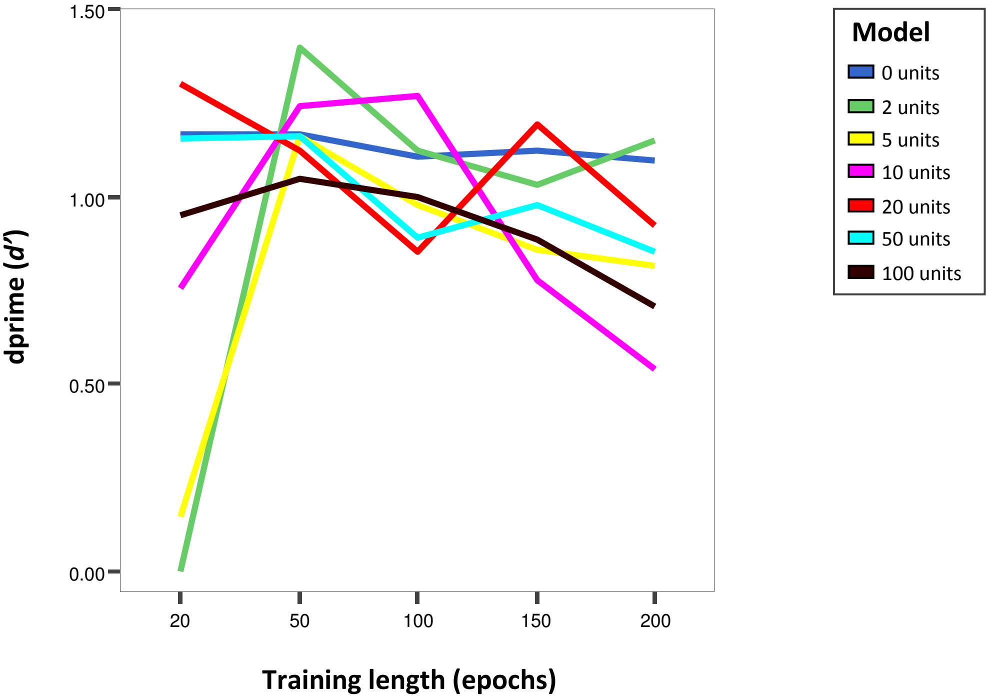


Figure 5.2. Performance of models tested on withheld cases

Figure 5.2 indicates that the optimal stopping point for connectionist modelling of the pilot data may be earlier than 200 epochs. A decline in performance with further training is seen after 50 epochs in the majority of the models. In view of the increase in performance between 50-100 epochs seen in Figure 5.1, the corresponding decline in Figure 5.2 may be indicative of overfitting or memorising the patterns of the training data at the expense of the minority criterion group (re-offenders). This may explain the substantial reduction in 'hits' and the increase in 'false alarms' seen between Tables 5.1 and 5.2. Since the withheld cases do not perfectly resemble the training cases the model is prone to increasing errors in its predictions as it over-learns the training data. A better stopping point may be at around 50 epochs; all bar two models seemed to peak in performance after this amount of training.

The performance of each connectionist model is also a function of the starting weights (randomly) assigned. The variation in starting weights can influence the results produced by a model. To ascertain the effect of this in the present data, the two best performing models were then subjected to multiple simulations. Each simulation began its training with a new random initialisation. Twenty different simulations of the present data were performed with the 2 unit model and the 0 unit model. Each model was trained to 200 epochs with testing at regular intervals. Figure 5.3 shows the average performance of the models over the multiple simulations.

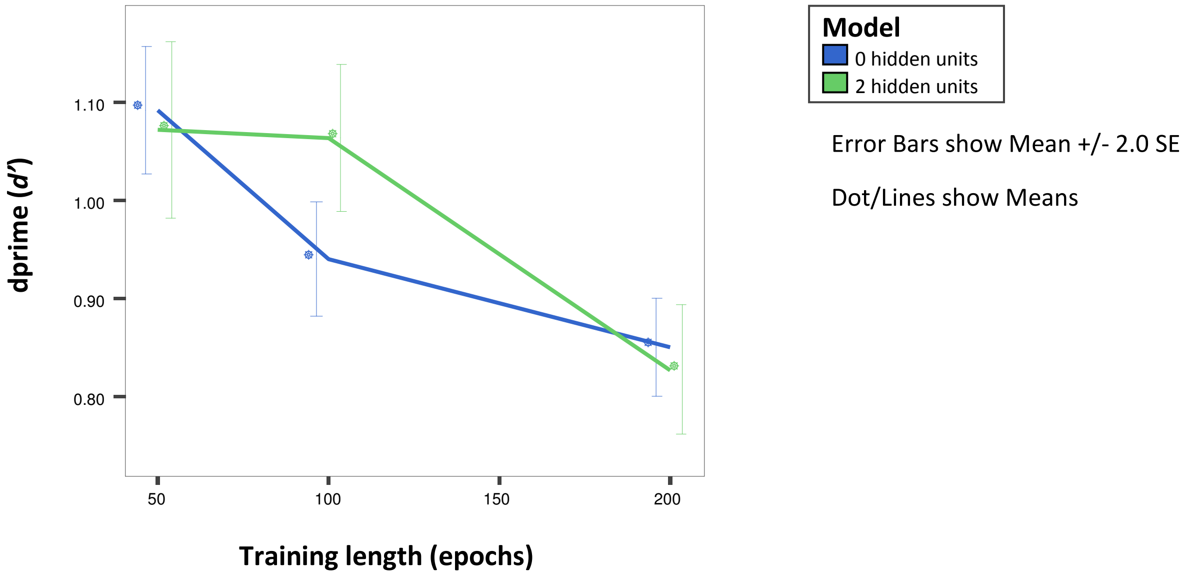


Figure 5.3. Performance of models tested on withheld cases over 20 simulations

Figure 5.3 supports the earlier results showing a decline in accuracy rates with further training. The best performances are early on in training, although there is greater deterioration in the average performance of the 0 unit model than in that of the 2 unit model between 50 and 100 epochs. In the 2 unit model *d′* only changes from 1.07 to 1.06 between the two training intervals. After this point it seems clear that performance worsens in both models.

5.3.2 Comparison with conventional statistical model.

DFA was applied to the data with re-offending as the criterion variable and the offender characteristics as predictor variables. The leave-one-out classification procedure was enabled to allow a direct comparison with the testing procedure for the connectionist models. A single discriminant function was calculated and the value of this function was significantly different for re-offenders and non re-offenders (χ^2=191.6, df=126, *p*<0.001). Overall the DFA model successfully predicted outcome for approximately 70% of cases, with accurate predictions being made for approximately 49% of offenders who re-offended. Results are shown in Table 5.3 alongside those for the best performing connectionist model (2 hidden units, 50 epochs of training).

Table 5.3

Accuracy of Traditional Statistical Technique Alongside that of the Best Connectionist Model

Model (*n* hidden units)	Overall accuracy	TPR	FPR	*d′*
Connectionist model:				
Train cases	.89	.59	.00	9.72
Test cases	.80	.31	.06	1.07
Discriminant Function Analysis:				
Train cases	.99	.97	.01	4.29
Test cases	.70	.49	.24	0.66

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); *d′*= dprime

Table 5.3 shows that the DFA model was able to match those cases on which it had been trained very well (99% correct). Although the DFA model was more accurate in identifying re-offenders it had been trained upon, on test cases it was less accurate, particularly in making negative predictions (i.e., non re-offenders). Non re-offenders were correctly identified 76% of the time by the DFA model, compared to 94% true negative classification in the connectionist model (1-FPR). This may therefore indicate a cautious approach to the classification of the minority class, re-offenders, by the connectionist model.

Performance on unseen cases during the cross-validation of both models reduced compared to that observed on the training cases. This is consistent with the results in Table 5.2 previously with a variety of connectionist models. Table 5.3 shows that the size of this shrinkage was greater in the connectionist model than in the DFA model although the sensitivity of the technique in discriminating the re-offender signal from the background noise remained stronger in the connectionist model compared to the DFA model (*d′*=1.07 and *d′*=0.66, respectively). As mentioned, this could be related to the improved accuracy of the connectionist model in correctly identifying non re-offenders.

The pattern recognition abilities of connectionist models may be such that they are liable to learn and generalise artificial patterns as accurately as genuine patterns. This would be problematic: whereas genuine patterns are presumably related to underpinning processes (social, psychological, and neurological), artificial patterns randomly generated have no such systematic underlying process. Connectionist models should therefore be unable to detect an underlying function or learning rule with which to correctly classify new artificial cases. To test this possibility the learning baseline was manipulated by testing the model against randomised patterns of the data.

5.3.3 Manipulation of the learning baseline.

Table 5.4 below shows the cross-validation performance for the random baseline model.

Table 5.4

Accuracy of Connectionist Models on Cross-Validation Against Random Baseline after 200 Epochs of Training

	Overall accuracy	TPR	FPR	*d′*
Mean (SD)	.66 (.05)	.21 (.05)	.21 (.07)	-0.01 (.32)

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); *d′*= dprime

Results in Table 5.4 relating to randomised patterns, are somewhat different to those in Table 5.2 where genuine data was the subject of model training. Here the TPR values are equivalent to those for FPR, and *d′* is noticeably lower than seen in the genuine data models. The mean *d′* is close to zero indicating no differentiation of re-offender signals from other cases. The average values for accuracy, hit rate, and false alarms could have been reached by model assessment of the prior probabilities in terms of the frequency of criterion responders to non responders (average proportion of dominant class = 76%). Hence the models trained against a random baseline were classifying at a rate similar to but beneath the default strategy of deferring to the dominant class. *d*ʹ is most indicative of model learning: Figure 5.4 below shows how average values changed with length of training.

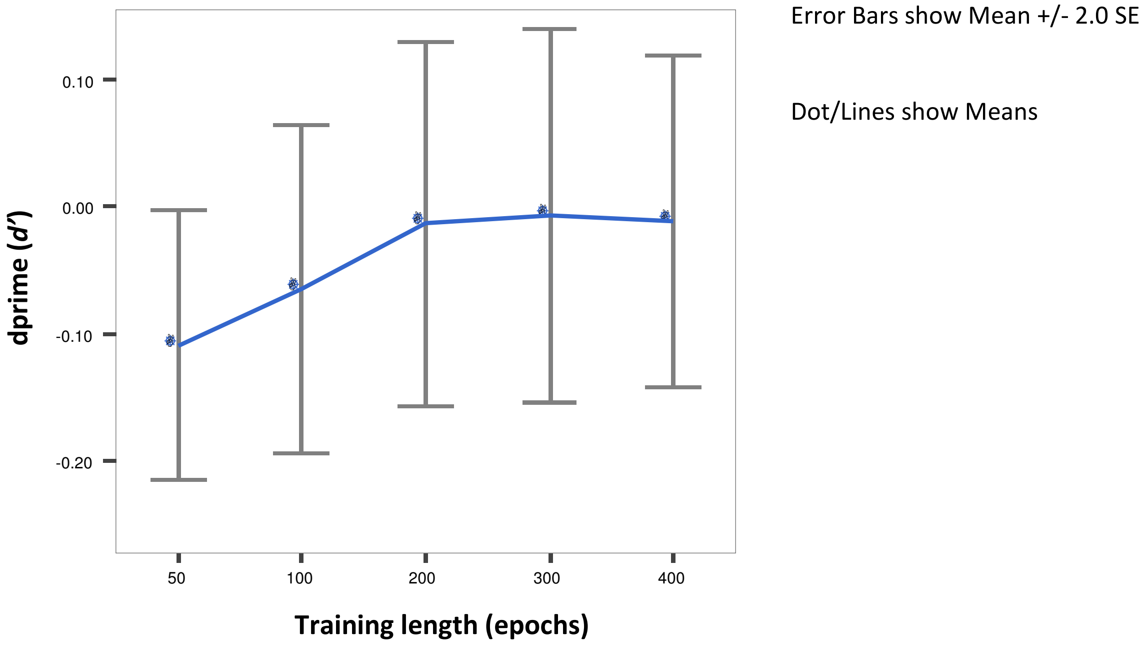


Figure 5.4. Mean performance of 20 models tested against random baseline: variation with training

As shown in Figure 5.4 the average *d'* remained around 0, consistent with chance performance. The error bars show that the high level of variability within models did not decrease. This may suggest a) the various random starting weights appear to be having a great effect upon the solution each time, and the related issue b) that the models had not learned a function with which to classify unseen cases.

5.4 Discussion

This chapter aimed to implement a connectionist model on a sub-sample of cases in order to assist learning about model development in preparation for larger simulations using the total sample of probation cases. The sub-sample selected, comprising all sexual offenders, was discrete and distinct from general offender populations.

Innovations of connectionist modelling relative to traditional statistical models include the hidden layer resources and the training length, and these were tested on the small sample. In light of the danger of overfitting associated with these parameters, highlighted after Chapter 3, performance of connectionist models and that of a traditional statistical model (DFA) on the training data was compared to that on unseen cases. As a further test of the potential for overfitting, the present chapter compared connectionist model performance on genuine data against that arising from testing the same model using randomised patterns.

Results of variation to hidden layer resources showed that best performances were found in models with fewer hidden units. Although the *d′* values of all models declined with training time, the models with fewer hidden nodes fared best in maintaining their performance. This may relate to the fact that the models were already highly complex with a low ratio of data points (154) to predictor variables (251). Marshall and English (2000) previously attributed the performance of their connectionist model to the additional processing available in the hidden layer. Their successful model however had an architecture with five input nodes, three hidden nodes and 9,084 data points. The present study had a greater number of predictor variables than data points which encourages a model to over-learn the peculiarities of the training data, and this was exacerbated by additional hidden units as reflected by the shrinkage in performance under cross-validation. This problem was also evident in the DFA model, which made many more false positive predictions on unseen cases than did the connectionist model. Over-estimation on the training sample using linear statistical analysis is not surprising given the high ratio of predictors to cases, where the absolute minimum ratio recommended for prediction tasks is 1 to 10 (Miller & Kunce, 1973).

In explaining the over-fitting in both models it is also important to look at the method used to cross-validate, leave-one-out. This method of sampling was used in response to exploratory work indicating that in the low base-rate circumstances there were problems in the representation of re-offenders within training and testing samples when a split sample design was used. A split sample design would therefore require a very large number of random data splits in order to reduce the variance. The advantage of leave-one-out on the other hand was that it includes all available cases (bar one) and therefore keeps the variance to a minimum by making maximum use of the available information. The disadvantage of leave-one-out is that when the stopping criterion for training is linked to the test prediction error, network learning at each iteration is based entirely upon an individual case. In these circumstances the undue impact of the individual case on network learning renders the model vulnerable to over- or under-training (Tourassi & Floyd, 1997). Over-training will occur if the testing pattern and another training pattern are similar or the same: the network will reduce its test error not because it has learned the underlying function but because it has memorised the particular training patterns that are similar to the test case that is left out. Tourassi and Floyd (1997) showed that this could be addressed by selecting an early stopping point in training. Although in the present study model learning was unrelated to the test case, uniformity in the training sample may explain why models had arrived at their best performances relatively quickly, generally by 50 epochs of training.

Offender data often contain patterns that are identical for re-offenders and non re-offenders (Caulkins et al., 1996). The offending of the specific sub-sample studied, sexual offenders, may be more likely to remain undetected than is the case for general offenders. Sexual offenders also often have fewer previous convictions and lower levels of the problems screened by OASys than do general samples of offenders (Howard et al., 2006). This lack of variation in the input data, and similarity between re-offender signals and background noise, increases the chances that the connectionist model would resort to the prior probabilities in the data and learn to reduce its error by predicting the more frequent outcome. Tourassi and Floyd (1997) found that leave-one-out was liable to do this using a medical data-set in which many similar cases had opposing target outputs. This might explain why in the current study the best performances of the models were not sufficiently different from the base-rate (20% incorrect compared to a 23% base-rate of re-offending). Results were however different to those produced using a random baseline. Models trained using randomised patterns were unable to learn a decision rule with which to classify new cases. Models using genuine data were able to classify new cases at a level better than chance. This suggests that, rather than fitting to any data pattern, connectionist modelling of the data was able to detect relationships in the genuine data to enable accurate classification of unseen cases. This gives confidence in the concept examined in this thesis.

The evidence from the current study gives rise to a number of recommendations for implementation in the next stage of applying connectionist models to offender data. Some of these recommendations are dependent on each other. The most apparent aspect of the current study was the small data sample. Although this was intentional at the outset since it facilitated learning on a manageable sample, this also caused problems related to the number of re-offenders in the sample and the ratio of predictor variables to cases. The first requirement for further application of connectionist modelling is therefore the use of a larger and more heterogeneous sample. Not only will a larger sample allow better modelling of interactions within the data, but there is also likely to be a greater level of variation between cases. These are aspects that may be suited to connectionist modelling of offender data. This may also impact on the accuracy of predictions of re-offending by sexual offenders. A wider model may alternatively obscure some of the specific learning to do with sexual offenders.

The limited number of re-offenders available for training meant that the method of sampling selected for cross-validation was leave-one-out. Important learning about this method has been gained from the current work including the propensity for the method quickly to over-fit where there is an imbalance in the number of cases at each level of the target output. An obvious solution to this problem would be to select a larger sample with a more equal proportion of re-offenders to non re-offenders and then to re-test using the leave-one-out procedure. Since testing in this procedure at each iteration is based on a single case, a better alternative might be *k*-fold cross-validation (Stone, 1974). In this procedure training is on n – *k* (e.g., withholding 10 cases sequentially selected at the start of each training process) and testing is on the *k*-fold. Leave-*n*-out should be compared to the split sample approach to verify whether there are differences in results depending on the method adopted. *k*-fold would also be a quicker means of cross-validation due to less computational costs.

The current study intentionally used all available predictor variables to avoid eliminating variables that may be important in sub-models within the data. With a larger and more varied sample of data this will be an important strategy to maintain. There may however be other connectionist models with fewer input parameters that perform as well or better. Eliminating unimportant variables is likely to reduce further the risk of over-fitting; however a starting point for future simulations will be to increase the sample size thereby allowing retention of all potentially important predictor variables.

CHAPTER 6

6. Results I: Parameter Variation

6.1 Background

The review of recidivism risk assessment in Chapter 2 concluded, with reference to a recent meta-analytic review (Yang, Wong, et al., 2010), that the field had reached a plateau in predictive efficacy with most current measures achieving no more than 'moderate' accuracy. Comparison studies involving more than one instrument show that each generally performs at between .56-.71 in terms of the area under the receiver operating characteristic curve (AUC) with performance rarely exceeding .75 (Coid et al., 2009; Farrington et al., 2008; Kroner & Mills, 2001; Snowden et al., 2007; Yang, Wong, et al., 2010). A 'glass ceiling' effect is therefore in operation which may relate to the method of combining predictor variables (Yang, Wong, et al., 2010). Instruments, principally combining variables through weighting schemes such as least-squares regression or simple summed points scores (e.g., Burgess, 1928; Nuffield, 1982), are seen as essentially interchangeable once methodological differences including type of sample, length of follow-up, and definition of the criterion variable have been controlled (Kroner et al., 2005; Schwalbe, 2007; Yang, Wong, et al., 2010). Equivalent predictive accuracy between 'second generation' static factor instruments and 'third generation' instruments also incorporating dynamic items suggests that existing methods are all drawing from the same pool of variance that can be captured by static historical indices, tapping long-term anti-social orientation, with little incremental validity by addition of current dynamic predictors (Yang, Wong, et al., 2010).

Coid et al. (2011) have recently recognised the limited influence of dynamic factors in a study examining the predictive ability of measures' constituent items. In each of the three leading risk assessment measures studied, a minority of items were independently predictive and in each case sub-scales comprised of these items were no more predictive than the overall risk assessment measure. A 'super-instrument' comprised of a combination of the independently predictive items from each risk assessment predicted recidivism with an AUC of .72, only slightly better than the original instruments (Coid et al., 2011). Other than 'negative attitudes' the super-instrument contained all static risk factors. Coid et al. suggest that measures incorporating static factors may be impossible to improve further; possibly due to the unreliable nature of dynamic risk factors, especially when measured prior to custodial release. Poor reliability can relate to dynamic factors' inherent ability to change in response to intervention, or the impact of subjective clinical judgement on their assessment. Inability to make full use of dynamic factors under existing methods would explain the field's limited impact in advancing prediction beyond accuracy achieved by models based on static factors only.

Limited predictive accuracy may also be due to the impact of statistical issues on generalisation performance with criminal justice data. Noise on this data also occurs on the criterion variable due to variation associated with police practice and detection rates, as well as how recidivism is defined in the research (e.g., recidivism within six months versus within two years). These variations can cause confusion between criterion cases and non-criterion cases. Such noise impacts on accurate generalisation by encouraging a statistical model to capitalise on chance variation in the construction sample (Gottfredson & Moriarty, 2006). This may explain existing methods' limited predictive accuracy on new samples since different samples may require a different combination of items yet conventional statistical methods are designed to find a single optimum solution for discriminating all cases. For example one of the leading measures currently is a static actuarial assessment named OGRS (Copas & Marshall, 1998; Taylor, 1999). Although in a large sample of prisoners this predicted subsequent recidivism better than any other measure with an AUC of .76 for male offenders, this reduced to .68 for female offenders (Coid et al., 2009). Since male gender was associated with recidivism overall in the OGRS construction sample, and logistic regression generates one solution that best discriminates recidivists, the prediction for female offenders is compromised. The field therefore needs to move beyond methods of combining variables that focus on 'main effects', and begin to explore alternative methods that can account for interactions between predictor variables.

Chapter 3 introduced connectionist models as pattern recognition systems capable of modelling interaction effects. The chapter explored the performance of such models on data that were characterised by low-base rates, noise, and numerous predictor variables, comparing the approach to regression analysis across a range of fields including medicine, business, and transport. Results showed that connectionist approaches have frequently performed better under cross-validation in these conditions. Applications to offender data have been few in number, and inconsistent in their findings, some supporting the value of the approach (Brodzinski et al., 1994; Palocsay et al., 2000), others inconclusive (Caulkins et al., 1996; Yang, Liu et al., 2010) and one finding that the approach generalised poorly between training and testing cases (Grann & Langstrom, 2007). This raises the need for further research using connectionist models. Pilot testing using a sub-set of the present data (Chapter 5) raised questions about the relationship between generalisation, sample size and cross-validation methodology. The bias of the trained model to individual data points and the variability associated with the testing sample must both be low enough to enable accurate generalisation. Chapter 3 indicated the importance of sample size with an association between larger training samples and improved performance. Pilot testing in Chapter 5 therefore sought to maximise the training sample by using leave-one-out cross validation. Although this appeared to reduce the variance relative to split sample testing, connectionist model performance remained equivalent to that of DFA. Increasing the model complexity by adding hidden layer units only increased model overfitting and reduced performance. However the pilot testing model was already complex, with the number of predictor variables (251) outnumbering the number of cases (154). In addition, the base-rate of recidivism in the particular sub-sample was low, meaning that the model had few opportunities to learn patterns for re-offenders, despite the use of the leave-one-out sampling procedure.

The current chapter therefore took forward the learning from pilot testing by applying a connectionist model to the total sample. This reflects better the clinical prediction task in which all cases are subject to recidivism risk assessment. The present chapter therefore moves from piloting into the development of a useable predictive model. Since the pilot sample, comprising sexual offenders, had a lower base-rate than does general offending (Chapter 2), widening the data-sample from the specific sub-sample to all offenders may allow the model better opportunity to learn a more general function for predicting recidivism.

Due to the computational costs of leave-one-out, requiring long and expensive training time, the current chapter also explored any difference in performance by also using 10-fold cross validation (Stone, 1974), discussed in previous chapters. There is some evidence that this can improve predictive accuracy by making model learning less focussed on individual cases (Kohavi, 1995; Tourassi & Floyd, 1997) as well as reducing training time.

The present chapter also sought to explore the impact of the length of training and hidden layer processing capacity by systematically varying these parameters within the larger sample model. Three studies using offender data varied these parameters (Caulkins, et al., 1996; Grann & Langstrom, 2007; Palocsay, et al., 2000). These variously found effects of training length and learning rate (Caulkins et al., 1996; Palocsay et al., 2000), and hidden layer size (Grann & Langstrom, 2007; Palocsay et al., 2000). The study by Palocsay et al. and that by Grann and Langstrom, found opposite effects of hidden layer size although the latter study did not present data relating to the manipulation of this parameter. Variation between these studies illustrates that the network parameters depend on the data at hand, even within offender samples, reinforcing a more general theme in the literature regarding the application of connectionist models.

6.1.1 Aims of the chapter.

The present chapter therefore aims to develop a connectionist model for the prediction of recidivism. The predictor variables will include all routinely collected static and dynamic predictor variables from OASys (Howard et al., 2006; see Chapter 4 for a description), including the OGRS measure. Variables will be combined using connectionist and linear statistical models. The main aim of this chapter is therefore to verify the performance of a connectionist model against existing approaches to see whether this makes any better use of the available dynamic variables.

In developing an ‘optimum’ connectionist model parameters subject to variation will be the training length and number of hidden layer units. The optimum model will also be trained and tested with a faster and more efficient data sampling procedure (described below) to check for any differences in performance. Notwithstanding model optimisation it is not clear whether the complexity of connectionist models militates against generalisation performance.

6.2 Method

6.2.1 Design and analysis of models.

Models were designed and analysed as described in the General Method chapter (Chapter 4). Thus the input / predictor variables were the 236 offender characteristics from OASys and these were used as inputs to the models. The criterion of any re-offending within the follow-up period (minimum 30 months) was used as the model output. Discriminant function analysis (DFA) was used as the comparison linear statistical model in addition to a network without a hidden layer (a two-layer model). OGRS-II was selected as the applied model, as used in clinical practice within NOMS. OGRS was the only model not developed on the present data; however it was developed on UK prison/probation data (see Chapter 4).

6.2.2 Parameter variation.

In developing a connectionist model for the present offender data, the stopping time for training and the number of hidden units were each varied. Training time followed that described in the General Method chapter (Chapter 4, section 4.5.4). Thus models were trained up to 1,000 epochs and were tested at each 100 epoch increment to examine the effect of training length. Complexity in the hidden layer of the connectionist models was examined by varying the number of hidden units as follows: 25, 50, 100. The best performing model was taken forward in the examination of data sampling.

6.2.3 Variation to data sampling.

Models were evaluated using the procedure for training and testing described in Chapter 4, namely leave-one-out (Efron, 1983). This method was applied to the DFA, the two-layer model, and the connectionist model.

The same data were also evaluated using 10-fold cross-validation, an example of *k*-fold cross-validation (Stone, 1974). In *k*-fold cross-validation the dataset is randomly split into *k* mutually exclusive partitions of the data of roughly equal size. The model is then trained on *k*-1 sub-sets and tested on the withheld sub-set. As with leave-one-out, after testing the withheld sub-set is returned to the training data and a second sub-set is then withheld. The rotation continues until each partition has been subject to testing and the cross-validation estimate of accuracy is the overall number of correct classifications divided by the number of instances in the data-set. This is a less intensive means of cross-validation than leave-one-out because the model is trained on a smaller set of patterns overall. Partitioning the data into 10 folds is expected to retain the low levels of bias resulting from leave-one-out, while reducing the variability seen when that method is used in smaller samples (Kohavi, 1995). Kohavi (1995) suggests that 10-fold cross validation should be used instead of leave-one-out with real-world data, even if computational power allows the use of more folds.

To minimise the probability that re-offenders were not disproportionately located in one or two 'folds', five different randomisations of the data were performed. Five different models were therefore implemented for evaluation using 10-fold cross-validation, with the mean accuracy rates taken as representing the performance under this method of data sampling. In the present chapter, 10-fold cross-validation was applied to the connectionist model and the two layer model but there was no automatic function for 10-fold cross-validation in SPSS for DFA. Thus DFA was evaluated with leave-one-out sampling.

6.3 Results

6.3.1 Optimising the connectionist model.

6.3.1.1 *Parameter variation*.

Results of variation to training time in each model are shown in Figure 6.1, and accuracy rates at the end of training are detailed in Table 6.1. Performance on test cases was assessed using leave-one-out and measured using dprime (*d'*) as described in the General Method chapter (Chapter 4).

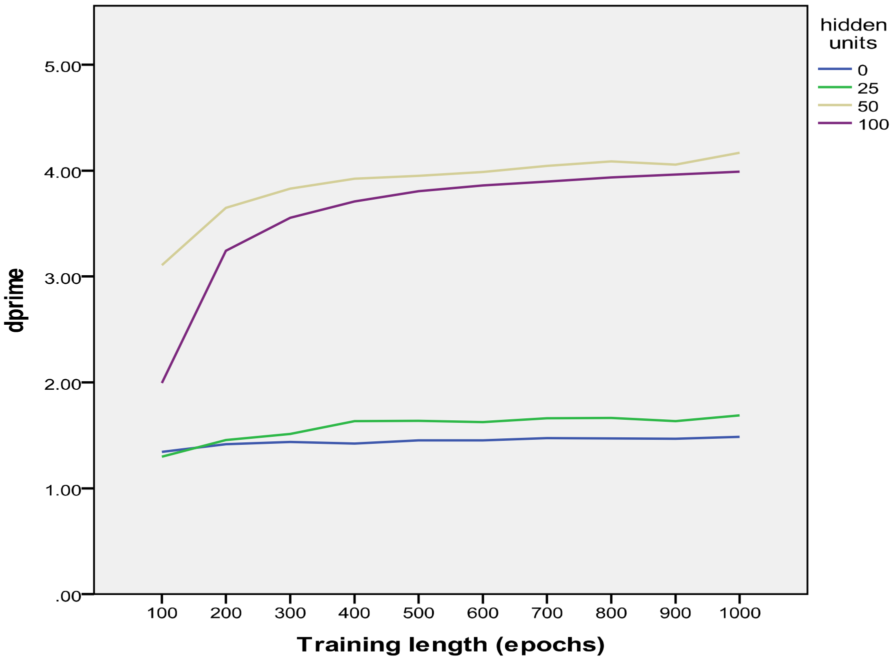


Figure 6.1. Performance of models by length of training and hidden layer size.

Table 6.1

Classification Accuracy of the Different Models After 1,000 Epochs of Training

Model complexity (hidden units)	TPR	FPR	Overall accuracy	(95% CI)	*d′*
100	.98	.03	.98	(.97-.98)	4.01
50	.99	.04	.98	(.98-.98)	4.17
25	.48	.04	.68	(.66-.69)	1.69
0	.83	.29	.78	(.77-.79)	1.49

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); 95% CI = 95 percent confidence interval

Figure 6.1 shows an effect of hidden layer complexity and, in the models with more resources, an effect of training length. Models with 50 and 100 hidden units were shown to outperform models with 25 or fewer hidden units, with the higher levels of *d′* in the more complex models increasing further with training and then reaching asymptote. Table 6.1 shows that the discrimination ability of the best models in terms of the distance between the means for recidivists and non recidivists, was at least four times as large as the standard deviations of the two distributions (*d′* column). The 50 unit model was marginally superior to that with 100 hidden units, and had the advantage of being more parsimonious. The model trained with only 25 hidden units performed only slightly better than the model with no hidden units (two layer model), suggesting that the present data contain multiple sub-models and interactions.

