ACTUARIAL ASSESSMENT OF VIOLENCE RISK

To Weigh or Not to Weigh?

MARTIN GRANN
NIKLAS LÅNGSTRÖM
Centre for Violence Prevention, Karolinska Institute, Stockholm, Sweden

The assigning of different weights to risk factors in actuarial formulas for the assessment of violence risk in criminal offenders has been debated. The authors explore the predictive validity of an index with 10 well-established risk factors for criminal recidivism with respect to violent reconvictions among 404 former forensic psychiatric examinees in Sweden. Four different weighting conditions are tested experimentally, including Nuffield's method, bivariate and multivariate logistic regression, and an artificial neural network procedure. Simpler weighting techniques do not improve predictive accuracy over that of a nonweighted reference, and the more complex procedures yield a statistical shrinkage effect. The authors hypothesize that the general lack of causal risk factors in prediction models may contribute to the observed low utility of weighting techniques.

Keywords: violence; risk assessment; prediction; forensic psychiatry; artificial neural network; methodology; VRAG; HCR-20

Growing public concern for violence perpetrated by offenders with mental disorders has called for refined means to assess and manage risk. Furthermore, because of the strong criticism previously raised concerning mental health professionals' inability to correctly assess dangerousness (Ennis & Litwack, 1974; Monahan, 1984), many legal systems currently require that assessment procedures are structured, transparent, and empirically validated to be admissible as expert evidence in court. Various actuarial checklists for risk assessment represent one response to these demands. In this context, actuarial means the assessment is systematized, performed according to fixed rules, and translated into numbers. The typical actuarial instrument comprises risk factors proposed by the empirical literature to be predictors of violence. The assessor systematically rates the presence of the risk factors in a particular offender by assigning a number to each checklist item, usually a 0 for a nonpresent risk factor and a positive integer value for a present risk factor. A higher number reflects a higher degree of presence of the risk factor. Finally, checklist items are summed up to a total score, and the higher this risk score, the higher the risk is hypothesized to be (Hanson, 1997; Quinsey, Rice, Harris, & Cormier, 1998). There are several unresolved issues in checklist-based risk assessment that warrant further theorizing and

AUTHORS' NOTE: *This study was supported by the Bank of Sweden Tercentenary Foundation, the Söderström-Königska and the Vårdal Foundations, the Swedish Council for Social Research, and the Bror Gadelius' Memorial Foundation. The authors gratefully acknowledge the helpful comments on an earlier draft of this article by Kevin S. Douglas and two anonymous reviewers. Correspondence concerning this article should be addressed to Martin Grann, Centre for Violence Prevention, Karolinska Institute, P.O. Box 23000, SE-104 35 Stockholm, Sweden; fax: +46 8 307298; e-mail: Martin.Grann@cvp.se.*

CRIMINAL JUSTICE AND BEHAVIOR, Vol. 34, No. 1, January 2007 22-36
DOI: 10.1177/0093854806290250

empirical scrutiny (see Litwack, 2001), one of which is the assignment of different weights to risk factors depending on the strength of their statistical association with violence risk. Below, two Canadian risk assessment devices published during the 1990s (one with and the other without a weighting formula) will be briefly reviewed.

THE VIOLENCE RISK APPRAISAL GUIDE (VRAG)

The VRAG (Quinsey et al., 1998) is an actuarial instrument that uses a weighting procedure with individual item scores to yield a weighted total score. The 12 risk factors of the VRAG and their weights were determined from studies of a sample of 618 mentally disordered Canadian offenders (Harris, Rice, & Cormier, 2002; Harris, Rice, & Quinsey, 1993; see also Webster, Harris, Rice, Cormier, & Quinsey, 1994). Depending on how the presence of a risk factor affected the base rate of violent failure on the group level, the corresponding VRAG item was weighted with positive or negative integer values. A weight of 0 was assigned if participants with the risk factor had the same (+/– 5%) recidivism base rate as the total study population (Nuffield, 1982, as cited in Quinsey et al., 1998; Webster et al., 1994). For each full deviation of 5% in base rate associated withthe presence of the item, it was assigned a weight of 1. For example, psychopathy assessed with Hare's Psychopathy Checklist–Revised (PCL-R; Hare, 1991), a robust predictor of violence was included as Item 1 of the VRAG. Thus, using the Nuffield weighting procedure, Item 1 of the VRAG will range from –5 for very low PCL-R scores to 12 for very high PCL-R scores. The presence of a *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed.; American Psychiatric Association, 1980) diagnosis of personality disorder (VRAG Item 2), on the other hand, would be scored either 3 (personality disorder present) or –2 (personality disorder absent). When rating individual risk for violence, the scores on each of the 12 weighted items are summed up, and the total score determines participant placement into one of nine risk categories (bins) assumed to reflect the probability of recidivism (Quinsey et al., 1998).

The VRAG has subsequently been tested in subsamples of the population from which it was derived and in independent samples of offenders (see Quinsey et al., 1998, for a review; Harris et al., 2002). The VRAG is an example of an empirically driven approach according to which only risk factors with well-established predictive validity should be considered when assessing risk for violence. Because the validity of unstructured clinical judgment is usually poor, the empiricist's view is that risk assessment should be based solely on actuarial techniques. Quinsey et al. (1998) concluded that the literature "does not make one sanguine about the prospect that intuitive clinical judgment can increase the accuracy of actuarially devised instruments, even when the independent judgments of clinicians are averaged to increase their reliability" (p. 65). They further stated: "What we are advising is not the addition of actuarial methods to existing practice, but rather the complete replacement of existing practice with actuarial methods" (p. 171). However, prediction models must always be tested in samples outside of that used for their calibration (Bleeker et al., 2003). Douglas Hart, Dempster, and Lyon (1999) attempted to replicate these findings in an independent sample. They concluded that the VRAG behaved quite differently than in the original VRAG studies and that observed probabilities of future violence within the VRAG nine-bin categorization of risk were not useful for risk prediction. Furthermore, Tengström (2001) tested the VRAG categorization of absolute probabilities for recidivism

