The Journal of Forensic Psychiatry & Psychology
Vol 15 No 1 March 2004 145–164

BrunnerRoutledge Taylor & Francis healthsciences

Differences between clinicians and researchers in assessing risk of violence in forensic psychiatric patients

VIVIENNE DE VOGEL and CORINE DE RUITER

ABSTRACT Do clinicians and researchers differ in their violence risk assessment of the same patient? In this study, the Dutch version of the HCR-20 was coded by two independent researchers and two independent clinicians (treatment supervisor and group leader) for 60 patients admitted to a Dutch forensic psychiatric hospital. The aim of the study was threefold: (1) to establish the interrater reliability of the Dutch HCR-20; (2) to gain insight into differences between researchers and clinicians in coding the HCR-20; and (3) to examine the relationship between clinicians' feelings towards their patients and their risk judgment. Overall, the interrater reliability of the HCR-20 was good. The group leaders gave significantly lower HCR-20 scores than the researchers. There were no significant differences between the mean HCR-20 scores of treatment supervisors and researchers, but there was a significant difference in the interpretation of the scores: treatment supervisors had more 'low risk' judgments than researchers. Furthermore, it was found that feelings of clinicians towards their patients were associated with their risk judgment. Feelings of being controlled and manipulated by the patient were related to higher HCR-20 scores, whereas positive feelings (helpful, happy, relaxed) were related to lower risk judgments.

Keywords: risk assessment, violence, HCR-20, interrater reliability, countertransference feelings

The Journal of Forensic Psychiatry & Psychology
ISSN 1478-9949 print/ISSN 1478-9957 online © 2004 Taylor & Francis Ltd
http://www.tandf.co.uk/journals
DOI: 10.1080/14789940410001655916

INTRODUCTION

The assessment of risk of violence is an important task of mental health professionals working in forensic practice. A carefully conducted risk assessment before a probationary leave, parole decision, or termination of (mandatory) treatment can help to appraise the risk of recidivism in an adequate way and thereby assist in risk management (Douglas and Webster, 1999a). To date, the best known and most widely used method in practice, at least in the Netherlands, is the unstructured clinical judgment approach which is exclusively based on the professional expertise of the clinician. However, research has revealed some important limitations of unstructured clinical judgment, such as poor reliability and validity (see for a discussion of these disadvantages Quinsey, Harris, Rice and Cormier, 1998, pp. 55–72). The employment of more structured risk assessment procedures is highly recommended (Borum, 1996; Webster, Douglas, Eaves and Hart, 1997a).

One of the most promising risk assessment instruments at the moment is the Historical, Clinical, Risk management-20 (HCR-20; Webster, Douglas, Eaves and Hart, 1997b). This instrument consists of 20 items representing risk factors for violence in the past, present and future. Research in various psychiatric and penitentiary settings in different countries has demonstrated good interrater reliability and predictive validity for the HCR-20 total score (Belfrage, 1998; Belfrage, Fransson and Strand, 2000; Douglas, Ogloff, Nicholls and Grant, 1999; Strand, Belfrage, Fransson and Levander, 1999). Furthermore, it was demonstrated that changes in risk during clinical intervention can be measured with the HCR-20 (Belfrage and Douglas, 2002). However, an important disadvantage of many studies into the HCR-20 is their retrospective design. So far, only a few prospective studies have been conducted into the predictive validity of the HCR-20 (e.g. Belfrage *et al.*, 2000; Douglas *et al.*, 1999).

Another disadvantage of many studies into the HCR-20 concerns the ecological validity, i.e. its relevance to actual clinical risk assessment practice. In most published studies, the HCR-20 was coded by independent researchers, not by practicing clinicians. Generally, the researchers did not know the patient personally and coded the HCR-20 solely on the basis of file information. When using the HCR-20 in clinical practice – the assessment of risk of future violence and the use in leave decisions – coding by experienced clinicians is required (Webster *et al.*, 1997b). Also, in clinical practice it is usually the treating clinicians who are responsible for leave decisions (Dernevik, Falkheim, Holmqvist and Sandell, 2001). However, there is some doubt about the objectivity of clinicians, especially clinicians who are closely involved in the treatment of the patient (Dernevik *et al.*, 2001; Litwack and Schlesinger, 1999). Are treating clinicians who know the patient well and who have invested a lot in his treatment capable of putting aside their personal feelings towards the patient when they assess their risk of violence? Several authors have argued that these feelings inevitably lead to subjectivity and as a consequence the roles of forensic assessor and treating clinician are irreconcilable (Ackerman, 1999; Greenberg and Schuman, 1997; Litwack and Schlesinger, 1999). These authors suggest that more distant assessors, not directly involved in the treatment, should conduct forensic evaluations. On the other hand, it can be argued that the treating clinician has the most detailed knowledge on the patient which is necessary to perform forensic evaluations. Also, the argument is raised whether these feelings of countertransference are truly irrelevant or irrational or may actually contribute to a more accurate risk assessment. Furthermore, it can be reasoned that because of the structured guidelines for risk assessment, such as the HCR-20, countertransference feelings are less likely to interfere. From all of these arguments it can be concluded that it is important to examine what exactly is the role of feelings towards the patient and if there are differences between researchers and clinicians in performing risk assessments, as well as in the accuracy of these assessments.