6.3.1.2 *Variation to data sampling*.

The model with the best configuration after leave-one-out, with 50 hidden units, was taken forward to examine the different method of data sampling. Results of the 10-fold cross-validation across the five different randomisations of the data are shown in Figure 6.2 below, alongside the previously reported results using leave-one-out.

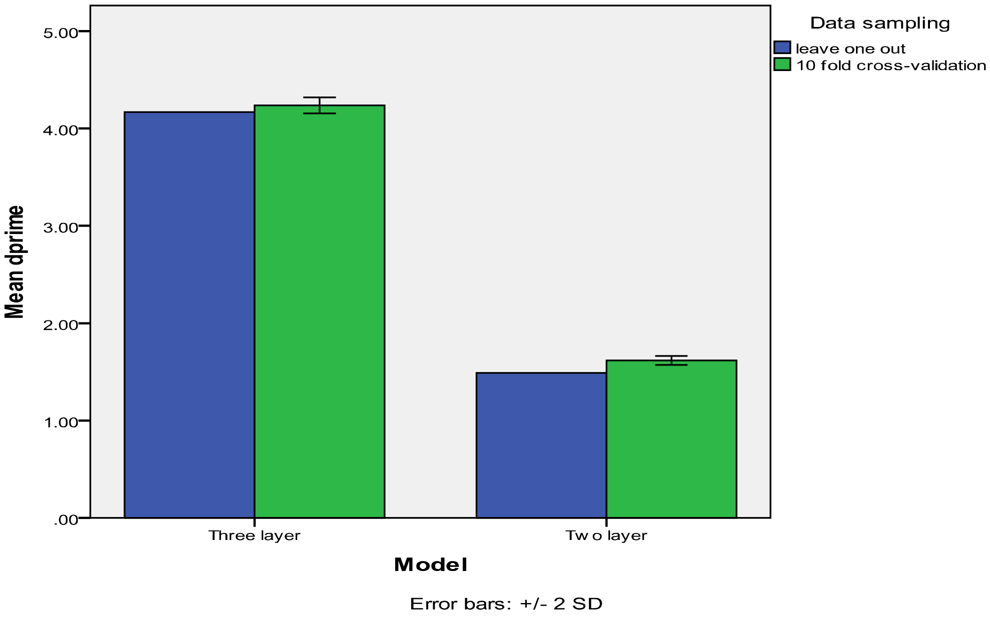


Figure 6.2. Mean accuracy (*d'*) of models evaluated using 10-fold cross-validation compared to using leave-one-out

Table 6.2

Mean Classification Accuracy of the Three Layer and Two Layer Models Using 10-fold Cross-Validation

Model	TPR	FPR	Overall accuracy	95% CI	*d'*
Three layer, *M* (SD)	.99 (.00)	.02 (.00)	.98 (.00)	.98-.99	4.24 (.04)
Two layer, *M* (SD)	.86 (.00)	.30 (.00)	.80 (.00)	.78-.81	1.62 (.02)

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); 95% CI = 95 percent confidence interval; *d'* = dprime.

Figure 6.2 illustrates that the mean accuracy figures are consistent with those generated using leave-one-out. Mean values in Table 6.2 are thus similar to those in Table 6.1 (50 hidden unit model). There was a stable level of performance across each of the different randomisations of the data, suggesting that fluctuations in the training and test sets did not have a profound effect on the model's classification.

6.3.2 Comparison of optimised connectionist model to alternative models.

The optimised connectionist model, with 50 units in the hidden layer and trained to asymptote, was selected for comparison with the alternative models. Predictive accuracy was validated using leave-one-out; results for each of the models is given in Table 6.3, and illustrated in Figure 6.3.

Table 6.3

Classification Performance of Models on Test Cases

Model	TP	TN	FP	FN	Overall accuracy (95% CI)	TPR	FPR	*d′*
Three layer	2369	1598	61	20	.98 (.98-.98)	.99	.04	4.17
Two layer	1973	1177	482	416	.78 (.77-.79)	.77	.29	1.49
DFA	1777	1232	427	612	.74 (.73-.76)	.74	.26	1.31
OGRS	1739	1119	540	650	.71 (.69-.72)	.73	.33	1.06

Note. TP= True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives; 95% CI = 95% confidence interval; TPR = True Positive Rate; FPR = False Positive Rate; *d′* = dprime

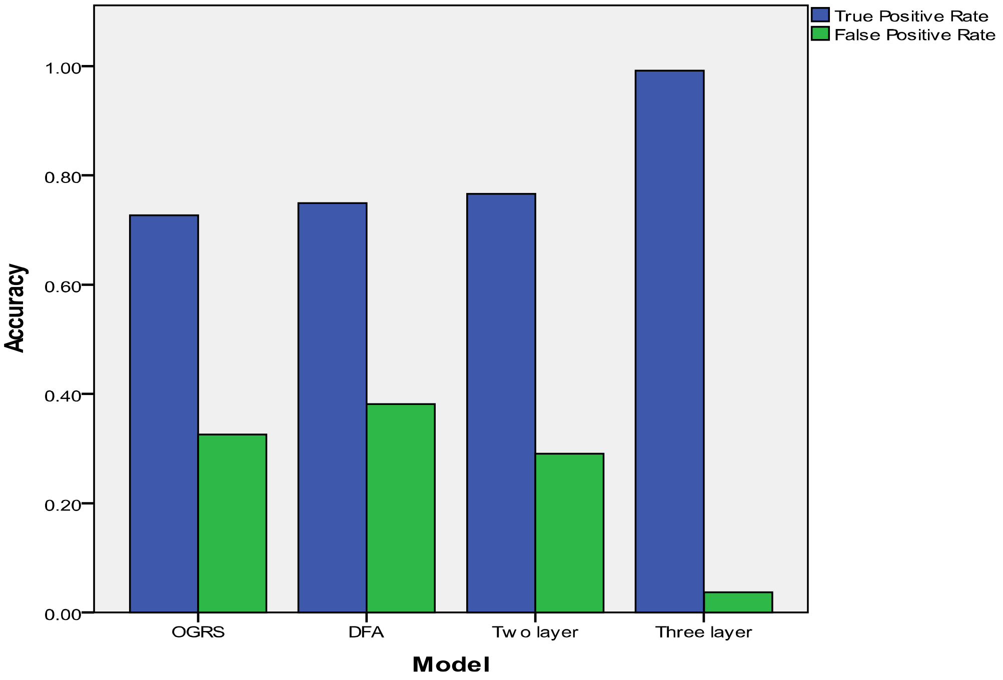
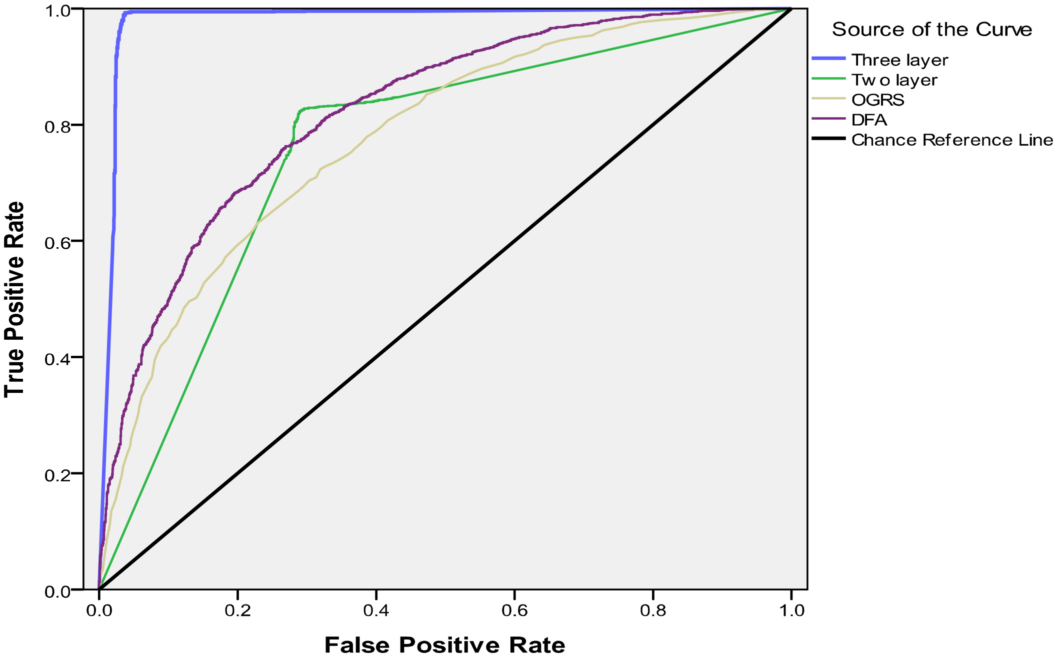


Figure 6.3. Rate of true positive ('hit') and false positive ('false alarm') predictions in each model

The three layer connectionist model stands out in Table 6.3 in terms of the identification of recidivists (TPR) and non-recidivists (1-FPR), reflected in the score for *d'*. The regression-based models, regardless of the incorporation of dynamic factors are all similar in terms of their level of accuracy. Thus DFA shows only a small improvement on OGRS. False positive rates are also similar to one-another in the regression-based models, wrongly labelling non-recidivists as recidivists at a rate of approximately .30 (Figure 6.3).

Since performance can depend on the cut-off selected for identification of the criterion cases, the accuracy of predictions of recidivism were further assessed using the area under the receiver operating characteristic (ROC) curves as described in Chapter 4. These plot the true positive and false positive rates at different classification-threshold settings (Figure 6.4). The diagonal line in Figure 6.4 represents chance classification.



Model	ROC / curve area (AUC)	Standard Error	p	(95% CI)
Three layer	.98	.003	.000	(.98-.99)
Two layer	.76	.008	.000	(.74-.78)
DFA	.82	.007	.000	(.81-.84)
OGRS	.78	.007	.000	(.76-.79)

Figure 6.4. ROC curve results on test cases (withheld from training)

The area under the ROC curve (AUC) values show that each model is better than chance in discriminating signals from noise, as shown previously with the *d'* value. The precision of the estimates for different samples in the same population is good as indicated by the small standard error values and narrow confidence intervals, particularly for the connectionist (three layer) model. The non overlapping confidence intervals associated with the AUC values for the three layer model compared to the conventional statistical models show that the difference in predictive accuracy using this approach is statistically significant.

6.4 Discussion

This chapter aimed to take forward learning from pilot testing and develop a useable connectionist model based on static and dynamic variables routinely collected by probation officers. The performance of the model was optimised and then compared to alternative regression-based models, including an applied model, to verify the existence of any benefit from connectionist modelling of the data. An alternative means of data sampling in cross-validation was also tried, to facilitate practical use of connectionist models.

The results of the chapter were clear in showing that the connectionist model far outperformed the conventional approaches. The performance of OGRS and OASys variables combined using linear regression, represented here by the DFA model, with AUCs of .78 and .82 respectively, is consistent with the literature in which OGRS scores approximately .76 (e.g., Coid et al., 2009; Snowden et al., 2007) and linear OASys scores .76, or .79 when OGRS information is also included (Howard et al., 2006). Using linear regression it is not therefore clear that the dynamic variables included by OASys lead to any improvement relative to the static historical variables included by OGRS. This supports suggestions that current models emphasising overall between-group differences overlook underlying patterns within the data (Brodzinski et al., 1994; Coid et al., 2011; Yang, Wong, et al., 2010). Thus although main effects predominantly captured by static variables are detected by conventional statistical models, interaction effects as may be more likely with dynamic variables remain undetected. Dynamic factors may influence the timing of recidivism, rather than the long-term probability, and therefore may be critical risk enhancing or risk reducing factors. This may explain the step increase in accuracy brought about by combining the same set of variables using a connectionist model thus leading to levels of predictive accuracy, AUC=.98, not previously seen in predictions of recidivism. Development of connectionist models may therefore have important practical implications both in terms of advancing psychological insight into the important risk factors for assessment and treatment of offender risk, as well as the direction of scarce public resources by offender management service providers.

Supporting the proposal that the connectionist model makes use of multiple sub-models within the data, in optimising the connectionist model there was a clear benefit from increasing the number of hidden units to 50. This architecture responded better to training than did the network with 25 or no hidden nodes. Doubling the number of hidden units to 100 did not improve predictive accuracy and therefore unnecessarily increased the model's complexity. The present work in optimising the model therefore accords with Palocsay et al. (2000) whose nine variables of offender data required just 26 hidden units to out-perform the logistic regression model. It disagrees with Grann and Langstrom (2007) where 84 hidden units were applied to ten variables and 404 cases of offender data, and found the shrinkage in performance between training and testing was larger in the connectionist model than in the bivariate regression model. Using the same heuristic as used by Grann and Langstrom (2007) on the current data would have led to a model incorporating 182 hidden units which was more than necessary for optimal performance.

The current chapter also established that using a slightly less rigorous sampling design for cross-validation did not harm model performance. In fact, a minor improvement was seen supporting previous observations (Kohavi, 1995). As suggested by Kohavi, this may be due to a marginal reduction in model bias, relative to leave-one-out, by reducing the variance between the model and the prediction on the test case. Even if performance were equivalent, or marginally inferior, 10-fold cross-validation would be of practical value, given the faster speed of training and testing. To put this in perspective, with the data structure of the optimised connectionist model, leave-one-out required an average of two months for training while 10-fold required little more than 24 hours for training and testing.

In the context of the pilot testing of the data (Chapter 5), the current findings suggest that the inclusion of a larger sample on which to train the models has been important. This adds evidence to suggestions that connectionist models respond better to large and heterogeneous samples (Yang, Liu, et al., 2010; see also Chapter 3). This was expected given the complexity of a connectionist model containing a highly populated input space (236 variables). In the pilot testing sample the number of inputs increased the variance by making the model highly specified to the training sample. This was minimised using leave-one-out, but the training sample remained relatively small and the problem was compounded by a low-base rate of target cases for model learning (recidivists). In the present chapter the size of the training sample and the low bias inherent to connectionist models seemed to provide the right conditions for accurate generalisation. One cannot however be certain that the difference is not due to a high but balanced base-rate (59%); performance might drop disproportionately with a more imbalanced outcome. This was not found in other applications of connectionist models, reviewed in Chapter 3, where connectionist models were often able to deal with low-base rate outcomes better than regression models (e.g., Marshall & English, 2000). This is of practical importance given that managing offenders requires differentiation of the expected timing of re-offending. Consequently this will be investigated in Chapter 7, addressing the proposal that better use of dynamic factors improves the prediction of time to recidivism.

A second practical issue concerns the important variables associated with the predictions. The accuracy achieved with the optimised model did not require a reduction to the input space and therefore it is not clear which dynamic variables must be manipulated in order to reduce the chances of the predicted outcome (in the case of re-offenders), or maintain the predicted probability of non re-offending. Given its importance clinically, the identification of key variables relevant to re-offenders and desisters will be considered in Chapter 8.

The present chapter found that a connectionist model can predict recidivism with very high accuracy (AUC=.98) after trialling different configurations of some of the available parameters. The chapter therefore supports one of the limited number of previous applications of connectionist models to offender data, in finding that predictive accuracy using the approach depends on the choice of network parameters relevant to the data at hand (Palocsay et al., 2000). It is therefore recommended that connectionist models should be more widely applied and developed for the prediction of recidivism. This must be mindful of the inherent problems of overfitting the training data, as seen when piloting a sub-sample of the present data.

CHAPTER 7

7. Results II: Predicting Time to Re-offending

7.1 Background

Chapter 2 reviewed the characteristics of offenders that have been found in research to be associated with recidivism. A consistent pattern emerged associating persistent offenders with high criminal history, anti-social associates and a deviant lifestyle. These features and underlying neurological and cognitive deficits may discriminate ‘life-course persistent’ from time-limited offenders (Moffitt, 1993). There is thus reason to believe that criminal attributes exist that are associated with chronicity and may lead to quicker relapse into offending behaviour following sentencing or custodial release. Howard (2011) recently found that generally criminal ‘versatile’ offenders had the highest hazards for recidivism. These prolific offenders are thus important to identify early for crime prevention and public protection. The ability of a connectionist model to predict recidivism across time intervals is the focus of the present chapter.

The performance of traditional actuarial instruments were found in Chapter 2 to be limited in identifying (non) re-offenders, rarely exceeding area under the curve (AUC) accuracy of .75 (Campbell et al., 2009; Coid et al., 2009; Farrington, et al., 2008; Gendreau et al., 1996; Kroner & Mills, 2001; Yang, Wong, et al., 2010). Moreover, accuracy of long-term predictions is not significantly different from that of short-term predictions due to high false positive error (Dahle, 2006; Mossman, 1994; Otto, 1992; Snowden et al., 2007). In Snowden et al. for example, 20% of offenders who were in the highest risk category using OGRS (Copas & Marshall, 1998) re-offended within 6 months although the instrument predicted that 80-100% of such individuals would be reconvicted.[12] Even at the two year follow-up, only 64% of these high risk individuals had been reconvicted. OGRS achieved an overall AUC of .74 over 6 months which rose to .78 over two years (Snowden et al., 2007). Current measures incorporating dynamic factors do not significantly improve upon OGRS predictions (Coid et al., 2009), including the OASys measure used by NOMS (Howard et al., 2006; see Chapter 4 for a description).

[12] OGRS predicts recidivism within 2 years thus may be expected to over-classify on six month projections.

Alternative means of statistically combining predictors are therefore needed to break through the 'glass ceiling' of accuracy (Yang, Wong, et al., 2010). This plateau in risk predictive accuracy using current methods, including those incorporating dynamic items, may relate to their failure to consider multiple interaction effects. This would explain the findings in Chapter 6 where connectionist modelling of offender data achieved strong accuracy, including low false positive rates, in the prediction of any re-offending over 30 months. However, identification of the most chronic offenders requires accuracy over shorter time-frames; in conditions where a minority of recidivists has yet failed. These conditions may be problematic in models that do not make full use of dynamic factors since these may help determine the timing of re-offending. Connectionist models may add value under these conditions due to automatic detection of interaction effects, unusual patterns, or any non-linearity in the data (Marshall & English, 2000).

Sensitivity to different speeds of offending within a sample requires a predictor to be able to reduce misclassifications above chance levels at different base-rates. Consideration of the base-rate is seen as critical in the development of a risk prediction model (Gottfredson, 1987; Gottfredson & Moriarty, 2006). Since a 'chance' classification strategy involves classifying all cases as belonging to the dominant outcome class, the difficulty of reliably predicting recidivism increases as the base-rate differs from .5 (Blackburn, 1993; Meehl & Rosen, 1955). Meehl and Rosen (1955) originally suggested that the base-rate ratio of recidivists to non recidivists must be greater than the ratio of false positives to true positives in order for a positive prediction to be more likely true than false. Even at a base-rate of .59 many current actuarial risk assessments are imprecise for many participants (Dahle, 2006). Dahle (2006) found, from an analysis of three current prominent measures, that predictive information added little to base-rate classification in up to two-thirds of cases with the majority of cases being classified in the unspecific middle group ('moderate' risk). At a base-rate of .14 'high' risk was consistently over-predicted however with false positive rates of .70-.80 (Dahle, 2006).

The tendency of current actuarial measures to produce average predicted probabilities for the majority of cases may not be surprising at low base-rates where disproportionately stronger predictor-criterion associations are required (Curtis, 1971). Curtis demonstrated that while small correlations may positively impact on misclassification errors at base-rates of .5, they have no impact at lower/higher base-rates. Even with an association of *r*=.4 as seen among the best available composite risk scales such as OGRS (Copas & Marshall, 1998; Taylor, 1999) and the LS/CMI (Andrews et al., 2004) reviewed in Chapter 2, the blanket prediction strategy may result in fewer misclassifications unless the base-rate is between .3 and .7 (Curtis, 1971). As surmised by Blackburn (1993) 'beating' a low base-rate therefore calls for predictors whose correlation with the criterion is greater than is typically found in clinical or actuarial prediction (p.325).

The performance of connectionist models in low-base rate circumstances across a range of fields was reviewed in Chapter 3. Some of these have shown positive results relative to linear regression (e.g., Das et al., 2003; Marshall & English, 2000; Mobley et al., 2000; Palocsay et al., 2000), while others have produced more equivocal or negative effects (Flaherty & Patterson, 2003; Grann & Langstrom, 2007). Positive results appear to associate with the use of methods to control over-training the model such as monitoring performance during training on an internal validation file (Mobley et al., 2000; Palocsay et al., 2000). Controlling over-training is key on problems with low base-rates to avoid the model learning to use the base-rate to reduce its error and thus becoming biased to the dominant outcome class.

Chapter 3 also discussed the ability of connectionist models to model survival over multi-step time periods (Alon et al., 2001; Lundin et al., 1999; Poulakis et al., 2004). Lundin and colleagues dealt with time dependencies in the data in which the outcome was more closely associated with certain time intervals, by developing separate connectionist models for 5, 10, and 15 year intervals. Results were consistent with other studies in indicating that connectionist models were as capable of predicting outcomes for the different intervals as were rival approaches (Alon et al., 2001). Alon et al. also found that in the time periods which were more turbulent or 'noisy' the connectionist model outperformed traditional time series modelling methods. This may be important in criminal justice data, given the propensity for noise due to measurement error on the predictor and criterion variables (Gottfredson, 1987; Gottfredson & Moriarty, 2006). Thus one recidivist may be identified as a re-offender after 12 months when they had also offended prior to 6 months, while another case may be incorrectly coded as a recidivist due to accepting responsibility for an offence on behalf of an acquaintance thus introducing noise in both cases. The problem of indistinct signals has previously been held responsible for null findings regarding the benefit of connectionist models with offender data (Caulkins et al., 1996). Chapter 6 showed that using the present data and any re-offending within 30 months, models could differentiate recidivists from other cases; this chapter sought to extend this to examine the ability of models to differentiate time to reoffending.

7.1.1 Aims of the chapter.

The present chapter therefore aimed to take forward the predictive model developed in Chapter 6 by training and testing a connectionist model on data with different follow-up intervals. This provides a test of models' sensitivity to speed of offending, an important practical consideration in the management of offender risk. Due to lower base-rates of re-offending in the different time periods, and the problem of mixed signals, it is uncertain to what extent the predictive accuracy seen in Chapter 6 can be maintained at earlier intervals. Given that those offenders that more quickly re-offend are likely to be recognised as marginally higher risk in terms of the scores on predictor variables including criminal history, and pro-criminal attitudes, it is possible that the model will be able to recognise these consistent, albeit subtle, data patterns. This may be more challenging for linear statistical models which emphasise variables associated with statistically significant between-group differences rather than general patterns within the data.

Since the chapter aims to classify recidivists within and across time periods, a second goal is to explore the ability to do this within a single connectionist model. Thus the chapter attempts to answer whether a connectionist model can predict offending in different time intervals within a single model, or whether separate outcomes are better predicted by a number of discrete models each trained on a unique criterion. This would be of benefit in practical implementation of the connectionist model, since a single model is more computationally efficient than is the use of multiple models.

7.2 Method

7.2.1 Design and analysis of models.

The General Method, Chapter 4, gave full details on the design and analysis of the models. Thus the predictor variables were the 236 offender characteristics and these were used as inputs to the models. The number of hidden units was set to 50, following the empirical findings of Chapter 6. The model output was varied for the present study (see below).

Discriminant function analysis (DFA) was used as the statistical comparison model in addition to a network without a hidden layer, i.e., a two-layer model. OGRS-II (Taylor, 1999) was selected as the applied model, as it is the major determinant of resourcing decisions within NOMS. This follows from the predictive accuracy of OGRS[13] in identifying high-risk offenders (Coid et al., 2009; Gray et al., 2004; Lloyd et al., 1994; Snowden et al., 2007; see also Chapter 2). OGRS was the only model not developed on the present data; however it was developed on UK prison/probation data (see Chapter 4).

Models were evaluated using 10-fold cross-validation (Stone, 1974), introduced previously. This was done for practical reasons given that previous work on the data did not suggest that the approach was associated with a reduction in accuracy (see Chapter 6). In the present chapter, 10-fold cross-validation was applied to all connectionist models and the two layer model but not the DFA as there was no automatic function for 10-fold cross-validation in SPSS. Therefore, as previously DFA was evaluated using leave-one-out (Efron, 1983).

7.2.2 Variation to target output.

The time elapsed from assessment to the re-offence date was coded into discrete categories: 6 months, 12 months, 18 months, 24 months, and 30 months. The rate of re-offending naturally increased cumulatively with each follow-up. The observed re-offending within 6 months of assessment was 7%, and this accumulated to 13%, 20%, 29%, and 59% at the respective follow-up intervals. Considering all cases with a known time to re-offending (i.e., excluding non re-offenders), the mean time to failure was

[13] OGRS and OGRS-II. OGRS is used to refer to both in this thesis, although only OGRS-II was used in the present research.

approximately twenty-four months (*M*=.24, SD=.12) similar to other studies of time to recidivism (e.g., Huebner & Berg, 2011). The distribution of time to re-offending was slightly but significantly negatively skewed (*D* [2393] = 0.07, *p*<.001). Thus time to re-offending among recidivists was non-normally distributed, with more cases re-offending later than earlier.

Each offender had a binary target output value (0/1) representing their re-offending up to each follow-up point (6, 12, 18, 24, or 30 months). Thus if an offender *X* re-offended at 13 months, he/she would have target output values 0, 0, 1, 1, 1, for the respective follow-up points. Five separate 'single output' models were created with each follow-up output assigned to a different model. After training, each time to re-offending model was tested in matching its output.

7.2.3 Temperature output model.

In addition, a single connectionist model was built with five output units for each time to re-offending follow-up. Since the activation of time to re-offending outputs was sequential across the five outputs, the number of units turned on for each offender represented speed of re-offending (analogous to an offending 'thermometer'). For example, activation values of .14, .43, .66, .68, .89 across the five respective outputs would correspond to the model's assessment of the likelihood of recidivism for each follow-up. As previously the model was considered to have made a correct classification if the activation value was greater than 0.5 where the target output was 1, and less than 0.5 where the target output was 0 (see Chapter 4, section 4.5.4). Therefore in this case the model predicts that the offender will not have re-offended at 6 months but will have re-offended within 18 months.

Thermometer coding in this way has been used previously to represent continuous input-output mappings (Jeon & Choi, 1999). Jeon and Choi suggested that using multiple nodes to represent a continuous value makes better use of the connectionist model's sigmoidal activation function than when using a single output, thus improving accurate classification without the need for increased hidden unit resources. A temperature output model was therefore tried on the present data to provide one standalone model for the prediction of time to reoffending. Like the other connectionist models the temperature output model was evaluated using 10-fold cross-validation.

7.3 Results

7.3.1 Single output models.

Accuracy rates for each of the models trained and cross-validated on each of the outcomes in single output models are summarised in Figure 7.1 below. More extensive details of the accuracy rates of each model are given in tables in Appendix B.

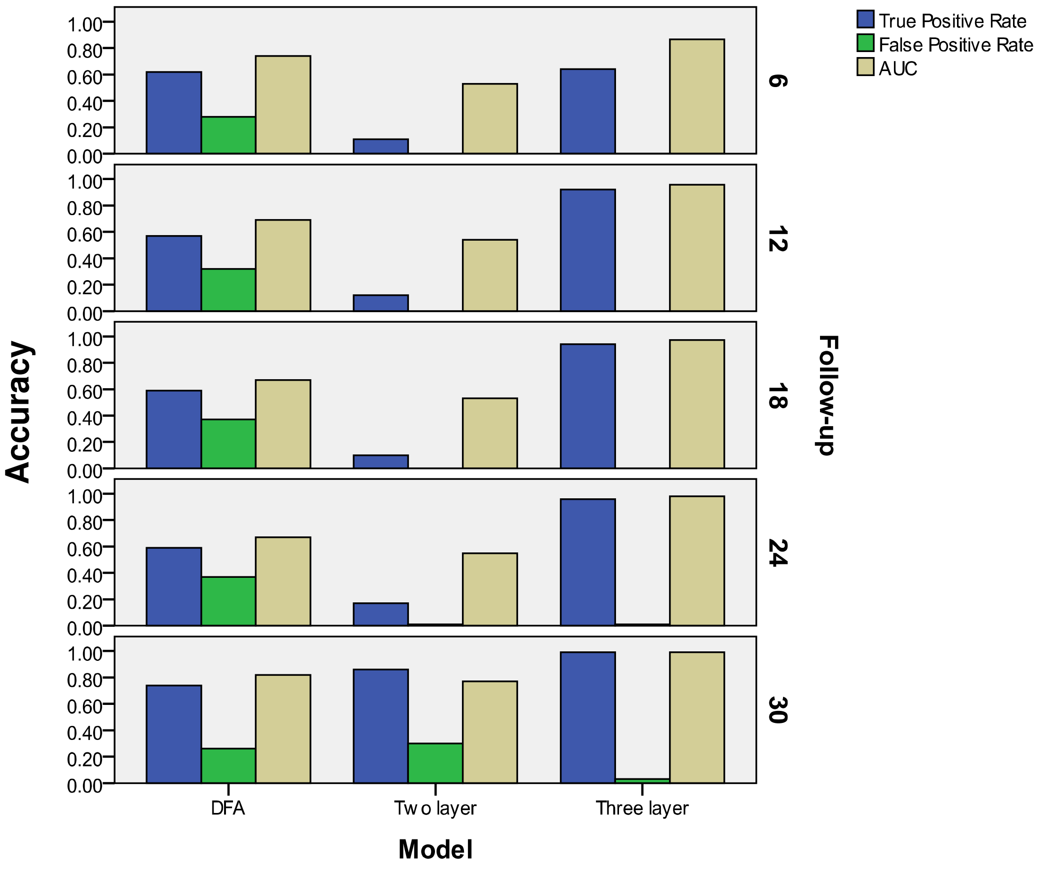


Figure 7.1. Predictive accuracy of models on cross-validation at each follow-up interval

Figure 7.1 illustrates the difference in results between models in which the three layer connectionist model maintained high levels of true positive (hit) rates and low levels of false positive (false alarm) rates at each follow-up interval. The two layer model however achieved consistently poor hit rates while the DFA model showed modest hit rates and sustained false alarms ranging between .26-.37 (see also Table B1 of Appendix

B). DFA signal detection rates did not alter using prior probabilities of group membership instead of equal group sizes for calculation of the discriminant equation.

Base-rates of re-offending had an impact on model performance. All models reduced in accuracy at lower base-rates of the criterion, although performance was better than chance (AUC=.5) at each follow-up in all models[14] with the exception of the two layer model at the 6 month interval. The three layer model achieved an AUC of .87 using the 6 month criterion, despite a low base-rate of re-offenders on which to train (7%). Hit rates at 6, 12, 18, 24, and 30 month intervals were .64, .92, .94, .96, and .99 respectively using the three layer models trained with single output nodes (see Table B2, Appendix B).