TABLE 1: Mean Scores, Standard Deviations, and Correlations With Postdetainment Violent Recidivism for Items in the H-10 Subscale of the HCR-20 Among 404 Mentally Disordered Violent Offenders

Item	M	SD	*Violent Recidivism (Pearson's* r*)*
H1. Previous violence	1.66	.47	.12*
H2. Young age at first violent incident	1.30	.64	.25**
H3. Relationship instability	1.18	.80	–.03
H4. Employment problems	1.18	.67	.15**
H5. Substance use problems	1.35	.82	.13*
H6. Major mental illness	1.40	.92	.07
H7. Psychopathy	0.81	.81	.32**
H8. Early maladjustment	0.73	.76	.17**
H9. Personality disorder	1.66	.75	.10*
H10. Prior supervision failure	0.94	.97	.26**
Total H-10 Score	12.21	3.72	.32**

Note. HCR-20 = Historical, Clinical, Risk 20.
*$p < .05$. **$p < .01$.

among Swedish forensic psychiatric examinees. The VRAG risk algorithm did not do particularly well, although Tengström's sample closely resembled the original sample used to develop the VRAG in terms of demographic and clinical characteristics.

HISTORICAL, CLINICAL, RISK 20 (HCR-20)

In contrast to the VRAG, the 20-item HCR-20 (Webster, Douglas, Eaves, & Hart, 1997) is an approach with less emphasis on statistical or empirical optimization of predictive accuracy, but more so on clinical utility. It comprises static and dynamic items, and the assessor reviews past (Historical), concurrent (Clinical), and future-oriented (Risk Management) risk factors. The 10 items of the HCR-20's Historical subscale (the H-10, see Table 1) bear close resemblance to those included in the VRAG, except the former (as well as the C-5 and the R-5) are not weighted but coded uniformly on a 3-point scale, with 0 = *definitely absent*, 1 = *partially or maybe present*, and 2 = *definitely present*. The authors prefer to refer to the HCR-20 as a structured professional judgment approach.[1] According to this methodology, the actuarial part provides but the basis for the assessment, and clinical judgment is invited to adjust the overall risk estimates. Clinicians are encouraged to also consider risk factor(s) not included in the HCR-20 that they might find suitable in the specific, individual case. The authors concluded that "for research purposes, it is possible to treat the HCR-20 as an actuarial scale and simply sum the numeric item codes" (Webster et al., 1997, p. 21), although they continued: "For clinical purposes, it makes little sense to sum the number of risk factors present in any given case, and then use fixed, arbitrary cutoffs to classify the individual" (p. 22).

The HCR-20 has been evaluated in forensic psychiatric (Strand, Belfrage, Fransson, & Levander, 1999; Whittemore, 1999), civil psychiatric (Douglas, Ogloff, Nicholls, & Grant, 1999), and correctional settings (Belfrage, Fransson, & Strand, 2000; Douglas & Webster, 1999; see also Douglas, 2004, for an overview).

TO WEIGH OR NOT TO WEIGH?

The factors included in the VRAG and those of the historical part of the HCR-20 (the H-10) are similar, save that the VRAG allocates different weights to the risk factors. Grann, Belfrage, and Tengström (2000) tested and compared the predictive abilities of the VRAG and the H-10 in a sample of 404 mentally disordered offenders. The sample was Swedish and as such truly independent of the Canadian populations studied in the original work on the VRAG and the HCR-20. The data indicated that the H-10 did better in forecasting future violence than did the VRAG, despite the weight optimization of the latter. Some authors have argued that weighting procedures may lead to suboptimal prediction models if the data are affected by high measurement error rates, the data are highly population specific, or the problem domain is complex (Cohen, 1990; Dawes, 1979). In conclusion, there is a need for studies comparing the relative efficacy of different statistical risk prediction procedures. The purpose of the present study was to experimentally explore potential benefits with weighted, as compared to nonweighted, algorithms for the actuarial assessment of risk for future violence among mentally disordered offenders.

METHOD

PARTICIPANTS

We investigated 404 mentally disordered violent offenders, previously reported on by Grann et al. (2000). The sample consisted of violent offenders clinically diagnosed according to guidelines established by the International Statistical Classification of Diseases and Related Health Problems (World Health Organization, 1976). Participants were diagnosed with either personality disorder or schizophrenia at a presentence forensic psychiatric evaluation in Sweden from 1988 to 1993. The cohort was subsequently followed for an average of 8 years through the forensic psychiatric and correctional systems and out into the community. The average age at baseline in the personality disorder subsample was 32 years (SD = 10.4; range = 17 to 72). Ten percent (n = 36) were women, and 32% (n = 115) were born abroad (i.e., were first-generation immigrants). Sixty-two percent (n = 222) suffered from concurrent substance abuse or dependence. A total of 293 (82%) of the personality disorder subsample's initial 358 violent offenders were available for postdetainment follow-up, with follow-up times of at least 2 years. Among 202 violent offenders diagnosed with schizophrenia, the average age at baseline was 33 years (SD = 9.1; range = 16 to 67). All these offenders were male, and 30% (n = 61) were born abroad. The prevalence of concomitant substance abuse or dependence was 50% (n = 101). One hundred and eleven of the offenders with schizophrenia (55%) were available for postdetainment follow-up for 2 years or more. Thus, the final sample consisted of 404 participants diagnosed with either personality disorder or schizophrenia.