Dernevik *et al.* (2001) studied the relationship between clinicians' feelings as measured by the Feeling Word Checklist (FWC; Whyte, Constantopoulos and Bevans, 1982) and HCR-20 scores in a forensic psychiatric institution in Sweden. Forty experienced nurses who had attended a one-day training workshop in risk assessment, coded the HCR-20 for one of eight patients. Every patient was thus judged by five nurses. The risk of these eight patients was also assessed by a number of independent experts in risk assessment. The nurses gave a significantly higher mean HCR-20 total score than the experts. Furthermore, this study demonstrated 'the nurses' scores on the HCR-20 assessments to be influenced by their scores on the FWC' ($R = 0.66$, $R^2 = 0.43$; Dernevik *et al.*, 2001, p. 94). Feeling close and accepting towards the patient was associated with a higher HCR-20 score and feelings of helpfulness and autonomy were associated with a lower score. Douglas and Belfrage (2001) commented on this study that it cannot be concluded that the FWC scores actually *caused* differences in HCR-20 total scores rather than merely being correlated with them. However, they acknowledged the possible effect of bias and considered this research a reminder for assessors. Recently, Dernevik and Douglas (2002) conducted a follow up on this study and demonstrated that after a period of 2 years, the nurses showed good predictive validity for inpatient violence, but not for (violent) recidivism after discharge from the hospital. On the contrary, the experts were more accurate in their predictions of violence in the long term, that is, (violent) recidivism after discharge.

In this article, results are presented of a prospective study which started in January 2001. The authorized Dutch version of the HCR-20 (Philipse, de Reiter, Hildebrand and Bouman, 2000) was coded for 60 patients admitted to the Dr. Henri van der Hoeven Kliniek, a Dutch forensic psychiatric hospital, by both clinicians and independent researchers. The aim of the study was to establish the interrater reliability of the HCR-20, to gain insight into differences between clinicians and researchers in coding the HCR-20, and to examine if clinicians' feelings towards their patients as measured by the FWC were related to the risk assessments. Results into the predictive validity of the Dutch HCR-20 as well as differences between clinicians and researchers in risk assessment accuracy are not presented in the present study, but can be expected within 1 or 2 years.

METHOD

Setting

This study was conducted at the Dr. Henri van der Hoeven Kliniek, a 114-bed forensic psychiatric hospital in the Netherlands. Patients are admitted under the judicial measure *terbeschikkingstelling* (tbs) which is translated as 'disposal to be treated on behalf of the state'. The tbs-order is imposed by court on offenders who have committed a serious offense and are considered to have diminished responsibility because of severe psychopathology. The tbs-order is of indefinite duration; every 1 or 2 years the court re-evaluates the patient to determine whether the risk of recidivism is still too high and treatment needs to be continued.

The Dr. Henri van der Hoeven Kliniek was founded in 1955 and is one of 13 inpatient forensic psychiatric institutions in the Netherlands. The hospital provides a varied treatment program which includes job training, education, sports, creative arts, and psychotherapy. The treatment model of the hospital is cognitive-behavioral with an emphasis on relapse prevention. The 'no cure but control' principle dominates (Laws, Hudson and Ward, 2000). The emphasis of treatment is not on changing the personality of the offender, but on reducing/managing risk factors for recidivism. An important phase in treatment is the transmural phase. During this resocialization phase, the patient lives outside the hospital, but is still undergoing treatment and is supervised by a specialized treatment team from the hospital.

Subjects

The current sample comprised 53 men and seven women. The mean age at the time of the risk assessment was 36.6 years (SD = 8.0, range 22 – 54). Seventy percent of the subjects had been convicted before their tbs-order, with an average number of 4.9 (SD = 13.0, range 0 – 15) prior convictions.

The index-offenses were: 57% (attempted) homicide, 20% sexual offenses, 17% other violent offenses and 7% arson. In 8% of the patients a DSM-IV (APA, 1994) Axis I disorder was diagnosed, 52% met the criteria for one or more Axis II disorders, and in 35% there was comorbidity of Axis I and II disorders. Axis I diagnoses were lifetime clinical diagnoses based on consensus between four raters (see Hildebrand and de Ruiter, in press), Axis II disorders were diagnosed with the Structured Interview for DSM-IV Personality (SIDP-IV; Pfohl, Blum and Zimmerman, 1995).

Instruments

HCR-20

The HCR-20 is a clinical guideline (checklist) designed for the assessment of risk of future violence in adult offenders with a history of violent behavior and/or a major mental disorder or personality disorder. The instrument was developed from a thorough consideration of the empirical literature and the forensic expertise of a number of clinicians. The HCR-20 consists of 20 items, divided into three subscales: Historical scale, Clinical scale and Risk management scale, that relate to risk factors in the past, the present and the future, respectively. Table 1 presents the items of the HCR-20. The Historical items are static, largely unchangeable factors, whilst the Clinical and Risk management factors are considered to be dynamic and changeable, for instance, as a result of clinical intervention.