7.3.2 Temperature output model.

A series of temperature output models were trialled starting with a model with the same configuration in the middle layer as the single output three layer model (50 hidden units) and trained for the same length of time (1,000 epochs). The accuracy of this model was similar but inferior to the single output three layer model. Adding hidden units without increasing the training time or reducing the learning rate did not improve accuracy rates. The optimal temperature output model required alteration to each of these parameters, and the final model was structured with 100 hidden units, with a learning rate of 0.1, and trained to 2,000 epochs. Cross-validated results for this 'best' temperature output model at each follow-up interval are given in Table 7.1 below.

[14] Wilk's lambda for the DFA model showed that the value of the discriminant function was not statistically significant in the 6 month model (χ^2=221.29, df=221, p=.482) nor the 12 month model (χ^2=250.018, df=221, p=.088).

Table 7.1

Predictive Accuracy of the Temperature Output (Three Layer) Model at Each Time to Re-Offending Follow-up

Follow-up (months)	Overall accuracy (95% CI)	TPR	FPR	AUC (95% CI)	*d'*
6	.99 (.99-.99)	.85	.00	.98 (.97-1.00)	4.39
12	.98 (.98-.99)	.88	.00	.99 (.98-1.00)	4.79
18	.98 (.97-.98)	.88	.00	.99 (.99-1.00)	4.78
24	.97 (.96-.97)	.89	.00	.99 (.99-1.00)	4.80
30	.93 (.92-.93)	.89	.02	.98 (.98-.99)	3.29

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); 95% CI = 95 percent confidence interval; AUC = Area Under the receiver operating characteristic Curve; *d'* = dprime.

Table 7.1 shows that detection accuracy rates are slightly improved relative to the single output three layer model, particularly at the 6 month follow-up interval. While both models made only one false positive prediction for recidivism within 6 months (FPR=.00), the temperature output model made more true positives (TPR=.85) compared to the single output connectionist model (TPR=.65). The temperature output model had better accuracy in all but the longest follow-up interval, where the single output model correctly identified more re-offenders (TPR=.99) compared to the temperature output model (TPR=.89) with only a small one percentage point increase in false alarms. The differences between the two models at each follow-up are evident in Figure 7.2 below. Figure 7.2 is scaled from AUC .80 to AUC 1.0 to enable focus on small differences at such high accuracy thresholds.

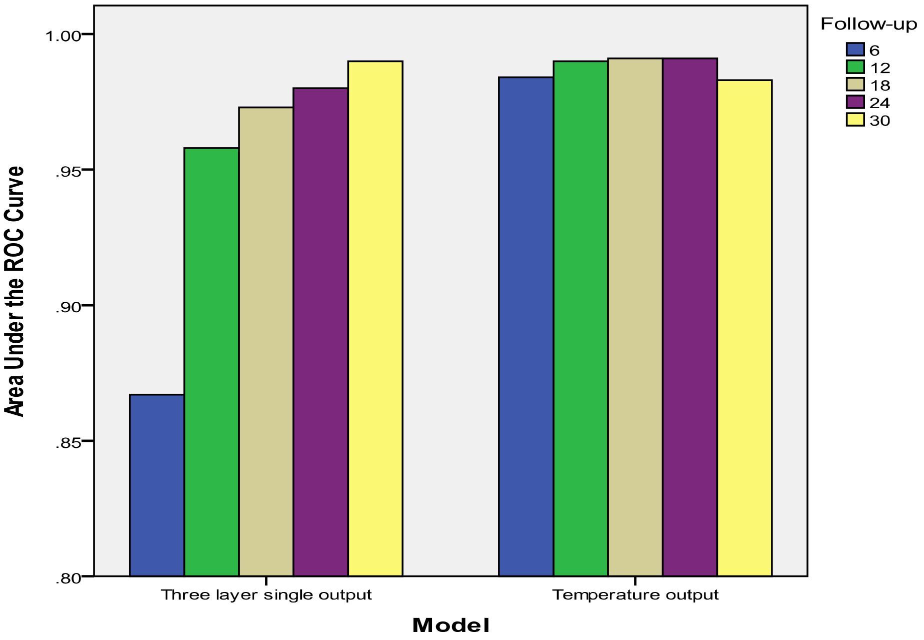
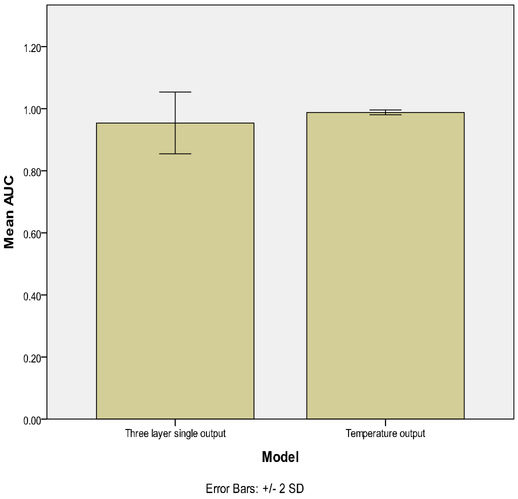


Figure 7.2. AUC accuracy of three layer models on cross-validation at each follow-up interval

Mean AUC and *d'* for the single output models, across follow-up intervals, and the same for the temperature output model are shown in Figures 7.3 and 7.4 respectively for each type of connectionist model.



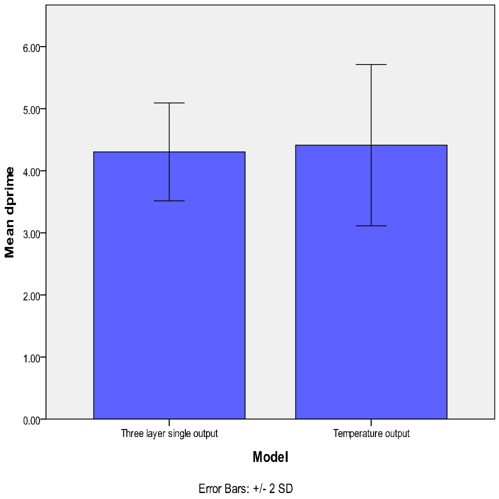


Figure 7.3. Mean AUC accuracy on cross-validation of connectionist models across follow-up intervals

Figure 7.4. Mean *d'* accuracy on cross-validation of connectionist models across follow-up intervals

Figures 7.3 and 7.4 show a small improvement, on average, associated with the temperature output model. The single output connectionist model showed a mean AUC of .95 (SD=.05) and *d′* of 4.30 (SD=.39), while the equivalent statistics for the temperature output model were .99 (SD=.00) and 4.41 (SD=.65). Thus error bars in Figure 7.3 show that the AUC values are more consistent using the temperature output model. However Figure 7.4 shows greater variation in *d′* in this model which is associated with the observed decrease in predictive accuracy at the 30 month follow-up interval (Figure 7.2). Its general improvement between 6-24 months, particularly at the earlier 6 month interval represents a possible benefit of the temperature output approach in the prediction of time to re-offending relative to single output modelling.

7.3.3 Comparison of best connectionist model to practice models.

The temperature output model was used as the 'best' connectionist model for predicting time to re-offending, and this was duly compared to the practice models OGRS[15] and OASys (represented by DFA). Figure 7.5 shows the variation in performance of these practice models by follow-up interval, compared to the three layer temperature output model.

[15] OGRS was designed to predict re-offending within 24 months and therefore is being used for an alternative purpose at shorter follow-ups. In addition OGRS was not constructed on the present data.

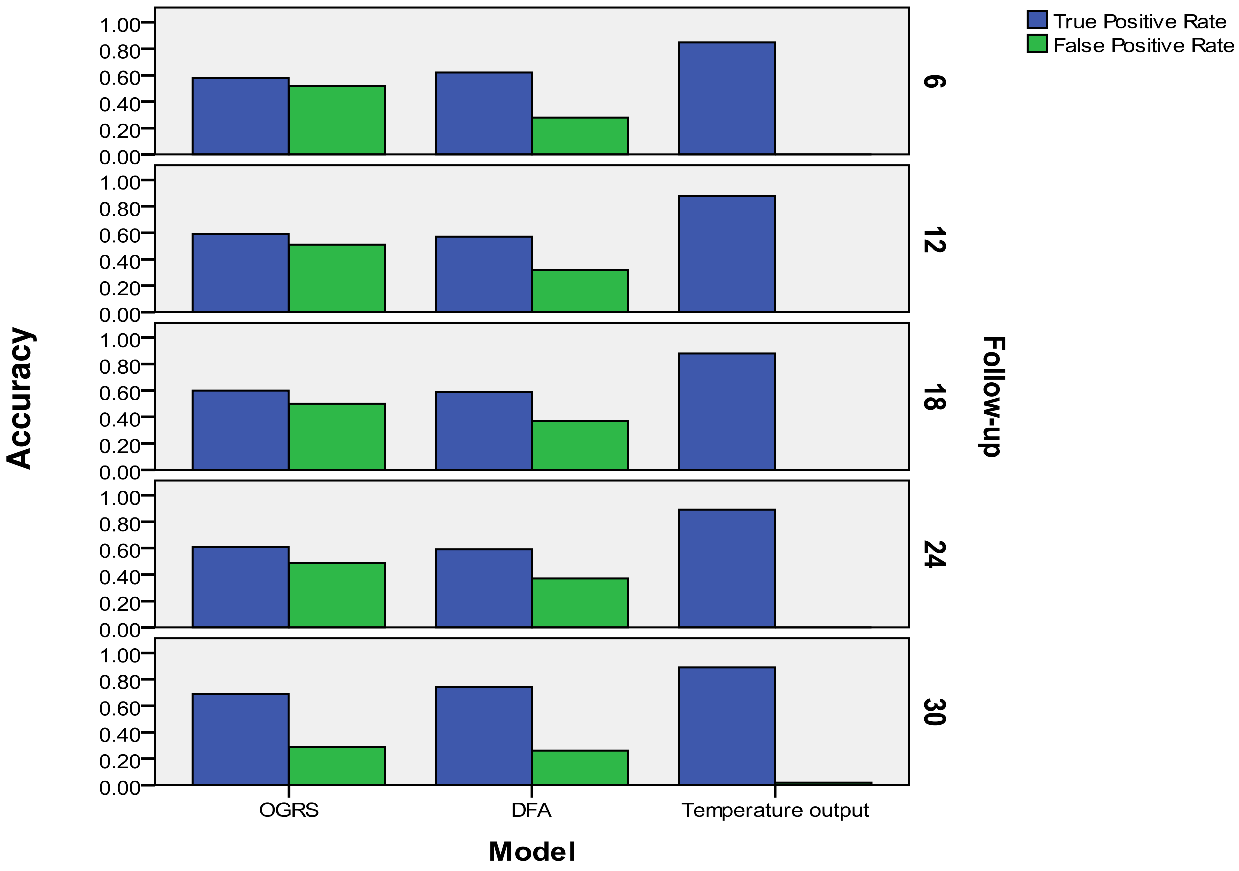


Figure 7.5: Predictive accuracy of practice models and temperature output model on cross-validation at each follow-up interval

It is apparent from Figure 7.5 that the practice models incurred a high proportion of false positive predictions, particularly at shorter follow-up intervals. By contrast the connectionist model using temperature output virtually eliminated false alarms, other than for re-offending by 30 months (FPR=.02). The difference in AUC accuracy at each follow-up is illustrated in Figure 7.6 below and Table 7.2 provides summary statistics for all models.

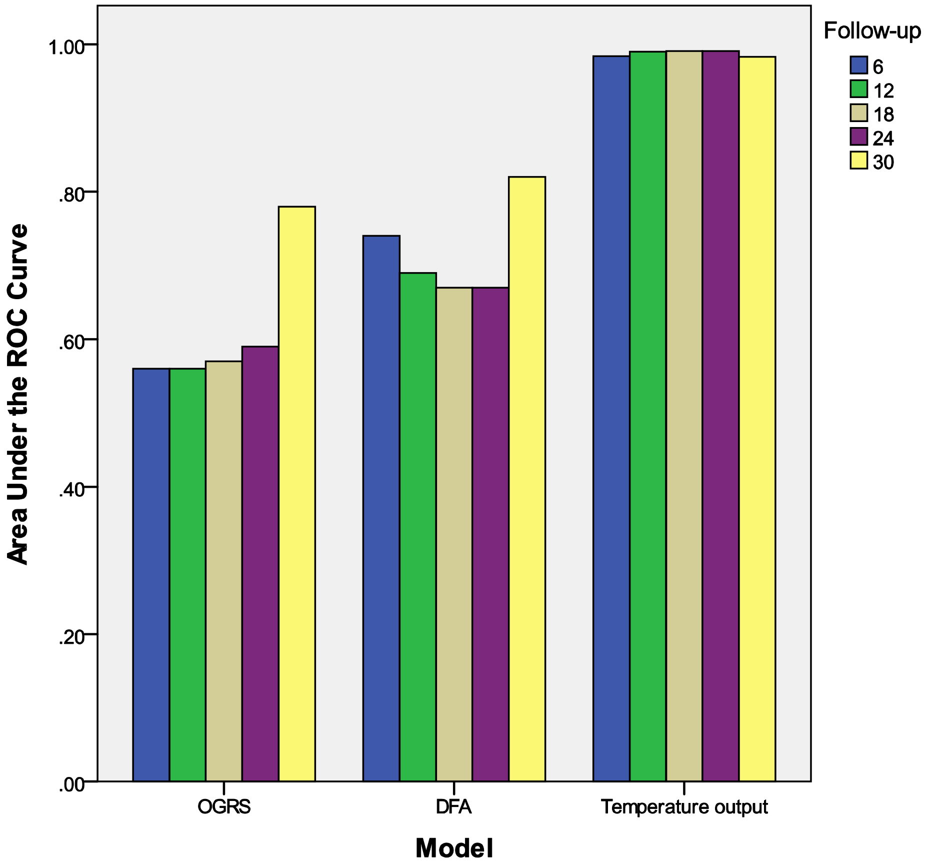


Figure 7.6. AUC accuracy of practice models and temperature output model on cross-validation at each follow-up interval

Table 7.2

Mean Accuracy Rates of Each Model Across Follow-up Periods

Model	Overall accuracy *M* (SD)	TPR *M* (SD)	FPR *M* (SD)	AUC *M* (SD)	*d′* *M* (SD)
OGRS	.55 (.09)	.61 (.04)	.46 (.10)	.61 (.09)	.39 (.37)
DFA	.67 (.06)	.62 (.07)	.32 (.05)	.72 (.06)	.80 (.31)
Two Layer	.84 (.07)	.27 (.33)	.06 (.13)	.58 (.10)	1.99 (.42)
Three Layer	.99 (.01)	.89 (.14)	.01 (.01)	.95 (.05)	4.30 (.39)
Temperature output	.97 (.02)	.88 (.02)	.00 (.01)	.99 (.00)	4.41 (.65)

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); AUC = Area Under the receiver operating characteristic Curve; *d′* = dprime.

To determine if there were significant differences in the mean results obtained, a one-way ANOVA was carried out between model and each summary measure of accuracy (i.e., excluding TPR and FPR). This showed statistically significant effects of model on each accuracy measure (overall accuracy: $F_{(4,20)}=55.433$, $p<.001$; AUC value: $F_{(4,20)}=33.931$, $p<.001$; *d'*: $F_{(4,20)}=90.513$, $p<.001$). Multiple comparison post-hoc tests using Bonferroni adjustment for AUC and *d'* measures, and the Tamhane test for 'overall accuracy',[16] confirmed both three layer models yielded statistically significant differences (improvements) on each measure of accuracy relative to all other models but not compared to each-other.

7.4 Discussion

Identification of time to recidivism is an important prediction task both in terms of the sensitivity of a prediction model, but also regarding public protection and the direction of resources within the criminal justice system. Existing methods for actuarial prediction produce 'moderate' detection of re-offenders with high false positive errors and modest true positive rates (e.g., Mossman, 1994; Yang, Wong, et al., 2010). Thus in identifying priorities for community risk, a high proportion of non re-offenders are wrongly identified while many actual re-offenders slip through the net. This may be due to theoretical and/or statistical issues relating to offender recidivism, providing the motivation for the application of a connectionist model. Chapter 6 indicated that using connectionist models high accuracy (AUC=.98) could be achieved over a 30 month follow-up time. The present chapter aimed to verify whether this predictive accuracy could be extended to the task of differentiating the speed of offending within the population.

Models with the same configuration as that applied to the 'any' re-offending (30 months) outcome in Chapter 6 were applied to data relating to offending within four additional follow-up intervals: 6, 12, 18, and 24 months in separate models with each outcome as the criterion variable in each case. Results showed that the three layer model outperformed all of the other models at each follow-up interval. In all models performance was best on the 30 month follow-up criterion and in each model was

[16] The Tamhane comparison was applied here because Levene's test on the 'overall accuracy' measure showed that equality of variance could not be assumed (*p*>.05).

equivalent to that seen in Chapter 6. At 30 months the recidivism base-rate was .59, a reasonably balanced outcome and not a good test of discrimination from chance. At the shorter follow-up periods, 6-24 months, AUC performance dropped in the linear statistical models, ranging from .67-.74 in the DFA, from .56-.59 in the OGRS model, and from .53-.55 in the two layer model. The OGRS and two layer models gave their worst performance using a 6 months recidivism criterion, and in the two layer model the predictions were not statistically different from chance (AUC=.53, 95% CI: .49-.57). Thus, in predicting recidivism within 6 months these models were little better than a blanket prediction strategy. This may bear out Blackburn's (1993) observation that predictor-criterion relationships are not strong enough in current actuarial prediction to beat a low base-rate. The DFA model appeared to do better on the 6 months criterion (AUC=.74) although again worse relative to its performance on the 30 month re-offending criterion (AUC=.82) despite its inclusion of dynamic factors. Accuracy could not be improved by allowing the DFA model to make use of the prior probabilities in the data. Although this helped the model avoid false positives the effect on the AUC and *d'* was nil due to reduced sensitivity in identifying re-offenders.

While the linear statistical models struggled to identify re-offenders (or made high false positive errors) at 6-24 months, the three layer model maintained AUC levels above .95 for all intervals except 6 months (AUC=.87). True positive predictions were similar to the linear models at 6 months (TPR=.62) but in contrast to the other models, false positive predictions were virtually eliminated using the three layer model. In an attempt to model time variations more successfully a 'temperature output' model was implemented. After optimisation this generally improved prediction of recidivism within each follow-up interval, with the exception of the 30 month outcome criterion which was slightly inferior compared to the single output connectionist model. Overall, AUC values were smoother using the temperature output suggesting that the thermometer coding in the output layer assisted in model learning for speed of re-offending. Contrary to Jeon and Choi (1999) this required additional resources in the connectionist model structure and longer training time, however there was no sign of over-fitting given the high performance under cross-validation. Yet longer training time may be required to improve prediction further on the 30 month criterion where the distinction between later recidivists and non recidivists may be less clear.

The improvement seen in these results relative to those found using traditional actuarial measures may relate to statistical issues associated with noise on the data, including unreliable predictor and criterion variables (Gottfredson & Moriarty, 2006). Such problems may be exacerbated in prediction of time to re-offending given stochastic variations within the data, e.g., when the offender happens to be apprehended, or whether a dynamic predictor has been accurately picked up by the assessor. This explains the reduction in accuracy among the conventional statistical methods in the present results for accuracy at earlier time intervals. Meanwhile the connectionist approach maintained its accuracy, in line with the review of connectionist model applications (Chapter 3) where connectionist models responded very well to noisy or degraded data including in conditions of time dependent or rare outcome probabilities (Alon et al., 2001; McMillen & Henley, 2001). This is consistent with the theory that random noise helps connectionist models avoid getting trapped in local minima prior to convergence on the best solution (Patterson, 1996). Accuracy in predicting speed of re-offending averaging AUC=.99 as seen in the present temperature output model surpasses that seen in the criminological literature using survival time models on offender data (e.g., Schmidt & Witte, 1989) or predictors using models based on logistic regression (e.g., Snowden et al., 2007).

The current plateau in predictive accuracy may also be due to theoretical issues related to the strength of the predictor-criterion relationships. Many dynamic items are likely to have low correlations with recidivism, but correlate better with more predictive static items. Traditional regression-based methods emphasising significant between-group differences may therefore overlook certain important patterns within the data. These dynamic variables may best discriminate time to re-offending and their weak influence within conventional statistical models may explain the finding that short-term predictions are no more accurate than longer-term predictions of recidivism (Dahle, 2006; Snowden et al., 2007). Further, this also extends to prediction more generally given that existing actuarial assessments including dynamic factors have not predicted recidivism consistently better than purely static measures such as OGRS (e.g., Coid et al., 2009; Farrington et al., 2008; Gendreau et al., 1996; Yang, Wong, et al., 2010). In contrast, the pattern recognition ability of connectionist modelling allows modelling of multiplicative interactions among variables many of which may not be independently predictive of the outcome. This may affect our theoretical understanding of which are the key predictor variables which in turn would affect practice. In practice the value of predictive ability depends on understanding the mediators of recidivism, thereby assisting with intervention efforts. Consequently the origin of the improvement will be explored in the next chapter (Chapter 8).

Accurate prediction of time to recidivism nevertheless has important practical benefits as it allows the criminal justice system to focus on the highest risk offenders at early stages of supervision when limited resources are more available. The minimal levels of false positive predictions observed here using the connectionist methodology might help criminal justice agencies avoid needlessly allocating those scarce resources to offenders who are likely to desist from offending for some time. The fact that precise connectionist model predictions can be achieved in one ‘time to recidivism temperature output’ model increases the practical benefit of the technique by reducing the implementation demands associated with matching cases to predictions from separate models.

CHAPTER 8

8. Results III: Explaining the Performance

8.1 Background

Studies of the predictors of adult offender recidivism have consistently highlighted a core set of variables, including young age, male gender, high criminal conviction history, and deviant lifestyle (Bonta et al., 1998; Farrington, 1995; Gendreau et al., 1996; see Chapter 2 for a review). Thus, with the exception of deviant lifestyle, the predictors of recidivism are predominantly static relating to criminal history and personal demographic variables. Deviant lifestyle 'criminogenic needs', including pro-criminal attitudes and associates, unstable employment, and problems with alcohol and/or drugs, have been shown to predict recidivism marginally more frequently than criminal history variables (Gendreau et al., 1996). However when both are included together in one risk instrument there is no clear evidence that dynamic factors show incremental predictive validity over static risk factors (Coid et al., 2009, 2011; Dempster & Hart, 2002; Doyle & Dolan, 2006; Gray et al., 2004). This may explain the so called 'glass ceiling' of predictive accuracy currently observed in recidivism risk assessment (Yang, Wong, et al., 2010).

Conventional statistical models may focus exclusively on static factors as these are the most reliable and independently predictive (Coid et al., 2011). Using three leading prediction instruments Coid and colleagues found that most of the predictive ability was provided by a small number of static factors focussing on early onset of behavioural problems and criminal versatility. This supported earlier work by Coid et al. (2009) in which a measure derived from a simple sum of the number of previous convictions for each offender demonstrated higher AUC values than all of the established risk measures, with the exception of one whose values were not statistically different (OGRS: Copas & Marshall, 1998; Taylor, 1999). Similar results have been seen in previous chapters of the present application of connectionist modelling to offender data. In Chapter 5 a conventional statistical model incorporating dynamic factors did not improve greatly upon a purely static model, OGRS, while connectionist modelling of the same static and dynamic factors showed a statistically significant increase in accuracy. This indicated the possibility that the connectionist model was better suited to the data, and/or made different use of the available predictor variables. The present chapter therefore sought to investigate which variables are employed most by the connectionist model and thereby attempt to increase understanding regarding the origin of the observed improvement in predictive accuracy. This may lead to major gains for clinical practice since the literature on the current measures suggests that the weighting given to static factors essentially defies efforts to demonstrably reduce offenders' risk of recidivism.

Previous applications of connectionist models to offender data have been limited in number and inconsistent in outcome (Brodzinski et al., 1994; Caulkins et al., 1996; Grann & Langstrom, 2007; Palocsay et al., 2000; Yang, Liu, et al., 2010). Caulkins et al. (1996) found no significant differences in accuracy over traditional statistics by using a connectionist model, with nearly identical performance for each combination of the 18 variables. This may be due to the linear stepwise method by which those variables were selected. Stepwise variable selection limits the input layer to those adding unique variance to the model, potentially excluding low level interacting variables. Yang, Liu, et al. (2010) similarly found no significant differences between traditional and connectionist approaches. These authors created four separate models each using variables from one of four recognised risk assessment measures described in Chapter 2 (e.g., Webster et al., 1997). Although there were no differences between connectionist and conventional models using each of the measures' variable sets, when the study was extended to explore the addition of eleven 'crime motivation' variables the connectionist model's accuracy improved while that of the conventional model remained the same. Yang et al. concluded that connectionist models are worthy of further research given their ability to detect additional effects with larger numbers of predictor variables. This may be due to their focus on pattern recognition, thus using all variables, rather than main effects of variables emphasising a small number of variables producing overall between-groups differences (Brodzinski et al., 1994). This conclusion also supports findings from the review of connectionist models in Chapter 3, where better performance compared to conventional statistical models was observed when trained on greater numbers of predictor variables but differences narrowed when using lower numbers of inputs (e.g., Price et al., 2000; Song et al., 2004).

Locating the origin of connectionist model performance is known to be difficult due to the distributed nature of the information stored in the network connections. Thus the learning or network weights are expected to be distributed over many neurons and interconnections, rather than a few units (Patterson, 1996). This has the advantage that the network is robust to noise or damage to individual neurons, much like the human brain. Research using connectionist models has used a variety of methods to identify the key variables responsible for changes in performance, including sensitivity analysis and iterative pruning (e.g., Naguib, Robinson, Neal, & Hamdy, 1998; Peng & Peng, 2008; see Chapter 3). These methods seek to quantify the effect of the input neurons on model performance, in terms of the model's ability to compensate for them when absent. Unless the analysis extends to systematically omitting pairs of variables, it does not however identify the strongest combinations of variables that may explain performance differences. Nevertheless, such an approach does begin to look inside the 'black box' of connectionist model performance and is thus used in the present chapter.

8.1.1 Aim of the chapter.

This chapter aimed to use the predictive model generated in the previous two chapters to examine the reasons behind the observed high predictive accuracy. Given that this may relate to the statistical methodology behind connectionist modelling and/or to its different use of predictor variables, the contribution of the predictor variables will be explored in an attempt to identify key factors. This is motivated by a desire to locate as far as possible the origins of recidivism predictions by the connectionist model, thereby informing practical interventions.

A second aim of the chapter was to examine redundancy in the connectionist model as part of an assessment of the extent to which the model employs the range of available predictor variables. Thus the chapter explores the impact on accuracy of refining the predictors by applying more parsimonious linear and connectionist models. The process of refinement was addressed theoretically as well as empirically, as described below.

8.2 Method

8.2.1 Design and analysis of models.

Models were designed and analysed as described in the General Method (Chapter 4). Thus the predictor variables for the connectionist model and the linear statistical model initially comprised the full 236 offender characteristics. However, the number and nature of the variables in the input layer was varied for the present purposes (see below). Since the present chapter aimed to explain the performance seen in earlier chapters, the parameters associated with the optimised connectionist model were all fixed accordingly. Thus the number of hidden units, the learning rate, and the length of training were all identical to the optimum parameters after Chapter 6. The model output was the criterion of any re-offending within the total follow-up period (minimum 30 months). The linear statistical model used for comparison with the connectionist model was DFA.

8.2.2 Impact of individual variables.

The connectionist model was trained up to 1,000 epochs as for previous models and then the weights of the trained network were saved. Then the input layer of the model was subjected to leave-one-out testing, with each variable omitted sequentially and the network tested in the absence of each predictor. The impact of the variable was examined by comparing the difference in performance on the test data compared to when all variables were included. This 'switch off variables' analysis is similar to approaches taken previously in the connectionist model literature in the examination of the impact of individual predictor variables (e.g., Naguib et al., 1998; Peng & Peng, 2008).

8.2.3 Pruning the models.

To examine whether the same performance could be achieved with fewer predictor variables, models were pruned empirically according to their impact on model accuracy in the switch off variables analysis. Thus the model was pruned down to the top 10 most influential predictor variables and then subjected to 10-fold cross-validation. Results were compared to a DFA model using the same variables and cross-validated using leave-one-out.[17] DFA performance was also measured after training and testing using the smallest set of predictors as determined by stepwise variable selection within SPSS (SPSS Inc., 2010). Connectionist and DFA model accuracy with less restrictive pruning was also examined by training and testing separate models using the top 20, 50, 100, and 150 input variables.

[17] As described in the General Method (Chapter 4), data sampling for testing the DFA model was done using leave-one-out rather than 10-fold cross-validation as the latter was not available as an automatic function within the options for DFA within SPSS v19 (2010).

The above empirical pruning was supplemented with an examination of theoretically motivated models. Thus predictor variables were assigned as either static factors or dynamic factors according to whether they are open to change by intervention (see corresponding labels in Appendix A). Two theoretically pruned models were then created, one using only the static factors (*n*=77) and the other using only the dynamic factors (*n*=151).

8.3 Results

8.3.1 Impact of individual variables.

The magnitude of the impact of individual variables, measured by the extent to which the various indicators of accuracy reduced when the variable was not available during the switch off variables analysis, is shown in Table 8.1 for the top 10 variables in order of change in *d′*. A more extensive table containing the top 50 predictors with corresponding values for the change in *d′*, TPR and FPR is included in Appendix C.

Table 8.1 shows that the biggest impact on *d′* was linked to the dichotomous variable ‘understands the importance of completing programmes’. In the absence of this variable, model predictions reduced from a *d′* of 4.31 (all variables) to 2.21. This variable impacted both on true positives and false positives, with a greater impact on false positives when omitted than seen after omission of any other variable. Without knowledge of whether the offender understood the importance of completing programmes the model increased its false positive rate from .02 (all variables) to .17, an increase of .15 (Table 8.1). This factor may be protective against recidivism within the model, discriminating non re-offenders, although its impact on true positives shows that it also helps identify re-offenders. Figure 8.1 further below illustrates the five variables impacting most on the TPR (hit rate) and their corresponding values on the FPR (false alarm rate), and vice-versa for the top five variables impacting on the FPR.

Table 8.1

Variables Impacting Most Upon Accuracy when Omitted During the Switch Off Variables Analysis

Predictor omitted	Change[a] in TPR	Change[a] in FPR	Change[a] in Overall accuracy	Change[a] in *d’*
Understands importance of completing programmes	-.09	.15	-.12	-2.10
Gender	-.10	.12	-.11	-1.99
White ethnicity	-.09	.12	-.10	-1.94
Coder type	-.07	.11	-.08	-1.78
OGRS	-.31	-.01	-.18	-1.69
No current serious offence	-.08	.07	-.08	-1.67
Risk category	-.08	.06	-.07	-1.55
Motivation to address offending	-.07	.06	-.07	-1.52
No previous significant risk event	-.05	.08	-.06	-1.51
Emotional well-being problems	-.04	.10	-.06	-1.51

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); *d’* = dprime; [a] Baseline TPR, FPR, Overall accuracy and d’ values were .99, .01. .98, and 4.31 respectively.