PROCEDURE

We used the 10 risk factors included in the Historical subscale of the HCR-20 as predictor variables (Webster et al., 1997; see Table 1). They were weighted according to algorithms described in detail below and thereafter summed up to H-10 risk scores that were tested for

predictive validity. Any reconviction for a violent crime during follow-up (within a fixed time frame of 2 years of time at risk from release or discharge) was chosen as criterion variable. Violent crime was defined as any conviction of attempted or completed homicide, assault, rape, or robbery. Data were reliably coded from forensic psychiatric evaluation files and official agency registers (Grann, Långström, Tengström, & Stålenheim, 1998; Långström et al., 1999).

Estimation of predictive validity. The accuracy of a risk assessment tool can be expressed in different ways. Traditional indices—sensitivity, specificity, and positive and negative predictive values (see Hart, Webster, & Menzies, 1993)—are useful but have been criticized for instability with varying predictor base rates (Rice & Harris, 1995). A state-of-the-art method for the estimation of predictive validity of a continuous risk measure (e.g., a VRAG or H-10 risk score) for a dichotomous outcome (recidivism) is receiver operating characteristic analysis (ROC; Henderson, 1993; Mossman, 1994; Rice & Harris, 1995; see also Hanley & McNeil, 1982, for a comprehensive mathematical account). ROC analysis is much less dependent on base rates than other measures of predictive accuracy. By plotting the hit rate (the rate of true positives) against the false-alarm rate (the rate of false positives) for all observed predictor values, the ROC curve graphically depicts the tradeoff in specificity that occurs as sensitivity is increased with lower cutoff scores and vice versa. The area under the curve (AUC) is the effect size estimate derived from the ROC analysis and ranges from 0.0 (perfect negative prediction) to 1.0 (perfect positive prediction). There is no strong consensus as to the proper interpretation of AUC estimates for predictive validity. It has been proposed that AUCs should be interpreted conservatively as follows: below 0.60 = low accuracy, 0.60 to 0.70 = marginal accuracy, 0.70 to 0.80 = modest accuracy, 0.80 to 0.90 = moderate accuracy, and greater than 0.90 = high accuracy (Sjöstedt & Grann, 2002). In the studies on the predictive accuracy of the VRAG and the HCR-20 cited above, the AUC was typically between 0.70 and 0.80. The AUC was chosen as an estimate of predictive accuracy in the present study.

SUBSETS AND SEEDS

A major problem when fitting a prediction model to a particular data set is that the models tend to be highly population specific. Applying the prediction algorithm to the population from which the algorithm was derived in the first place will always yield an excellent model fit, given that there is any relationship whatsoever between the included predictor variables and the predicted outcome. Therefore, cross-validating the algorithm in independent populations is essential (Kleinbaum, Kupper, Muller, & Nizan, 1998). However, one cross-validation is usually not sufficient, especially if the sample size is small (Cohen, 1990) or the ratio between sample size and number of predictor variables (the subject to variable ratio, S:V ratio) is low. An S:V ratio of at least 5:1 is usually recommended (Cicchetti, 1992; Fletcher, Rice, & Ray, 1978). Thus, deriving a prediction algorithm from one sample and testing it in only one different sample may result in high model fit values by pure coincidence. To counteract the effect of such stochastic variations, the population in the present study was further divided into five subsets. These subsets were in turn combined into five different seeds. The predictive validity (AUC) was estimated for each of the five seeds and then averaged into a combined estimate.

Subsets. Each of the 404 participants was allotted a random value between 0 and 1. The participants were then divided into five subgroups denoted with letters A through E depending on the participant's random number: 0.00 to 0.19 (Subset A, $n = 79$ [20%]), 0.20 to 0.39 (B, $n = 93$ [23%]), 0.40 to 0.59 (C, $n = 73$ [18%]), 0.60 to 0.79 (D, $n = 83$ [21%]), and 0.80 to 1.00 (E, $n = 76$ [19%]). Combinations of letters presented from hereon refer to subgroups constituted by such combined subsets (see Table 2).

Seeds. The subsets were combined in a predefined manner to allow both for construction and validation, with each unique combination referred to as a seed (far left column of Table 2). The seed designation defines which subsets were combined and used for derivation of statistical weights and which were used for cross-validation of these weights, respectively.

Weighting conditions. Five different procedures commonly used for data analysis in various prediction tasks were applied. First, a nonweighted model was included for reference purposes, and the nonweighted H-10 represents the H-10 score coded according to the HCR20 manual (Webster et al., 1997). The four weighting procedures representing increasing complexity (with level of complexity within brackets) were as follows: one bivariate and rough (simple; the Nuffield procedure), one bivariate but "fine grained" (intermediate; the bivariate logistic regression), one multivariate (complex; the multivariate logistic regression), and one multivariate nonlinear (highly complex; the artificial neural network).