The items have to be coded on a three point scale: '0' item does not apply according to the available information, '1' the item probably or partially applies, and '2' the item definitely applies. The Psychopathy Checklist-Revised (PCL-R; Hare, 1991) is used to code Item 7 'Psychopathy': a PCL-R score of 30 justifies a code 2, and a PCL-R score between 20 and 29 represents a code 1. Aside from the 20 items, the HCR-20 offers the possibility to code 'other considerations', i.e. case-specific risk factors that do not fit within the descriptions of the 20 items. The HCR-20 has to be coded by an experienced forensic clinician, who should use all available information on the offender, preferably from different sources and gathered with different methods, for example, criminal records/police files, psychological reports, interviews with significant others, and staff observations.

The final overall risk judgment has to be indicated as low, moderate, or high and is only valid for a specific time period, e.g. during a specific treatment phase. The risk judgment not only depends on the simple summation of the item scores, but also on specific combinations of risk factors or other considerations. In some cases, only one or two items may be sufficient to justify the judgment 'high-risk', for example, when a patient has active psychotic symptoms (e.g. auditory hallucinations that instruct the

Table 1 HCR-20 Items

Historical items	
1.	Previous violence
2.	Young age at first violent incident
3.	Relationship instability
4.	Employment problems
5.	Substance use problems
6.	Major mental illness
7.	Psychopathy
8.	Early maladjustment
9.	Personality disorder
10.	Prior supervision failure
Clinical items	
1.	Lack of insight
2.	Negative attitudes
3.	Active symptoms of major mental illness
4.	Impulsivity
5.	Unresponsive to treatment
Risk management items	
1.	Plans lack feasibility
2.	Exposure to destabilizers
3.	Lack of personal support
4.	Noncompliance with remediation attempts
5.	Stress

Note. From Webster *et al.* (1997b).

patient to commit homicide). The final risk judgment can be considered a structured clinical judgment that is arrived at through the process of coding the checklist and integrating all available information.

FWC

The FWC was administered to map the clinicians' feelings towards their patients. This list was originally developed by Whyte *et al.* (1982) and adapted by Holmqvist and Armelius (1994) and Holmqvist and Fogelstam (1996). Table 2 shows the 30 items of the instrument relating to 'feelings' that have to be coded on a four point scale according to the extent to which the clinician has the specific feeling towards the patient whose risk is being assessed. The items are grouped into eight subscales which can be viewed as four continuous dimensions: (1) Helpful vs. Unhelpful; (2) Accepting vs. Rejecting; (3) Close vs. Distant; (4) Autonomy vs. Controlled.

Table 2 Feeling Word Checklist

When I think about … I feel:			
1.	Helpful	16.	Surprised
2.	Happy	17.	Tired
3.	Angry	18.	Threatened
4.	Enthusiastic	19.	Receptive
5.	Anxious	20.	Objective
6.	Strong	21.	Overwhelmed
7.	Manipulated	22.	Bored
8.	Relaxed	23.	Motherly
9.	Cautious	24.	Confused
10.	Disappointed	25.	Embarrassed
11.	Indifferent	26.	Interested
12.	Affectionate	27.	Aloof
13.	Suspicious	28.	Sad
14.	Sympathetic	29.	Inadequate
15.	Disliked	30.	Frustrated

Note. From Whyte *et al.* (1982). Subscale items: Helpful: 1, 2, 8; Unhelpful: 3, 10, 15, 18, 24; Close: 12, 16, 21, 23; Distant: 9, 11, 27; Accepting: 4, 14, 19, 26; Rejecting: 13, 17, 22, 30; Autonomy: 6, 20; Controlled: 5, 7, 25, 28, 29.

Procedure

First, all raters in this study were trained in coding the HCR-20 during a one-day workshop given by a senior clinical psychologist and the first author. During this workshop, relevant literature was discussed and the HCR-20 coding procedure was practiced on the basis of file information and videotapes of actual cases. All raters were instructed to use the HCR-20 manual and all available file information in all cases they had to rate.

