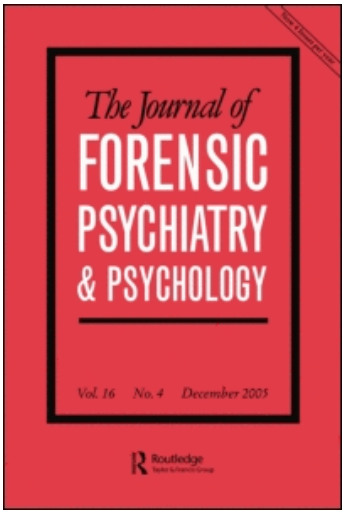

Journal of Forensic Psychiatry & Psychology
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t714592861

Most items in structured risk assessment instruments do not predict violence

Jeremy W. Coid[a]; Min Yang[b]; Simone Ullrich[a]; Tianqiang Zhang[a]; Steve Sizmur[c]; David Farrington[d]; Robert Rogers[e]
[a] Forensic Psychiatry Research Unit, Queen Mary University of London, London, UK [b] School of Community Health Sciences, University of Nottingham, Nottingham, UK [c] Picker Institute, Oxford, UK [d] Institute of Criminology, University of Cambridge, Cambridge, UK [e] Department of Psychiatry, Warneford Hospital, University of Oxford, Oxford, UK

First published on: 18 January 2011

To cite this Article Coid, Jeremy W. , Yang, Min , Ullrich, Simone , Zhang, Tianqiang , Sizmur, Steve , Farrington, David and Rogers, Robert(2011) 'Most items in structured risk assessment instruments do not predict violence', Journal of Forensic Psychiatry & Psychology, 22: 1, 3 — 21, First published on: 18 January 2011 (iFirst)
To link to this Article: DOI: 10.1080/14789949.2010.495990
URL: http://dx.doi.org/10.1080/14789949.2010.495990

The Journal of Forensic Psychiatry & Psychology
Vol. 22, No. 1, February 2011, 3–21

Most items in structured risk assessment instruments do not predict violence

Jeremy W. Coid[a]*, Min Yang[b], Simone Ullrich[a], Tianqiang Zhang[a], Steve Sizmur[c], David Farrington[d] and Robert Rogers[e]

[a]*Forensic Psychiatry Research Unit, Queen Mary University of London, London, UK;* [b]*School of Community Health Sciences, University of Nottingham, Nottingham, UK;* [c]*Picker Institute, Oxford, UK;* [d]*Institute of Criminology, University of Cambridge, Cambridge, UK;* [e]*Department of Psychiatry, Warneford Hospital, University of Oxford, Oxford, UK*

(*Received 1 April 2009; final version received 1 April 2010*)

The predictive ability of static risk assessment instruments may be explained by a limited number of their items. This study examined the independent predictive accuracy of individual items in the Psychopathy Checklist-Revised (PCL-R), Violence Risk Appraisal Guide (VRAG) and Historical-Clinical-Risk Management-20 (HCR-20) for violent reconvictions following release among 1353 male prisoners in England and Wales. The study found most items in the three instruments were not independently predictive. Items not independently predictive were removed and all significant items in the original three instruments were combined, resulting in negligible gains in predictive accuracy for the VRAG and HCR-20, but a small improvement in the PCL-R. The study demonstrated that the predictive power of the PCL-R, VRAG and HCR-20 are based on a small number of their items. This may partly explain the 'glass-ceiling' effect beyond which further improvement cannot be achieved. Instruments lack outcome-specificity for violence, and independently predictive items include measures of general criminality.

Keywords: risk assessment; predictive items; violence

Introduction

It is accepted that structured risk assessment instruments outperform clinical judgement in the accuracy of prediction of violent and sexual behaviour (Dawes, Faust, & Meehl, 1989; Grove & Meehl, 1996). However, most achieve an accuracy around a moderate area under curve (AUC) value of 0.7 in comparison studies between more than one instrument (Belfrage, Fransson, & Strand, 2000; Coid et al., 2009; de Vogel, de Ruiter, van Beek,

*Corresponding author. Email: j.w.coid@qmul.ac.uk

ISSN 1478-9949 print/ISSN 1478-9957 online

DOI: 10.1080/14789949.2010.495990
http://www.informaworld.com

& Mead, 2004; Douglas, Ogloff, Nicholls, & Grant, 1999; Douglas, Yeomans, & Boer, 2005; Doyle & Dolan, 2006; Doyle, Dolan, & McGovern, 2002; Glover, Nicholson, Hemmati, Bernfeld, & Quinsey, 2002; Grann, Belfrage, & Tengström, 2000; Gray et al., 2003, 2004; Grevatt, Thomas-Peter, & Hughes, 2004; Kroner & Loza, 2001; Kroner & Mills, 2001; Mills, Kroner, & Hemmati, 2007; Morrissey et al., 2007; Nicholls, Ogloff, & Douglas, 2004; Snowden, Gray, Taylor, & MacCulloch, 2007; Stadtland et al., 2005; Tengström, 2001; Warren et al., 2005), which suggests a 'glass-ceiling' effect beyond which few instruments can improve. Furthermore, instruments tend to show 'shrinkage' of their results when subsequently applied to populations on which they were not originally standardised. Part of the reason for these observations may be that the developmental origins, and in certain cases the original purposes, of these instruments are very different. For example, the Psychopathy Checklist-Revised (PCL-R, Hare, 2003) was originally developed as a diagnostic instrument to measure the degree to which an individual offender matches a prototypical psychopathic personality. Subsequent research indicated that it can predict both violent and general recidivism (Hemphill, Hare, & Wong, 1998; Salekin, Rogers, & Sewell, 1996; Serin 1996). The Violence Risk Appraisal Guide (VRAG), an actuarial instrument, was specifically developed to predict violence among patients discharged from a maximum security hospital. It can also predict general recidivism (Glover et al., 2002). In contrast, the Historical-Clinical-Risk Management-20 (HCR-20) includes a structured professional judgement approach to risk assessment for violence, effectively an attempt to bridge the gap between clinical and actuarial approaches by combining both into professional guidelines for clinical practice (Webster, Muller-Isberner, & Frannson, 2002). The instrument also predicts general offending (Coid et al., 2009), adding further evidence that structured risk assessment instruments lack instrument-outcome specificity. It has been argued that they may be limited to measuring a general construct of criminal risk rather than specific tendencies to violence as originally intended (Gendreau, Goggin & Smith, 2002; Glover et al., 2002; Kroner & Mills, 2001).