The Nuffield approach was included because it had previously been employed for the derivation of item weights in the VRAG (Quinsey et al., 1998). As described by Nuffield (1982, as cited in Quinsey et al., 1998; Webster et al., 1994), this procedure allots one unit weight for each full 5% deviation from the base rate of violent recidivism. For example, if the recidivism base rate were 24% in the entire derivation sample and individuals with PCL-R psychopathy (Item H7 of the H-10) partially present (coded 1) exhibit a base rate of 35%, the weight for a 1 on H7 would be set to 2 (because 35% is more than 10 percentage units but less than 15 percentage units above 24%). The same procedure is used if the base rate is lower among those with the risk factor present, except that the assigned weight would be negative. Bivariate logistic regression analyses were used to derive crude regression coefficients for each risk factor. The 10 factors of the H-10 were analyzed one by one with recidivism as the dependent variable. An absent risk factor (coded 0) was used as reference category. Weights were assigned to each individual H-10 item according to the corresponding regression coefficient (Beta weight). In the multivariate logistic regression condition, all the 10 risk factors of the H-10 were entered together as covariates into one single logistic regression model with violent recidivism as dependent variable. In each seed, the predicted probability of violent recidivism for each individual was derived from the logistic formula [$p = 1 / (1 - e^{z})$]. These predicted probabilities were used as predictor measures on validation. An artificial neural network is a type of artificial intelligence software designed to mimic the problem-solving process of the human brain. Neural networks aid decision making and solve classification problems by means of pattern recognition. In many respects, artificial neural network modeling differs from traditional statistics. It is a technique known to perform well also with sparse information. Neural networks may outperform other statistical approaches with data that suffer from many missing values, large measurement error, when the causal mechanisms between predictors and outcome are unknown, and for otherwise complex pattern recognition problems. It is beyond the scope of this article to fully describe the fundamentals of neural network modeling, but several

TABLE 2: Predictive Accuracy Estimates for Violent Recidivism for H-10 Scores Obtained With Five Different Weighting Conditions Across Five Predefined Combinations of Subjects Among 404 Mentally Disordered Violent Offenders

	Area Under the Curve of the Receiver Operating Characteristic (95% Confidence Interval)				
Seed (a)	*Nonweighted H-10*	*Nuffield Approach*	*Bivariate Logistic Regression Model*	*Multivariate Logistic Regression Model*	*Artificial Neural Network*
AB(C) / DE	.71 (.63 to .78)	.73 (.66 to .80)	.72 (.65 to .79)	.70 (.63 to .77)	.72 (.65 to .79)
BC(D) / EA	.78 (.71 to .84)	.79 (.71 to .85)	.77 (.70 to .83)	.70 (.62 to .77)	.60 (.52 to .68)
CD(E) / AB	.77 (.70 to .83)	.75 (.68 to .81)	.75 (.68 to .82)	.73 (.66 to .80)	.73 (.66 to .79)
DE(A) / BC	.67 (.60 to .74)	.64 (.57 to .71)	.66 (.58 to .73)	.62 (.54 to .69)	.51 (.43 to .59)
EA(B) / CD	.67 (.59 to .74)	.66 (.58 to .73)	.66 (.58 to .73)	.63 (.55 to .70)	.62 (.54 to .70)
M	.72 (.65 to .79)	.71 (.64 to .78)	.71 (.64 to .78)	.68 (.60 to .73)	.64 (.56 to .71)

Note. (a) Letters A to E denote five subsets of the total sample of 404 mentally disordered offenders released from prison or discharged from forensic psychiatric treatment in Sweden. Each participant was randomly allocated to one of the subsets (A, $n = 79$; B, $n = 93$; C, $n = 73$; D, $n = 83$; E, $n = 76$). Subsets were combined to form subpopulations (e.g., DE [$n = 159$] or EA [$n = 155$]). These subpopulations were in turn used to derive weights (left side of slash) and for the validation of weighting algorithms (right side of slash), respectively (letter within brackets refers to the subset used as a test set specifically for the neural network condition).

comprehensive overviews and introductory texts are available (e.g., Caulkins, Cohen, Gorr, & Wei, 1996; Guerriere & Detsky, 1991; Kartalopoulos, 1995; Lawrence, 1993; Mulsant, 1990; Ripley, 1996).

A few previous studies have used artificial neural networks for the modeling of criminal recidivism risk. Palocsay, Wang, and Brookshire (2000) reported on an investigation of more than 10,000 offenders released from prisons in North Carolina in the United States from 1978 to 1980. The authors used a set of nine predictor variables to compare the performance of three-layer back-propagation neural networks with multivariate logistic regressions. The overall rate of correct classification varied between 60% and 69% for both methods, with the neural network models slightly but consequently outperforming logistic regression modeling. The authors concluded that although neural network performance depends heavily on network topology, they might be superior to multivariate logistic regression for prediction. In addition, Palocsay et al. (2000) underlined the need for a continued search for predictive variables and the development of new models. Caulkins et al. (1996) used three-layer back-propagation neural networks to model risk for recidivism with 3,400 offenders released from prison in the United States from 1970 to 1972. The predictor variables reflected previous and current criminality as well as social and intrainstitutional adjustment. However, the authors did not find neural networks to perform better than multivariate linear regression or nonweighted predictor summary scores.[2] To conclude, findings reporting on the effectiveness of artificial neural networks to model recidivism risk relative to logistic regression procedures are inconsistent, and future studies are clearly warranted.

We chose a three-layer back-propagation neural network architecture, the most commonly used network architecture for the modeling of recidivism risk with the 10 historical variables included in the HCR-20. Figure 1 schematically outlines this architecture. Each "neuron" of the input layer is connected to each of the neurons in the second layer, which are all in turn connected to the output neuron. The links between the neurons are referred to as "axons," and the connection strengths are the weights of the network. The 10 variables of the H-10 were represented by 10 input neurons, respectively, all linearly scaled from –1 to 1. These neurons made up the first neural network layer. One output neuron was assigned a logistic activation function to represent the criterion variable (a violent reconviction within 2 years after release or discharge). This function constituted the third layer of the neural network. For the second (or hidden) layer of neurons, 84 neurons, also with logistic activation functions, were assigned.[3] In artificial neural network modeling, the process of obtaining an algorithm is called training. Training involves the repeated presentation of each of the cases in the same training set to the network. Weights associated with the network's interneural connections are iteratively adjusted during the process to closer reflect the characteristics of the training data set. The goal with training is to find a set of weights that minimizes the total error based on observed neural network output values and the desired values with the training data set. Thus, the cases in the training subset are randomly[4] "fired" one by one through the net, and the error is back-propagated to update the weights of the axons connecting the input and hidden layers, and the hidden layer and the output, respectively. During training, the network is exposed to the entire training set of participant data a large number of times. How many is a matter of trial and error (Kartalopoulos, 1995). The neuronal or axonal weights (which are set to vary from –1 to 1) are set to an arbitrary value (.30) when the training begins. As the training proceeds, axonal weights are updated so that