During treatment, a number of specific phases can be distinguished in which the liberties of a patient expand and therefore the risk of violence needs to be (re-)evaluated. These phases are when a patient has his first leave from the hospital without supervision and when a patient is about to enter a resocialization phase called transmural treatment. During this transmural phase, the patient lives outside the hospital, but is still treated and supervised by a specialized team from the hospital. Since January 2001, the HCR-20 was coded for all patients who were in the above two phases, and for all patients who were already in the transmural phase. In addition, the HCR-20 was coded for all patients who were newly admitted to the hospital to assess the risk of inpatient violence. Two researchers, a group leader and the treatment supervisor independently coded the HCR-

20 for each case. When the patient was a sex offender, the Dutch version of the Sexual Violence Risk-20 (SVR-20; Boer, Hart, Kropp and Webster, 1997; authorized Dutch version: Hildebrand, de Ruiter and Van Beek, 2001) was coded in addition to the HCR-20.[1] All raters had access to file information that, in general, consisted of psychological reports, reports to the court regarding treatment progress and recommendations for termination or prolongation, treatment plans and evaluations. Subsequently, the two researchers agreed upon a consensus score in a case conference, and one of the researchers took this to a joint meeting with the two clinicians. The group leader and treatment supervisor independently filled out the Feeling Word Checklist before this consensus meeting. In general, the consensus meetings lasted about 1 hour and were considered very useful by both the researchers and clinicians. The results of the risk assessments were not communicated to the patients because we still consider the Dutch HCR-20 as a research instrument with insufficiently established predictive validity.

Assessors

Forty-four assessors participated in the present study: five researchers, seven treatment supervisors and 32 group leaders. The researchers were all clinical psychologists of the Research department, which is responsible for psychological assessment and empirical research in the hospital. The researchers are not in a treatment relationship with patients. The treatment supervisors were all senior clinicians, mostly clinical psychologists or psychotherapists. The professional background of the group leaders was diverse, but most of them had relevant higher vocational or academic training (e.g. in nursing, social work, psychology).

Statistical analyses

The interrater reliability of the HCR-20 was examined by means of the Intraclass Correlation Coefficient (ICC), using the two-way random effect variance model and consistency type (McGraw and Wong, 1996). Critical values for single measure ICCs were: $\text{ICC} \geqslant 0.75$ = excellent; $0.60 \leqslant \text{ICC} < 0.75$ = good; $0.40 \leqslant \text{ICC} < 0.60$ = moderate; $\text{ICC} < 0.40$ = poor (Fleiss, 1986).

The *F*-test was used to examine differences between researchers, group leaders and treatment supervisors on HCR-20 subscales and total scores. For differences in HCR-20 risk judgments we used Chi-square tests. The relationship between FWC subscales and HCR-20 total scores/risk judgments were analyzed using Pearson product moment correlations and stepwise multiple regression analyses.

RESULTS

Risk assessments

The mean HCR-20 total score as agreed upon during the final consensus meetings was 26.1 (SD = 6.5, range = 10 – 37), the mean score for the Historical scale 14.6 (SD = 3.3, range = 6 – 19), for the Clinical scale 5.3 (SD = 2.2, range = 0 – 9) and for the Risk management scale 6.1 (SD = 2.1, range = 2 – 10). On average, 3.3 (SD = 1.8, range = 0 – 8) other considerations were identified.

The risk of 17 patients was judged as low (mean total score = 20.8, range = 10 – 28), of 24 as moderate (mean total score = 24.7, range = 16 – 34), and of 19 as high (mean total score = 32.4, range = 23 – 37). The differences in mean total scores of the low, moderate and high categories were significant ($F = 31.6$, $p < 0.01$). The overlap in ranges of total scores for the risk judgments low, moderate and high was rather large.

Other considerations

Frequently coded other considerations were sadistic fantasies, social desirability/elusiveness, social isolation, financial problems and lack of prospects.

Interrater reliability HCR-20

Table 3 shows the single measure ICCs for the HCR-20 subscales, total score and risk judgment for the different groups of assessors. Overall, the interrater reliability of the HCR-20 subscales and total score was good. Particularly for the Historical scale we found excellent interrater reliability. However, two Historical items showed poor interrater reliability: Previous Violence and Early maladjustment. Good interrater reliability was demonstrated for the Clinical items, except for the item Impulsivity. The Risk management scale revealed moderate interrater reliability. Exposure to destabilizers, Noncompliance with remediation attempts, and Stress were items with poor interrater reliability.

We found some differences in interrater reliability between the different assessor groups. Overall, the interrater reliability among the researchers was excellent. The interrater reliability among the researchers and group leaders, as well as among the researchers and treatment supervisors was good, except for the Risk management items. In contrast, the interrater reliability among group leaders and treatment supervisors was moderate for both the Clinical and Risk management items.

Table 3 Intraclass Correlation Coefficients (ICCs) single measure

	H items	*C items*	*R items*	*Total score*	*Risk judgment*
Researcher 1 and 2	.85	.75	.70	.85	.76
Researchers and treatment supervisors	.83	.67	.56	.79	.65
Researchers and group leaders	.78	.70	.58	.82	.68
Treatment supervisors and group leaders	.86	.54	.58	.77	.63
Researchers, treatment supervisors and group leaders	.82	.64	.57	.79	.65

Note. All ICCs above 0 ($p < 0.001$). Researchers = consensus researcher 1 and 2. H items = Historical items. C items = Clinical items. R items = Risk management items.