Inherent in the study of the accuracy of risk assessment instruments is the investigation of their aggregate scores. Few studies provide information on the predictive ability of individual items. However, it is possible that certain items within these instruments do not have predictive ability, and that bivariate correlations with violence as an outcome are merely the result of a strong correlation with other items, which are truly predictive. This may be especially the case for instruments designed for other purposes. It is also possible that certain items could even reduce the overall predictive efficiency of an instrument. This might explain first, the limitation in accuracy of these instruments, and second, why their predictive ability can generalise to other outcomes.

The aim of the study was to examine the independent predictive ability of individual items within three structured risk assessment instruments for violent and general criminal recidivism. To do this, we investigated items from a structured risk assessment guide (HCR-20), a personality assessment (PCL-R) and an actuarial instrument (VRAG). Our second aim was to examine whether it was possible to improve their predictive accuracy by including only items that were independently predictive and comparing their aggregate score with aggregate scores measured using the original three instruments.

Method

We carried out a prospective study of a cohort of 1353 male prisoners released from prisons in England and Wales between 14 November 2002 and 7 October 2005. Participants were interviewed during the 6–12-month period before their expected date of release by trained interviewers using a battery of clinical and risk assessment measures for violent and other criminal behaviour. The dependent variable was the proportion of participants who were or were not reconvicted within different categories of offending behaviour (including violence) derived from criminal records. These were measured following their release into the community over a mean follow-up of 1.97 years (range 6 days to 2 years 11 months) with over 94.5% of sample followed up over 1 year.

Sample

The sample was generated from the Prison Service Inmate Information System if they met the following criteria: (1) serving a prison sentence of 2 years or more for a sexual or violent principal offence (excluding life sentence prisoners), (2) aged 18 years and over, (3) having 1 year left to serve. Information was provided on previous criminal history using the Home Office Offenders Index on all prisoners in England and Wales meeting these criteria. On the basis of their current and previous convictions, a stratified sample was identified with over-selection of prisoners from ethnic minority groups, prisoners from younger age groups and potentially high-risk offenders selected using the highest scoring 10% on the Offenders Group Reconviction Scale (OGRS: Copas & Marshall, 1998; Taylor, 1999).

A total of 1396 prisoners were interviewed by 12 research assistants, psychology graduates, who visited and carried out interviews with the participants, usually spending a day with each participant, initially reading and extracting data from prison files and carrying out the interview, which lasted for 3–4 hours. The interview initially established the criminal history and nature of the index offence. The Structured Clinical Interview for DSM-IV Axis II disorder was administered to establish diagnoses of personality

disorder, and modules from the Diagnostic Interview Schedule were administered to measure the presence of current or lifetime schizophrenia or delusional disorder, depressive disorder, drug and alcohol use disorders or dependence, followed by the risk assessment instruments. Forty-three participants who were not released from prison during the follow-up period or who could not be identified using criminal records were excluded. Outcome data were derived from reconvictions recorded up to the date 13 October 2005 in the Police National Computer (PNC), an operational police database containing criminal histories of all offenders in England, Wales and Scotland. This source had a lower failure rate than the Home Office Offenders Index for non-identification and is updated more regularly (Howard & Kershaw, 2000).

Measures

A semi-structured interview was developed to collect all relevant data using the battery of risk instruments, together with diagnostic data. Rating the risk assessment instruments required exploration with the participant of their criminal history, which was made available to the researcher before the interview. Researchers were trained in the administration and scoring of all risk assessment instruments.

The Psychopathy Checklist Revised (PCL-R, Hare, 2003) was used to assess psychopathy and administered as part of a comprehensive clinical assessment of personality. It consists of 20 items that are scored 0, 1 or 2 based on a clinical interview and review of file information. Item scores are summed to create a total score, ranging from 0 to 40 and reflect an estimate of the degree to which an individual matches the prototypical psychopath at a cut-off of 30. Although not originally developed as a risk assessment instrument, two meta-analyses have demonstrated that the PCL-R is a strong predictor of violent recidivism (Hemphill et al., 1998; Salekin et al., 1996). This has resulted in psychopathy, as measured by the PCL-R, being included as a risk factor within other risk assessment instruments such as the HCR-20 and VRAG (see below). Inter-rater reliability for researchers using the PCL-R involved scoring of six reliability cases over a 2-day assessment. The intraclass correlation was 0.85.