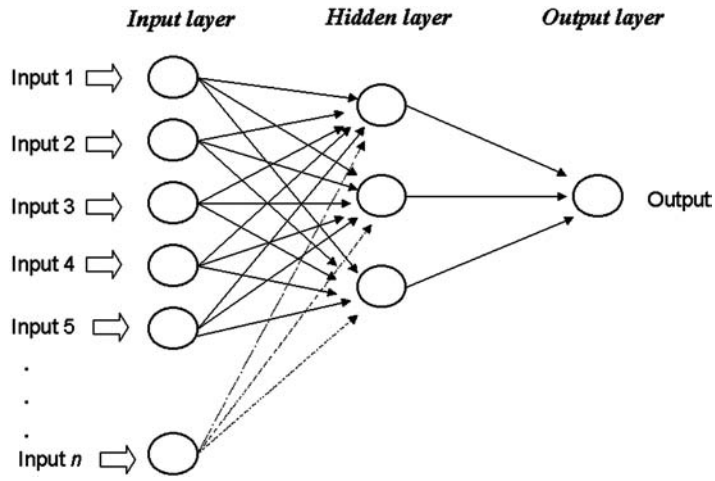

Figure 1: Architecture of a Three-Layer Back-Propagation Artificial Neural Network

the weighting algorithm approaches a tighter and tighter fit around data in the training set. An insufficiently trained network will not have its weights set optimally to produce a good fit with data. However, if training continues for too long, the network fit with training data will be too tight, and it will generalize poorly when exposed to cases that it has not encountered before.

This major problem in network modeling is called overlearning and may be described as if the network has completely memorized all the cases in the training set. In other words, in the population where the network has been trained, the training set, the predictive performance is likely to be excellent. However, when confronted with cases representing an unknown population, the verification set, the artificial neural network will generally perform poorly. In this study, one third of the derivation subpopulation was used as a test set. This means the network was trained with the training cases and the test set cases used for periodical checks)[5] of network performance. Training was interrupted as soon as the network no longer seemed to improve in the test set. In each seed, the subpopulation resulting from the combination of three subsets (e.g., ABC in the first seed) was used to derive the weighted prediction algorithm across all five weighting conditions. The algorithm was then applied to the subpopulation made up by the two remaining subsets (e.g., DE in the first seed) to cross-validate its predictive accuracy for each weighting condition.

STATISTICS

Random assignment of individuals to subsets, calculations of H-10 raw scores, and Nuffield weights were performed using a Microsoft Excel 97 spreadsheet. Regression coefficients were calculated with Statistical Package for the Social Sciences (1998) Version 8.5. For artificial neural network analysis, we used Neuroshell 2 software (Ward Systems Group, 1996). Finally, ROC analyses, AUC estimates, and corresponding 95% confidence intervals (95% CIs) were computed with MedCalc for Windows 95 (Schoonjans, 1998).

RESULTS

The base rate of recidivism among the 404 mentally disordered offenders was 23% (n = 91). In the different subsets, recidivism base rates 19 out of 79 individuals, 24% (subset A); 18 of 96, 19% (B); 18 of 73, 25% (C); 20 of 83, 24% (D); and 16 of 76, 21% (E). The nonweighted H-10 summary score in the total sample ranged from 4 to 19 points, with an arithmetic mean of 11.78 (SD = 3.60, mdn = 12.00). Means and standard deviations for individual risk factor scores are presented in Table 1. Bivariate risk factor correlations (Pearson's r) to violent recidivism ranged from r = –.03 ($p > .05$) for Relationship Instability to r = .32 ($p < .01$) for Psychopathy.

Nonweighted H-10 total scores predicted recidivism with an average area under the ROC curve of .72 (95% CI = .65 to .79). In the different subsets of offenders, the AUCs ranged from .67 (95% CI = .59 to .74) in subpopulation CD to .78 (95% CI = .71 to .84) in subpopulation EA (see Table 2). When the H-10 items were weighted according to the Nuffield procedure, a small shrinkage effect was observed. Similarly, the application of weights derived from bivariate logistic regression coefficients to the H-10 items resulted in a small shrinkage effect. The AUCs of models weighted with bivariate logistic regression were comparable to those of the Nuffield condition. When individual beta values for each H-10 item derived from multivariate logistic regression were used as weights, a clear shrinkage effect was observed. For the neural network models, finally, a dramatic shrinkage effect was seen.

DISCUSSION

This study reports an attempt to explore experimentally the potential benefits of using weighting algorithms for actuarial assessment of risk for violence. The findings suggest that applying weights does not improve predictions but rather results in statistical shrinkage effects. Furthermore, the more sophisticated the weighting algorithm, the greater the shrinkage effect tended to become. Another finding was that the AUC estimate varied markedly between the different combinations of randomly drawn subsets of the population. For example, for the nonweighted H-10, the AUC ranged from .67 to .78 across seeds. This may illustrate empirically the extent to which stochastic variations could affect AUCs and hence the relative (un)importance of a .10 difference in AUCs when comparing estimates across assessment methods and samples.