Differences between assessors in risk assessments

Table 4 presents the mean HCR-20 scores and overall risk judgments of the different assessor groups. Group leaders gave significantly lower scores on the Historical items, Risk management items, other considerations and total scores. There were no significant differences in mean HCR-20 scores between the researchers and the treatment supervisors, except for the number of other considerations. However, there was a significant difference in risk judgments: treatment supervisors more often judged patients as 'low risk' compared to researchers.

At the end of the consensus meetings, the assessors were asked how much time they had spent coding the HCR-20, and which information they had used in their risk assessment. Researchers said they had spent on average 120 minutes per risk assessment, group leaders 30 minutes and treatment supervisors 15 minutes. Besides, the researchers stated they based their risk assessments predominantly on file information whereas the group leaders and treatment supervisors mostly relied on their personal experiences with the patient. Several treatment supervisors stated that they did not need to read the file information, because they were already familiar with it and sometimes they had written the information themselves, for instance, a treatment plan or report to the court.

Clinicians' feelings towards their patients and the risk assessments

First, Pearson product moment correlations were computed between the FWC subscales scores and the HCR-20 total scores/risk judgments (see Table 5). The subscales Unhelpful, Distant, Rejecting and Controlled showed significant positive correlations with HCR-20 total scores and risk

Table 4 Risk assessments ($n = 60$)

	Mean scores					*Risk judgments*		
	H items	*C items*	*R items*	*Total*	*Other*	*Low*	*Moderate*	*High*
Researchers	14.5^{a}	5.2	6.2^{c}	25.9^{e}	2.4^{g}	8^{j}	28	24
Treatment supervisors	13.8	5.0	5.8	24.6	0.8^{h}	18^{k}	21	21
Group leaders	13.2^{b}	4.5	5.0^{d}	22.7^{f}	0.5^{i}	12	28	20
Consensus	14.6	5.3	6.1	26.1	3.3	17	24	19

Note. $^{a} > ^{b}$, $p < 0.05$. $^{c} > ^{d}$, $p < 0.01$. $^{e} > ^{f}$, $p < 0.01$. $^{g} > ^{h} > ^{i}$, $p < 0.05$. $^{k} > ^{j}$, $p < 0.05$. H items = Historical items. C items = Clinical items. R items = Risk management items. Other = other considerations.

Table 5 Pearson product moment correlations FWC subscales and HCR-20 total score and risk judgment

	Helpful	*Unhelpful*	*Close*	*Distant*	*Accepting*	*Rejecting*	*Autonomy*	*Controlled*	*HCR-20 total score*
Helpful	–								
Unhelpful	-0.45^{**}	–							
Close	0.19^{*}	0.07	–						
Distant	-0.48^{**}	0.50^{**}	-0.10	–					
Accepting	0.72^{**}	-0.49^{**}	0.36^{**}	-0.42^{**}	–				
Rejecting	-0.35^{**}	0.69^{**}	0.07	0.55^{**}	-0.44^{**}	–			
Autonomy	0.24^{**}	0.05	0.21	-0.03	0.11	-0.02	–		
Controlled	-0.40^{**}	0.63^{**}	0.27^{**}	0.42^{**}	-0.24^{**}	0.56^{**}	-0.06	–	
HCR-20 total score	-0.28^{**}	0.38^{**}	0.12	0.20^{*}	-0.19^{*}	0.34^{**}	0.04	0.46^{**}	–
HCR-20 risk judgment	-0.34^{**}	0.33^{**}	0.19^{*}	0.40^{**}	-0.23^{*}	0.34^{**}	-0.12	0.37^{**}	0.56^{**}

Note. $^{*}p < 0.05$ (two-tailed). $^{**}p < 0.01$ (two-tailed).

judgments. In contrast, the subscales Helpful and Accepting demonstrated significant negative correlations with HCR-20 total scores and risk judgments. Next, stepwise multiple regression analyses were conducted. Feelings of being controlled or manipulated by the patient significantly predicted high HCR-20 total scores: 21% of the variance in HCR-20 total scores was explained by the subscale Controlled (see Table 6; F (1) = 31.1, $p < 0.001$). None of the other FWC subscales yielded significant prediction for the HCR-20 total score. Three subscales were significant predictors for the HCR-20 risk judgments: the subscales Distant and Close predicted high risk judgments, whereas the subscale Helpful predicted low risk judgments. Together, these three subscales explained 23% of the variance in HCR-20 risk judgments (see Table 7; F (3) = 13.0, $p < 0.001$).

DISCUSSION

The present study demonstrated good interrater reliability for the Dutch HCR-20, provided insight into differences between researchers and clinicians in coding the HCR-20 and showed that clinicians' feelings towards their patient were related to their risk assessments. The mean HCR-20 scores we found resemble those found in previous studies conducted in

Table 6 Stepwise multiple regression FWC subscales as predictor variables for HCR-20 total scores

Variable FWC	*B*	*SE B*	*β*	*sign. T*	R^2	*Adjusted* R^2
Controlled	5.8	1.0	0.46	< 0.01	0.21	0.21

Table 7 Stepwise multiple regression FWC subscales as predictor variables for HCR-20 risk judgments

Variables FWC	*B*	*SE B*	*β*	*sign. T*
Distant	0.46	0.14	0.31	< 0.01
Close	0.35	0.11	0.26	< 0.01
Helpful	− 0.34	0.13	− 0.24	< 0.01
Multiple *R*	0.50			
R^2	0.25			
Adjusted R^2	0.23			

forensic psychiatric institutions in other countries (e.g. Belfrage, 1998; Strand *et al.*, 1999).