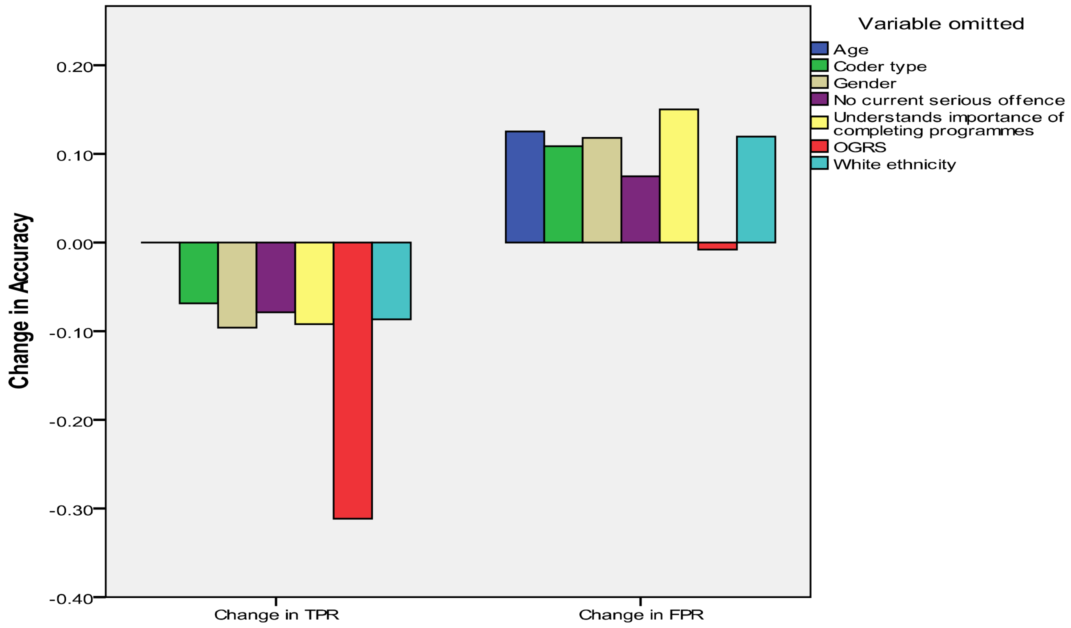


Figure 8.1. Variables impacting most on TPR and FPR when omitted from the connectionist model

As illustrated in Figure 8.1 the greatest impact on the TPR was provided by the variable OGRS. The reduction of .31 was much greater than the reduction of the next most influential variable, Gender. However OGRS' overall impact on *d'* was limited by its tendency to over-prediction of reoffending. While hit rates deteriorated in the absence of the OGRS variable, FPR values actually improved (Figure 8.1). OGRS therefore acts as a risk promoting factor within the overall model. Tendency to over-predict risk is consistent with OGRS' intended use as a screening tool to target further assessment at higher risk groups of offenders (Taylor, 1999).

Other than offender understanding of the importance of completing programmes, the greatest impact on the FPR was due to offender age at assessment. Offender age impacted on the FPR but only ranked twentieth in terms of its overall impact on *d'* due to negligible impact on the TPR. Thus age may mitigate risk within the model, perhaps tempering the risk promoting effect of other variables. Since it was not possible to control for all of the possible interactions and sub-models open to the connectionist model, it is hard to define the precise role of individual variables in isolation. It is clear from Table 8.1 and Figure 8.1 that Gender, White ethnicity and Coder type (grade of assessor), were all important in increasing the TPR and simultaneously minimising the FPR, suggesting that there may be different sub-models underpinning the relation between these variables and recidivism. For instance, either level of gender may be risk promoting under some circumstances but risk reducing under others depending on interactions with other variables.

8.3.2 Pruning the models.

8.3.2.1 *Empirical pruning*.

Results of pruning the connectionist model according to the rank of each predictor after the switch off variables analysis are shown in Figure 8.2. This showed a clear relationship between pruning and accuracy, with model performance being limited with fewer variables. Predictive accuracy with 10 variables was characterised by a high FPR (.42) which reduced sharply with an increase in variables included. Best connectionist model performance was with all 236 variables, although improvements with more than

100 variables were only minor maintaining a TPR of approximately .98 and an FPR of around .03.

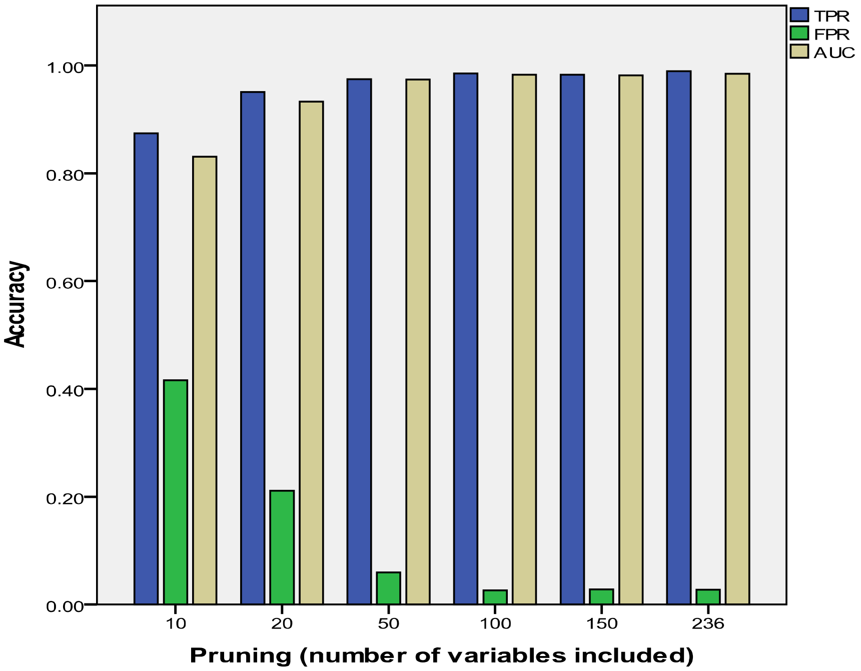


Figure 8.2. Impact of empirical pruning on the connectionist model

Different results were observed using the same variables when pruning the DFA model (Figure 8.3). Regardless of the number of variables included DFA performance was unchanged, maintaining TPR levels of around .74 and an FPR of around .27. Thus predictive accuracy was not improved by reducing/increasing the number of predictor variables.

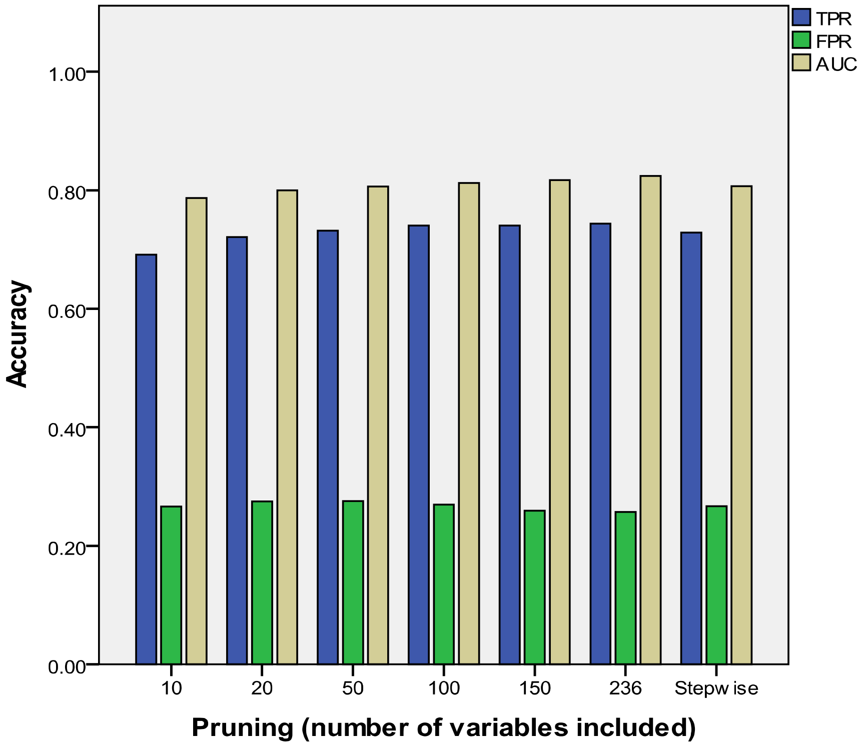


Figure 8.3. Impact of empirical pruning on the DFA model

Since the variables selected for omission were those that were of least benefit to the connectionist model, Figure 8.3 also shows the results after variable refinement using a linear stepwise procedure. The final model included 19 variables (given in Table D1 of Appendix D) of which only 6 had been selected in the 'top 20' pruning based on the switch off variables analysis, and only 11 in the 'top 100'. There was no impact on *d'* however, suggesting that DFA performance accounts for similar levels of variance regardless of the variables selected.

8.3.2.2 ***Theoretical pruning.***

Results of pruning the connectionist model based on whether variables were 'static' or 'dynamic' are shown in Figure 8.4. The model based on static variables achieved very good predictive accuracy with TPR, FPR, and *d'* values of .84, .10, and 2.27 respectively. These were not as good as equivalent values for the dynamic model however: .96, .07, and 3.23. Predictive accuracy of the dynamic model (AUC=.97, 95% CI: .97-.98) was significantly better than that achieved by the static model (AUC=.94, 95% CI: .93-.95).

However the best model comprised all variables, static and dynamic, producing an AUC of .985 (95% CI: .98-.99) (see Chapter 6).

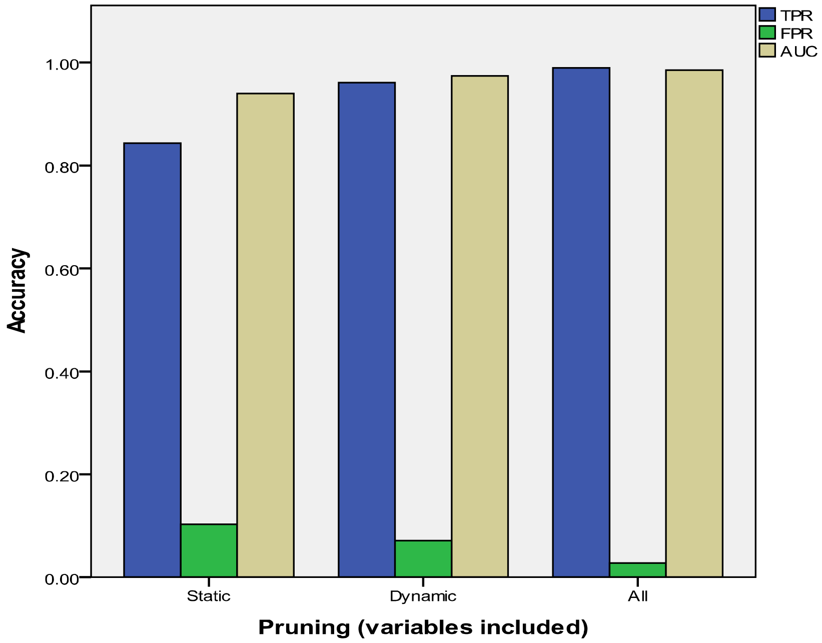


Figure 8.4. Impact of theoretical pruning on the connectionist model

Results of pruning the DFA model according to the static or dynamic nature of the predictor variables is shown in Figure 8.5. This shows that accuracy varied little according to whether static or dynamic variables were included. The static model showed scores of .72, .27, and 1.22 for TPR, FPR and *d'* respectively which were each marginally better than the corresponding values of .69, .29, and 1.06 achieved by the dynamic model. The AUC associated with the static model (AUC=.80, 95% CI: .79-.81), was also marginally higher than that achieved by the dynamic model (AUC=.77, 95% CI: .76-.79). Although the un-pruned model containing all variables was the most accurate (AUC=.82, 95% CI: .81-.83), this was not significantly different to the model comprising purely static factors.

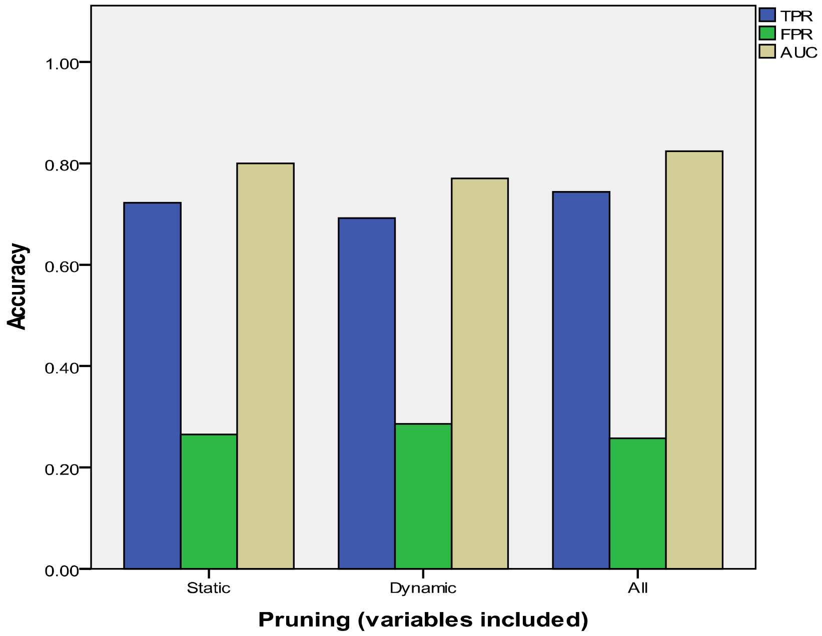


Figure 8.5. Impact of theoretical pruning on the DFA model

8.4 Discussion

The present chapter sought to understand better the reasons behind the excellent performance of the connectionist model in recidivism predictions explored in previous chapters of this thesis. This is theoretically important in terms of explaining how the model arrives at its prediction, but more specifically in terms of enabling criminal justice agencies to respond practically by targeting resources at the source of risk. Existing methods of risk prediction are largely driven by static risk indicators, thus preventing demonstrable changes in risk.

The extent to which each variable impacted on the connectionist model's performance across all cases was estimated using a 'switch off variables' analysis. This showed that dynamic factors were being used by the model to increase *d'*, while a linear stepwise procedure identified static factors as the strongest influences on the DFA model. This was confirmed by empirical and theoretical pruning: while the connectionist model's predictions improved by inclusion of further dynamic information, the DFA model's performance was unchanged. This explains the previous finding that the DFA model including all variables showed no appreciable improvement over the purely static OGRS model (Chapter 6). It also supports research suggesting that the field may not be able to improve upon models based on static factors using conventional statistical methods (Coid et al., 2011). In the present chapter the conventional model comprising only static factors performed better than the model comprising only dynamic factors using the same method. This contrasted with theoretical pruning of the connectionist model which showed the reverse: the dynamic factors model was more effective than the model limited to static factors. This demonstrates that accurate risk prediction is possible based on the assessment of dynamic risk, and opens the way to future work aimed at exploring the link between practical intervention efforts and changes in risk probability as determined by the connectionist model.

The fact that the connectionist model's accuracy was similar to that of the DFA model when restricted to 10 variables may help explain the inconsistent findings in previous applications of connectionist models to offender data (e.g., Caulkins et al., 1996; Grann & Langstrom, 2007; Yang, Liu, et al., 2010). These studies have all employed 20 variables or fewer, using data-sets smaller than the current sample. The results of the present study showed that advances in AUC performance were most evident with 50 or more variables. The present study therefore adds evidence to the suggestion by Yang, Liu, and colleagues that connectionist models may require more data, cases and variables, to achieve improvements in predictive accuracy over conventional models.

The improvement in accuracy with increasing numbers of predictor variables indicates that the model uses the additional information to classify patterns that it was unable to classify on the basis of fewer variables. Thus different individual or groups of cases may activate different patterns in the knowledge stored in the network's connections. This means for instance that the network may have memorised multiple 'recidivist' patterns potentially with contrasting values on some individual variables. The switch off variables analysis was able to show those variables that were of most general importance to the connectionist model, but unless the impact was clearly either on the hit rate or the false alarm rate, it was not able to show whether each variable was used specifically to classify re-offenders or non re-offenders, much less which level of the variable was used for each classification. For example, the most influential variable was whether the offender understands the importance of completing programmes. When omitted from the model this variable impacted negatively on both the hit rate and the false alarm rate, reducing the former and increasing the latter, suggesting that the model uses the variable to identify both re-offenders and non re-offenders. This variable had a very weak negative coefficient within the DFA model, suggesting that it acted protectively against recidivism, but that the independent outcome relationship was not consistently strong. The role of offender insight into the importance of completing programmes in classifying re-offenders and non re-offenders thus depends on its relationship with other variables in the data. Offender insight is a very important variable clinically that has been attributed to the non completion of accredited intervention programmes and thus implicated in recidivism risk (Olver et al., 2011). Due to its clinical utility it has also been included in prominent risk assessment schemes (e.g., Webster et al., 1997), although its contribution to predictive validity is reportedly weak (Coid et al., 2011; Doyle & Dolan, 2006). The current findings suggest that, using a connectionist model, offender insight does have a strong bearing on risk predictions although its precise role in individual cases is hard to delineate due to the possibility of multiple interactions.

The same can be said of many of the other key variables emerging from the switch off variables analysis, including gender, white ethnicity, coder type and no current serious offence. The relationship between male gender and recidivism is expected to be positive in general, and accordingly instruments such as the OGRS ascribe a negative weighting to female gender (Copas & Marshall, 1998). A uniform relationship between gender and recidivism is unlikely however and OGRS' accuracy among female offenders has been shown to be inferior to that for men (Coid et al., 2009). Coid et al. suggest that different factors may be involved in the risk presented by some female offenders. This is borne out in the present chapter: while the connectionist model made strong use of gender to improve its identification of re-offenders and non re-offenders, the same predictor did not feature in the top ten variables contributing to the DFA model. Thus the connectionist model is likely to be making use of different relationships in the data for each level of gender. More surprising was the importance of white ethnicity in the model given that over 95% of the sample came from this ethnic group (Table 4.1, Chapter 4). Minority ethnicity has been a consistent predictor of adult offender recidivism which is expected to be due to its association with other unmeasured variables (Gendreau et al., 1996). Due to the predominance of white ethnicity in the North-East of England (Dobbs, Green, & Zealey, 2006), the role of white ethnicity in the connectionist model may represent a similar, albeit inverse, finding. Thus the presence of this ethnic category appears to improve predictive accuracy of the model and this may be due to less consistency in the data for the remaining ethnic groups.

Coder type and the seriousness of the index offence may well have variable connections with recidivism. One might expect coder type to be generally risk-promoting since higher grades of probation officer supervise higher risk offenders. However in some offenders it may be expected to be risk-reducing since these higher graded staff have generally benefited from more intensive training. 'No current serious offence' may also sometimes operate counter-intuitively: Bonta et al.'s (1998) meta-analysis of the predictors of recidivism among mentally disordered offenders for instance found that a violent index offence was negatively related to recidivism. Thus not having a serious index offence may relate to ongoing risk in some offenders perhaps due to lower levels of supervision and risk management. The same can be said for the influential variable 'no previous significant risk event', where the absence of this factor may protect against offending due to subsequent community practices, but its presence may be associated with lower levels of supervision given that this must be targeted at risk.

The way in which the model used the variables OGRS and 'age' appeared to be clearer. The impact on hits and false alarms suggested that OGRS helped identify re-offenders whereas age helped identify non re-offenders. The role of OGRS in identifying re-offenders was expected given that it has been consistently shown to be valid in predicting recidivism (Coid et al., 2009; Farrington et al., 2008; Gray et al., 2004; Lloyd, et al., 1994; Snowden et al., 2007). Its negative impact on the false alarm rate was also unsurprising given that OGRS is known to over-predict risk (e.g., Snowden et al., 2007; see also Chapters 6 and 7). Inclusion of this risk predictor in the connectionist model's input layer has clearly helped overall prediction effectiveness. The influence of offender age in the identification of non re-offenders is also consistent with prior research indicating a general 'ageing out' of crime (Bowles & Florackis, 2007; Hirschi & Gottfredson, 1983; Lloyd et al., 1994). Thus increasing age is likely to be associated with a decrease in recidivism risk among most, but not all, offenders. Age may be well-suited to connectionist modelling since its relationship with recidivism may not be linear, i.e., certain later age periods may be associated with an increase in risk, before risk decreases again (Bowles & Florackis, 2007). Overall it appeared that offender age helped avoid false positive predictions, thus assisting in reliably identifying non re-offenders.

Other variables in the top 20 after the switch off variables analysis included thinking skills deficits, employability needs, alcohol misuse needs, and criminal associates. These dynamic 'criminogenic need' areas are all components of a deviant lifestyle that has been consistently associated with persistent offending (e.g., Andrews & Bonta, 2010; Gendreau et al., 1996). These factors all helped identify re-offenders and non re-offenders again suggesting that they may either promote or reduce risk depending on the pattern of other variables in individual cases. The DFA model also highlighted the importance of these variables, but only weakly. The probability of risk identified by combining variables using a conventional statistical model may therefore be harder to alter due to the weaker influence of dynamic factors within the model.

8.4.1 State of the network for individual cases.

The above discussion of the complexity of identifying the role of individual variables can be illustrated using screen-shots of the trained connectionist model's activation when presented with new individual cases. Figure 8.6 below shows two re-offender cases that were correctly predicted by the connectionist model. Case A on the left was also correctly predicted by the DFA model, while it falsely labelled case B on the right as a non re-offender. Although both cases were classified as recidivists by the connectionist model it is clear that the activation in the middle layer was different in each case. This reflects the fact that the two cases' characteristics were very different. Case A had high static and dynamic risk as evident from 28 previous convictions, young age, a high OGRS score, eight criminogenic needs, and poor understanding of the importance of completing programmes. Case B meanwhile was older with just two prior convictions, and low OGRS. However, case B also had a high number of criminogenic needs (six) and poor insight into the benefit of completing programmes. The connectionist model appears to have prioritised the dynamic risk information over the static risk, while the DFA model's linear prediction has been forced to separate all cases with one decision rule, thus overlooking the risk presented by case B.

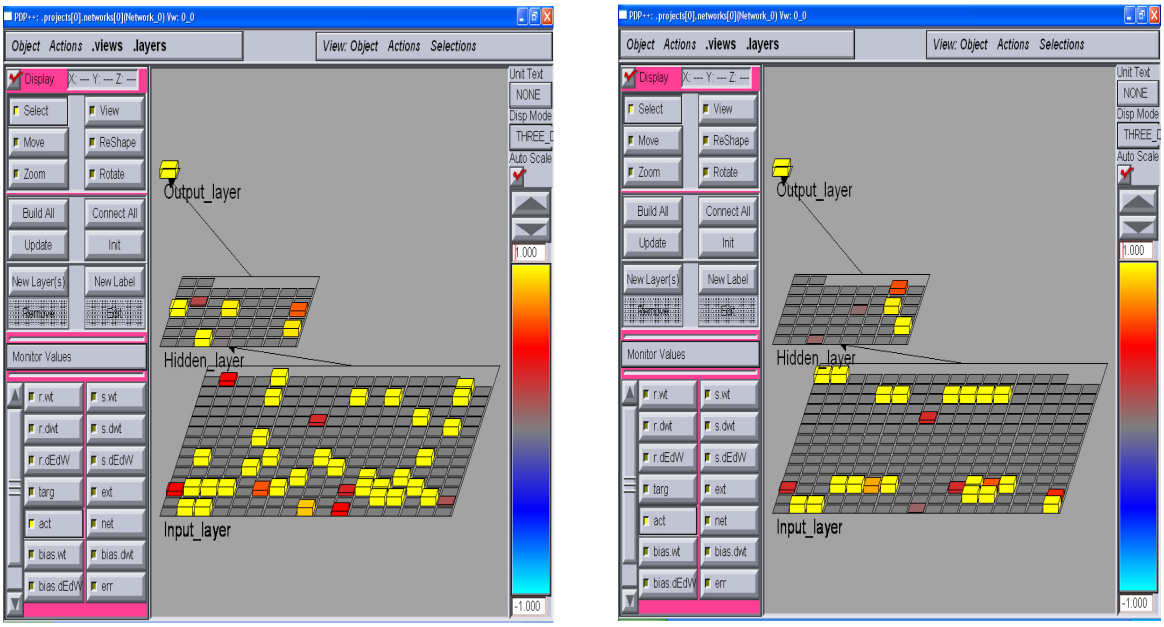


Figure 8.6. Connectionist model network activation for re-offenders in two cases: one where the model agreed with the DFA model (left, case A) and another where the DFA model made a false negative prediction (right, case B).

Figure 8.7 shows the network activations for two non re-offenders correctly classified by the connectionist model. Again it is clear that different units in the middle layer were activated by the connectionist model, even though both cases were non re-offenders. Both offenders were approximately 25 years of age, but case C on the left of Figure 8.7, was low OGRS risk with just one previous conviction and few dynamic risks. Case D meanwhile was high OGRS risk with five criminogenic needs. Thus both models agreed in classifying case C, but the DFA model made a false positive prediction on case D.

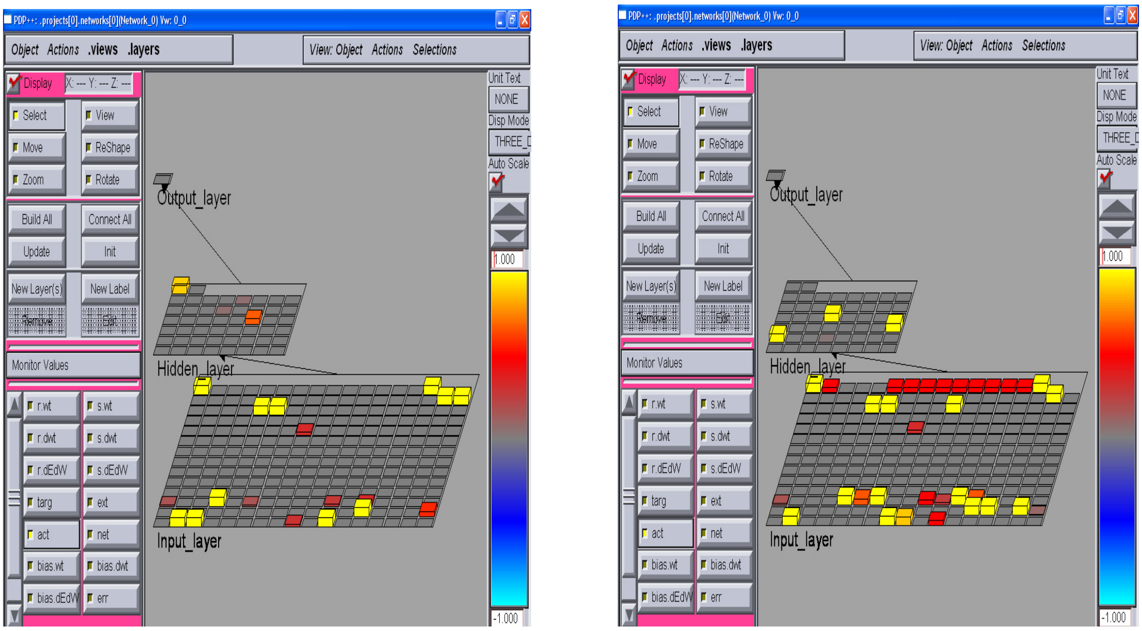


Figure 8.7. Connectionist model network activation for non re-offenders in two cases: one where the model agreed with the DFA model (left, case C) and another where the DFA model made a false positive prediction (right, case D).

Although both offenders had poor understanding of the importance of completing programmes, the connectionist model correctly labelled Case D as a non re-offender. A key difference in the cases was that Case D was not of white ethnicity (ethnicity 'not stated') and had mental health problems (emotional well-being). Both of these factors were shown in Table 8.1 to be important to the connectionist model in reducing the false positive rate.

Figure 8.8 shows the network activation in two cases where the connectionist model made false predictions. Case E (left) was aged 29 with no previous convictions, low OGRS, and problems with alcohol, relationships and mental health who subsequently re-offended. Case F (right) was aged 32 with twenty-three previous convictions, high OGRS, and problems with drugs, finance, employment, deviant peers and mental health who did not re-offend. In both cases, models appeared to give priority to the static factors and thus the connectionist model unfortunately agreed with the DFA model. Interestingly the connectionist model comprising only dynamic factors correctly classified both of these cases, and thus may be an important reference where clinical information disputes the full connectionist model prediction.

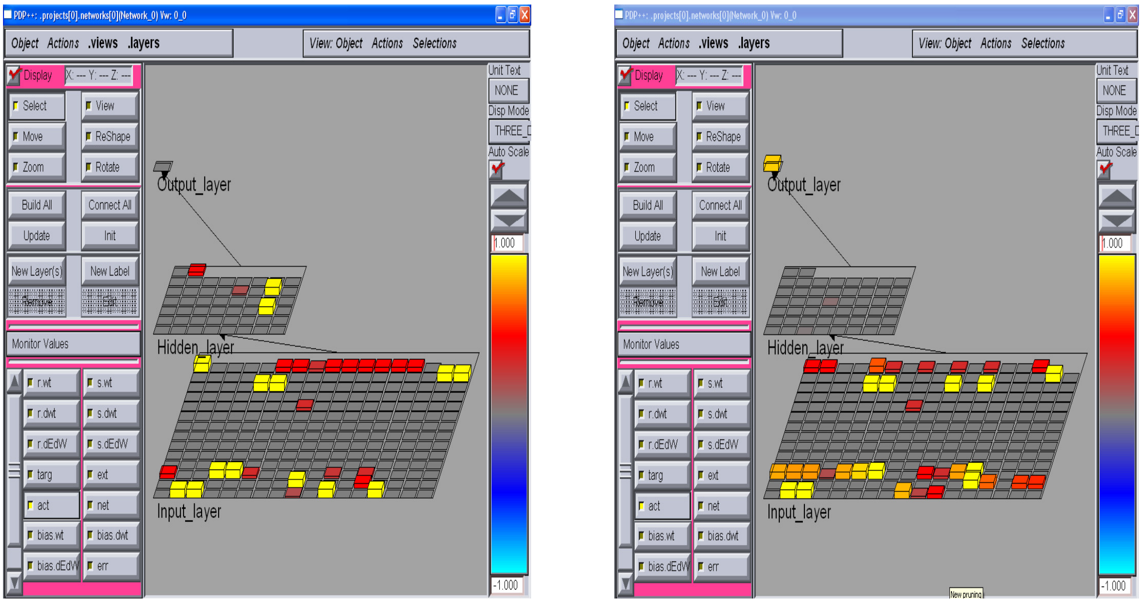


Figure 8.8. Connectionist model network activation in re-offender (left, case E), and non re-offender (right, case F) cases where it made false predictions.