The VRAG (Quinsey, Harris, Rice, & Cormier, 1998) is a 12-item actuarial instrument developed from the files of male criminal offenders and forensic patients with attributed integer weights, with scores ranging from – 26 to + 38. The instrument was designed for use with forensic populations, and items require rating of the index offence, psychopathy, alcohol use, and past non-violent crime.

The HCR-20 (Webster, Douglas, Eaves, & Hart, 1997) is a structured risk assessment guide and a composite of 20 risk factors for future violence in adult offenders with a violent history and/or a major mental disorder or

personality disorder. The instrument is divided into three sub-scales with 10 historical factors relating to past, relatively stable violence risk factors; five clinical items reflecting current, dynamic correlates of violence, which are thought to be changeable; five risk management items focusing on situational factors that might aggravate or mitigate risk. In this study, the clinical and risk management items were rated at an interview prior to the release on the basis of clinical presentation and anticipated situational factors. The HCR-20 measures included total and sub-scale scores in subsequent analyses. Inter-rater reliability tests were carried out, with the researchers achieving intraclass correlations of 0.98 for total, 0.98 for historical, 0.80 for clinical and 0.87 for risk scores.

Ethical Committee approval was obtained from South East Multi-Centre Research Ethics Committee, Kent & Medway Strategic Health Authority. Participants gave written informed consent for the interview and searching of their criminal records.

Statistical analysis

The analysis was carried out in three steps, including descriptive analysis, assessing differences between the full scales of the instruments and their sub-scales based on independently predictive items, and examining the predictive effect of the pooled predictive items from the three instruments.

The first step included calculating descriptive statistics of the risk assessment instruments and their overall observed accuracy in predicting violence reoffending. The conventional AUC for the receiver operating characteristic (ROC) measure was calculated. In addition, a more direct statistic for the association between the continuous score of the instrument and the binary outcome of reoffending, the biserial correlation coefficient, was presented. A higher AUC should be accompanied by a higher biserial correlation. Furthermore, the predictive effects of individual items for each instrument were examined by means of simple Kendall's τ-c correlation coefficient for two ordinal scales. The dichotomous reoffending outcome was treated as a special form of ordinal scale. This analysis provided simple statistics to compare the magnitude of predictive effects among the instruments and to separate individual items into two groups. Group 1 contained those items that did not contribute to the prediction of violent reoffending singly, and Group 2 contained those items that were significantly associated with the outcome singly in this sample.

The next step examined independent predictive effects of individual items in Group 1 and Group 2, respectively, but adjusted for each of the other items within each group. Logistic regression analysis was carried out. The z-score, defined as the regression coefficient divided by its standard error, was used as an effect size measure. The main advantages of z-scores in this context are two-fold. First, statistical significance can be assessed directly,

based on its face value. For example, a *z*-score of 1.96 indicates a *p* value equal to 0.05 for two-tailed test. Second, the magnitude of effect for each item can be compared directly because the *z*-score is standardised for all items within each of the three instruments in the study. For each instrument, all items in Group 1 were entered in the same regression model to examine the possibility that non-significant items by their own might have certain predictive power when they were pooled with others. The aggregated effect on the summary score of all items in Group 1 was estimated for each instrument. Meanwhile, the AUC measure was also presented for each item individually and all items combined.

For items in Group 2, the analysis took into account dependence between them and used stepwise regression procedure to select the most effective items. *Z*-scores for their independent effects, based on the final model and AUC value, are presented. Both forward and backward stepwise procedures were used for each instrument to ascertain the validity of the most predictive items. Only those items selected by both procedures were considered valid to be included in a new sub-scale for each instrument.

To take into account interactive effects between individual items, we also ran stepwise regression analysis for all items for each scale. If the two-group regression and the all-item regression yielded different results, we examined our results carefully, but favoured the latter.

Step 3 validated each sub-scale by comparing its predictive effects with those of the full scale, based on the aggregated summary scores. In this analysis, we calculated the standardised *z*-score of each scale when considering the different score ranges of the full scale and the sub-scale. This ensured a direct comparison of *z*-scores. The overall AUC values were also compared. Furthermore, the direct discriminant regression analysis for the dichotomous outcome was carried out for the full scale, including all items of the instrument and for the new sub-scale including only significant items. This analysis provided the percentages of correctly classified violent and non-violent reoffenders. Combining percentages of the two categories yielded the overall classification accuracy. Cross-validation for all cases was performed by leaving one subject out from each analysis. Conventional χ^2 tests were used to compare the percentages of classification accuracy between the full- and sub-scales. An optimised sub-scale should show either no difference in its predictive effects or outperform the full scale. All analyses were carried out using SPSS v12.0.

Results

The released sample consisted of 1353 participants with a mean age of 30.7 years (SD = 11.4 years, range = 17–75 years). One thousand and sixty five (78.7%) were of white British origin, 204 (15.1%) of black Caribbean or black African origin, 41 (3.0%) of Asian origin and 43 (3.2%) of other

ethnic origins. The mean length of sentence completed prior to release was 2.4 years (SD = 2.1 years, range = 0–18). Most participants had DSM-IV personality disorder (1004, 74.2%), 106 (7.8%) had a lifetime history of schizophrenia/schizophreniform disorder, 37 (2.7%) delusional disorder, 411 (30.4%) depressive disorder, 524 (38.7%) drug dependence and 276 (20.4%) alcohol use disorder (Coid et al., 2009).

At a mean follow-up of 1.97 years, 607 (44.9%) of the total released sample had been convicted of a further offence, including 178 (13.2%) violent offences.