When reviewing the performance of the different weighting paradigms, aspects other than the predictive validity should also be considered. One such aspect is the transparency of the model, that is, whether it is possible to go back and reconstruct what went wrong in cases where the model failed. This appears particularly important for models for which usefulness in the context of actual clinical and legal decision making has been claimed, such as the VRAG (Quinsey et al., 1998). With a nonweighted model, the level of transparency depends on how clearly defined the included risk factors are and the stringency with which these factors have been operationalized into actuarial items. In practice, this part is straightforward, and even laypersons are expected to be able to follow rating guidelines and understand in principle how individual risk factors have been assessed and coded. Likewise, with a weighting procedure in which crude (bivariate) statistical relationships are

used to derive the weights, such as the Nuffield or unadjusted odds ratio approaches used in this study, it is relatively easy to pursue backward the trail from risk score to actual observations in any given case. With a traditional multivariate statistical method, such as the logistic regression, it is still possible to reconstruct how a predicted probability estimate was calculated, although the mathematics behind these algorithms is complex. Indeed, to get a clear picture of colinearity in the data set, one needs to explore the relationship between predictors by entering them pairwise into regression models and to introduce interaction terms into the models as well (Kleinbaum, 1994).

With neural network models, however, it is very difficult to reconstruct the relationship between raw data and model output. The operative structure of the trained neural network is the combination of weight values, that is, the axonal connection strengths. With a backpropagation architecture, it is possible to inspect the so-called contribution factors for each of the input variables and thereby get a picture of the relative importance of each factor. However, contribution factors can only be used to compare input variables within the same neural network and do not generalize even across the seeds of a study (Tam & Kiang, 1992).

TO WEIGH OR NOT TO WEIGH?

Several arguments exist against the use of weights in statistical algorithms for the assessment of risk for violence. As noted above, evidence in support of complex procedures such as artificial neural networks has been inconsistent. Cohen and Cohen (1983), while experimenting with a data set on college faculty members' salaries and four independent variables (gender, years since PhD, number of publications, and number of citations, respectively), illustrated how little, if any, informational value beta weights from regression have over simple, unit weights. Only with very large *n*s for calibration and cross-validation did the goodness of fit improve, and then only trivially. In the present study, prediction model goodness of fit as expressed with area under ROC curve did not improve with logistic regression; instead, a shrinkage effect was seen.

Obviously, the results of this study were strongly determined by the predictor variables used, as risk factors are the raw material from which the prediction algorithm is built.[6] Even if the weighting techniques applied in this study showed poor results, this need not necessarily be the case should predictor variables be improved, exchanged, or complemented with other variables. Future research is essentially bound to find new factors or refine those already identified. With new predictor variables of improved heuristic value, the assigning of weights reflecting the actual strength of the causal relationship between predictor and outcome may become useful. In agreement with Caulkins et al. (1996), we believe that theory building to delineate behavioral mechanisms and contextual influences involved in criminal recidivism should be prioritized over development of complex statistical prediction models. Meanwhile, at least judging from the results of the present study, nonweighted approaches seem to outperform weighted ones for the actuarial assessment of risk for future violence in mentally disordered offenders.

RATIONALE FOR RISK FACTOR IDENTIFICATION

From our reading, three general approaches for the identification of risk factors can be traced in the literature: empirical, theoretical, and clinical. From an extreme empiricist

viewpoint, the only interesting aspect of a predictor or risk factor is the technical reliability and validity. In other words, with what precision can the predictor be measured reliably, and what is the strength of the statistical relationship between the risk factor and recidivism in follow-up studies? In medicine, this overemphasis on "predictionism" (Grann, 1998, p. 52) has been called "black-box epidemiology" (Susser & Susser, 1996). In the context of risk assessment for forensic mental health purposes, Hart (2002, p. 126) referred to the same phenomenon as the "passive prediction paradigm." He argued against the dead-end position reached when reducing the true task clinicians face with mentally disordered offenders, that is preventing violence and not simply predicting it, into a statistical exercise. Hart advised that only the integration of assessment of risk with the management of risk could truly bridge the gap between research and practice in the field (see also Dernevik, Johansson, & Grann, 2002; Grann et al., in press; Gunn, 1996; Hart, 1998; Heilbrun, 1997; Rogers, 2000; Webster et al., 1997).

Second, from the downright clinical perspective, the sole interest in a predictor variable is whether its status could be changed (treated) and, if so, to what extent the required intervention is ethically sound, practically possible to administer, and reasonably inexpensive. In some instances, empirical data may be of little relevance from a clinical point of view. For example, structural factors such as poverty or segregation have well-established theoretical value and are consistently associated with criminal behavior (e.g., Loeber & Farrington, 1998). There should be little doubt that true primary prevention through social policy making and social medicine would have great importance for combating violence as a public health problem. However, for the clinician expected to provide expert testimony on future risk in an actual, individual case, information about structural factors is of limited value. The clinician is bound to rely on factors that apply to the individual and the individual's environment here and now (see also Silver, 2000). Therefore, clinicians generally take the greatest interest in dynamic factors such as those identified by the Structured Outcome Assessment and Community Risk Monitoring checklist (Grann et al., in press; Sturidsson, Haggård-Grann, Lotterberg, Dernevik, & Grann, 2004) or the clinical and risk management factors (C and R subscales) of the HCR-20 (Webster et al., 1997).

Third, the theorist would judge the value of a predictor only by how well the causal relationship of that risk factor to future violence has been delineated and to what extent theory can account for its effects. The heuristic value of a risk factor is thus of higher priority than any statistical index of predictive validity. From a theoretical perspective, it is meaningless to consider a risk factor only because of a statistical relationship. Obviously, with a large enough sample size, any factor can be shown to be a significant predictor of any outcome (e.g., Cohen, 1994; Kraemer et al., 1997). An extreme theorist would insist on risk factor concepts even if empirical data were inconsistent with the hypothesis postulated by theory.