In conclusion, the present chapter sought to examine the reasons behind the high predictive accuracy observed in previous chapters of this work. Pruning the model showed that the connectionist model makes use of the full range of variables, unlike the DFA model where similar predictive accuracy was achieved whether all variables were included or whether limited to a few static factors. Although both models appeared to use the same key variables as top factors contributing to their decision functions, the way in which these variables were employed appeared to be different. The conventional statistical model emphasised a few variables contributing to between-groups differences, while the connectionist model spread its activation across a wide range of variables thus incorporating more dynamic variables that are inter-related but do not independently show statistically significant outcome relationships. For example the connectionist model identified a strong role for the assessment of whether the offender understands the importance of completing programmes, which was not identified by the DFA model. This has important clinical relevance in terms of working with offenders to increase insight into the problems causing their offending behaviour. It enables criminal justice agencies not only to focus on higher risk offenders (better defined), but also to measure the impact on model predictions of changes in dynamic risk where such changes have previously made little difference to risk estimates.

CHAPTER 9

9. General Discussion

9.1 Aim of the Chapter

This chapter aims to review and discuss the results that have emerged from this thesis. First it will provide a summary of the results and their theoretical implications before progressing to discuss some of the practical implications for criminal justice agencies tasked with working with offenders to protect the public and reduce re-offending. Finally consideration will be given to the limitations of the present research and future directions for recidivism risk prediction using connectionist models.

9.2 Results and Theoretical Implications

The main finding of this thesis was the statistically significant positive impact on the accuracy of recidivism prediction by combining predictor variables using a different statistical methodology, connectionist modelling. Using this approach resulted in cross-validated AUC accuracy rates of .98, far exceeding those found using conventional statistical models to combine the same data. The DFA and the logistic two layer model produced AUC values of .82 and .76 respectively. The difference using the connectionist approach was apparent equally in the identification of recidivists and non recidivists (1-FPR), with accuracy rates of .99 and .96 respectively compared to .77 or below on either target using the conventional models. Importantly, the DFA model comprising also the dynamic OASys items represented only a small improvement on the purely static OGRS model which was found to identify recidivists and non recidivists with .77 and .69 accuracy respectively for an overall AUC of .78. This replicates a finding by the UK Home Office where a logistic regression model containing the OASys variables showed no improvement relative to the predictive accuracy of OGRS on its own (Howard et al., 2006).

The improvement using connectionist modelling supports the hypothesis that the 'glass ceiling' of predictive accuracy currently experienced may be penetrated by consideration of alternative approaches to those that emphasise key static risk factors responsible for large but insensitive between-group differences (Coid et al., 2011; Yang, Wong, et al., 2010). AUC accuracy levels of .98 represent a step-increase compared to those seen typically in the recidivism prediction literature where they rarely exceed .75 (Yang, Wong, et al., 2010). The substantial increase observed here using a fully developed connectionist model indicates that such models may offer the field the opportunity to break through the glass ceiling currently experienced as a result of reliance on multiple regression models.

The difference between the connectionist and conventional models relates to processing capacity available to connectionist models in the hidden layer. The present results showed that in the absence of sufficient hidden layer resources performance was similar to the DFA model (Chapter 6). This processing capacity allows the connectionist model to discover automatically any important interactions between variables. Given 236 predictor variables and the number of potential interactions, it is perhaps not surprising that at least 50 hidden units were required with this data. Previous work using connectionist models has suggested that the number of hidden units required depends on the characteristics of the data sample; too few units may fail to exploit the value of the connectionist approach (Palocsay et al., 2000), while too many may make the model liable to over-fit the training data with a decrement in subsequent performance under cross-validation (Grann & Langstrom, 2007). Grann and Langstrom used the HCR-20 variables and a high proportion of hidden units. Thus the low number of variables and the high number of hidden units probably conspired to produce sub-optimal cross-validation performance. In the present study, a second feature associated with the optimised connectionist model was longer training time; but only in models that contained at least 50 hidden units. Thus more iterations over the data increased model accuracy where a greater ability to detect multiple interactions was present.

The ability to model interaction effects and rare patterns represents a difference to existing ARA and other risk measures which, due to their linear statistical basis, are constrained to assign fixed weights according to main effects of differences between recidivists and non recidivists. This explains the limited accuracy among women, for instance, in measures such as OGRS which perform well on male offenders (Coid et al., 2009). Thus existing ARAs are unable to differentiate or keep independent sub-components of risk, such as different theoretical dimensions or recidivism pathways (Doren, 2002). Doren describes an example in which an offender might score high for a relevant drive towards offending such as deviant sexual interest, and low on another dimension such as anti-social behavioural history. In these circumstances existing ARAs would be forced to find a single solution that reflected the majority of cases in the training sample, or to average across the two dimensional measures for a 'moderate' risk rating. Neither solution is adequate for high accuracy, leading to problems with individual predictions, and requiring practitioners to use a second risk measure with all the concomitant problems of how to combine the measures reliably (Vrieze & Grove, 2010). Case information presented in Chapter 8 illustrated that the connectionist model was able to classify offenders at high dynamic but low static risk and vice versa. In theory connectionist models are able to respond to the existence of multiple risk pathways (see Caulkins et al., 1996), and the ability both to model interactions and keep independent sub-models may explain the improvement seen consistently in the present results relative to conventional statistical methods.

The difference between the results of the full model and those from pilot testing the method on a specific sub-sample in Chapter 5, with TPR values of .98 and .49 respectively on the same cases, suggests that this ability may relate to the size and heterogeneity of the sample population (Schwalbe, 2007; Yang, Liu, et al., 2010). Uniformity in the sub sample, comprising only sexual offenders, may have precluded the pattern recognition capabilities seen when the model was exposed to a more diverse sample. Alternatively this difference may merely reflect that the smaller model over-fitted the data due to a higher number of predictors than cases in the pilot model training sample, particularly in the context of an imbalanced criterion output. Given that the criterion was any recidivism and the pilot sample were sexual offenders, it didn't seem surprising that the model for general recidivism was better served by a large heterogeneous training sample. This also supports findings in the wider connectionist modelling literature suggesting that their advantages over conventional models are exposed better when trained using larger samples (Marshall & English, 2000; Song et al., 2004; see Chapter 3).

The full data may have better allowed the model to use the range of available predictor variables to discriminate risk from the sample base-rate. This conclusion follows from results showing good accuracy at low base-rates when the connectionist model was restricted to short follow-up intervals (Chapter 7). The three layer model identified recidivists and virtually eliminated false positives at the earlier follow-up intervals, while the DFA and OGRS models considerably over-predicted recidivism. While the OGRS model was not constructed to predict short-term risk, the DFA model was, yet its discriminant function was nevertheless non significant in separating the groups at the shortest risk prediction intervals. This confirms research indicating that not only the static measures such as the OGRS, but also the dynamic measures such as the LSI-R and HCR-20 produce high FPRs at lower recidivism base-rates (Dahle, 2006; Snowden et al., 2007). This should be surprising given that dynamic factors might improve understanding of the imminence of recidivism, thus explaining the low base-rate at short intervals. This improvement is what was seen in the three layer temperature output model where false positives were totally eliminated in base-rates below .25. A similar although less pronounced finding was evident with the standard single output connectionist model. This is consistent with findings from the literature on connectionist models which suggest they can perform as well or better than conventional statistical methods on low base-rate problems despite the overfitting risk (e.g., Baxt & Skora, 1996; Marshall & English, 2000; Price et al., 2000).

The proposal that the connectionist model makes better use of the dynamic factors was also explored directly (Chapter 8). Model pruning identified a clinically assessed dynamic factor ‘understands the importance of completing programmes’ as most important to discrimination performance, while the static item OGRS was overwhelmingly the strongest contributor to the DFA model. However, when models were restricted to inclusion of the most important variables according to the results of model pruning, an interesting effect emerged. First, when restricted to the ten most influential variables, the connectionist model’s accuracy was only slightly better than the equivalent ten variable DFA model (AUCs of .83 and .79 respectively). As more variables were included however, the connectionist model’s accuracy increased until 100 variables whereafter improvements were only marginal. Meanwhile including more and more variables in the DFA model made virtually no difference to its performance, nor did allowing the model to select the smallest set of independent contributors via a stepwise procedure. This indicated that the connectionist model made use of a large amount of subtle information among the predictors while the DFA model appeared to capitalise on a small set of variables associated with general between-group differences. This corroborates other work in which no differences between the connectionist model and rival approaches were apparent until further variables were added (Price et al., 2000; Song et al., 2004; Yang, Liu, et al., 2010). It suggests that connectionist models may be of benefit in finding additional effects by incorporation of more variables rather than being limited to those that add unique variance.

The suggestion that conventional statistical models merely emphasise criminal history at the expense of the range of other predictors finds support in the recidivism prediction literature (Coid et al., 2009, 2011; Kroner et al., 2005; Yang, Wong, et al., 2010). Kroner et al. (2005) for example found equivalent and limited accuracy regardless of the items included in risk measures. In Chapter 8 of the present work, pruning according to the theoretical distinction between static or unchanging predictors and those that are dynamic, showed that better prediction was achieved using the DFA model when restricted to static variables than when confined to dynamic variables, while the reverse held when using the connectionist approach. That dynamic factors are more frequently related to recidivism than static factors is strongly held both theoretically and empirically (Andrews & Bonta, 2010; Gendreau et al., 1996). These reviews suggest an additional contribution to predictive validity from criminogenic needs, particularly anti-social attitudes and behaviour relating to education and employment, yet this has not been confirmed in studies of prediction instruments that include these factors (e.g., Farrington et al., 2008; Gray et al., 2004; Yang, Wong, et al., 2010). For example in Farrington et al. a pure static instrument had the strongest predictive validity but the authors were forced to recommend a measure with consistent but weaker evidential support to allow for the assessment of treatment related dynamic factors.

This may be due to the subjective and noisy nature of dynamic variables which are able to vary depending on either the true state of the variable which might be higher or lower at different times, or the ability of the assessor to identify this accurately. Recidivism studies regularly find unexpected effects among dynamic items such as an increase in perception of the positive consequences to crime among desisters (Brown et al., 2009) or a positive relationship between community contacts and violent recidivism (Klassen & O'Connor, 1988). Such findings attest to the inconsistent meanings of these variables for different individuals whereby the same score may not reflect the same risk. Since connectionist models are not dependent on one piece of information they are tolerant to high levels of noise or degradation that invalidate the role of those variables in conventional statistical models (Brodzinsky et al., 1994; Gonzales & DesJardins, 2002; Gottfredson & Moriarty, 2006; McMillen & Henley, 2001; Nolan, 2002; see Chapter 3). Brodzinsky et al. for instance found that high accuracy could be achieved using subjective data that could not enter into their DFA model due to not producing significant overall between-group differences. Even when the variables can be included noise on the data mean that interaction effects are often overlooked by classical statistical methods (Dawes & Corrigan, 1974) but not by connectionist models (e.g., McMillen & Henley, 2001). The present study supports the effectiveness of connectionist models in processing a greater number and complexity of variables than appears possible using conventional methods.

The ability to exploit the information provided by dynamic variables allows the connectionist model to incorporate assessment of the impact of criminogenic needs that are deemed important in general personality and cognitive social learning perspectives on offending behaviour (Andrews & Bonta, 2010). Thus while conventional models appear to be influenced only by indicators of long-term anti-social orientation (static factors), the present connectionist model showed incremental validity by incorporation of measures designed to tap some of the short- and medium-term variations in anti-social potential (dynamic factors). The impact of stable dynamic factors such as ‘understands the importance of completing programmes’ on the present connectionist model’s performance, suggests the importance of combining current attitudinal measures with criminal history in assessment of recidivism risk. The pattern recognition ability of the connectionist model is able to detect interactions between these dynamic characteristics and longer-term static indicators, thus accounting for time to recidivism.

9.3 Practical Implications

Improved accuracy in identifying recidivists and non recidivists carries a number of practical implications including reducing the costs of false classifications, increasing the scope for targeting risk factors for intervention, and understanding if these are impacting on an individual’s recidivism risk. These areas are now briefly discussed in relation to the application of a fully developed connectionist model.

9.3.1 Reduced decision costs.

Tonry (1987) describes that the benefit of improving the accuracy of predictions lies in better policy decisions due to an improved ability to target resources on dangerous offenders, to extend greater leniency to non-dangerous offenders, to reduce prison populations, and thereby achieve greater crime control at less financial cost. Decisions by judges regarding sentence type and length, parole boards regarding the timing of prisoners' release, recommendations by probation and prison staff, and their follow-up actions, are all done on the basis of an offender's apparent risk or dangerousness. The limited accuracy of current risk measures means that these authorities are more inclined to make decisions in an intuitive way, contrary to evidence-based practice, leading to divergence in the decisions reached (e.g., Grove & Meehl, 1996). An effective justice system decision policy therefore depends on the relative likelihood of recidivism and this is where traditional ARAs have been of most value (Glaser, 1987). Attempts to limit costs to the system thus relate directly to the measure's accuracy.

In the prediction of recidivism, one cannot afford to make errors of omission: ruling out recidivism risk when in fact the offender will re-offend. This results in costs to society including substantial police, court, and prison costs not to mention the economic and psychological effects of crime on its victims. Thus the minimisation of false negatives is of utmost importance. It is not cost-effective however to respond to this by being biased towards ruling-in risk due to the professional and administrative costs to criminal justice agencies of supervising cases that will not re-offend and therefore do not require intervention. Moreover, low risk offenders may be negatively affected by levels of supervision that are higher than necessary due to defiance reactions (Sherman, 1993) or deviant peer contagion (Dodge, Dishion, & Lansford, 2006).

The apparent strong ability of fully optimised connectionist models to identify non re-offenders, whilst also improving on the identification of re-offenders, represents a major advance on existing risk measures where moderate accuracy has necessitated a widening of the net by lowering the selection threshold (Campbell, 2003). A by-product of this is the inclusion of many non re-offenders. The near perfect accuracy observed in the current application of connectionist models, including false alarm rates below 5%, reduces organisational decision costs associated with prediction errors and thus alleviates the need to search for the optimal cut point (Harris & Rice, 2007). Although choosing a cut-off lower than 0.5 would increase certainty that a negative classification was not a recidivist, at all cut-points between .90 and .12 prediction errors were below 5%. Wollert (2006) proposed that measures should have positive predictive power (PPP) better than .50 so that professionals have confidence in the classification of an offender as a potential recidivist every time a positive prediction is made. It is therefore reassuring that the PPP of OASys using the connectionist model was .97. Furthermore the negative predictive power was .99 meaning that professionals can also be assured that negative predictions, which are possibly deemed more important by the general public, are strongly reliable too.

This increased accuracy avoids some of the related ethical objections that have been raised to the use of current ARAs (Campbell, 2003; Silver & Miller, 2002). Silver and Miller point out that 'unacceptably high' (p.142) false alarm rates of existing actuarials have hampered their widespread adoption despite limited evidence to justify prioritisation of alternative approaches (e.g., Grove et al., 2000; see Chapter 2). Silver and Miller further contend that the use of actuarial tests that discriminate aggregate groups of individuals along demographic lines also leaves criminal justice providers open to the charge of being agents of social control. Such allegations are mitigated when relevant variables are combined using a connectionist approach due to i) the influence of dynamic and protective factors within these models, and ii) the ability of these models to identify unique combinations of risk factors rather than relying on a few key variables that discriminate the groups.

9.3.2 Interventions: Potential to link with reductions in risk.

There would be little point in accurately identifying re-offenders if recidivism was not readily treatable or preventable. There is good evidence for the ability to reduce recidivism by completion of cognitive-behavioural treatment (e.g., Andrews et al., 1990; Lipsey et al., 2001; Tong & Farrington, 2006). The ability of connectionist models to make use of dynamic variables in prediction, rather than relying on static factors opens up the possibility of better identifying the core factors for treatment to reduce recidivism risk. Due to evidential support for treatment programmes the effect of their completion is considered 'very relevant' to the interpretation of actuarial test results (Harris & Rice, 2007, p.1653). However existing ARAs are insensitive to changes in dynamic characteristics due to the overpowering influence of static factors in the underpinning statistical models. Inability to modify actuarial risk estimates in an explicit way is said to make conventional statistical tests an improper basis for public policy (Lilford & Braunholtz, 1996). Since the additional information relating to treatment progress must then be handled subjectively, without empirical criteria, it is much less useful in defending decisions.

The improved accuracy when clinical assessments of dynamic factors are combined using connectionist modelling lends support to the measurement of these variables within OASys, particularly the offender's understanding of the importance of completing programmes, but also the assessment of a range of criminogenic needs such as problem-solving skills and employability. The importance of this in reducing risk is substantial, particularly given recommendations that general recidivism can be used as a proxy for violent recidivism given the low base-rates for violence (Snowden et al., 2007). This recommendation follows from the overlap in factors associated with general and violent recidivism and proposals that violent offences are committed at random by prolific offenders in the course of criminal careers rather than by violence specialists (e.g., Farrington, 1995; see Chapter 2). Notwithstanding, the ability to impact on risk estimates should encourage probation offender managers to spend more supervision time on underpinning dynamic criminogenic needs; direction of resources in this way has been shown to be productive in reducing reconvictions (Bonta, Rugge, Scott, Bourgon, & Yessine, 2008). Furthermore, criminal justice providers in the UK are now contracted to deliver services to reduce re-offending and will be paid according to their success in achieving reductions (National Offender Management Service, 2005). Effective measurement of the impact of interventions on risk probability therefore provides an important basis for understanding and demonstrating effectiveness.

9.3.3 Accuracy for individuals.

A criticism of current ARA measures is that they do not generate highly individualised risk profiles nor take advantage of unique combinations of risk factors (Clements, 1996; Hart et al., 2007). While conventional statistical models derive risk estimates from scores on variables associated with overall between-group differences, connectionist models recognise patterns in the variables of individual cases while retaining the ability to account for main effects. Connectionist models therefore address the notion that all re-offenders are alike, distinct from non re-offenders, whilst also responding to the possibility that some re-offenders have unique qualities.

Disagreement exists over whether ARAs are appropriate in relation to predictions for individual cases. Since confidence intervals around a prediction relate to the sample size, then their application to an individual case will range across all categories of risk (Hart et al., 2007). This is considered an inapposite use of the confidence interval however since it should be used as a probability statement about the population rather than about any one case within it (Hanson & Howard, 2010; Scurich & John, 2011). Scurich and John suggest that the key issue relates to the precision of a specific estimate, rather than the general estimated interval. They propose that users of actuarials should recognise the conditional probabilities within measures' estimates and interpret them accordingly. Thus a risk estimate is based on combining prior knowledge about the sample base-rate with information about its relationship with individual characteristics from new observations in the population (see also Lilford & Braunholtz, 1996). Since clinical judgement methods are poor at making use of prior probabilities, e.g., the base-rate of recidivism among offenders of a certain age, gender and criminal history (Grove et al., 2000), they are also at a disadvantage for individual predictions (Harris & Rice, 2007).

Although the connectionist model is also dependent on the sample's base-rate, the precision of the estimate provided by the connectionist model is increased relative to conventional models. The present results suggested that this occurred by provision of a moderately large, heterogeneous sample and with longer training. Similar to the MacArthur risk studies (Steadman et al., 2000; see Chapter 2), precision increased with the number of iterations of training (Chapter 6). Although this increases the bias to individual data points, this was minimised by maintaining a large number of observations in the training sample. The accuracy of individual risk predictions on new cases may therefore be contingent on updating the training sample with new complete data patterns, i.e., with known recidivism outcomes.

As well as being more precise, predictions resulting from connectionist models are expected to be more relevant to individual cases, than are predictions resulting from regression models. Current ARAs based on these regression models require an evaluator to refer to the relationship between the scores and the group data in the construction sample, e.g., ‘The majority of cases scoring between .75 and .99 reoffended within 30 months. This case scored .89’. The evaluator has then to decide if the case is more like the majority with the same score that re-offended, or the minority with the same score that did not. This dilemma is not required in interpreting estimates made by connectionist models since they refer to patterns of data rather than a single set of fixed scores that best separate the groups. This is enhanced using the leave-one-out procedure for training and testing where predictions are specifically made for individual cases based on the maximum sample available from the screening setting. The estimate nonetheless needs to be couched in probabilistic terms, e.g., ‘The probability is .89 that an individual with these characteristics will re-offend within 30 months’. Reporting a continuous risk estimate in this way is seen as good practice in mental health expert testimony (Schopp, 1996; Steadman & Monahan, 1994). Further, there is evidence that practitioners can understand risk framed in this way even though the advice cannot be presented in absolute terms (Hilton, Harris, Rawson, & Beach, 2005).

At an individual offender level the effect of a change on a dynamic variable within a connectionist model is arguably more meaningful than within a conventional statistical model since social and psychological risk factors can be expected to be inter-related. For instance it makes little sense to attend to substance misuse independently of the social and personal conditions that made it attractive in the first place. Factors such as poor neighbourhood, lack of legitimate access to paid employment and negative role models are likely to maintain risk in some cases regardless of substance misuse treatment. However, it is possible that the enhancement of protective factors may moderate or mediate exposure to risk (Moore, forthcoming; Rutter, 1985). The pattern recognition capability of the connectionist model, when presented with the individual case, should recognise whether a change on substance misuse needs has reduced the probability of recidivism based on the relationship between the updated pattern of factors and the previous behaviour of other similar cases in the population. Fully trained connectionist models might be used in this way by probation workers, e.g., when formulating risk management plans. This provides an evidence-based test of risk enhancing or risk reducing factors as may be proposed after an assessment using structured professional judgement (e.g., within the OASys risk of serious harm analysis). In any case the prediction arising from the connectionist model should be put together with up-to-date information on the state of dynamic risk factors, consistent with recommendations for ethical risk assessment (Campbell, 2003).

9.3.4 Implementation of connectionist models and impact on service delivery.

Given the observed high predictive accuracy, the use of connectionist modelling is likely to be of interest to practitioners and managers in the criminal justice system. This invokes consideration of implementation costs. Take up of a new method might be hampered if it were considered costly in terms of implementation requirements such as the time needed to input data. However, the risk factors driving the connectionist model were collected during the course of operational delivery using OASys (Home Office, 2002) the dominant framework employed by NOMS prison and probation services across the UK. Thus the data required are being collected as part of standard practice. Given that OASys is not owned by specific prisons or probation trusts responsible for its delivery, unless connectionist modelling is taken on at national level there may still be a need for double entry of the variables into an alternative electronic environment with which the connectionist model can interface. Clearly this would increase the costs of implementing the innovation in practice, and may negate its use if information is needed promptly such as in the context of pre-sentence recommendations for court. Nevertheless the current work is a demonstration that it is possible to link a connectionist model to data in a delivery setting, with excellent results. Although this would need to be embedded in routine practice to allow the model to remain up-to-date with the characteristics of the population, the benefit to the justice system is expected to far outweigh any costs associated with data entry.

A tangible benefit to service delivery associated with the connectionist model is the possibility demonstrated of combining different assessments such as OGRS and the probation offender manager's clinical assessment. Thus clinical assessment is fully incorporated but the model is used to decide which items are relevant and to what extent; without being affected by emotional content or other sources of evaluator bias. The proficiency of the connectionist model in finding the optimal way to combine these measures avoids the confusion that has been experienced by OASys users faced with a number of different assessment tools (Fitzgibbon, 2008; Kemshall, 2003; Moore, 2007b).

This occurs principally when the measures are discordant thus raising suspicion about each of the scales and presenting a dilemma about how to combine them, or which to prioritise. A climate of defensiveness following independent reviews of high profile further offences (HM Inspectorate of Probation, 2006a, 2006b) is likely to mean that practitioners are inclined to default to whichever measure gives the highest risk prediction, often neglecting person-specific factors (Fitzgibbon, 2008). This might be minimised by provision of a fully developed connectionist model, adding confidence to clinical decision-making, even at short follow-up intervals. If cases referred to the model for prediction are from within the same setting as used for the training sample, and this sample has been updated recently, the resulting estimate should be valid. Nevertheless the case should be reviewed as to its similarity to the model's training sample, so that predictions are made only in appropriate cases.

9.4 Limitations

Despite the strong performance of the connectionist model in the current thesis, it is recognised that the study lacked transparency in identifying the relationships between the variables and groups of cases, e.g., differences among variables in females versus males, or in violent offenders versus non-violent cases.[18] This was due to the way in which connectionist models make use of patterns of data which prevented definition of which variables routinely promoted risk and which consistently reduced risk. Thus 'understands the importance of completing programmes' was recognised as an influential variable in the model, but it was not possible to make a statement regarding its effect on sub-groups of cases. This might have been explored by withdrawing pairs of variables sequentially from the model and then testing the effect on model accuracy compared to when each pair was present. Using the formula in Table 3.3 the number of paired interactions for testing would have been 27,730 however. Moreover this would not have provided a complete answer since connectionist models also capitalise on higher order interactions where necessary. Attesting to the difficulty in this area, Marshall and English (2000) found that they could not identify the origin of their model's superiority over

[18] The Risk of Serious Harm screening section of OASys records whether a current or previous serious offence is on record. However, specific offence types were not collected within OASys and therefore were not available to the present study.

logistic regression, despite entering multiple interaction terms into the regression equation. This inability to fully inspect what is inside the 'black box' therefore appears to be a limitation of connectionist modelling as currently applied (Ripley, 1996).

A second limitation, not unique to the current study, was the extent to which the criterion measure of reconviction was a reliable index of re-offending. Re-offences resulting in a reconviction may represent only a small minority of such offences due to the attrition that occurs between the crime occurring and a sentence being secured (Hall, 1987; Lloyd et al., 1994). Much of the attrition is due to differing police practices and performance including detection and prosecution rates, and this may at least have been standardised in the current thesis by use of a single police force area. In addition, the attrition may be reduced among known offenders since conviction outcomes may be pursued more vigorously than for unknown first time offenders (Grubin & Wingate, 1996). In any event there are few alternatives to the use of reconvictions, since offender self-reports are subject to self-presentation bias. Moreover, they are not systematically collected by offender managers or anywhere else in the criminal justice system. Despite the lack of alternatives it is acknowledged that reconviction is an imperfect proxy measure of offender recidivism.

9.5 Future Directions

Although the current application of connectionist modelling demonstrated very high accuracy in the classification of recidivists, the importance of correct classifications is such that even small improvements on this would be of benefit by increasing confidence in decision-making and reducing costs to the system. In addition the same justifications apply to the importance of replicating the present study as discussed further below.

The temperature output connectionist model is a promising area of development, not only because of its ability to model shorter recidivism intervals, but also due to the accuracy levels which were above those seen in the single output models. The only exception to this was at the longest follow-up interval where the single output model had better sensitivity with no difference in false alarms. This may relate to the difficulty in discriminating late re-offenders from non re-offenders. Longer training time than was provided may raise accuracy levels to those seen in the 30 month single output connectionist model, if not further above.

Another way of attempting to increase performance, or at least customise it for the present concerns, might be to change the model's learning rule for adjusting its weights. Instead of generally minimising the sum of squared errors, the error criterion for training might be *d'* in order that the model minimises false predictions according to a pre-determined cost function determined by criminal justice policy. The model's learning would therefore respond to the priorities of the system, by reinforcing weights on units associated with reducing false negatives more than those associated with reducing false positives (or vice versa).

If false positives were eliminated, leaving non re-offenders perfectly classified, then a second stage of training might be implemented to classify the cases into more and less serious recidivists. This is similar to the method used in the MacArthur risk studies (Monahan et al., 2000; Steadman et al., 2000; see Chapter 2) where cases that were successfully classified by the iterative classification tree were withdrawn and the remaining cases were then subjected to recursive partitioning. One-fifth of the previously unclassified cases were then successfully classified due to the model being subjected to a different data distribution. Thus in a similar way a second stage of connectionist model training could be directed at model learning to discriminate violent from non violent recidivists. Alternatively, given high TPR levels, non violent recidivists could be withdrawn and the model trained on a clearer distinction between violent re-offenders and desisters. Future research in this area is needed due to the importance of minimising false negatives in the prediction of violent recidivism, although this was not the goal of the present research.

Connectionist modelling of a full offender caseload does not allow one to isolate the risk factors that contribute to the predictions in sub-groups of cases. Thus in the present study it was not clear to what extent non re-offenders had been protected from recidivism by interventions provided by the criminal justice system, e.g., effective supervision, or accredited treatment programmes. A natural extension of this project would therefore be to test the effect of withdrawing interventions from the cases that are predicted to desist from further offending. Reducing the intensity of supervision for low risk offenders has been shown under experimental study to do no harm (Barnes et al., 2010) but this requires confidence in the risk assessment. If this were implemented successfully the resources saved could then be re-directed to the supervision of higher risk offenders. Future work using connectionist models could therefore evaluate the benefit of organisational strategies such as the introduction or withdrawal of interventions in response to the model's prediction, compared to 'blind' cases in which practice is unaffected by the model's prediction.

The most important future direction in taking forward the application of the present connectionist model, however, is to independently validate the model on new cases from the population. Since the connectionist model prediction is strongly related to the relationships between the data in the construction sample, this may cause problems for independent validation. Even a small change across a single risk factor can potentially change the model. Leave-one-out is a good means of reducing the variance between the model and the testing sample data but it may not be a good test of external validity because the test cases may represent the construction sample too perfectly. Although this is desirable for the local area providing the data for the model, it may be problematic for external applications which may contain different base-rates. Thus the model may over-predict in areas with lower base-rates, and under-predict in areas with higher base-rates. Even in the local area the current model will need to be reinforced and updated with more recent cases, as discussed earlier, to support the excellent predictive accuracy linked to its optimisation in the present thesis.

9.6 Conclusions

This thesis showed a dramatic improvement in accuracy of recidivism risk prediction by application of connectionist modelling, a different method of combining the predictor variables than used traditionally. AUC accuracy levels of .98 as produced by the connectionist model are far higher than previously seen in recidivism prediction where extant methods have reached a 'glass ceiling' of accuracy of around an AUC of .75 regardless of the predictor variables included (Coid et al., 2011; Kroner et al., 2005; Yang, Wong, et al., 2010). This upper limit of predictive accuracy may be interpreted in light of the fact that recidivism risk prediction in criminological psychology has been dominated by a single statistical approach, based on multiple regression equations, in which predictive factors are combined to emphasise average or main effects. Such methods are at a disadvantage where there are a large number of sub-models responsible for additional smaller effects. By contrast connectionist models are able to uncover such sub-models and interactions automatically even under conditions of complex, noisy, or otherwise degraded data.

The advantages of connectionist models in the prediction of recidivism that have been discussed in this thesis are therefore that their outcomes are influenced by multiple factors rather than a few conspicuous items; that they are robust to data problems that negatively affect conventional models; that they represent a reliable means of constructing a composite measure from multiple clinical tests; and that their results can apply to each individual rather than to broad statistical groups. Connectionist models therefore promote individualised clinical assessment but help minimise the subjectivity involved in selecting and emphasising interpretative material (e.g., Grove & Meehl, 1996).

More effective use of the available predictor variables, including treatment-related dynamic factors, supports an influential review attesting to the predictive value of these factors (Gendreau et al., 1996). The under-use of dynamic factors by conventional models also explains the anomalous finding that third generation risk measures incorporating criminogenic needs do not markedly improve on the predictive accuracy of second generation static risk measures (Coid et al., 2009; Farrington et al., 2008; Kroner & Mills, 2001; Yang, Wong, et al., 2010). The accuracy with which static and dynamic factors were combined by the connectionist model, predicting recidivism at short and longer follow-up intervals, indicates the ability of such models to exploit dynamic aspects of recidivism risk. Justice system professionals may therefore focus more effort on reducing offenders’ anti-social attitudes and lifestyle with the aim of demonstrably altering risk probability. This carries the potential to reduce system costs significantly by improving the targeting of supervision or security levels, and monitoring changes to recidivism risk estimates; thereby promoting defensible decision-making.