Table 1 demonstrates that the aggregated PCL-R, VRAG, HCR-20 total and H, C and R sub-scales each significantly predicted violent recidivism, although none exceeded a moderate level of predictive ability according to their AUC values. The VRAG had the highest level of predictive ability according to AUC values, but because the 95% CIs of the AUCs for the instruments overlapped, there were no statistical differences between them in their accuracy in predicting violent reoffending. The AUCs of the PCL-R and R-5 overlapped, but at a lower level than the others, suggesting that they were generally less predictive than the VRAG and H-10 scales.

Table 2 demonstrates that seven items in the PCL-R, three in the VRAG and three in the HCR-20 were not significantly correlated with the outcome of violent reconviction using simple correlations (see Appendix).

Table 3 lists the independently predictive items selected from both forward and backward stepwise regression analysis for each instrument (see Appendix). For the PCL-R, the two-group regression selected four items, including stimulation/boredom, poor behavioural controls, juvenile delinquency and criminal versatility. The all-item regression selected four somewhat different items from the original 20 items, including stimulation/boredom, conning/manipulative, poor behavioural controls and criminal versatility. However, the item conning/manipulative had significantly negative predictive ability. We therefore used the latter items after

Table 1. Score distribution and correlation with violence reconviction.

					Violence reconviction	
	Valid N	Score range	Mean	SD	r^a	AUC (95% CI)[b]
PCL-R	1347	0–37	18.1	7.6	0.26***	0.63 (0.59–0.67)
VRAG	1351	−18 to 37	11.7	10.9	0.36***	0.70 (0.66–0.73)
HCR20-total	1271	0–39	19.1	7.8	0.31***	0.67 (0.63–0.71)
H-10	1281	0–20	11.1	4.6	0.31***	0.66 (0.63–0.70)
C-5	1339	0–10	3.3	2.2	0.26***	0.64 (0.60–0.68)
R-5	1337	0–10	4.5	2.6	0.18***	0.59 (0.54–0.63)

Note: [a]Biserial correlation coefficient of the instrument score and violence reconviction (reconvicted or not); [b]All AUC areas are significantly greater than 0.5 at $p < 0.001$; ***$p \leq 0.001$.

Table 2. Simple correlation between risk items and violence reconviction.

PCL-R		VRAG		HCR-20	
Item	τ^a	Item	τ^a	Item	τ^a
P1	0.02	V1	0.12***	H1	0.07***
P2	0.03	V2	0.10***	H2	0.12***
P3	0.14***	V3	0.08***	H3	0.05*
P4	0.03	V4	0.10***	H4	0.07***
P5	−0.02	V5	0.02	H5	0.09***
P6	0.06**	V6	0.08***	H6	0.02
P7	0.06**	V7	0.10***	H7	0.08***
P8	0.04*	V8	0.05**	H8	0.12***
P9	0.10***	V9	−0.003	H9	0.08***
P10	0.12***	V10	−0.02	H10	0.11***
P11	−0.02	V11	0.09***	C1	0.06**
P12	0.08***	V12	0.08***	C2	0.11***
P13	0.06***			C3	−0.003
P14	0.07***			C4	0.10***
P15	0.08***			C5	0.07***
P16	−0.003			R1	0.07***
P17	0.03			R2	0.07***
P18	0.13***			R3	0.02
P19	0.09***			R4	0.07***
P20	0.14***			R5	0.06***

Note: [a]Kendall's τ-c statistic for correlation between two ordinal scales from SPSS v12; *$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$.

excluding conning/manipulative for the next-stage analysis. For the VRAG and HCR-20, both the two-group stepwise regression and all-item stepwise regression yielded the same items with independently significant predictive effects. For the VRAG, only 5 of the original 12 items were independently predictive: Psychopathy, Younger age at Index Offence, Non-violent offence score, History of alcohol problems and Not female victim. For the HCR-20, only 8 of the original 20 items were independently predictive: Young age at first violence, Substance use problems, Early maladjustment, Prior supervision failure, Negative attitudes, Impulsivity, Exposure to destabilisers, and Non-compliance with remediation attempts.

We then examined how well a new sub-scale (consisting of an aggregate of the independently significant items) for each instrument performed when compared to the aggregate scores of all items within the original instruments in predicting violent re-offending. In the case of the PCL-R, item P5 was not included. Table 4 compares the effects of these new sub-scales and the original instrument in predicting violent reoffending: for the PCL-R, the sub-scale showed a small improvement over the full scale in its predictive ability, measured by both the *z*-score and AUC value. The sub-scale of the

Table 3. Independent predictive effects of significant items, selected from stepwise logistic regression analysis.

PCL-R			VRAG			HCR-20		
Item	*z*-score	AUC	Item	*z*-score	AUC	Item	*z*-score	AUC
P3	2.21*	0.61***	V1	3.10***	0.63***	H2	4.05***	0.64***
P5	−3.13**	0.48	V4	3.52***	0.61***	H5	2.15*	0.59***
P10	2.48***	0.63***	V7	3.47***	0.60***	H8	2.90**	0.62***
P20	2.94**	0.65***	V11	3.00***	0.60***	H10	2.18*	0.61***
			V12	2.33*	0.59***	C2	4.44***	0.62***
						C4	3.79***	0.61***
						R2	2.40*	0.58***
						R4	2.48*	0.58***

Note: *$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$.
Note: Stepwise regression procedures of both backwards and forwards were used to select these items from all items except for those listed in Table 3. For HCR20 this analysis was carried out for items of H, C and R separately.