With regard to the weighting of risk factors in actuarial formulas for the assessment of recidivism risk, we argue that statistical weighting is an expression of empiricism. We believe weighting risk factors is a premature exercise as too little is still known about the proximal causes and true mechanisms behind violence. In addition, we agree with Kraemer et al. (1997) that the generic use of the term *risk factors* is confusing and that the roles of specific factors as "fixed markers," "variable markers," or "causal risk factors," respectively, needs further elucidation. With this terminology, H-10 factors such as "early adjustment problems" and "young age at first violent incident" may be defined as fixed markers for risk of violence, as fixed markers do not change within a participant. "Relationship instability" and

"employment problems" are variable markers—factors that can be demonstrated to change spontaneously within a participant or to be changed by intervention. A true, causal risk factor is one that "can be shown to be manipulable and, when manipulated, can be shown to change the risk of the outcome" (Kraemer et al., 1997, p. 340). This is indeed a very strict definition, and hardly any one of the items included in the HCR-20 or the VRAG have yet received the empirical support to qualify as such true, causal factors.

To conclude, we recommend against using weighted actuarial models in clinical practice. Risk assessment research in the forensic mental health field should continue to take input from theoretical and clinical perspectives and not be restricted to empirical data. Although it is a complex and demanding task, we believe the development of tools that succeed to optimally balance empirically demonstrated predictive validity with theoretical heuristic value, well-elucidated causal mechanisms, and clinical utility may improve risk assessment practice.

NOTES

1. In the latest version of the HCR-20, the authors have even changed the coding legends from 0, 1, and 2 to *no*, *partially*, and *yes*, respectively, to further "deactuarialize" their checklist.

2. Brodzinski, Crable, and Scherer (1994, as cited in Palocsay et al., 2000) reportedly obtained very promising results with neural network modeling of criminal recidivism in a sample of almost 800 juvenile probation participants. In addition, Grann (1998) reported on the application of a three-layer back propagation neural network for the prediction of recidivism in a population partly overlapping the one used in the present study. However, no comparisons with logistic regression were done in either of these two studies.

3. A common rule of thumb in assigning the number of neurons in the hidden layer is to take half the sum of input and output neurons plus the square root of the number of cases (participants) used for training (Ward Systems Group, 1996). Previous pilot studies on parts of the same data used in this study (Grann, 1998) suggested that using a greater number of neurons in the hidden layer increased network performance somewhat. The number set (84 neurons) was approximately 4.5 times that suggested by the thumb rule.

4. During training of neural networks in this study, "random pattern selection" was employed. That is, the cases in the training set were fired through the net at random instead of being fired in a given sequence.

5. In this case, every 100th event.

6. "If noise goes in, noise comes out" is a commonly referred devise in applied prediction modeling.

REFERENCES

American Psychiatric Association. (1980). *Diagnostic and statistical manual of mental disorders* (3rd ed.). Washington, DC: Author.

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.

Bleeker, S. E., Moll, H. A., Steyerberg, E. W., Donders, A. R. T., Derksen-Lubsen, G., Grobbee, D. R., et al. (2003). External validation is necessary in prediction research: A clinical example. *Journal of Clinical Epidemiology*, *56*, 826-832.

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

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*, 227-240.

Cicchetti, D. V. (1992). Neural network and diagnosis in the clinical laboratory: State of the art. *Clinical Chemistry*, *38*, 9-10.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, *45*, 1304-1312.

Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, *49*, 997-1003.

Cohen, J., & Cohen, P (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed). Hillsdale, NJ: Lawrence Erlbaum.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision models. *American Psychologist*, *34*, 571-582.

Dernevik, M., Johansson, S., & Grann, M. (2002). Violent behaviour in forensic psychiatric patients: Risk assessment and different risk-management levels using the HCR-20. *Psychology, Crime & Law*, *8*, 93-112.

Douglas, K. S. (2004). *HCR-20 violence risk assessment scheme: Overview and annotated bibliography*. Vancouver, British Columbia, Canada: Simon Fraser University, Department of Psychology.

Douglas, K. S., Hart, S. D., Dempster, R. J., & Lyon, D. J. (1999, July). *The Violence Risk Appraisal Guide (VRAG): Attempt at validation in a maximum security forensic psychiatric sample*. Poster session presented at the joint meeting of European Association of Psychology & Law/American Psychology-Law Society, Dublin, Ireland.

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.

Ennis, B. J., & Litwack, T. R. (1974). Psychiatry and the presumption of expertise: Flipping coins in the courtroom. *California Law Review*, *62*, 693-752.

Fletcher, J. M., Rice, W. J., & Ray, R. M. (1978). Linear discriminant function analysis in neuropsychological research: Some uses and abuses. *Cortex*, *14*, 564-577.

Grann, M. (1998). *Personality disorder and violent criminality: A follow-up study with special reference to psychopathy and risk assessment* (doctoral dissertation). Stockholm, Sweden: Karolinska Institutet.

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

Grann, M., Långström, N., Tengström, A., & Stålenheim, E. G. (1998). The reliability of file-based retrospective ratings of psychopathy with the PCL-R. *Journal of Personality Assessment*, *70*, 416-426.

Grann, M., Sturidsson, K., Haggård-Grann, U., Hiscoke, U. L., Alm, P.- O., Dernevik, M., et al. (in press). Methodological development: Structured Outcome Assessment and Community Risk Monitoring (SORM). *International Journal of Forensic Psychiatry*.

Guerriere, M. R. J., & Detsky, A. S. (1991). Neural networks: What are they? *Annals of Internal Medicine*, *115*, 906-907.

Gunn, J. (1996). Let's get serious about dangerousness. *Criminal Behaviour and Mental Health*, *6*, 51-64.