APPENDIX A: List of Variables for Modelling

	Field	Description	Type
1	UserRef4	ID	
2	PERSISTENT_OFFENDER	Prolific offender status	Static
3	GENDER_CODE	Gender	Demographic
4	ETHNICITY_WHITE	Ethnicity	Demographic
5	ETHNICITY_MIXED	Ethnicity	Demographic
6	ETHNICITY_BLACK	Ethnicity	Demographic
7	ETHNICITY_ASIAN	Ethnicity	Demographic
8	ETHNICITY_CHINESE	Ethnicity	Demographic
9	ETHNICITY_notstated	Ethnicity	Demographic
10	ogrspct	OGRS-II score	Static
11	totalprecons	Total number of prior convictions	Static
12	MOTIVAT	Is the offender motivated to address offending?	Dynamic
13	S13Q3_CPO_REL	Religious or cultural issues affect suitability for community payback	Dynamic
14	S13Q3_CPO_ED	Education / training issue affect suitability for community payback	Dynamic
15	S13Q3_CPO_EMPL	Employment issues affect suitability for community payback	Dynamic
16	S13Q3_CPO_ALC	Alcohol misuse issues affect suitability for community payback	Dynamic
17	S13Q3_CPO_DRUG	Drug misuse issues affect suitability for community payback	Dynamic
18	S13Q3_CPO_CARE	Childcare / domestic responsibilities affect suitability for community payback	Dynamic
19	S13Q3_EM_REL	Religious or cultural issues affect suitability for electronic monitoring	Dynamic
20	S13Q3_EM_ED	Education / training issue affect suitability for electronic monitoring	Dynamic
21	S13Q3_EM_EMPL	Employment issues affect suitability for electronic monitoring	Dynamic
22	S13Q3_EM_ALC	Alcohol misuse issues affect suitability for electronic monitoring	Dynamic
23	S13Q3_EM_DRUG	Drug misuse issues affect suitability for electronic monitoring	Dynamic
24	S13Q3_EM_CARE	Childcare / domestic responsibilities affect suitability for electronic monitoring	Dynamic
25	S13Q3_P_REL	Religious or cultural issues affect suitability for programmes	Dynamic
26	S13Q3_P_ED	Education / training issue affect suitability for programmes	Dynamic
27	S13Q3_P_EMPL	Employment issues affect suitability for programmes	Dynamic
28	S13Q3_P_ALC	Alcohol misuse issues affect suitability for programmes	Dynamic
29	S13Q3_P_DRUG	Drug misuse issues affect suitability for programmes	Dynamic

30	S13Q3_P_CARE	Childcare / domestic responsibilities affect suitability for programmes	Dynamic
31	S13Q4	Understands importance of completing programmes	Dynamic
32	S1_2_WGT	Criminal history	Static
33	S3_WGT	Accommodation needs	Dynamic
34	S4_WGT	Employability needs	Dynamic
35	S5_WGT	Financial Management needs	Dynamic
36	S6_WGT	Relationships needs	Dynamic
37	S7_WGT	Lifestyle and Associates problems	Dynamic
38	S8_WGT	Drug Misuse	Dynamic
39	S9_WGT	Alcohol Misuse	Dynamic
40	S10_WGT	Emotional wellbeing problems	Dynamic
41	S11_WGT	Thinking and Behaviour deficits	Dynamic
42	RISK_RECON	Risk category	Dynamic
43	S1Q2_RESENTENCE_FOR_BREACH	Current sentence for breach	Static
44	sex_off	Sexual offender (y/n)	Static
45	disposal_length	Disposal Length	Static
46	ageass	Age at assessment	Static
47	totneeds_new1	Total number of needs	Dynamic
48	likelihoodrecon1	Risk score	Dynamic
49	curcon_suicide	Current suicide concerns	Dynamic
50	IH	Sentence planning code: HARM	Dynamic
51	IHa	Risk to children	Dynamic
52	Ihb	Risk to prisoners	Dynamic
53	Ihc	Risk to staff	Dynamic
54	Ihd	Risk to public	Dynamic
55	Ihe	Self-harm issues	Dynamic
56	Ihf	Bullying (as perpetrator)	Dynamic
57	Ihg	Partner abuse	Dynamic
58	Ihh	Escape / abscond risk	Dynamic
59	I2	Sentence planning code: ANALYSIS OF OFFENCES	Dynamic
60	I2a	Attitude to victim	Dynamic
61	I2b	Racist attitudes	Dynamic
62	I3	Sentence planning code: ACCOMMODATION	Dynamic
63	I3a	Need for housing / improved housing / more suitable housing	Dynamic
64	I4	Sentence planning code: EDUCATION, TRAINING AND EMPLOYMENT	Dynamic
65	I4a	Problems with literacy / numeracy	Dynamic
66	I4b	Work related skills	Dynamic
67	I4c	Attitude / motivation	Dynamic
68	I5	Sentence planning code: FINANCIAL MANAGEMENT	Dynamic
69	I5a	Money management	Dynamic
70	I6	Sentence planning code: RELATIONSHIPS	Dynamic
71	I6a	Relationships	Dynamic
72	I6b	Domestic Violence issues	Dynamic
73	I6c	Experience of childhood	Dynamic
74	I7	Sentence planning code: LIFESTYLE AND ASSOCIATES	Dynamic

75	I7a	Community integration	Dynamic
76	I7b	Gambling	Dynamic
77	I7c	Recklessness / risk taking behaviour	Dynamic
78	I8	Sentence planning code: DRUG MISUSE	Dynamic
79	I8a	Drug misuse (including motivation to tackle it)	Dynamic
80	I9	Sentence planning code: ALCOHOL MISUSE	Dynamic
81	I9a	Violent behaviour related to alcohol (including drink driving)	Dynamic
82	I9b	Alcohol misuse (inc motivation to tackle alcohol misuse)	Dynamic
83	I10	Sentence planning code: EMOTIONAL WELL-BEING	Dynamic
84	I10a	Difficulties coping	Dynamic
85	I10b	Psychological problems	Dynamic
86	I10c	Social isolation	Dynamic
87	I10d	Psychiatric problems	Dynamic
88	I11	Sentence planning code: THINKING AND BEHAVIOUR	Dynamic
89	I11a	Interpersonal skills	Dynamic
90	I11b	Aggressive / controlling behaviour	Dynamic
91	I11c	Temper control	Dynamic
92	I11d	Problem solving skills	Dynamic
93	I11e	Understands others views	Dynamic
94	I12	Sentence Planning code: ATTITUDES	Dynamic
95	I12a	Motivation to address offending (inc attitudes to staff / supervision)	Dynamic
96	I12b	Discriminatory attitudes	Dynamic
97	I13	Sentence planning code: OTHER	Dynamic
98	I13a	Make constructive use of time	Dynamic
99	I13b	Other	Dynamic
100	ART	Aggression Replacement Training	Dynamic
101	ASRO	Addressing Substance Related Offending	Dynamic
102	CALM	Controlling Anger and Learning to Manage it	Dynamic
103	CSCP	Cognitive Self Change Programme	Dynamic
104	CSCBooster	Cognitive Self Change Booster	Dynamic
105	DID	Drink Impaired Drivers	Dynamic
106	ETS	Enhanced Thinking Skills	Dynamic
107	FOV	Focus On Violence	Dynamic
108	McGuire	McGuire Problem Solving	Dynamic
109	POTO	Priestley One to One	Dynamic
110	PRISM	Substance Misuse	Dynamic
111	RAPt	Rehabilitation for Addicted Prisoners trust	Dynamic
112	RandR	Reasoning and Rehabilitation	Dynamic
113	SexOffender_Booster	Sex Offender Booster programme	Dynamic
114	SexOffender_RP	Sex Offender Relapse Prevention programme	Dynamic
115	SOTP	Sex Offender core Treatment Programme	Dynamic

116	Extended_SOTP	Extended Sex Offender Treatment Programme	Dynamic
117	Rolling_SOTP	Rolling Sex Offender Treatment Programme	Dynamic
118	Substance_Misuse	Substance Misuse	Dynamic
119	TVSOGP	Thames Valley Community Sex Offender Groupwork programme	Dynamic
120	TF	Think First	Dynamic
121	CSOGP	Community Sex Offender Groupwork Programme (W.Midlands)	Dynamic
122	Womens_Programmes	Women's Programmes	Dynamic
123	Other_Acc_Prog	Other accredited programme	Dynamic
124	Accommodation_Advocacy	Advice and Support	Dynamic
125	Employment_Advocacy	Advice and Support	Dynamic
126	Finance_Advocacy	Advice and Support	Dynamic
127	SubstanceAbuse_Advocacy	Advice and Support	Dynamic
128	Health_Advocacy	Advice and Support	Dynamic
129	FamilyIssues_Advocacy	Advice and Support	Dynamic
130	CommunityIntegration_Advocacy	Advice and Support	Dynamic
131	Other_Advocacy	Advice and Support	Dynamic
132	Debt_Counselling	Counselling	Dynamic
133	Addiction_Counselling	Counselling	Dynamic
134	Victims_Counselling	Counselling	Dynamic
135	OffenderCentred_Counselling	Counselling	Dynamic
136	Family_Counselling	Counselling	Dynamic
137	Other_Counselling	Counselling	Dynamic
138	Psychiatric_Interv	Specialist Intervention	Dynamic
139	Psychological_Interv	Specialist Intervention	Dynamic
140	MentalHealth_Inter	Specialist Intervention	Dynamic
141	Therapeutic_Community	Specialist Intervention	Dynamic
142	RMO_work	Specialist Intervention	Dynamic
143	DTTO	Specialist Intervention	Dynamic
144	Other_Specialist_Interv	Specialist Intervention	Dynamic
145	Basic_Skills	Skills	Dynamic
146	Budgetary_Skills	Skills	Dynamic
147	Lifeskills	Skills	Dynamic
148	Thinking_Skills	Skills	Dynamic
149	Thinking_Skills_Sexual	Skills	Dynamic
150	Thinking_Skills_Violence	Skills	Dynamic
151	Relapse_Prevention	Skills	Dynamic
152	Work_Skills	Skills	Dynamic
153	Citizenship	Skills	Dynamic
154	Other	Skills	Dynamic
155	emptyssp	No Interventions planned	Dynamic
156	RSSP1_ACCEPTABLE_ABSENCES1	Number of 'acceptable' absences	Review (Dynamic)
157	RSSP1_UNACCEPTABLE_ABSENCES1	Number of 'unacceptable' absences	Review (Dynamic)
158	RSSP1_FORMAL_WARNING	Formal warning	Review (Dynamic)
159	RSSP1_BREACH_ACTION	Breach	Review (Dynamic)
160	RSSP1_WARNINGS_BEHAVIOUR	Any warnings for behaviour	Review (Dynamic)

161	RSSP3_MOTIVATION_CHANGED	Any change in motivation	Review (Dynamic)
162	RSSP3_HOW_MUCH_MOTIVATED1	How motivated	Review (Dynamic)
163	RSSP3_CAPACITY_CHANGED	Has the capacity to change and reduce re-offending changed	Review (Dynamic)
164	RSSP6_HARM	Any changes to risk of harm	Review (Dynamic)
165	R1Q2_MURDER_CURR1	Current offence type	Static
166	R1Q2_MURDER_PREV1	Previous offence type	Static
167	R1Q2_WOUNDING_CURR1	Current offence type	Static
168	R1Q2_WOUNDING_PREV1	Previous offence type	Static
169	R1Q2_CHILD_SEX_OFF_CURR1	Current offence type	Static
170	R1Q2_CHILD_SEX_OFF_PREV1	Previous offence type	Static
171	R1Q2_ADULT_SEX_OFF_CURR1	Current offence type	Static
172	R1Q2_ADULT_SEX_OFF_PREV1	Previous offence type	Static
173	R1Q2_SCHEDULE1_CURR1	Current offence type	Static
174	R1Q2_SCHEDULE1_PREV1	Previous offence type	Static
175	R1Q2_AGG_BURGLARY_CURR1	Current offence type	Static
176	R1Q2_AGG_BURGLARY_PREV1	Previous offence type	Static
177	R1Q2_ARSON_CURR1	Current offence type	Static
178	R1Q2_ARSON_PREV1	Previous offence type	Static
179	R1Q2_DAMAGE_WITH_INTENT_CURR1	Current offence type	Static
180	R1Q2_DAMAGE_WITH_INTENT_PREV1	Previous offence type	Static
181	R1Q2_KIDNAPPING_CURR1	Current offence type	Static
182	R1Q2_KIDNAPPING_PREV1	Previous offence type	Static
183	R1Q2_FIREARM_CURR1	Current offence type	Static
184	R1Q2_FIREARM_PREV1	Previous offence type	Static
185	R1Q2_RACIAL_CURR1	Current offence type	Static
186	R1Q2_RACIAL_PREV1	Previous offence type	Static
187	R1Q2_ROBBERY_CURR1	Current offence type	Static
188	R1Q2_ROBBERY_PREV1	Previous offence type	Static
189	R1Q2_WEAPONS_CURR1	Current offence type	Static
190	R1Q2_WEAPONS_PREV1	Previous offence type	Static
191	R1Q2_OTHER_SERIOUS_CURR1	Current offence type	Static
192	R1Q2_OTHER_SERIOUS_PREV1	Previous offence type	Static
193	R1Q2_NONE_CURR1	No Current Serious offence	Static
194	R1Q2_NONE_PREV1	No Previous Serious offence	Static
195	R1Q3_ASSAULTED_STAFF_CURR1	Current significant event	Static
196	R1Q3_ASSAULTED_STAFF_PREV1	Previous significant event	Static
197	R1Q3_ASSAULTED_OTHERS_CURR1	Current significant event	Static
198	R1Q3_ASSAULTED_OTHERS_PREV1	Previous significant event	Static
199	R1Q3_VIOLENT_TO_FAMILY_CURR1	Current significant event	Static
200	R1Q3_VIOLENT_TO_FAMILY_PREV1	Previous significant event	Static
201	R1Q3_MEDICATION_CURR1	Current significant event	Static
202	R1Q3_MEDICATION_PREV1	Previous significant event	Static
203	R1Q3_FOUR_YEAR_PLUS_CURR1	Current significant event	Static
204	R1Q3_FOUR_YEAR_PLUS_PREV1	Previous significant event	Static
205	R1Q3_HIGH_RISK_CURR1	Current significant event	Static
206	R1Q3_HIGH_RISK_PREV1	Previous significant event	Static
207	R1Q3_LIFE_CURR1	Current significant event	Static
208	R1Q3_LIFE_PREV1	Previous significant event	Static

209	R1Q3_SUBJECT_S90_92_CURR1	Current significant event	Static
210	R1Q3_SUBJECT_S90_92_PREV1	Previous significant event	Static
211	R1Q3_SECTION41_CURR1	Current significant event	Static
212	R1Q3_SECTION41_PREV1	Previous significant event	Static
213	R1Q3_EXTENDED_SENTENCE_CURR1	Current significant event	Static
214	R1Q3_EXTENDED_SENTENCE_PREV1	Previous significant event	Static
215	R1Q3_STALKER_CURR1	Current significant event	Static
216	R1Q3_STALKER_PREV1	Previous significant event	Static
217	R1Q3_OBSESSIVE_CURR1	Current significant event	Static
218	R1Q3_OBSESSIVE_PREV1	Previous significant event	Static
219	R1Q3_BIZARRE_RITUALISTIC_CURR1	Current significant event	Static
220	R1Q3_BIZARRE_RITUALISTIC_PREV1	Previous significant event	Static
221	R1Q3_NONE_CURR1	No Current significant event	Static
222	R1Q3_NONE_PREV1	No Previous significant event	Static
223	ISSP7_ASSESSOR_POSITION1	Coder type	Static
224	s12q8score1	Is the offender motivated to address offending (0/1/2)	Dynamic
225	R1Q3_HATE_BASED_CURR1	Current significant event	Static
226	R1Q3_HATE_BASED_PREV1	Previous significant event	Static
227	S12_WGT_NEW1	Attitudes	Dynamic
228	R10Q6_CHILDREN_COMM1	Community Risk	Review (Dynamic)
229	R10Q6_CHILDREN_CUST1	Custody Risk	Review (Dynamic)
230	R10Q6_PUBLIC_COMM1	Community Risk	Review (Dynamic)
231	R10Q6_PUBLIC_CUST1	Custody Risk	Review (Dynamic)
232	R10Q6_ADULT_COMM1	Community Risk	Review (Dynamic)
234	R10Q6_ADULT_CUST1	Custody Risk	Review (Dynamic)
235	R10Q6_STAFF_COMM1	Community Risk	Review (Dynamic)
236	R10Q6_STAFF_CUST1	Custody Risk	Review (Dynamic)
237	R10Q6_PRISONERS_CUST1	Custody Risk	Review (Dynamic)
238	SENTENCE_GRP1	Prison and Probation or Probation only (1 / 0)	Static
239	Reoff_in_30m_or_more	Longest first re-offending 30 months, maximum follow-up 55 months	Criterion / Target

APPENDIX B: Cross-Validation Results of Single Output Models Predicting Time to Re-Offending

Table B1

Predictive Accuracy of the DFA Model at Each Time to Re-Offending Follow-Up

Follow-up (months)	Overall accuracy (95% CI)	TPR	FPR	AUC (95% CI)	*d′*
6	.72 (.70-.73)	.62	.28	.74 (.71-.77)	.90
12	.67 (.65-.68)	.57	.32	.69 (.66-.71)	.66
18	.62 (.61-.64)	.59	.37	.67 (.65-.69)	.57
24	.62 (.60-.63)	.59	.37	.67 (.66-.69)	.57
30	.74 (.73-.76)	.74	.26	.82 (.81-.84)	1.31

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); 95% CI = 95 percent confidence interval; AUC = Area Under the receiver operating characteristic Curve; *d′* = dprime.

Table B2

Predictive Accuracy of the Three Layer Model at Each Time to Re-Offending Follow-Up

Follow-up (months)	Overall accuracy (95% CI)	TPR	FPR	AUC (95% CI)	*d′*
6	.98 (.97-.98)	.64	.00	.87 (.84-.90)	3.70
12	.99 (.98-.99)	.92	.00	.96 (.94-.97)	4.58
18	.99 (.98-.99)	.94	.00	.97 (.96-.98)	4.72
24	.99 (.98-.99)	.96	.01	.98 (.97-.99)	4.31
30	.98 (.98-.99)	.99	.03	.99 (.98-.99)	4.21

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); 95% CI = 95 percent confidence interval; AUC = Area Under the receiver operating characteristic Curve; *d′* = dprime.

Table B3

Predictive Accuracy of the Two Layer Model at Each Time to Re-Offending Follow-Up

Follow-up (months)	Overall accuracy (95% CI)	TPR	FPR	AUC (95% CI)	*d'*
6	.94 (.93-.95)	.11	.00	.53 (.49-.57)	2.40
12	.88 (.87-.89)	.12	.00	.54 (.51-.57)	2.47
18	.82 (.81-.83)	.10	.00	.53 (.51-.55)	1.90
24	.76 (.74-.77)	.17	.01	.55 (.53-.57)	1.58
30	.80 (.78-.81)	.86	.30	.77 (.76-.79)	1.61

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); 95% CI = 95 percent confidence interval; AUC = Area Under the receiver operating characteristic Curve; *d'* = dprime.

Table B4

Predictive Accuracy of the OGRS Model at Each Time to Re-Offending Follow-Up

Follow-up (months)	Overall accuracy (95% CI)	TPR	FPR	AUC (95% CI)	*d'*
6	.49 (.47-.50)	.58	.52	.56 (.52-.59)	.15
12	.50 (.49-.52)	.59	.51	.56 (.53-.58)	.20
18	.52 (.50-.53)	.60	.50	.57 (.55-.59)	.25
24	.54 (.53-.56)	.61	.49	.59 (.57-.61)	.32
30	.70 (.68-.71)	.69	.29	.78 (.76-.79)	1.05

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); 95% CI = 95 percent confidence interval; AUC = Area Under the receiver operating characteristic Curve; *d'* = dprime.

APPENDIX C: Impact of Individual Variables on Connectionist Model

Table C1

Variables Impacting Most Upon Accuracy when Omitted During the Switch Off Variables Analysis (in order of change in d')

Rank	Predictor omitted	Change [a] in TPR	Change [a] in FPR	Change [a] in Overall accuracy	Change [a] in *d'*
1	Understands importance of completing programmes	-.09	.15	-.12	-2.10
2	Gender	-.10	.12	-.11	-1.99
3	White ethnicity	-.09	.12	-.10	-1.94
4	Coder type	-.07	.11	-.08	-1.78
5	OGRS	-.31	-.01	-.18	-1.69
6	No current serious offence	-.08	.07	-.08	-1.67
7	Risk category	-.08	.06	-.07	-1.55
8	Motivation to address offending	-.07	.06	-.07	-1.52
9	No previous significant risk event	-.05	.08	-.06	-1.51
10	Emotional well-being problems	-.04	.10	-.06	-1.51
11	Thinking and behaviour deficits	-.05	.07	-.06	-1.49
12	Employability needs	-.04	.09	-.06	-1.49
13	No previous serious offence	-.04	.10	-.06	-1.45
14	Alcohol misuse needs	-.04	.08	-.06	-1.40
15	No current significant risk event	-.03	.09	-.06	-1.36
16	Sentence group	-.03	.07	-.05	-1.25

Table C1 (continued)

Variables Impacting Most Upon Accuracy when Omitted During the Switch Off Variables Analysis

Rank	Predictor omitted	Change [a] in TPR	Change [a] in FPR	Change [a] in Overall accuracy	Change [a] in *d'*
17	Lifestyle and associates problems	-.03	.04	-.04	-1.09
18	Previous significant event: Assaulted others	-.03	.05	-.04	-1.07
19	OASys criminal history	-.02	.06	-.03	-.97
20	Age at assessment	-.00	.13	-.05	-.94
21	Motivation to stop offending	.01	.05	-.03	-.89
22	Current significant event; Assaulted others	-.02	.04	-.03	-.86
23	Total number of criminogenic needs	-.02	.03	-.02	-.81
24	Accommodation needs	-.02	.03	-.02	-.74
25	Relationships needs	-.02	.02	-.02	-.66
26	Financial management needs	-.01	.02	-.02	-.61
27	Motivation to address offending (review stage)	.00	.04	-.02	-.48
28	Drug misuse needs	-.01	.01	-.01	-.46
29	Risk score	-.02	.00	-.01	-.45
30	Need for citizenship skills	-.01	.02	-.01	-.44
31	Disposal length	.00	.02	-.01	-.43
32	Risk of harm to the public in the community (review stage)	.00	.02	-.01	-.43
33	Current sentence for breach	-.01	.02	-.01	-.42
34	Previous significant offence: wounding	.00	.02	-.01	-.42

Table C1 (continued)

Variables Impacting Most Upon Accuracy when Omitted During the Switch Off Variables Analysis

Rank	Predictor omitted	Change [a] in TPR	Change [a] in FPR	Change [a] in Overall accuracy	Change [a] in *d′*
35	Risk of harm to children in the community	.00	.03	-.01	-.41
36	Risk of harm to a known adult in the community	.00	.02	-.01	-.41
37	Pro-criminal attitudes	-.01	.01	-.01	-.41
38	Formal warning (review stage)	-.01	.01	-.01	-.38
39	Current concerns for suicide	-.01	.01	-.01	-.37
40	Employment issues affect programme suitability	-.01	.01	-.01	-.37
41	Previous significant event: violent to family	-.01	.00	-.01	-.35
42	Current significant event: violent to family	.00	.02	-.01	-.34
43	Sentence planning objective: Risk to public	-.01	.01	-.01	-.33
44	Current significant offence: wounding	.00	.02	-.01	-.32
45	Employment issues affect suitability for electronic monitoring	-.01	.01	-.01	-.32
46	Employment issues affect suitability for community unpaid work	.00	.01	-.01	-.32
47	Sentence planning objective: Attitude to victim	-.01	.01	-.01	-.30
48	Change in capacity to reduce re-offending	.00	.01	-.01	-.30
49	Risk of harm to staff in the community	.00	.02	-.01	-.26
50	Sentence planning objective: Other	.00	.01	-.01	-.24

Note. TPR = True Positive Rate; FPR = False Positive Rate (1-specificity); *d′* = dprime; [a] Baseline TPR, FPR, Overall accuracy and *d′* values were .99, .01. .98, and 4.31 respectively.

APPENDIX D: Variables Remaining in the DFA Model after Stepwise Variable Refinement

Table D1

Variables Selected by the DFA Stepwise Method (in order of coefficient weighting)

Rank	Predictor	Standardised Discriminant Function Coefficient (β)
1	OGRS	.574
2	Age at assessment	-.331
3	Total number of criminogenic needs	.311
4	No current serious offence	.160
5	No previous significant risk event	-.096
6	Motivation to address offending (review)	-.090
7	Sentence plan objective: Interpersonal skills	-.088
8	Sentence plan objective: Addressing Substance Related Offending programme	.086
9	Sentence plan objective: Thinking Skills programme	.081
10	Previous significant risk event: Violent to family	.079
10	Previous significant risk event: Four years (or more) imprisonment	.079
12	Sentence group	-.074
13	Sentence plan objective: Risk of harm to prisoners	-.073
14	Gender	.071
14	Sentence plan objective: Self harm issues	.071
16	Sentence plan objective: Controlling Anger and Learning to Manage it programme	-.070
17	Previous significant risk event: Obsessive behaviour	-.066
18	Childcare / domestic responsibilities affect programme suitability	-.061
18	Specialist intervention: Mental health	-.061

Glossary

APD — Anti-social Personality Disorder. An enduring pattern of behaviour marked by a history of irresponsible and antisocial acts beginning in childhood or early adolescence and continuing into adulthood.

ARA — Actuarial Risk Assessment. Used to describe prediction measures whose items are scored based on empirical criteria rather than clinical judgement.

AUC — The Area Under the ROC Curve. The common measure of accuracy of risk assessment in legal and criminological psychology. Provides a number between 0 and 1 representing the probability that a case that goes on to re-offend will have a higher score on the measure than a case that subsequently stays offence free. 0.5 represents chance classification and 1 represents perfect positive prediction.

d′ — dprime. Measures a test's ability to detect re-offenders from other cases. The difference is measured by transforming the scores from the two samples (to neutralise their variability) and then measuring the standard deviation units between the means for re-offenders and non re-offenders. Values of 0.00 indicate an inability to detect re-offenders, and larger values indicate correspondingly greater signal detection. A value of 2.00 indicates that the distance between the means is twice as large as the standard deviations of the two distributions.

DFA — Discriminant Function Analysis. A variety of MRA that identifies linear combinations of the predictor variables that best discriminate the different levels of the criterion variable.

FPR — False Positive Rate. The fraction of non re-offenders wrongly predicted to re-offend. Calculated by FP/(TN+FP) i.e., dividing the number of false alarms by the sum of the number of true negatives and false alarms.

HCR-20 Historical, Clinical and Risk management 20 (Webster et al., 1997). A structured clinical judgement approach designed for the prediction of violent offending using a pre-defined set of static and dynamic items from the research literature on violence.

(I)CT (Iterative) Classification Tree. Structured sequence of yes/no answers that lead to the classification of a case as high/low risk. Iterative trees repeatedly run the sequence over unclassified cases to improve overall classification.

LR Logistic Regression. A version of MRA used when the criterion variable is dichotomous rather than continuous.

LSI-R Level of Service Inventory – Revised (Andrews et al., 1995). A statistically derived risk/needs assessment instrument. It is an example of a ‘third generation’ ARA due to its inclusion of theoretically informed static and dynamic items.

MDO Mentally Disordered Offender. Label given to offenders identified as having a clinical syndrome (e.g., schizophrenia, manic depression, major depression), or a personality disorder (e.g., APD) using an accepted diagnostic system.

MRA Multiple Regression Analysis. A statistical technique used to investigate linear relationships between three or more variables. It indicates the extent to which a continuous criterion variable can be explained by one or more of the predictor variables.

NPP Negative Predictive Power. Proportion of cases identified as low risk that are in fact low risk. Calculated by TN/(FN+TN), i.e., dividing the number of correct rejections by the sum of the number of misses and correct rejections.

NOMS National Offender Management Service. An executive agency of the UK Ministry of Justice responsible for commissioning criminal justice services from prisons and probation trusts.

OASys Offender Assessment System (Home Office, 2002). Risk assessment and management framework used across NOMS. OASys is an example of a 'fourth generation' ARA due to its inclusion of static and dynamic items and multiple time point assessment.

OGRS Offender Group Reconviction Scale (Copas & Marshall, 1998). Brief risk assessment device completed on the basis of offender demographics and criminal history. An example of a 'second generation' ARA (based on static items).

PCL-R Psychopathy Checklist – Revised (Hare, 2003). Psychometric tool for assessing personality constructs relevant to violent recidivism. Incorporated in the HCR-20 and the VRAG.

PPP Positive Predictive Power. Proportion of cases designated as a risk that are in fact a risk. Calculated by TP/(FP+TP), i.e., dividing the number of hits by the sum of the number of hits and false alarms.

ROC Receiver Operating Characteristic. Describes graphically the sensitivity of a test in detecting risk at ever decreasing levels of caution (i.e., as false alarms rise). Thus a test's TPR is plotted as a function of its FPR to give the AUC statistic.

TPR True Positive Rate. The fraction of actual re-offenders correctly predicted to reoffend. Also referred to as 'sensitivity'. Calculated by TP/(FN+TP), i.e., the number of hits divided by the sum of hits and misses.

VRAG Violence Risk Appraisal Guide (Quinsey et al., 1998). A 'second generation' ARA measure designed to predict violent offending from a combination of static risk factors, including the PCL-R.

References

Entries marked with an asterisk (*) were included in the systematic review (Chapter 3).

*Abe, H., Ashizawa, K., Li, F., Matsuyama, N., Fukushima, A., Shiraishi, J., … & Doi, K. (2004). Artificial neural networks (ANNs) for differential diagnosis of interstitial lung disease : results of a simulation test with actual clinical cases1. *Academic Radiology, 11*(1), 29-37.

*Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services, 8* (3), 147-156.

*Alonso-Betanzos, A., Mosqueira-Rey, E., Moret-Bonillo, V., & del Rio, B.B. (1999). Applying statistical, uncertainty-based and connectionist approaches to the prediction of fetal outcome: a comparative study. *Artificial Intelligence in Medicine, 17*(1), 37-57.

American Psychiatric Association (1987). *Diagnostic and statistical manual of mental disorders* (3rd ed. Rev.). Washington, DC: Author.

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of mental disorders* (4th ed.). Washington, DC: Author.

Andrews, D.A. (1980). Some experimental investigations of the principles of differential association through deliberate manipulation of the structure of service systems. *American Sociological Review, 45,* 448-462.