PCL-R was less accurate in predicting non-violent reoffending but more accurate in predicting violent reoffending than the original PCL-R. This difference resulted in an overall lower accuracy in the prediction of any offending using the sub-scale. However, χ^2 tests did not demonstrate the differences in the probabilities of positive and negative prediction between the two scales, indicating that these apparent differences could be due to sampling error or chance.

For the VRAG, the differences between the sub-scale and original instrument for all effect measures were very small, with no significant differences. Similarly, non-significant differences were also observed for the HCR-20 total score and the name sub-scale derived from the independently predictive items.

The predictive accuracy of the three instrument scales for violent reoffending either by their full scale or sub-scale was further ascertained from the cross-validation analysis. It suggested that the same results based on our cohort sample would have been observed in the population where the sample was drawn.

Can risk prediction be optimised using independently predictive items?

In our final analysis, we pooled all items from the sub-scales of the three instruments into a new, 'super instrument' to test its overall predictive effects. The 17 items with significant independent predictive ability (see Table 3 and Appendix) were entered into the regression model. Both forward and backward stepwise selection procedures yielded the same

Table 4. Comparison between the full scale and sub-scale of instrument for violent reoffending.

	PCL-R			VRAG			HCR-20		
	Full scale	Sub-scale	p^a	Full scale	Sub-scale	p^a	Full scale	Sub-scale	p^a
Number of items	20	3		12	5		20	8	
Valid cases	915	1337		1343	1347		1253	1290	
Z-score	6.03***	7.81***		8.11***	8.21***		6.75***	7.68***	
AUC	0.63***	0.68***		0.70***	0.71***		0.67***	0.69***	
Correct classification non-violence offending: *n* (%)	504 (60.4)	663 (57.1)	0.108	695 (59.6)	659 (56.4)	0.114	651 (60.2)	637 (57.2)	0.155
Correct classification violence: *n* (%)	86 (73.5)	133 (75.6)	0.469	127 (71.3)	135 (75.8)	0.380	125 (73.1)	129 (73.3)	0.968
Correct classification any offending: *n* (%)	590 (62.0)	796 (59.5)	0.227	822 (61.2)	794 (58.9)	0.222	776 (61.9)	766 (59.4)	0.188
Correct classification from cross-validation: *n* (%)	558 (58.7)	796 (59.5)	0.452	806 (60.0)	794 (58.9)	0.572	769 (60.6)	757 (58.7)	0.166

Note: [a]From χ^2 test; ***$p \leq 0.001$.

seven items that significantly predicted violent re-offending. These included Criminal versatility from the PCL-R; Younger age at Index offence, History of alcohol problems and Not female victim from VRAG; Young age at first violence, Early maladjustment and Negative attitudes from HCR-20. The new scale ranged from −7 to 14, constituting the sum of the seven items selected. A total of 1295 individuals had valid scores on this new instrument, which had a *z*-score value 8.33, AUC value 0.72 (SE = 0.019) and correctly predicted violent re-offending in 71.6% and non-offending in 63.0% of cases. This demonstrated a small improvement over the three original instruments (see Table 1). However, if the AUC and *z*-values of the new instrument are compared with those observed in Table 4, they show that the actual improvement in predictive power is very small compared to the sub-scales of each instrument that we had previously developed.

Discussion

What is the limit to predictive ability?

Our study confirmed that certain items had no independent predictive ability for our sample of released prisoners in the PCL-R, VRAG and HCR-20. We demonstrated that it is still possible to achieve similar accuracy in risk assessment for violent reoffending by removing the items with no predictive ability. However, any gains we achieved were small. The largest gain was achieved with the PCL-R, where removing most items and leaving a sub-scale containing only three items raised its predictive value to above a moderate level measured using AUC values, and with an increase in its positive predictive value. However, reducing the VRAG and HCR-20 to include only their independently predictive items achieved very small gains. Similarly, our further attempts (not reported here) to add predictive items from the C and R scales of the HCR-20 to the PCL-R and VRAG sub-scales achieved little or no improvement. Finally, an attempt to create a 'superinstrument', incorporating the strongest independent predictors from the three instruments, also achieved minimal additional improvement.

Taken together, these findings suggest that a limiting process is in operation, or a 'glass ceiling' effect, on risk instruments which incorporate historical (or static) risk factors. It may ultimately prove impossible to achieve further improvement in predictive accuracy above a certain limit using these measures. In this study, the limit appeared to be an AUC value of 0.72 and no permutation of significantly predictive items appeared able to exceed this. We have previously observed that those studies that compared more than one risk assessment instrument to predict violence have also demonstrated an optimum level of prediction around an AUC value of 0.7. Only three included instruments, which achieved AUC values of 0.8 or

above, but these employed retrospective methods (de Vogel et al., 2004; Douglas et al., 2005; Gray et al., 2004).

One of the many possible explanations for the ceiling effect could be that other events intervened following release, some of which may be protective. For example, participants may subsequently have found work, stable relationships or even experienced illnesses and injury that limited their criminal and violent behaviour. On the other hand, many may have offended or behaved violently during their follow-up, but these events were not recorded. Studies using criminal records require convictions and participants may deny actual violence using self-report measures.

Should non-predictive items be removed?

We confirmed that an aggregate of items with no independent predictive ability did not have any predictive effects. Our approach differed from that of Kroner, Mills and Reddon (2005) who found aggregate scores from four randomly generated 'instruments', derived from four original instruments (PCL-R, VRAG, HCR-20, LSIR), predicted as well as the original instrument. The important difference in their study was that each 'instruments' contained 14 items randomly picked from a pool of all the original items in the four instruments. It is probable that even a random selection of 14 items will include some with independent predictive ability. This would suggest that only a small number of independently predictive items are necessary to achieve moderate predictive ability.