Overall, the interrater reliability of the HCR-20 was good and this corresponds to previous findings (Belfrage, 1998; Douglas and Webster, 1999b). Moreover, the differences in interrater reliability between the subscales – excellent for the Historical scale, good for the Clinical scale, and moderate for the Risk management scale – were also demonstrated before (Belfrage, 1998). Some individual items, however, showed poor interrater reliability. We suggest three possible causes for this. First, the inexperience of some assessors in coding standardized instruments such as the HCR-20 could have led to the low interrater reliability for these items. For example, the item Previous violence was erroneously rated by some clinician assessors: they did not rate the index offense as an instance of previous violence. However, during the training workshop it had been emphasized that previous violence refers to all violence prior to assessment. The researchers, on the contrary, who were more experienced in the use of standardized instruments, obtained an almost perfect interrater reliability on this item. Second, some HCR-20 item descriptions are unclear or so global that they are easily open to multiple interpretations. A number of assessors argued that the items Impulsivity and Exposure to destabilizers are items with a rather broad and imprecise definition. Third, differences in degree of clinical experience and personal attitudes possibly contributed to poorer interrater reliability. A recurrent discussion during the consensus meetings was if the patient's problems were serious enough to warrant a code of '2' instead of '1', for instance, when coding the items Early maladjustment and Stress. More experienced clinicians, such as the treatment supervisors, tended to view the problems of the patients as less serious, because they compared them to even more severely disturbed patients they had treated before.

The present study revealed some important differences between researchers and clinicians, not only in their HCR-20 scores, but also in their way of coding, i.e. the time taken to code the HCR-20 and the information used for the coding. Clinicians usually relied on their personal experiences with the patient and made almost no use of file information, whereas researchers predominantly relied on file information, naturally because they did not know the patient well enough personally. Future research will have to establish whether these differences impact on the accuracy of the risk assessments. We also found some differences between the three assessor groups. First, although researchers and treatment supervisors did not significantly differ in their mean HCR-20 scores, there was a substantial difference in their risk judgments. Treatment supervisors more often judged the overall risk as 'low', and thus seemed more optimistic in their interpretation of HCR-20 scores compared to researchers. A possible reason is that treatment supervisors experience pressure to let patients pass to the transmural phase as soon as possible.[2] It could be that – despite the structured way of assessing risk – treatment supervisors are susceptible to cognitive distortions/biases, such as the tendency to correlate information intuitively rather than by laws of probability ('conjunction fallacy') or the tendency to view unrelated events as correlated ('illusory correlation bias'; see for further explanation Dernevik *et al.*, 2001). The researchers were more distant from the patients and their treatment, and not involved in leave decisions and were therefore probably less susceptible to these biases.

A second difference was that group leaders gave lower HCR-20 scores compared to researchers, and – albeit not significantly – also compared to treatment supervisors. A possible reason is the group leaders' daily interaction with the patients; continuous awareness of the risk these patients pose would probably get in the way of a therapeutic interaction. In addition, the frequent interaction can induce emotional ties and involvement with the patient and as a consequence more access to the 'nicer sides' of the patient, which the group leaders then took into account when assessing risk. On the contrary, researchers may have emphasized the negative aspects of patients because file information usually focuses on the risks and problems of patients. Thus far, only one other study has been conducted that compared HCR-20 ratings of clinicians and researchers. Contrary to our findings, Dernevik *et al.* (2001) found a higher mean HCR-20 total score for nurses compared to independent experts (26.3 vs. 22.7). However, these nurses are hardly comparable to our group leaders because of substantial differences in the mean number of work years in the forensic institution (nurses: 24.6 vs. group leaders: 6.2).

A third difference lies in the interrater reliabilities of the assessor groups. Notable is the merely moderate agreement for the Clinical and Risk management items among treatment supervisors and group leaders. A possible cause is the different roles in treatment they fulfill: treatment supervisors have a supervising and planning role whilst the group leaders conduct the daily and practical supervision and spend most of the time with the patients. However, these differences in item scores did not seem to interfere substantially with the complete risk assessment since the mean scores of the treatment supervisors and group leaders did not differ significantly.