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. (1997). *The development of a brief actuarial scale for sexual offense recidivism* (User Report No. 1997-04). Ottawa, Ontario, Canada: Department of the Solicitor General of Canada.

Hare, R. D. (1991). *The Hare PCL-R: Rating booklet*. Toronto, Ontario, Canada: Multi-Health Systems.

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*, *4*, 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. (1998). The role of psychopathy in assessing risk for violence: Conceptual and methodological issues. *Legal and Criminological Psychology*, *3*, 123-140.

Hart, S. D. (2002, September). *Communicating about violence risk*. Workshop at the Safer Society Conference, Glasgow, Scotland.

Hart, S. D., Webster, C. D., & Menzies, R. J. (1993). A note on portraying the accuracy of violence predictions. *Law and Human Behavior*, *17*, 695-700.

Heilbrun, K. (1997). Prediction versus management models relevant to risk assessment: The importance of legal decision-making context. *Law and Human Behavior*, *21*, 347-360.

Henderson, A. R. (1993). Assessing test accuracy and its clinical consequences: A primer for receiver operating characteristic curve analysis. *Annals of Clinical Biochemistry*, *30*, 521-539.

Kartalopoulos, S. V. (1995). *Understanding neural networks and fuzzy logics. Basic concepts and applications*. Piscataway, NJ: Institute of Electrical and Electronics Engineers Press.

Kleinbaum, D. G. (1994). *Logistic regression: A self-learning text*. New York: Springer.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied regression analysis and other multivariate methods*. Pacific Grove, CA: Duxbury.

Kraemer, H. C., Kazdin, A. E., Offord, D. R., Kessler, R. C., Jensen, P. S., & Kupfer, D. J. (1997). Coming to term with the terms of risk. *Archives of General Psychiatry*, *54*, 337-343.

Långström, N., Grann, M., Tengström, A., Lindholm, N., Woodhouse, A., & Kullgren, G. (1999). Data extraction for file-based forensic psychiatric research: Some methodological considerations. *Nordic Journal of Psychiatry*, *53*, 61-67.

Lawrence, J. (1993). *Introduction to neural networks*. Nevada City, CA: California Scientific Software.

Litwack, T. R. (2001). Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and Law*, *7*, 409-443.

Loeber, R., & Farrington, D. P. (Eds.). (1998). *Serious and violent juvenile offenders: Risk factors and successful intervention*. Thousands Oaks, CA: Sage.

Monahan, J. (1984). The prediction of violent behavior: Toward a second generation of theory and policy. *American Journal of Psychiatry*, *141*, 10-15.

Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, *62*, 783-792.

Mulsant, B. H. (1990). A neural network approach to clinical diagnosis. *M.D. Computing*, *7*, 25-36.

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

Quinsey, V. L., Rice, M. E., Harris, G. T., & Cormier, C. A. (1998). *Violent offenders: Appraising and managing risk*. Washington, DC: American Psychological Association.

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

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

Rogers, R. (2000). The uncritical acceptance of risk assessment in forensic practice. *Law and Human Behavior*, *5*, 595-605.

Schoonjans, F. (1998). *MedCalc. Statistics for biomedical research* (Version 5) [Computer software]. Mariakerke, Belgium: MedCalc Software.

Silver, E. (2000). Race, neighborhood disadvantage, and violence among persons with mental disorder: The importance of contextual measurement. *Law and Human Behavior*, *24*, 449-456.

Sjöstedt, G., & Grann, M. (2002). Risk assessment: What is being predicted by actuarial "prediction instruments"? *International Journal of Forensic Mental Health*, *1*, 179-183.

Statistical Package for the Social Sciences. (1998). *Statistical Package for the Social Sciences* (Version 8.5) [Computer software]. Chicago: Author.

Strand, S., Belfrage, H., Fransson, G., & Levander, S. (1999). Clinical and risk management factors in risk prediction of mentally disordered offenders: More important that actuarial data? *Legal and Criminological Psychology*, *4*, 67-76.

Sturidsson, K., Haggård-Grann, U., Lotterberg, M., Dernevik, M., & Grann, M. (2004). Clinicians' perceptions of which factors increase or decrease risk of violence among forensic outpatients. *International Journal of Forensic Mental Health*, *3*, 23-36.

Susser, M., & Susser, E. (1996). Choosing a future epidemiology: II. From black box to Chinese boxes and eco-epidemiology. *American Journal of Public Health*, *86*, 674-677.

Tam, K. Y., & Kiang, M. Y. (1992). Managerial applications of neural networks: The case of bank failure predictions. *Management Science*, *20*, 879-888.

Tengström, A. (2001). Long-term predictive validity of historical factors in two risk assessment instruments in a group of violent offenders with schizophrenia. *Nordic Journal of Psychiatry*, *55*, 243-249.

Ward Systems Group. (1996). *Neuroshell 2* [Computer software]. Frederick, MD: Author.

Webster, C., Douglas, K., Eaves, D., & Hart, S. (1997). *HCR-20. Assessing risk for violence* (Version 2). Vancouver, British Columbia, Canada: Simon Fraser University and Forensic Psychiatric Services Commission of British Columbia.

Webster, C. D., Harris, G. T., Rice, M. E., Cormier, C., & Quinsey, V. L. (1994). *The violence prediction scheme. Assessing dangerousness in high risk men*. Toronto, Ontario, Canada: University of Toronto, Center of Criminology.

Whittemore, K. E. (1999). *Releasing the mentally disordered offender: Disposition decisions for individuals found unfit to stand trial and not criminally responsible*. Unpublished doctoral dissertation, Simon Fraser University, Vancouver, Canada.

World Health Organization. (1976). *International statistical classification of diseases and related health problems* (9th rev.). Geneva, Switzerland: Author.