Andrews, D.A. (1982). *The Level of Supervision Inventory (LSI): The first follow-up.* Ottawa, Canada: Carlton University, Department of Psychology.

Andrews, D.A. (1983). The assessment of outcome in correctional samples. In M.L. Lambert, E.R., Christensen, & S.S. DeJulo (Eds.), *The Measurement of Psychotherapy Outcome* (pp.160-201). New York, NY: Wiley.

Andrews, D.A., & Bonta, J. (1995). *LSI-R: The Level of Service Inventory-Revised.* Toronto, ON: Multi-Health Systems.

Andrews, D.A., & Bonta, J. (2010). *The psychology of criminal conduct* (5th ed.). New Providence, NJ: Lexis Nexis / Mathew Bender.

Andrews, D.A., Bonta, J., & Wormith, S.J. (2004). *The Level of Service/Case Management Inventory (LS/CMI).* Toronto, ON: Multi-Health Systems.

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

Andrews, D.A., Zinger, I., Hoge, R.D., Bonta, J., Gendreau, P., & Cullen, F.T. (1990). Does correctional treatment work? A clinically relevant and psychologically informed meta-analysis. *Criminology, 28,* 369-404.

*Andriulli, A., Grossi, E., Buscema, M., Festa, V., Intraligi, N. M., Dominici, P., … & Perri, F. (2003). Contribution of artificial neural networks to the classification and treatment of patients with uninvestigated dyspepsia. *Digestive and Liver Disease, 35* (4), 222-231.

Ang, R.P., & Huan, V.S. (2008). Predictors of recidivism for adolescent offenders in a Singapore sample. *Criminal Justice and Behavior, 35* (7), 895-905.

Armstrong, T.A., & Britt, C.L. (2004). The effect of offender characteristics on offense specialization and escalation. *Justice Quarterly, 21* (4), 843-876.

Aubrey, R., & Hough, M. (1997). *Assessing offenders' needs: Assessment scales for the probation service.* Home Office Research Study 166. London, UK: Home Office.

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

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

Barnes, G.C., Ahlman, L., Gill, C., Sherman, L.W., Kurtz, E., & Malvestuto, R. (2010). Low-intensity community supervision for low-risk offenders: A randomized, controlled trial. *Journal of Experimental Criminology, 6,* 159-189.

Barnett, A., Blumstein, A., & Farrington, D.P. (1987). Probabilistic models of youthful criminal careers. *Criminology, 25,* 83-107.

Barnett, G., Wakeling, H., & Howard, P. (2010). An examination of the predictive validity of the Risk Matrix 2000 in England and Wales. *Sexual Abuse: A Journal of Research and Treatment, 22* (4), 443-470.

*Bassoe, C-F., (1995). Automated diagnoses from clinical narratives: A medical system based on computerised medical records, natural language processing, and neural network technology. *Neural Networks, 8* (2), 313-319.

Baumer, E. (1997). Levels and predictors of recidivism: The Malta experience. *Criminology, 35* (4), 601-628.

*Baxt, W.G., & Skora, J. (1996). Prospective validation of artificial neural network trained to identify acute myocardial infarction. *The Lancet, 347,* 12-15.

Beech, A.R. (1998). A psychometric typology of child abusers. *International Journal of Offender Therapy and Comparative Criminology, 42,* 319-339.

Beech, A.R., Erikson, M., Friendship, C., & Ditchfield, J. (2001). A six year follow-up of men going through probation-based sex offender treatment programmes. Home Office Research Findings 144. London, UK: Home Office.

Belfrage, H., Fransson, G., & Strand, S. (2000). Prediction of violence using the HCR-20 risk. A prospective study in two maximum security correctional institutions. *Journal of Forensic Psychiatry, 11,* 167-175.

Benda, B.B., & Tollett, C.L. (1999). A study of recidivism of serious and persistent offenders among adolescents. *Journal of Criminal Justice, 27* (2), 111-126.

*Betechuoh, B.L., Marwala, T., & Manana, J.V. (2008). Computational intelligence for HIV modelling. INES 12th international conference on intelligent engineering systems, February 25-29, Miami, FL.

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

Blackburn, R. (1993). *The psychology of criminal conduct: Theory, research and practice.* Chichester, UK: Wiley.

Blumstein, A., Cohen, J., & Farrington, D.P. (1988). Criminal career research: Its value for criminology. *Criminology, 26,* 1-35.

Blumstein, A., Farrington, D.P., & Moitra, S. (1985). Delinquency careers: Innocents, desisters, and persisters. In M. Tonry and N. Morris (Eds.), *Crime and justice: An annual review of research,* Volume 6. Chigago, IL: University of Chicago Press.

Bonta, J. (1996). Risk-needs assessment and treatment. In A.T. Harland (Ed.), *Choosing correctional options that work: Defining the demand and evaluating the supply* (pp.18-32). Thousand Oaks, CA: Sage.

Bonta, J., Harman, W.G., Hann, R.G., & Cormier, R.B. (1996). The prediction of recidivism among federally sentenced offenders: A revalidation of the SIR scale. *Canadian Journal of Criminology, 38,* 61-79.

Bonta, J., Law, M., & Hanson, R.K. (1998). The prediction of criminal and violent recidivism among mentally disordered offenders: A meta-analysis. *Psychological Bulletin, 123,* 123-142.

Bonta, J., Rugge, T., Scott, T., Bourgon, G., & Yessine, A.K. (2008). Exploring the black box of community supervision. *Journal of Offender Rehabilitation, 47,* 248-270.

Bonta, J., & Yessine, A.K. (2005). *The national flagging system: Identifying and responding to high risk, violent offenders* [Report No. 2005-04]. Ottawa, ON: Department of Public Safety and Emergency Preparedness Canada.

Boothby, J.L., & Clements, C.B. (2000). A national survey of correctional psychologists. *Criminal Justice and Behavior, 27,* 715-731.

Borum, R. (1996). Improving the clinical practice of violence: Risk assessment, technology, guidelines, and training. *American Psychologist, 51,* 945-956.

*Bottaci, L., Drew, P. J., Hartley, J. E., Hadfield, M. B., Farouk, R., Lee, P.W. R., … & Monson, J.R.T. (1997). Artificial neural networks applied to outcome prediction for colorectal cancer patients in separate institutions. *The Lancet, 350*, 469-472.

Bowles, R.A., & Florackis, C. (2007). Duration of the time to reconviction: Evidence from UK prisoner discharge data. *Journal of Criminal Justice, 35,* 365-378.

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

Brennan, P., Mednick, S., & Richard, J. (1989). Specialization in violence: Existence of a criminal sub-group. *Criminology, 27,* 437-453.

Britt, C.L. (1996). The measurement of specialization and escalation in the criminal career: An alternative modeling strategy. *Journal of Quantitative Criminology, 12,* 193-222.

Brown, L.D. (1978). The development of a parolee classification system using discriminant analysis. *Journal of Research in Crime and Delinquency, 15,* 92-108.

Brown, S.L., St. Amand, M.D., & Zamble, E. (2009). The dynamic prediction of criminal recidivism: A three-wave prospective study. *Law and Human Behavior, 33,* 25-45.

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

*Bryce, T. J., Dewhirst, M. W., Floyd Jr, C. E., Hars, V., & Brizel, D. M. (1998). Artificial Neural Network Model of Survival in Patients Treated With Irradiation With and Without Concurrent Chemotherapy for Advanced Carcinoma of the Head and Neck. *International Journal of Radiation Oncology Biology Physics, 41*(2), 339-345.

Burgess, E.M. (1928). Factors determining success or failure on parole. In A.A. Bruce, E.W. Burgess, & A.J. Arn (Eds.), *The working of the intermediate sentence law and the parole system in Illinois* (pp. 205-249). Springfield, IL: State Board Parole.

*Buscema, M., Mazzetti di Pietralata, M., Salvemini, V., Intraligi, M., & Indrimi, M. (1998). Application of artificial neural networks to eating disorders. *Substance Use & Misuse, 33 (3),* 765-791.

Buss, D.M., & Craik, K.H. (1989). On the cross-cultural examination of acts and dispositions. *European Journal of Personality, 3,* 19-30.

*Buzatu, D.A., Taylor, K.K., Peret, D.C., Darsey, J.A., & Lang, N.P. (2001). The Determination of Cardiac Surgical Risk Using Artificial Neural Networks. *Journal of Surgical Research, 95* (1), 61-66.

Campbell, M.A., French, S., & Gendreau, P. (2009). The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. *Criminal Justice and Behavior, 36*, 567-590.

Campbell, T.W. (2003). Sex offenders and actuarial risk assessments: Ethical considerations. *Behavioral Sciences and the Law, 21* (2), 269-279.

Castillo, E.D., & Fiftal Alarid, L. (2011). Factors associated with recidivism among offenders with mental illness. *International Journal of Offender Therapy and Comparative Criminology, 55* (1), 98-117.

*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,* (3), 227-240.

*Cazzaniga, M., Borroni, G., Ceriani, R., Guerzoni, P., Casiraghi, M.A., & Salerno, F. (2008). Artificial neural network (ANN) to predict cirrhosis in chronic hepatitis C (CHC) comparison with a logistic regression (LG) model. *Journal of Hepatology, 48* (Suppl.), S270.

Centre for Reviews and Dissemination (CRD) (1996). *Undertaking systematic reviews of research on effectiveness.* York, UK: University of York.

*Cevenini, G., Barbini, E., Scolletta, S., Biagioli, B., Giomarelli, P., & Barbini, P. (2007). A comparative analysis of predictive models of morbidity in intensive care unit after cardiac surgery - Part II: An illustrative example. *BMC Medical Informatics and Decision Making, 7*, 36-48.

*Chang, L-Y. (2005). Analysis of freeway accident frequencies: negative binomial regression versus artificial neural network. *Safety Science, 43,* 541-557.

*Chun, F.K.H., Graefen, M., Briganti, A., Gallina, A., Hopp, J., Kattan, M.W., … & Karakiewicz, P.I. (2007). Initial biopsy outcome prediction - head to head comparison of a logistic regression-based nomogram versus artificial neural network. *European Eurology, 51,* 1236-1243.

*Ciampi, A., & Zhang, F. (2002). A new approach to training back-propagation artificial neural networks: empirical evaluation on ten data-sets from clinical studies. *Statistics in Medicine, 21,* 1309-1330.

Cleckley, H. (1976). *The mask of sanity,* (6th ed.). St. Louis, MO: Mosby.

Clark, D.A., Fisher, M.J., & McDougall, C. (1993). A new methodology for assessing the level of risk in incarcerated offenders. *British Journal of Criminology, 33* (3), 436-448.

Clements, C.B. (1996). Offender classification: Two decades of progress. *Criminal Justice and Behavior, 23,* 121-143.

Coid, J., Yang, M., Ulrich, S., Zhang, T., Roberts, A., Roberts, C., … & Farrington, D.P. (2007). Predicting and understanding risk of re-offending: The Prisoner Cohort Study. Ministry of Justice Research Summary, 6. London, UK: Ministry of Justice.

Coid, J., Yang, M., Ullrich, S., Zhang, T., Sizmur, S., Farrington, D., & Rogers, R. (2011). Most items in structured risk assessment do not predict violence. *The Journal of Forensic Psychiatry & Psychology, 22* (1), 3-21.

Coid, J., Yang, M., Ullrich, S., Zhang, T., Sizmur, S., Roberts, C., … & Rogers, R.D. (2009). Gender differences in structured risk assessment: Comparison of the accuracy of five instruments. *Journal of Consulting and Clinical Psychology, 77*, 337-348.

Collins, R.E. (2010). The effect of gender on violent and nonviolent recidivism: A meta-analysis. *Journal of Criminal Justice, 38,* 675-684.

*Collins, J., & Clark, M. (1993). An application of the theory of neural computation to the prediction of workplace behaviour: An illustration and assessment of network analysis. *Personnel Psychology, 46* (3), 503-524.

Concato, J, Feinstein, A.R., & Holford, T.R. (1993). The risk of determining risk with multivariable models. *Annals of International Medicine, 118,* 201-210.

*Connor, J.P., Symons, M., Feeney, G.F.X., Young, R. McD., & Wiles, J. (2007). Pilot study: The application of machine learning techniques as an adjunct to clinical decision-making in alcohol dependence treatment. *Substance Use & Misuse, 42,* 2193-2206.

Cook, D.J., Mylrow, C.D., & Haynes, R.B. (1997). Systematic reviews: Synthesis of best evidence for clinical decisions. *Annals of Internal Medicine, 126,* 376-380.

Cook, E., & Goldman, L. (1984). Empiric comparison of multivariate analytic techniques: Advantages and disadvantages of recursive partitioning analysis. *Journal of Chronic Disease, 37,* 721-731.

Cooke, D.J., & Mitchie, C. (2001). Refining the construct of psychopathy: Towards a hierarchical model. *Psychological Assessment, 13,* 171-188.

Cooke, D.J., Mitchie, 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 paper No. 5/2001. Edinburgh, UK: The Scottish Prison Service.

Cookson, H., & Clark, D.A. (1998). *Predicting reconviction rates from psychometric and self-report measures.* HM Prison Service, unpublished.

Copas, J., & Marshall, P. (1998). The Offender Group Reconviction Scale: The statistical reconviction score for use by probation officers. *Journal of the Royal Statistical Society , 47C*, 159-171.

Cornish, D., & Clarke, R.V., (1975). *Residential treatment and its effects on delinquency.* Home Office Research Unit Study No.32. Her Majesty's Stationery Office: London.

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

Craig, A.R., Frankin, J.A., & Andrews, G. (1984). A scale to measure locus of control of behaviour. *British Journal of Medical Psychology, 57,* 173-180.

Craig, L A., Thornton, D., Beech, A., & Browne, K. D. (2007). The relationship of statistical and psychological risk markers to sexual reconviction. *Criminal Justice and Behavior, 34,* 314–329.

Curtis, E.W. (1971). Predictive value compared to predictive validity. *American Psychologist, 26,* 908-914.

Dahle, K-P. (2006). Strengths and limitations of actuarial prediction of criminal re-offence in a German prison sample: A comparison study of LSI-R, HCR-20, and PCL-R. *International Journal of Law and Psychiatry, 29,* 431-442.

*Das, A., Ben-Manachem, T., Cooper, G.S., Chak, A., SivakJr, M.V., Gonet, J.A., & Wong, R.C.K. (2003). Prediction of outcome in acute lower-gastrointestinal hemorrhage based on an artificial neural network: internal and external validation of a predictive model. *The Lancet, 362,* 1261-1266.

*da Silva, L.A., Hernandez, E.D.M. & Rangayyan, R.M. (2008). Classification of breast masses using a committee machine of artificial neural networks. *Journal of Electronic Imaging, 17* (1), 1-10.

Dawes, R.M., & Corrigan, B. (1974). Linear models in decision making. *Psychological Bulletin, 81,* 95-106.

Dayhoff, J.E., & Deleo, J.M. (2001). Artificial neural networks – opening the black box. *Cancer, 91 (Suppl)*, 1615-1635.

de Vogel, de Ruiter, Hildebrand, M., Bos, B., & van den 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. *International Journal of Forensic Mental Health, 3* (2), 149-165.

Dempster, R., & Hart, S. (2002). The relative utility of fixed and variable risk factors in discriminating sexual recidivists and nonrecidivists. *Sexual Abuse: A Journal of Research and Treatment, 14* (2), 121-138.

Dingwall, R. (1989). Some problems about predicting child abuse and neglect. In O. Stevenson (Ed.), *Child abuse: Public policy and professional practice* (pp.28-53). Hemel Hempstead, UK: Harvester Wheatsheaf.

Dobbs, J., Green, H. & Zealey, L. (2006). Focus on ethnicity and religion . Office for National Statistics. London: Her Majesty's Stationery Office. Available from http://www.ons.gov.uk/ons/

Dodge, K.A., Dishion, T.J., & Lansford, J.E. (2006). *Deviant peer influences in programs for youth*. New York, NY: Guilford Press.

Doren, D.M. (2002). *Evaluating sex offenders: A manual for civil commitments and beyond.* Thousand Oaks, CA: Sage.

Douglas, K.S., Ogloff, J.R.P., Nicholls, T.L., & Grant, I. (1999). Assessing risk for violence among psychiatric patients: The HCR-20 violence risk assessment scheme and the Psychopathy Checklist: Screening Version. *Journal of Consulting and Clinical Psychology, 67,* 917-930.

Douglas, K.S., & Webster, C.D. (1999). The HCR-20 violence risk assessment scheme: Concurrent validity in a sample of incarcerated offenders. *Criminal Justice and Behavior, 26,* 3-19.

Doyle, M., & Dolan, M. (2006). Predicting community violence from patients discharged from mental health services. *British Journal of Psychiatry, 189,* 520-526.

*Dybowski, R., Weller, P., Chang, R., & Gant, V. (1996). Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm. *The Lancet, 347,* 1146-1150.

D'Zurilla, T., & Goldfried, M. (1971). Problem solving and behavior modification. *Journal of Abnormal Psychology, 78,* 107-126.

*Edwards, D.F., Hollingsworth, H., Zazulia, A.R., & Diringer, M.N. (1999). Artificial neural networks improve the prediction of intercerebral hemorrhage. *Neurology, 53,* 351.

Efron, B. (1983). Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American Statistical Association, 78* (382), 316-330.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science, 14,* 179-211.

English, D.J., Marshall, D.B., Brummel, S.C., & Coghlan, L.K. (1998). *Decision-making in child protective services: A study of effectiveness. Final Report, Phase 1: Quantitative analysis.* Olympia, WA: State of Washington Department of Social and Health Services.

*Fallah-Tafti, M. (2001). The application of artificial neural networks to anticipate the average journey time of traffic in the vicinity of merges. *Knowledge-Based Systems, 14,* 203-211.

Farrington, D.P. (1978). The family backgrounds of aggressive youths. In L. Hersov, M. Berger, & D. Shaffer (Eds.), *Aggression and Antisocial Behaviour in Childhood and Adolescence.* Oxford, UK: Pergamon.

Farrington, D.P. (1989). Early predictors of adolescent aggression and adult violence. *Violence and victims, 4*, 79-100

Farrington, D.P. (1992). Criminal career research in the United Kingdom. *British Journal of Criminology, 32* (4), 521-536.

Farrington, D.P. (1995). The development of offending and anti-social behaviour from childhood: Key findings from the Cambridge Study in Delinquent Development. *Journal of Child Psychology and Psychiatry, 360* (6), 929-964.

Farrington, D.P., Coid, J.W., Harnett, L.M., Jolliffe, D., Soteriou, N., Turner, R.E., & West, D. (2006). Criminal careers up to age 50 and life success up to age 48: New findings from the Cambridge study on delinquent development. *Home Office Research Study 299.* London, UK: HMSO.

Farrington, D.P., & Hawkins, J.D. (1991). Predicting participation, early onset, and later persistence in officially recorded offending. *Criminal Behaviour and Mental Health, 1,* 1-33.

Farrington, D.P., Jolliffe, D., & Johnstone, L. (2008). *Assessing violence risk: A framework for practice.* Glasgow: Scottish Risk Management Authority.

Farrington, D.P., Lambert, S., & West, D.J. (1998). Criminal careers of two generations of family members in the Cambridge Study in Delinquent Development. *Studies on Crime and Crime Prevention, 7,* 85-106.

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology, Research and Practice, 17,* 420-430.

Field, A. (2005). *Discovering statistics using SPSS.* 2nd Ed. Sage: London.

*Finne, P., Finne, R., Auvinen, A., Juusela, H., Aro, J., Maattanen, L.,...& Stenman, U-H. (2000). Predicting the outcome of prostate biopsy in screen-positive men by a multi-layer perceptron network. *Urology, 56* (3), 418-422.

Fitzgerald, S., Gray, N.S., Taylor, J., & Snowden, R.J. (2011). Risk factors for recidivism in offenders with intellectual disabilities. *Psychology, Crime & Law, 17* (1), 43-58.

Fitzgibbon, D.W. (2008). Fit for purpose? OASys assessments and parole decisions. *Probation Journal, 55* (1), 55-69.

*Flaherty, C.W. & Patterson, D.A. (2003). Predicting child physical abuse recurrence: Comparison of a neural network to logistic regression. *Journal of Technology in Human Services, 21* (4), 93-111.

Francis, B., Soothill, K., & Piquero, A.R. (2007). Estimation issues and generational changes in modeling criminal career length. *Crime & Delinquency, 53* (1), 84-105.

Friendship, C., Thornton, D., Erikson, M., & Beech, A. (2001). Reconviction: A critique and comparison of two main data sources in England and Wales. *Legal and Criminological Psychology, 6* (1), 121-129.

*Frize, M., Ennett, C. M., Stevenson, M., & Trigg, H. C. E. (2001). Clinical decision support systems for intensive care units: using artificial neural networks. *Medical Engineering & Physics, 23*(3), 217-225.

Frude, N., Honess, T., & Maguire, M. (1994). *CRIME PICS II*. Cardiff, UK: Michael and Associates.

*Fukushima, A., Ashizawa, K., Yamaguchi, T., Matsuyama, N., Hayashi, H., Kida, I., … & Hayashi, K. (2004). Application of an artificial neural network to high resolution CT: Usefulness in differential diagnosis of diffuse lung disease. *American Journal of Radiology, 183,* 297-305.

Gendreau, P., Goggin, C., & Smith, P. (2002). Is the PCL-R really the “unparalleled” measure of offender risk? A lesson in knowledge cumulation. *Criminal Justice and Behavior, 29,* 397– 426.

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology, 34* (4), 575-607.

*Gioftsos, G., & Grieve, D.W. (1996). The use of artificial neural networks to identify patients with chronic low-back pain conditions from patterns of sit-to-stand manoeuvres. *Clinical Biomechanics, 11* (5), 275-280.

Girard, L., & Wormith, J. (2004). The predictive validity of the Level of Service Inventory – Ontario Revision on general and violent recidivism among various offender groups. *Criminal Justice and Behavior, 31,* 150-181.

Glaser, D. (1987). Classification for risk. *Crime and Justice, 9,* 249-291.

Glass, G.V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5* (10), 3-8.

*Gonzales, J.M.B., & DesJardins, S.L. (2002). Artificial neural networks: A new approach to predicting application behaviour. *Research in Higher Education, 43* (2), 235-258.

Gottfredson, S.D. (1987). Prediction: An overview of selected methodological issues. In D.M. Gottfredson and M. Tonry (Eds.), *Prediction and classification: Criminal justice decision-making, crime and justice: A review of research* (Vol.9, pp. 21-53). Chicago, IL: Chicago University Press.

Gottfredson, M.R., & Hirschi, T. (1990). *A general theory of crime.* Stanford, CA: Stanford University Press.

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

Gottfredson, S.D., & Gottfredson, D.M. (1980). Screening for risk: A comparison of methods. *Criminal Justice and Behaviour, 7* (3), 315-330.

Gottfredson, S.D., & Gottfredson, D.M. (1985).Screening for risk among parolees: Policy, practice and method. In D.P. Farrington and R. Tarling (Eds.) *Prediction in criminology*. Albany, NY: State University of New York Press.

Grann, M., Belfrage, H., & Tengstrom, A. (2000). Actuarial assessment of risk for violence: Predictive validity of the VRAG and the historical part of the HCR-20. *Criminal Justice and Behavior, 27,* 97-114.

*Grann, M., & Langstrom, N. (2007). Actuarial assessment of violence risk: To weigh or not to weigh? *Criminal Justice and Behavior, 34* (1), 22-36.

Gray, N.S., Snowden, R.J., MacCulloch, S., Phillips, H., Taylor, J., & MacCulloch, M.J. (2004). Relative efficacy of criminological, clinical and personality measures of future risk of offending in mentally disordered offenders: A comparative study of HCR-20, PCL-SV, and OGRS. *Journal of Consulting and Clinical Psycholgy, 72,* 523-530.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York, NY: Wiley.

Grove, W.M., & Meehl, P.E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy and Law, 2* (2), 293-323.

Grove, W.M., Zald, D.H., Lebow, B.S, Snitz, B.E., Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12* (1), 19-30.

Grubin, D. (2008). *Validation of Risk Matrix 2000 for use in Scotland.* Glasgow: Scottish Risk Management Authority.

Grubin, D., & Wingate, S. (1996). Sexual offence recidivism: Prediction versus understanding. *Criminal Behavior and Mental Health, 6,* 349–359.

Gudjonnson, G.H., & Bownes, I. (1992). The reasons why suspects confess during custodial interrogation: data for Northern Ireland. *Medicine, Science, and the Law, 32,* 204-212.

Hall, H.V. (1987). *Violence prediction: Guidelines for the forensic practitioner.* Springfield, IL: Charles C. Thomas.

Hanley, J.A., & McNeil, B.J. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology, 143*, 29-36.

Hanson, R.K., & Bussière, M.T., (1998). Predicting relapse: A meta-analysis of sexual offender recidivism studies. *Journal of Consulting and Clinical Psychology, 66* (2), 348-362.

Hanson, R.K., & Harris, A.J.R. (2000). Where should we intervene? Dynamic predictors of sexual offense recidivism. *Criminal Justice and Behavior, 27,* 6-35.

Hanson, R.K., & Harris, A.J.R. (2001). A structured approach to evaluating change among sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 13,* 105-122.

Hanson, R.K., Harris, A.J.R., Scott, T-L., & Helmus, L. (2007). *Assessing the risk of sexual offenders on community supervision: The dynamic supervision project* [Report No. 2007-05]. Ottawa, ON: Department of Public Safety and Emergency Preparedness Canada.

Hanson, R.K., & Howard, P.D. (2010). Individual confidence intervals do not inform decision-makers about the accuracy of risk assessment evaluations. *Law and Human Behavior, 34*, 275–281.

Hanson, R.K., & Morton-Bourgon, K.E. (2005). The characteristics of persistent sexual offenders: A meta-analysis of recidivism studies. *Journal of Consulting and Clinical Psychology, 73* (6), 1154-1163.

Hanson, R.K., & Morton-Bourgon, K.E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment, 21* (1), 1-21.

Hanson, R.K., & Thornton, D. (2000). Improving risk assessment for sexual offenders: A comparison of three actuarial scales. *Law and Human Behavior, 24,* 119-136.

Hanson, R.K., & Wallace-Capretta, S. (2004). Predictors of criminal recidivism among male batterers. *Psychology, Crime & Law, 10* (4), 413-427.

Hare, R.D. (1991). *The Hare Psychopathy Checklist—Revised.* Toronto, ON: Multi-Health Systems.

Hare, R.D. (1980). A research scale for the assessment of psychopathy in criminal populations. *Personality and Individual Differences, 1,* 111-119.

Hare, R. D. (2003). *The Hare Psychopathy Checklist—Revised* (2nd ed.). Toronto, ON: Multi-Health Systems.

Hare, R.D., Clark, D., Grann, M., & Thornton, D. (2000). Psychopathy and the predictive validity of the PCL-R: An international perspective. *Behavioral Sciences and the Law, 18* (5), 623-645.

Harris, G.T., & Rice, M.E. (2007). Characterizing the value of actuarial violence risk assessments. *Criminal Justice and Behavior, 34* (12), 1638-1658.

Harris, G.T., Rice, M.E., & Cormier, C.A. (2002). Prospective replication of the Violence Risk Appraisal Guide in predicting violent recidivism among forensic patients. *Law and Human Behavior, 26,* 377-394.

Harris, G.T., Rice, M.E., & Quinsey, V.L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior, 20,* 315-335.

Hart, S.D., Laws, D.R., & Kropp, R.P. (2003). The promise and peril of sex offender risk assessment. In Ward, T., Laws, D.R., & Hudson, S.M. (Eds.). *Sexual deviance: Issues and controversies*. Thousand Oaks, CA: Sage.

Hart, S.D., Michie, C., & Cooke, D.J. (2007). Precision of actuarial risk assessment instruments: Evaluating 'margins of error' of group v. individual predictions of violence. *British Journal of Psychiatry, 190* (S49), s60-s65.

Hartwell, S.W. (2004). Comparison of offenders with mental illness only and offenders with dual diagnoses. *Psychiatric Services, 55* (2), 145-150.

Hathaway, S.R., & McKinley, J.C. (1967). *Minnesota Multiphasic Personality Inventory manual* (rev. ed.). New York, NY: Psychological Corporation.

Hautus, M. (1995). Corrections for extreme proportions and their biasing effects on estimated values of *d'*. *Behavior Research Methods, Instruments, & Computers, 27,* 46-51.

*Hayashi, Y., Hsieh, M-H., & Setiono, R. (2010). Understanding consumer heterogeneity: A business intelligence application of neural networks. *Knowledge-Based Systems, 23,* 856-863.

Haykin, S. (1999). *Neural networks: A comprehensive foundation* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Healy, D. (2010). Betwixt and between: The role of psychosocial factors in the early stages of desistance. *Journal of Research in Crime and Delinquency, 47* (4), 419-438.

Heil, P., Harrison, L., English, K., Ahlmeyer, S. (2009). Is prison sexual offending indicative of community risk? *Criminal Justice and Behavior, 36* (9), 892-908.

Hemphill, J.F., & Hare, R.D. (1998). Psychopathy Checklist factor scores and recidivism. *Issues in Criminological and Legal Psychology, 24,* 68-73.

Hemphill, J.F., & Hare, R.D. (2004). Some misconceptions about the Hare PCL-R and risk assessment: A reply to Gendreau, Goggin, and Smith. *Criminal Justice and Behavior, 31*(2), 203-243.

Hemphill, J.F., Hare, R.D., & Wong, S. (1998). Psychopathy and recidivism: A review. *Legal and Criminological Psychology, 3,* 139-170.

Hill, G. (1985). Predicting recidivism using institutional measures. In D.P. Farrington and R. Tarling (Eds.), *Prediction in criminology.* Chichester, UK: Wiley.

Hilton, N.Z., Harris, G.T., Rawson, K., & Beach, C.A. (2005). Communicating violence risk information to forensic decision makers. *Criminal Justice and Behavior, 32* (1), 97-116.

Hirschi, T., & Gottfredson, M. (1983). Age and the explanation of crime. *American Journal of Sociology, 89,* 552-584.

Hirschi, T., & Hindelang, M.J. (1977). Intelligence and delinquency: A revisionist review. *American Sociological Review, 42,* 571-587.

Hodgins, S. (1995). Major mental disorder and crime: An overview. *Psychology, Crime & Law, 2,* 5-17.

Hoffman, P.B. (1983). Screening for risk: A revised salient factor score (SFS 81). *Journal of Criminal Justice, 11,* 539-547.

Hoffman, P.B. (1994). Twenty years of operational use of a risk prediction instrument: The United States Parole Commission's Salient Factor Score. *Journal of Criminal Justice, 22,* 477-494.

Home Office (2002). Offender Assessment System: User Manual. London, UK: Author.

Howard, P.D. (2011). Hazards of different types of reoffending. Ministry of Justice Research Series 3/11. London, UK: Ministry of Justice.

Howard, P.D., Clark, D., & Garnham, N. (2006). An evaluation of the Offender Assessment System (OASys) in three pilots, 1999-2001. London, UK: Home Office.