The feeling of being controlled and manipulated by the patient was strongly related to high HCR-20 total scores. Since psychopaths (as measured by the PCL-R) in particular are known to be capable of evoking these types of feelings in clinicians (Hare, Clarke, Grann and Thornton, 2000; Lösel, 1998), this would be an interesting issue for further study. Negative feelings, such as frustration, disappointment, suspiciousness and rejection were related significantly to high HCR-20 total scores. Interestingly, both feelings of closeness and distance/aloofness were related to high risk judgments. A possible explanation is that these two feelings are both at extremes of a continuum, which could signify a lack of adequate professional distance towards the patient, which could subsequently impact on the risk judgment to a greater degree than is desirable. Positive feelings like helpfulness, happiness and relaxation were related to low risk judgments. The conclusion of the present study that feelings of clinicians towards their patients are related with the risk assessment is in line with the findings of Dernevik *et al.* (2001).

A number of limitations to the present study should be mentioned. First of all, the small sample size is a limitation. For example, because of the small sample size it was difficult to examine possible differences between the supervisors and group leaders in their feelings toward the patients. Second, the exact meaning and implications of the relationship between clinicians' feelings towards patients and their risk assessment remains unclear. The question to be addressed is if these feelings do actually interfere with the accuracy of the risk judgments. It is possible that patients who receive a high score on the HCR-20 also tend to evoke aversive feelings in the people around them, e.g. because of their psychopathic traits or (personality) disorder. Dernevik and Douglas (2002) found that clinicians were able to accurately predict inpatient violence, whereas experts showed good predictive validity for violence after discharge. Hence, the question who is most suitable to conduct risk assessments – the treating clinician or the independent researcher – cannot be answered at this moment. Another limitation is the large number of assessors who participated in this study, especially group leaders. Almost no group leader performed more than two risk assessments. Consequently, mistakes such as those described above with the item Previous violence were made quite regularly.

Based on the results and experiences of the present study, we would like to make some recommendations for the implementation of standardized risk assessment instruments in clinical practice. First, for clinical use, we recommend the use of consensus risk assessments with both clinicians and researchers in order to rule out possible effects of assessor bias, to discuss risk management strategies and to identify possible additional risk factors or protective factors. In the present study, the combination of the more distant, objective researcher and the treating clinicians who know the patient well seemed to benefit the discussion about risk factors. Second, a thorough and repeated training in coding risk assessment instruments is essential. The number of persons trained in such a workshop should not exceed 10, because this facilitates group discussion and attention to individual biases. The time period between the workshop and the first risk assessment should not exceed 6 months, otherwise the obtained knowledge and skills may have been lost. Continuous training can be achieved by organizing 'return days' to discuss questions and pitfalls and also to provide the assessors with feedback on the accuracy of their risk assessments, for instance, when a patient who is living outside the hospital has recidivated with a violent offense.

Future research will have to demonstrate the predictive validity of the Dutch HCR-20 and to examine if there are differences in predictive validity of HCR-20 assessments of researchers and clinicians, both in the short and longer term. In addition, the relationship between clinicians' feelings towards their patients and their risk judgment needs to be clarified, for example, by studying the relationship between FWC scores and recidivism or measures of inpatient violence. Prospective research is strongly recommended, although a number of problems might be encountered. The most important problem is that clinical goals of risk assessment (i.e. risk management) will interfere with prospective research into predictive validity. Hart (1998) stated that predictions of violence are not passive assessments, but decisions that influence services delivered to individuals: 'Clinicians are bound – morally, ethically, and legally – to try to prove themselves wrong when they predict violence and take every reasonable action to prevent violence' (Hart, 1998, p. 365). Thus, when clinicians perform HCR-20 risk assessments it is very likely that the outcome influences decisions concerning probationary leave or termination of (mandatory) treatment and high risk patients will not be let out of the hospital. In conclusion, continuous effort in research will be needed to clarify the processes underneath coding structured risk assessment instruments and to improve the accuracy of risk assessment procedures in clinical practice.

ACKNOWLEDGEMENTS

The authors wish to thank all clinicians and researchers who participated in this study. Special thanks go to Cécile Vandeputte-van de Vijver who functioned as workshop trainer together with the first author and also participated as a researcher in the study.

Vivienne de Vogel, MD, Dr. Henri van der Hoeven Kliniek, Netherlands Institute of Mental Health and Addiction (Trimbos Institute), Utrecht, the Netherlands.
E-mail: vdevogel@hoevenstichting.nl

Corine de Ruiter, PhD, University of Amsterdam, Amsterdam, the Netherlands

Correspondence to Vivienne de Vogel

NOTES

[1] Eighteen of the 60 patients were sex offenders. We consider this group too small and will not present results on the interrater reliability of the SVR-20 at this moment.

[2] There are capacity problems in the Dutch tbs-system, resulting in a strong call on the forensic psychiatric institutions to complete the treatment of patients within a specific time period.

REFERENCES

Ackerman, M. J. (1999) *Essentials of Forensic Psychological Assessment*. New York: Wiley.

American Psychiatric Association (1994) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn. Washington, DC: American Psychiatric Association.

Belfrage, H. (1998) 'Implementing the HCR-20 Scheme for Risk Assessment in a Forensic Psychiatric Hospital: Integrating Research and Clinical Practice'. *The Journal of Forensic Psychiatry* 9: 328–38.