Some items conveyed negative predictive ability and this may contribute to reducing aggregate scores, particularly in the case of the PCL-R. This would indicate that, if the sole purpose of these instruments is to attribute a numerical risk score with the intention of stratifying offenders into levels of risk, then they can be limited to those items that are actually predictive. But this would restrict conventional usage in terms of measuring psychopathy using the PCL-R and guiding clinical risk management using the HCR-20.

It would clearly be premature to recommend deletion of specific items from any of these risk scales. Items for actuarial scales are selected precisely because they have predictive properties, and what we may be witnessing is a failure for this capability to generalise across different populations or outcomes. Mills, Kroner and Hemmati (2007), for example, identified a different sub-set of items in the VRAG and HCR-20 as predictive of recidivism, but over a longer follow-up period. Furthermore, items in the HCR-20 that had no predictive ability for our sample might have in another, for example psychiatric patients, and might therefore retain applicability in risk management in clinical settings in contrast to prisons. A large element of caution is therefore needed in interpreting these findings, and more results from a variety of different studies will be needed before their significance becomes fully clear.

Can static instruments be improved?

Only the VRAG achieved an AUC value above a moderate level in our study, and the HRC-20-R scale was at a lower level of predictive ability than other instruments. However, lack of significant differences between instruments and sub-scales, as indicated by their confidence intervals, questions whether they were in fact measuring very similar constructs. Examining items in the three sub-scales that were independently predictive suggested that this was partly the case. Early behavioural problems appeared to be a significant predictor in each instrument, as indicated by P18 juvenile delinquency, V4 younger age at index offence, and both H2 young age at first violence and H8 early maladjustment. Criminal versatility (P20) and non-violence offence score (V7) may have measured similar constructs of general criminality in the PCL-R and VRAG. The PCL-R does not measure substance misuse, but this item appeared predictive in the HCR-20 (H5), and as alcohol problems (V11) in the VRAG. However, a generalised tendency to both violent and criminal behaviour may have been measured by all four significant items in the PCL-R (P3, P10, P18, P20); similarly, psychopathy (V1) and non-violent offence score (V7) in the VRAG; together with prior supervision failure (H10) and impulsivity (C4) in the HCR-20. It could also be argued that negative attitudes (C2), exposure to destabilisers (R2) and non-compliance with remediation attempts (R4) are merely characteristics and indeed outcomes among individuals with violent and criminal tendencies, despite these items being placed within different domains of risk according to the theoretical basis of the C and R scales of the HCR-20 as putative dynamic risk factors.

In a previous study, Mills et al. (2007) observed that more HCR-20 items were significantly predictive than VRAG items. They argued that this may have been due to the original method of item selection and construction of the two instruments, the VRAG having been developed using a more statistically rigorous selection procedure, but where the general ability of the scale would be relatively more limited to the sample on which it was originally standardised. However, this was not the case in our study, where 80% of PCL-R items were not predictive and where examination of the HCR-20 and VRAG indicated that a similar proportion of items (41.7%) were operating effectively. The most powerful items appeared to be those indicating early onset of behavioural problems, general criminality and possibly a tendency to violence manifested through impulsive lifestyle factors, possibly exacerbated by substance (primarily alcohol) misuse. Alternatively, substance misuse within the VRAG and HCR-20 sub-scales may have merely added to their aggregate scores, enabling them to match the predictive ability of the otherwise superior PCL-R sub-scale. Overall, there are strong indications from our study and others that the underlying construct being measured by these scales is not propensity for violence per

se, but a more general anti-social/criminal tendency that is sometimes manifested violently and sometimes not.

At the item level, although certain items appeared to measure the same constructs in more than one instrument, their predictive accuracy still differed. Some may not have accurately predicted violence because of the manner in which they had been constructed. Psychopathy as included in the HCR-20 (H7) contrasted with psychopathy (V1) as included in the VRAG. This could be due to the inclusion of more information within the VRAG where PCL-R scores are scaled, in contrast to the more limited scoring method employed in the HCR-20. Another example included impulsivity (P14 in the PCL-R and C4 in the HCR-20) where the measures differed both in their construction and application over a specified time span. This indicates that any future attempts to develop new instruments will require considerable caution over the manner in which individual items are constructed and subsequently presented for scoring.

Limitations

The prisoner cohort included a large sample, prospectively interviewed prior to release. Few prisoners declined the interview, and attrition was primarily due to delays in access, unexpected transfer or release of prisoners. However, the study outcome was limited to criminal convictions, and the base rate of violence would have been higher if additional measures had been collected. Furthermore, the follow-up period was at a mean of 1.97 years rather than all participants being measured at the 2-year stage. This meant that a sub-group may not have been in the community long enough to have recorded a violent conviction. However, considering that only 5.5% of participants were followed up after less than 1 year, with small changes in the base rate of violent re-offending and in correlations between the outcome and risk instrument scores observed after excluding those participants, the findings of the study were unlikely to be affected.

A further limitation is that the participants were subjected to varying degrees of supervision post-release that may have acted positively to prevent the outcomes predicted, but which were not included in this study.