Huebner, B.M., & Berg, M.T. (2011). Examining the sources of variation in risk for recidivism. *Justice Quarterly, 28* (1), 146-173.

*Iyer, S.R., & Sharda, R. (2009). Prediction of athletes' performance using neural networks: An application in cricket team selection. *Expert Systems with Applications, 36,* 5510-5522.

Jeon, Y., & Choi, C-H. (1999). Thermometer coding for multi-layer perceptron learning on continuous mapping problems. *Proceedings of the International Joint Conference on Neural Networks, 3*, 1685-1690.

Jones, L.F. (2004). Offence paralleling behaviour (OPB) as a framework for assessment and interventions with offenders. In A. Needs & G. Towl (Eds.), *Applying Psychology to Forensic Practice* (pp. 34-63). Oxford, UK: Blackwell and British Psychological Society.

Kempf, K.L. (1987). Specialization and the criminal career. *Criminology, 25* (2), 399-420.

Kemshall, H. (2003). *Understanding risk in criminal justice.* Maidenhead, UK: Open University Press.

*Kennedy, R.L., Harrison, R.F., Burton, A.M., Fraser, H.S., Hamer, W.G., MacArthur, D., … & Steedman, D.J. (1997). An artificial neural network system for diagnosis of acute myocardial infarction (AMI) in the accident and emergency department: evaluation and comparison with serum myoglobin measurements. *Computer Methods and Programs in Biomedicine, 52,* 93-103.

Klassen, D. & O’Connor, W. (1988). A prospective study of predictors of violence in adult male mental patients. *Law and Human Behavior, 12,* 143-158.

Kline, P. (1994). *An easy guide to factor analysis.* London, UK: Routledge.

Kroner, D.G., & Mills, J.F. (2001). The accuracy of five risk appraisal instruments in predicting institutional misconduct and new convictions. *Criminal Justice and Behavior, 28* (4), 471-489.

Kroner, D.G., Mills, J.F., & Reddon, J.R. (2005). A coffee can, factor analysis, and prediction of antisocial behavior: The structure of criminal risk. *International Journal of Law and Psychiatry, 28,* 360–374.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the International Joint Conference on Artificial Intelligence, 1137-1145. Available online at http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.48.529&rep=rep1&type=pdf

*Ladstätter, F., Garrosa, E., Badea, C., & Moreno, B. (2010). Application of artificial neural networks to a study of nursing burnout. *Ergonomics, 53* (9), 1085-1096.

Laub, J.H., & Sampson, R.J. (2003). *Shared beginnings, divergent lives: Delinquent boys to age 70.* Cambridge, MA: Harvard University Press.

*Lee, Y-C., Lee, W-J., Lee, T-S., Lin, Y-C, Wang, W., Liew, P-L., … & Chien, C-W. (2007). Prediction of successful weight reduction after bariatric surgery by data mining technologies. *Obesity Surgery, 17,* 1235-1241.

Lilford, R.J., & Braunholtz, D. (1996). For debate: The statistical basis of public policy: a paradigm shift is overdue. *British Medical Journal, 313,* 603-607.

Lipsey, M.W., Chapman, G.L., & Landenberger, N.A. (2001). Cognitive-behavioural programs for offenders. *The ANNALS of the American Academy of Political and Social Science, 578,* 144-157. DOI: 10.1177/000271620157800109.

Litwack, T.R. (2001). Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy and Law, 7,* 409-443.

Lloyd, C., Mair, C., & Hough, M. (1994). Explaining reconviction rates: A critical analysis. Home Office Research Study No. 136. London, UK: Her Majesty's Stationery Office.

Loeber, R. (1982). The stability of antisocial and delinquent child behaviour: A review. *Child Development, 53,* 1431-1446.

Loeber, R., & Dishion, T. (1983). Early predictors of male delinquency: A review. *Psychological Bulletin, 94* (1), 68-99.

Loza, W., & Green, K. (2003). The Self-Appraisal Questionnaire: A self-report measure for predicting recidivism versus clinician administered measures: A 5-year follow-up study. *Journal of Interpersonal Violence, 18,* 781-797.

*Lundin, M., Lundin, J., Burke, H.B., Toikkanen, S., Pylkkanen, L., Joensuu, H. (1999). Artificial neural networks applied to survival prediction in breast cancer. *Oncology, 57,* 281-286.

Lyons, P., Doueck, H.J., & Wodarski, J.S. (1996). Risk assessment for child protective services: A review of the empirical literature on instrument performance. *Social Work Research, 20* (3), 143-155.

MacKenzie, D.L. (1997). *Criminal justice and crime prevention*. In L. Sherman, D. Gottfredson, D. MacKenzie, J. Eck, P. Reuter, and S. Bushway (Eds.), *Preventing crime: What works, what doesn't, what's promising.* Washington, DC: National Institute of Justice.

Makarios, M., Steiner, B., & Travis III, L.F. (2010). Examining the predictors of recidivism among men and women released from prison in Ohio. *Criminal Justice and Behavior, 37* (12), 1377-1391.

Mannheim, H., & Wilkins, L.T. (1955). Prediction methods in relation to Borstal training. Home Office Studies in the Causes of Delinquency and the Treatment of the Offender, No. 1. London, UK: Her Majesty's Stationery Office.

*Maqsood, I., & Abraham, A. (2007). Weather analysis using ensemble of connectionist learning paradigms. *Applied Soft Computing, 7,* 995-1004.

*Marshall, D.B. & English, D.J. (2000). Neural network modelling of risk assessment in child protective services. *Psychological Methods, 5,* 102-124.

May, C. (1999). Explaining reconviction following a community sentence: the role of social factors. Home Office Research Study 192. London: Home Office.

May, C., Sharma, N., & Stewart, D. (2008). *Factors linked to re-offending: a one-year follow-up of prisoners who took part in the resettlement surveys 2001, 2003, and 2004.* London, UK: Ministry of Justice.

McAdams, D.P. (1997). A conceptual history of personality psychology. In R. Hogan, J. Johnson, and S. Briggs (Eds.), *Handbook of Personality Psychology.* San Diego, CA: Academic Press.

McCulloch, W.S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5,* 115-133.

McDougall, C., & Clark, D.A. (1991). A risk assessment model. In S. Boddis (Ed.), *Proceedings ofthe Prison Psychology Conference.* London: Her Majesty's Stationery Office.

McDougall, C., Clark, D.A., & Woodward, R. (1995). Application of operational psychology to assessment of inmates. *Psychology, Crime and Law, 2*, 85-99.

McGrath, R.J., Cumming, G., Livingstone, J.A., & Hoke, S.E. (2003). Outcome of a treatment program for adult sex offenders: From prison to community. *Journal of Interpersonal Violence, 18* (1), 3-17.

*McMillen, R., & Henley, T. (2001). Connectionism isn't just for cognitive science: Neural networks as methodological tools. *The Psychological Record, 51,* 3-18.

Meehl, P.E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194-216.

*Mecocci, P., Grossi, E., Buscema, M., Intraligi, M., Savere, R., Rinaldi, P., & Cherubini, A. (2002). Use of artificial neural networks in clinical trials: A pilot study to predict responsiveness to donepezil in Alzheimer's Disease. *Journal of the American Geriatrics Society, 50,* 1857-1860.

Meloy, J.R. (1996). Stalking (obsessional following): A review of some preliminary studies. *Aggression and Violent Behavior, 1* (2), 147-162.

Miller, D. E., & Kunce, J. T. (1973). Prediction and statistical overkill revisited. *Measurement and Evaluation in Guidance, 6*, 157-163.

Mills, J.F., & Kroner, D.G. (2003). Anger as a predictor of institutional misconduct and recidivism in a sample of violent offenders. *Journal of Interpersonal Violence, 18* (3), 282-294.

Mills, J.F., & Kroner, D.G. (2006). The effect of discordance among violence and general recidivism risk estimates on predictive accuracy. *Criminal Behavior and Mental Health, 16,* 155-166.

Ministry of Justice (2010a). Statistics on race and the Criminal Justice System 2008/09: A Ministry of Justice publication under section 95 of the Criminal Justice Act 1991. London, UK: Crown copyright

Ministry of Justice (2010b). Statistics on women and the Criminal Justice System: A Ministry of Justice publication under section 95 of the Criminal Justice Act 1991. London, UK: Crown copyright.

Minor, K.I., Wells, J.B., & Sims, C. (2003). Recidivism among federal probationers – predicting sentence violations. *Federal Probation, 67,* 31-36.

Mischel, W. (1968). *Personality and assessment.* New York, NY: Wiley.

Mischel, W., & Shoda, Y. (1995). A cognitive system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review, 102,* 246-268.

*Mobley, B.A., Schechter, E., Moore, W.E., McKee, P.A., & Eichner, J.E. (2000). Predictions of coronary artery stenosis by artificial neural network. *Artificial Intelligence in Medicine, 18,* 187-203.

Moffitt, T.E. (1993). Adolescence limited and life-course persistent anti-social behavior: A devlopmental taxonomy. *Psychological Review, 100,* 674-701.

Moffitt, T.E., Gabrielli, W.F., Mednick, S., & Schulslinger, F. (1981). Socioeconomic status, IQ, and delinquency. *Journal of Abnormal Psychology, 90,* 152-156.

Moore, R. (2007a). *Adult offenders' perceptions of their underlying problems: Findings from the OASys self-assessment questionnaire.* Home Office Research Findings 284. London: Home Office.

Moore, R. (2007b). The compatibility of Asset and OASys: How do the risk/needs assessment systems for young and adult offenders compare? *Vista, 11* (1), 2-13.

Moore, R. (forthcoming). Identifying offenders' positive, promotive and protective factors through the Offender Assessment System (OASys). Ministry of Justice Research Series. London, UK: Ministry of Justice.

*Moreno, L., Pineiro, J.D., Sanchez, J.L., Manas, S., Merino, J.J., Acosta, L., & Hamilton, A. (1995). Using neural networks to improve classification: Application to brain maturation. *Neural Networks, 8* (5), 815-820.

Mossman, D. (1994). Assessing predictors of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology, 62*, 783-792.

Moston, S., Stephenson, G., & Williamson, T.M. (1992). The effects of case characteristics on suspect behaviour during police questioning. *British Journal of Psychology, 32,* 23-40.

Monahan, J. (1981). *The clinical prediction of violence.* Beverly Hills, CA: Sage.

Monahan, J., Steadman, H.J., Applebaum, 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,* 312-320.

Monahan, J., Steadman, H.J., Silver, E., Applebaum, P.S., Robbins, P.C., Mulvey, E.P., Roth, L., Grisso, T., & Banks, S. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence.* New York, NY: Oxford University Press.

Mulvey, E.P., & Lidz, C.W. (1984). Clinical considerations in the prediction of dangerousness in mental patients. *Clinical Psychology Review, 4,* 379-401.

*Naguib, R.N.G., Robinson, M.C., Neal, D.E., & Hamdy, F.C. (1998). Neural network analysis of combined experimental and conventional prognostic markers in prostate cancer: A pilot study. *British Journal of Cancer, 78* (2), 246-250.

*Nakamura, K., Yoshida, H., Engelmann, R., MacMahon, H., Katsuragawa, S., Ishida, T., … & Doi, K. (2000). Computerized analysis of the likelihood of malignancy in solitary pulmonary nodules with use of artificial neural networks. *Radiology, 214,* 823-830.

National Offender Management Service (NOMS) (2005). Restructuring probation to reduce re-offending. London, UK: Home Office.

National Probation Directorate (2000). National standards for the supervision of offenders in the community. London, UK: Home Office.

National Probation Directorate (2002). Revision of national standards of supervision of offenders in the community. Probation Circular 7/02. London, UK: Home Office

*Nguyen, T., Malley, R., Inkelis, S. H., & Kuppermann, N. (2002). Comparison of prediction models for adverse outcome in pediatric meningococcal disease using artificial neural network and logistic regression analyses. *Journal of Clinical Epidemiology, 55* (7), 687-695.

*Nolan, J.R. (2002). Computer systems that learn: an empirical study of the effect of noise on the performance of three classification methods. *Expert Systems with Applications, 23,* 39-47.

Nuffield, J. (1982). *Parole decision-making in Canada: Research towards decision guidelines* (Cat. No. JS-22-65/1982E). Ottawa, ON: Minister of Supply and Services Canada.

Nuttall, C.P. (1960). *The parole risk predictor.* Home Office, Research and Planning Unit, London, UK: Home Office.

Olver, M.E., Stockdale, K.C., & Wormith, J.S. (2011). A meta-analysis of predictors of offender treatment attrition and its relationship to recidivism. *Journal of Consulting and Clinical Psychology, 79,* 6-21.

O'Reilly, R.C., Dawson, C.K., & McClelland, J.L. (1995). *PDP++ neural network simulation software.* Pittsburgh, PA: Carnegie Mellon University.

O'Reilly, R.C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: Massachusetts Institute of Technology.

Otto, R.K. (1992). Prediction of dangerous behavior: A review and analysis of ‘second generation’ research. *Forensic Reports, 5,* 103-133.

Ouston, J. (1984). Delinquency, family background and educational attainment. *British Journal of Criminology, 24,* 2-26.

*Palmer, A., Montano, J.J., & Franconetti, F.J. (2008). Sensitivity analysis applied to artificial neural networks for forecasting time-series. *Methodology, 4* (2), 80-86.

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

Patterson, D.W. (1996). *Artificial neural networks: Theory and applications*. Singapore: Prentice Hall.

Pearson, D.A.S. (2007, July). Exploring the potential value of connectionist modelling in the prediction of further offending of sex offenders. In J. Wood (Chair), *Establishing the role of forensic psychology and economics with multi-agency public protection cases and prolific offenders*. Symposium conducted at the Division of Forensic Psychology Conference, York, UK.

*Peng, S.Y., & Peng, S.K. (2008). Predicting adverse outcomes of cardiac surgery with the application of artificial neural networks. *Anaesthesia, 63,* 705-713.

*Peng, S.Y., Wu, K.C., Wang, J.J., Chuang, J.H., Peng, S.K., & Lai, Y.H. (2007). Predicting post-operative nausea and vomiting with the application of an artificial neural network. *British Journal of Anaesthesia, 98* (1), 60-65.

Petersilia, J. (2003). *When prisoners come home: Parole and prisoner re-entry*. New York, NY: Oxford University Press.

Philipse, M.W.G., Koeter, M.W.J., van der Staak, C.P.F., & van den Brink, W. (2006). Static and dynamic patient characteristics as predictors of criminal recidivism: A prospective study in a Dutch forensic psychiatric sample. *Law and Human Behavior, 30* (3), 309-327.

Piquero, A.R. (2008). Taking stock of developmental trajectories of criminal activity over the life course. In A. Liberman (Ed.), *Longitudinal research on crime and delinquency* (pp. 23-78). New York, NY: Springer.

Piquero, A.R., Brame, R., & Lynam, D. (2004). Studying criminal career length through early adulthood among serious offenders. *Crime & Delinquency, 50* (3)*,* 412-435.

Piquero, A.R., Sullivan, C.J., & Farrington, D.P. (2010). Assessing differences between short-term, high-rate offenders and long-term, low-rate offenders. *Criminal Justice and Behavior, 37* (12), 1309-1329.

Pollack, N. (1990). Accounting for predictions of dangerousness. *International Journal of Law and Psychiatry, 13,* 207-215.

*Poulakis, V., Witzsch, U., de Vries, R., Emmerlich, V., Meves, M., Altmannsberger, H-M., & Becht, E. (2004). Preoperative neural network using combined magnetic resonance imaging variables, prostate specific antigen, and Gleason score to predict prostate cancer recurrence after radical prostatectomy. *European Urology, 46,* 571-578.

*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), 40-51.

Powis, B., (2002). Offenders' risk of serious harm: A literature review. Home Office Occasional Paper, 81. London, UK: Home Office.

Quinsey, V.L., Harris, G.T., Rice, M.E., & Cormier, C.A. (1998). *Violent offenders: Appraising and managing risk.* Washington, DC: American Psychological Association.

*Ravdin, P.M., Clark, G.M., Hilsenbeck, S.G., Owens, M.A., Vendely, P., Pandian, M.R., & McGuire, W.L. (1992). A demonstration that breast cancer recurrence can be predicted by neural network analysis. *Breast Cancer Research and Treatment, 21,* 47-53.

Raynor, P. (2007). Risk and need assessment in British probation: The contribution of LSI-R. *Psychology, Crime and Law, 13,* 125-138.

Raynor, P., Kynch, J., Roberts, C., & Merrington, S. (2000). *Risk and need assessment in probation services: An evaluation.* Home Office Research Study 211. London, UK: Home Office.

Reisig, M.D., Holtfreter, K., & Morash, M. (2006). Assessing recidivism risk across female pathways to crime. *Justice Quarterly, 23* (3), 384-405.

Rettinger, J.L., & Andrews, D.A. (2010). General risk and need, gender specificity, and the recidivism of female offenders. *Criminal Justice and Behavior, 37* (1), 29-46.

Rice, M.E., & Harris, G.T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63* (5), 737-748.

Rice, M.E., & Harris, G.T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior, 29* (5), 615-620.

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

*Risser, R., Chaloupka, C, Grundler, W., Sommer, M., Hausler, J., & Kaufmann, C. (2008). Using non-linear methods to investigate the criterion validity of traffic-psychological test batteries. *Accident Analysis and Prevention, 40,* 149-157.

Rogers, S. (1981). *Factors related to recidivism among adult probationers in Ontario*. Ontario: Ministry of Correctional Services Planning and Research Branch.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organisation in the brain. *Psychological Review, 65,* 386-408.

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

Rosenthal, R. (1984). *Meta-analytic procedures for social research.* Beverly Hills, CA: Sage.

Rosenthal, R. (1991). *Meta-analytic procedures for social research.* Newbury Park, CA: Sage.

Ross, R.R., & Fabiano, E.A. (1985). *Time to think: A cognitive model of delinquency prevention and offender rehabilitation.* Johnson City, TN: Institute of Social Sciences and Arts.

Rojek, D.G., & Erickson, M.L. (1982). Delinquent careers: A test of the escalation model. *Criminology, 20,* 5-28.

Rumelhart, D.E., Hinton, G.E., & Williams, R.J. (1986). Learning representations by back-propagating errors. *Nature, 323,* 533-536.

Rutter, M. (1985). Resilience in the face of adversity: Protective factors and resistance to psychiatric disorder. *British Journal of Psychiatry, 147,* 598-611.

Salekin, R.T., Rogers, R., & Sewell, K.W. (1996). A review and meta-analysis of the Psychopathy Checklist-Revised: Predictive validity of dangerousness. *Clinical Psychology: Science and Practice*, *3,* 203–215.

Sampson, R., & Laub, J. (1995). *Crime in the making: Pathways and turning points through life.* Cambridge, MA: Harvard University Press.

*Santori, G., Fontana, I., & Valente, U. (2007). Application of an artificial neural network model to predict delayed decrease of serum creatinine in pediatric patients after kidney transplantation. *Transplantation Proceedings, 39,* 1813-1819.

Sargent, D.J. (2001). Comparison of artificial neural networks with other statistical approaches: Results from medical data sets. *Cancer, 91, 8,* 1636-1642.

Schmidt, P., & Witte, A.D. (1989). Predicting criminal recidivism using 'split population' survival time models. *Journal of Econometrics, 40*, 141-159.

Schopp, R.F. (1996). Communicating risk assessments: Accuracy, efficacy and responsibility. *American Psychologist, 51*, 939–944.

Schwalbe, C.S. (2007). Risk assessment for juvenile justice: A meta-analysis. *Law and Human Behavior*, *31*, 449-462.

Scurich, N., & John, R.S. (2011). A Bayesian approach to the group versus individual prediction controversy in actuarial risk assessment. *Law and Human Behavior.* Advance online publication. doi: 10.1007/s10979-011-9286-0

Sherman, L.W. (1993). Defiance, deterrence, and irrelevance: A theory of the criminal sanction. *Journal of Research in Crime and Delinquency, 30,* 445-473.

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

Silver, E. & Miller, L.L. (2002). A cautionary note on the use of actuarial risk assessment tools for social control. *Crime & Delinquency, 48* (1), 138-161.

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

Simourd, D.J. (1997). The Criminal Sentiments Scale-Modified and Pride in Delinquency: Psychometric properties and construct validity of two measures of criminal attitudes. *Criminal Justice and Behavior, 24 (1),* 52-70.

Simourd, L., & Andrews, D.A. (1994). Correlates of delinquency: A look at gender differences. *Forum on Corrections Research, 6,* 26-31.

Singleton, N., Meltzer, H., Gatward, R., Coid, J., & Deasy, D. (1998). Psychiatric morbidity among prisoners: Summary report. London, UK: Office for National Statistics.

Sjostedt, G., & Grann, M. (2002). Risk assessment: What is being predicted by actuarial "prediction instruments"? *International Journal of Forensic Mental Health, 1* (2), 179-183.

Sjostedt, G., & Langstrom, N. (2001). Actuarial assessment of sex offender recidivism risk: A cross-validation of the RRASOR and Static-99 in Sweden. *Law and Human Behavior, 25* (6), 629-645.

Sjostedt, G., & Langstrom, N. (2002). Assessment of risk for criminal recidivism among rapists: A comparison of four different measures. *Psychology, Crime, and Law, 8* (1), 25-40.

Smith, D.A., Visher, C.A., & Jarjoura, G.R. (1991). Dimensions of delinquency: Exploring the correlates of participation, frequency, and persistence of delinquent behavior. *Journal of Research in Crime and Delinquency, 28* (1), 6-32.

*Smith, E.R., & DeCoster, J. (1998). Knowledge acquisition, accessibility, and use in person perception and stereotyping: Simulation with a recurrent connectionist network. *Journal of Personality and Social Psychology, 74* (1), 21-35.

Snowden, R.J., Gray, N.S., Taylor, J., & MacCulloch, M.J. (2007). Actuarial prediction of violent recidivism in mentally disordered offenders. *Psychological Medicine, 37,* 1539-1549.

*Song, X., Mitnitski, A., MacKnight, C., & Rockwood, K. (2004). Assessment of individual risk of death using self-report data: An artificial neural network compared with a frailty index. *Journal of American Geriatrics Society, 52,* 1180-1184.

SPSS Inc. (1993). *SPSS for Windows CHAID (release 6.0).* Chicago, IL: Author.

SPSS Inc. (1997). *Neural connection 2.0 user's guide.* Chicago, IL: Author.

SPSS Inc. (2010). *SPSS (release 19.0)*. Chicago, IL: Author.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers, 31* (1), 137-149.

Stattin, H., & Magnussen, D. (1991). Stability and change in criminal behavior up to age 30. *British Journal of Criminology, 31,* 327-346.

Steadman, H.J., & Cocozza, J.J. (1974). *Careers of the criminally insane: Excessive social control of deviance.* Lexington, MA: Lexington Books.

Steadman, H.J., Fabisiak, S., Dvoskin, J., & Holohean, E.J. (1989). A survey of mental disability among state prison inmates. *Hospital and Community Psychiatry, 38,* 1086-1090.

Steadman, H.J., & Monahan, J. (1994). Toward a rejuvenation of risk assessment research. In J. Monahan, and H. Steadman (Eds.) *Violence and Mental Disorder: Developments in Risk Assessments* (pp.1-19). Chigago, IL: University of Chicago Press.

Steadman, H.J., Mulvey, E.P., Monahan, J., Robbins, P.C., Appelbaum, P.S., Grisso, T., Roth, L.H., & Silver, E. (1998). Violence by people discharged from acute psychiatric inpatient facilities and by others in the same neighborhoods. *Archives of General Psychiatry, 55*, 393-401.

Steadman, H.J., Silver, E., Monahan, J., Applebaum, 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 Behaviour, 24* (1), 83-100.

*Stephan, C., Xu, C, Finne, P., Cammann, H., Meyer, H, Lein, M, Jung, K., & Stenman, U., (2007). Comparison of two different artificial neural networks for prostate biopsy indication in two different patient populations. *Urology, 70* (3), 596-601.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, 36,* 111-147.

Strand, S., Belfrage, H., Fransson, G., & Levander, S. (1999). Clinical and risk management factors in risk prediction of mentally disordered offenders – more important than historical data. *Legal and Criminological Psychology, 4,* 67-76.

Swanson, J.W., Swartz, M.S., Van Dorn, R.A., Elbogen, E.B., Wagner, H.R., Rosenheck, R.A.,… & Lieberman, J.A. (2006). A national study of violent behavior in persons with schizophrenia. *Archives of General Psychiatry, 63* (5), 490-499.

Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285-1293.

Swets, J.A., Dawes, R.M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest: A Journal of the American Psychological Society, 1,* 1-26.

Taylor, R. (1999). Predicting reconvictions for sexual and violent offences using the revised Offender Group Reconviction Scale. Home Office Research Findings, 104. London, UK: Home Office.

Theobold, D., & Farrington, D.P. (2010). Policy implications of research on the effects of getting married on offending. *European Journal of Criminology, 7*, 239–247.

Thornberry, T.P., & Jacoby, J.E. (1979). *The criminally insane: A community follow-up of mentally ill offenders.* Chicago, IL: University of Chicago Press.

Thornton, D. (2002). Constructing and testing a framework for dynamic risk assessment. *Sexual Abuse: A Journal of Research and Treatment, 14,* 137-151.

Thornton, D., Mann, R., Webster, S., Blud, L., Travers, R., Friendship, C., & Erikson, M. (2003). Distinguishing and combining risks for sexual and violent recidivism. *Annals of the New York Academy of Sciences, 989,* 225-235.

Tittle, C.R., & Meier, R.F. (1990). Specifying the SES/delinquency relationship. *Criminology, 28,* 271-299.

Tittle, C.R., & Meier, R.F. (1991). Specifying the SES/delinquency relationship by social characteristics of contexts. *Journal of Research in Crime and Delinquency, 28,* 430-455.

Tong, L.S.J., & Farrington, D.P. (2006). How effective is the 'Reasoning and Rehabilitation' programme in reducing reoffending? A meta-analysis of evaluations in four countries. *Psychology, Crime and Law, 12,* 3-24.

Tonry, M. (1987). Prediction and classification: Legal and ethical issues. *Crime and Justice, 9,* 367-413.

Tourassi, G.D., & Floyd, C.E. (1997). The effect of data sampling on the performance evaluation of artificial neural networks in medical diagnosis. *Medical Decision Making, 17,* 186-192.

Tripodi, S.J., Kim, J.S., & Bender, K. (2010). Is employment associated with reduced recidivism? The complex relationship between employment and crime. *International Journal of Offender Therapy and Comparative Criminology, 54* (5), 706-720.

Tsai, C., & Wu, J. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications, 34,* 2639-2649.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124-1131.

Vrieze, S. I., & Grove, W. M. (2010). Multidimensional assessment of criminal recidivism: Problems, pitfalls, and proposed solutions. *Psychological Assessment, 22*, 382-395.

Wakeling, H.C., Howard, P.D., & Barnett, G.D. (2011). Comparing the validity of the RM2000 scales and OGRS3 for predicting recidivism by internet sexual offenders. *Sexual Abuse: A Journal of Research and Treatment, 23* (1), 146-168.

Walter, M., Wiesbeck, G.A., Dittmann, V., & Graf, M. (2011). Criminal recidivism in offenders with personality disorders and substance use disorders over 8 years time at risk. *Psychiatry Research, 186,* 443-445.

Walters, G.D. (1992). *Foundations of criminal science, vol. 2: The use of knowledge.* New York, NY: Praeger.

Walters, G.D. (1995). The psychological inventory of criminal thinking styles, part I: Reliability and preliminary validity. *Criminal Justice and Behavior, 22,* 307-325.

Walters, G.D. (2003). Predicting criminal justice outcomes with the Psychological Checklist and Lifestyle Criminality Screening Form: A meta-analytic comparison. *Behavioral Sciences and the Law, 21*, 89-102.

Walters, G.D. (2006). Risk appraisal versus self-report in the prediction of criminal justice outcomes: A meta-analysis. *Criminal Justice and Behavior, 33,* 279-304.

Walters, G.D. (2009). The Psychological Inventory of Criminal Thinking Styles and Psychopathy Checklist: Screening Version as incrementally valid predictors of recidivism. *Law and Human Behavior, 33,* 497-505.

*Wang, S.J., Ohno-Machado, L., Fraser, H.S.F., & Kennedy, R.L. (2001). Using patient-reportable clinical history factors to predict myocardial infarction. *Computers in Biology and Medicine, 31,* 1-13.

Ward Systems Group (1996) *Neuroshell 2* [computer software]. Frederick, MD: Author.

Webster, C.K., Douglas, D.E., Eaves, D., & Hart, D. (1997). *HCR-20 assessing risk for violence: Version II.* Burnaby, British Columbia, Canada: Mental Health, Law, & Policy Institute, Simon Fraser University.

Wehrman, M.W. (2010). Race, concentrated disadvantage, and recidivism: A test of interaction effects. *Journal of Criminal Justice, 38,* 538-544.

*Wichard, J.D., Cammann, H., Stephan, C., & Tolxdorff, T. (2008). Classification models for early detection of prostate cancer. *Journal of Biomedicine and Biotechnology,* S*pecial Issue of the FBIT 2007,* January 2008.

Wollert, R. (2006). Low base-rates limit expert certainty when current actuarials are used to identify sexually violent predators. *Psychology, Public Policy, and Law, 12,* 56-85.

*Wu, T., Huang, S., & Meng, Y. (2008). Evaluation of ANN and SVM classifiers as predictors to the diagnosis of students with learning disabilities. *Expert Systems with Applications, 34,* 1846-1856.

*Yamamura, S., Takehira, R., Kawada, K., Momose, Y., Nishizawa, K., Katayama, S., & Hirano, M. (2003). Application of artificial neural network modelling to identify severely ill patients whose aminoglycoside concentrations are likely to fall below therapeutic concentrations. *Journal of Clinical Pharmacy and Therapeutics, 28* (5), 425-432.

*Yang, M., Liu, Y.Y., & Coid, .J. (2010). Applying neural networks and other statistical models to the classification of serious offenders and the prediction of recidivism. Ministry of Justice Research Series 6/10. London, UK: Ministry of Justice.

Yang, M., Wong, S.C.P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136* (5), 740-767.

Zadek, L.A. (1965). Fuzzy sets. *Information Control, 8,* 338-354.

Zahedi, F. (1993). *Intelligent Systems for Business: Expert Systems with Neural Networks.* Belmont, CA: Wadsworth Publishing Company.

Zamble, E., & Porporino, F. (1988). *Coping, behavior, and adaptation in prison inmates.* New York, NY: Springer-Verlag.

Zamble, E., & Quinsey, V.L. (1997). *The process of criminal recidivism.* Cambridge, UK: Cambridge University Press.

Zinger, I. (2004). Actuarial risk assessment and human rights: A commentary. *Canadian Journal of Criminology and Criminal Justice, 46,* 607-621.

*Zou, Y., Shen, Y., Shu, L., Wang, Y., Feng, F., Xu, K., … & Liu, W. (1996). Artificial neural network to assist psychiatric diagnosis. *British Journal of Psychiatry, 169*(1), 64-67.