Clinical implications and future research

The key implication of this study is that clinicians should be aware of the limitations and be critical when using either an actuarial, structured clinical risk assessment instrument, or a personality disorder assessment instrument, if the intention is to carry out a comprehensive assessment of risk on which to base subsequent risk management or treatment interventions. The majority of the predictor items within these instruments will not be relevant to the risk posed by any single patient/client at a point of time. The clinician

must ultimately rely on more complex processes of deduction and formulation of risk based on clinical experience, training and knowledge of the relevant literature. Even when seven independently predictive items were drawn from three other instruments in the form of a 'super instrument' (criminal versatility, young age at index offence, alcohol problems, non-female victim, young age at first violence, early maladjustment and negative attitudes), our findings demonstrated that this cannot reflect the complexity of violent risk assessment and ultimately inform risk management.

The findings emphasise the importance of differentiating predictive ability and clinical utility of currently available instruments to assess risk. Although the addition of individual item ratings to generate HCR-20 total and sub-scale scores are necessary to investigate validity (the approach taken in this study), this is discouraged in clinical practice where the HCR-20 is not intended to be used as a risk prediction instrument. Furthermore, a number of items were intentionally included during its development because they were believed to have clinical utility in violent risk assessments, which superseded their predictive validity in terms of potential relevance. For example, items H6 major mental illness and C3 active symptoms of major mental illness are often noted to lack predictive validity in research, yet their presence and relative relevance in the individual case may be critical.

In this study, we effectively ignored the original purpose of two of our instruments and treated all three as actuarial instruments. This may be useful in stratifying patients/clients into levels of risk, for example low, medium, high and very high. However, the application of such measures prior to release from prison or discharge from hospital emphasises the limitations of such a procedure, and where accuracy in stratification is largely determined by static factors. An alternative approach in future research may be to include actual dynamic factors observed when in the community and then to prospectively study the interactive effects of these true dynamic measures and the original, aggregate scores on the static instruments. This approach is closer to clinical practice in risk management and may indicate future clinical interventions by identifying those dynamic factors that have large or even multiplicative effects on scores obtained from risk assessment instruments containing static items. In a study of patients with schizophrenia and delusional disorder discharged from UK medium secure services, four levels, or strata, of risk were attributed to participants using an actuarial measure based on static factors prior to discharge (Hickey, Yang, & Coid, 2008). The impact of subsequent dynamic factors, including failure to accept community supervision and take prescribed medication, further increased the likelihood of subsequent offending. Furthermore, the effects of these dynamic factors were greatest among patients within the strata of highest risk.

Our findings demonstrate the importance of separating the effects of static and dynamic factors when investigating future models of risk assessment for incorporation into clinical practice. A study of patients

discharged from mental health services concluded positively on the incremental validity of adding HCR-20 clinical and risk management items to other risk measures (Doyle & Dolan, 2006). However, closer examination of their findings indicated that the gains achieved were very small, the percentage of the sample correctly classified as violent increasing only by 2% when including C and R items in the aggregate. Their findings are therefore similar to our own observations when attempting to optimise risk assessment by adding static items from different instruments. HCR20-C and -R scales were developed to identify clinical items that might be amenable to intervention, primarily in patients with severe mental illness, and indicators of risk were the individual to be discharged or released into the community. However, empirical studies that have tested the predictive accuracy of the C and R scales usually rated these items whilst participants are detained in hospital or prison. This means that a proportion of these items may no longer apply once participants return to the community. Clinical intervention specifically as a result of these measures may have successfully reduced the risk. Future empirical research will need to add repeated, prospective measures of dynamic risk factors to static risk assessments if it is ultimately to reflect the complexity of clinical risk assessment and contribute to clinical risk management.

Acknowledgement

The Prisoner Cohort Study was funded by the Ministry of Justice (formerly Home Office). Jeremy Coid, Simone Ullrich and Tianqiang Zhang were supported by a Programme Grant PP-PG-6407-10500 from the National Institute of Health Research, UK (NIHR).

References

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

Coid, J., Yang, M., Ullrich, S., Zhang, T., Sizmur, S., Roberts, C., ... Roger, R.D. (2009). Gender differences in structured risk assessment: Comparing the accuracy of five instruments. *Journal of Consulting and Clinical Psychology*, *7*, 337–348.

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

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgement. *Science*, *243*, 1668–1674.

de Vogel, V., de Ruiter, C., van Beek, D., & Mead, G. (2004). Predictive validity of the SVR-20 and Static-99 in a Dutch sample of treated sex offenders. *Law and Human Behavior*, *28*, 235–251.

Douglas, K., Ogloff, J., Nicholls, T., & 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*, *6*, 917–930.

Douglas, K.S., Yeomans, M., & Boer, D. (2005). Comparative validity of multiple measures of violence risk in a sample of criminal offenders. *Criminal Justice and Behavior*, *32*, 479–510.

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

Doyle, M., Dolan, M., & McGovern, J. (2002). The validity of North American risk assessment tools in predicting inpatient violent behavior in England. *Legal & Criminological Psychology*, *7*, 141–154.

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.

Glover, A., Nicholson, D., Hemmati, T., Bernfeld, G., & Quinsey, V. (2002). A comparison of predictors of general and violent recidivism among high risk federal offenders. *Criminal Justice and Behavior*, *29*, 235–249.

Grann, M., Belfrage, H., & Tengström, 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.

Gray, N.S., Hill, C., McGleish, A., Timmons, D., MacCulloch, M.J., & Snowden, R.J. (2003). Prediction of violence and self-harm in mentally disordered offenders: A prospective study of the efficacy of HCR-20, PCL-R, and psychiatric symptomatology. *Journal of Consulting and Clinical Psychology*, *71*, 443–451.

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 Psychology*, *72*, 523–530.

Grevatt, M., Thomas-Peter, B., & Hughes, G. (2004). Violence, mental disorder and risk assessment: Can structured clinical assessments predict the short-term risk of inpatient violence? *Journal of Forensic Psychiatry & Psychology*, *15*(2), 278–292.

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*, 293–323.

Hare, R.D. (2003). *The Hare psychopathy checklist – revised* (2nd ed. Manual). Toronto, ON: Multi-Health Systems.

Hemphill, J.F., Hare, R.D., & Wong, S. (1998). Psychopathy and recidivism: A review. *Legal and Criminological Psychology*, *3*, 139–170.

Hickey, N., Yang, M., & Coid, J. (2009). The development of the Medium Security Recidivism Assessment Guide (MSRAG): An actuarial risk prediction instrument. *Journal of Forensic Psychiatry and Psychology*, *20*(2), 202–224

Howard, P., & Kershaw, C. (2000). *Using criminal career data in evaluation*. British Criminology Conference: Selected Proceedings. Retrieved from www.lboro.ac.uk/departments/ss/bsc/bccsp/vol03/howard.html

Kroner, D.F., & Loza, W. (2001). Evidence for the efficiency of self-report in predicting nonviolent and violent crime recidivism. *Journal of Interpersonal Violence*, *16*, 168–177.

Kroner, D.G., Mills, J.F., & Reddon, J.R. (2005). A coffee can, factor analysis, and prediction of antisocial behaviour: The structure of criminal risk. *International Journal of Law and Psychiatry*, *28*, 360–374.

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*, 471–489.

Mills, J.F., Kroner, D.G., & Hemmati, T. (2007). The validity of violence risk estimates: An issue of item performance. *Psychological Services*, *4*, 1–12.

Morrissey, C., Hogue, T., Mooney, P., Allen, P., Johnston, S., Hollins, C., ... Taylor, J.L. (2007). Predictive validity of the PCL-R in offenders with intellectual disability in a high secure hospital setting: Institutional aggression. *Journal of Forensic Psychiatry and Psychology*, *18*, 1–15.

Nicholls, T.L., Ogloff, J.R.P., & Douglas, K.S. (2004). Assessing risk for violence among male and female civil psychiatric patients: The HCR-20, PCL:SV, and VSC. *Behavioral Sciences and the Law*, *22*, 127–158.

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

Salekin, R., Rogers, R., & Sewell, K. (1996). A review and meta-analysis of the Psychopathy Checklist and Psychopathy Checklist – Revised: Predictive validity of dangerousness. *Clinical Psychology: Science and Practice*, *3*, 203–215.

Serin, R.C. (1996). Violent recidivism in criminal psychopaths. *Law & Human Behavior*, *20*, 207–216.

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.

Stadtland, C., Hollway, M., Kleindienst, N., Dietl, J., Reich, U., & Nedopil, N. (2005). Risk assessment and prediction of violent and sexual recidivism in sex offenders: Long-term predictive validity of four risk assessment instruments. *Journal of Forensic Psychiatry and Psychology*, *16*, 92–108.

Taylor, R. (1999). *Predicting reconvictions for sexual and violent offences using the Revised Offender Group Reconviction Scale*. Home Office Research Findings No.104. London: Home Office.

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.

Warren, J.I., South, S.C., Burnette, M.L., Rogers, A., Friend, R., Bale, R., & Van Patten, I. (2005). Understanding the risk factors for violence and criminality in women: The concurrent validity of the PCL-R and HCR-20. *International Journal of Law and Psychiatry*, *28*, 269–289.

Webster, C.D., Douglas, K.S., Eaves, D., & Hart, S.D. (1997). *HCR-20: Assessing risk of violence (version 2)*. Vancouver: Mental Health Law & Policy Institute, Simon Fraser University.

Webster, C.D., Muller-Isberner, R., & Frannson, G. (2002). Violence risk assessment: Using structured clinical guidelines professionally. *International Journal of Mental Health*, *1*, 185–193.

Appendix. Items in PCL-R, VRAG and HCR-20 instruments.

PCL-R		VRAG		HCR-20	
Glib/superficial	P1	Psychopathy	V1	Previous violence	H1
Grandiose	P2	Elementary school maladjustment	V2	Young age at first violence	H2
Stimulation/boredom	P3	DSM-IV PD	V3	Relationship instability	H3
Pathological lying	P4	Younger age at Index Offence	V4	Employment problems	H4
Conning/manipulative	P5	Lived with both parents to 16	V5	Substance use problems	H5
Lacks remorse/guilt	P6	Failure on prior conditional release	V6	Major mental problems	H6
Shallow affect	P7	Non-violent offence score	V7	Psychopathy	H7
Callous/empathy	P8	Never married	V8	Early maladjustment	H8
Parasitic lifestyle	P9	DSM-IV schizophrenia	V9	Personality disorder	H9
Poor behavioural controls	P10	Less severity of victim injury	V10	Prior supervision failure	H10
Promiscuous	P11	History of alcohol problems	V11	Lack of insight	C1
Early behavioural problems	P12	Not female victim	V12	Negative attitudes	C2
Lacks realistic goals	P13			Active symptoms of major mental illness	C3
Impulsivity	P14			Impulsivity	C4
Irresponsibility	P15			Unresponsive to treatment	C5
Failure accepts responsibility	P16			Plans lack feasibility	R1
Many relationships	P17			Exposure to destabilisers	R2
Juvenile delinquency	P18			Lack of personal support	R3
Revocation conditional release	P19			Non-compliance with remediation attempts	R4
Criminal versatility	P20			Stress	R5