Belfrage, H. and Douglas, K. S. (2002) 'Treatment Effects on Forensic Psychiatric Patients Measured with the HCR-20 Violence Risk Assessment Scheme'. *International Journal of Forensic Mental Health* 1: 25–36.

Belfrage, H., Fransson, G. and Strand, S. (2000) 'Prediction of Violence using the HCR-20: A Prospective Study in Two Maximum-Security Correctional Institutions'. *The Journal of Forensic Psychiatry* 11: 167–75.

Boer, D. P., Hart, S. D., Kropp, P. R. and Webster, C. D. (1997) *Manual for the Sexual Violence Risk-20. Professional Guidelines for Assessing Risk of Sexual Violence*. Vancouver: British Columbia Institute against Family Violence.

Borum, R. (1996) 'Improving the Clinical Practice of Violence Risk Assessment: Technology, Guidelines and Training'. *American Psychologist* 51: 945–56.

Dernevik, M. and Douglas, K. S. (2002) *The Role of Context and Training in the Accuracy of Risk Factors.* Paper presented at the 12th European Conference on Psychology and Law. Leuven, Belgium: September 14 – 17.

Dernevik, M., Falkheim, M., Holmqvist, R. and Sandell, R. (2001) 'Implementing Risk Assessment Procedures in a Forensic Psychiatric Setting: Clinical Judgement Revisited'. In Farrington, D. P. Hollin, C. R. and McMurran, M. (eds) *Sex and Violence: The Psychology of Crime and Risk Assessment*. London: Routledge, pp. 83 – 101.

Douglas, K. S. and Belfrage, H. (2001) 'Use of the HCR-20 in Violence Risk Management: Implementation and Clinical Practice'. In Douglas, K. S. Webster, C. D. Hart, S. D. Eaves, D. and Ogloff, J. R. P. (eds) *HCR-20 Violence Risk Management Companion Guide*. Vancouver: Mental Health, Law, and Policy Institute, pp. 41 – 58.

Douglas, K. S. and Webster, C. D. (1999a) 'Predicting Violence in Mentally and Personality Disordered Individuals'. In Roesch, R. Hart, S. D. and Ogloff, J. R. P. (eds) *Psychology and Law: The State of Discipline*. New York: Kluwer Academic, pp. 175–239.

Douglas, K. S. and Webster, C. D. (1999b) 'The HCR-20 Violence Risk Assessment Scheme. Concurrent Validity in a Sample of Incarcerated Offenders'. *Criminal Justice and Behavior* 26: 3–19.

Douglas, K. S., Ogloff, J. R. P., Nicholls, T. L. and 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: 917–30.

Fleiss, J. L. (1986) *The Design and Analysis of Clinical Experiments*. New York: Wiley.

Greenberg, S. and Shuman, D. (1997) 'Irreconcilable Conflict Between Therapeutic and Forensic Roles'. *Professional Psychology: Research and Practice* 28: 50–7.

Hare, R. (1991) *Manual for the Hare Psychopathy Checklist-Revised*. Toronto, Ontario: Multi-Health Systems.

Hare, R. D., Clark, D., Grann, M. and Thornton, D. (2000) 'Psychopathy and the Predictive Validity of the PCL-R: An International Perspective'. *Behavioral Sciences and the Law* 18: 623–45.

Hart, S. D. (1998) 'Psychopathy and the Risk for Violence'. In Cooke, D. J. Forth, A. E. and Hare, R. D. (eds) *Psychopathy: Theory, Research and Implications for Society*. Dordrecht: Kluwer, pp. 355 – 373.

Hildebrand, M. and de Ruiter, C. (in press) 'PCL-R Psychopathy and its Relation to DSM-IV Axis I and Axis II Disorders in a Sample of Male Forensic Psychiatric Patients in the Netherlands'. *International Journal of Law and Psychiatry*.

Hildebrand, M., de Ruiter, C. and van Beek, D. (2001) *SVR-20. Richtlijnen voor het beoordelen van het risico van seksueel gewelddadig gedrag* [SVR-20. Guidelines for the Assessment of Risk of Sexual Violence]. Utrecht: Forum Educatief.

Holmqvist, R. and Armelius, B. A. (1994) 'Emotional Reactions to Psychiatric Patients'. *Acta Psychiatrica Scandinavica* 90: 204–9.

Holmqvist, R. and Fogelstam, H. (1996) 'Psychological Climate and Countertransference in Psychiatric Treatment Homes'. *Acta Psychiatrica Scandinavica* 93: 288–95.

Laws, D. R., Hudson, S. M. and Ward, T. (2000) *Remaking Relapse Prevention: A Sourcebook*. London: Sage Publications.

Litwack, T. R. and Schlesinger, L. B. (1999) 'Dangerousness Risk Assessments: Research, Legal, and Clinical Considerations'. In Hess, A. K. and Weiner, I. B. (eds) *The Handbook of Forensic Psychology*. New York: Wiley, pp. 171–217.

Lösel, F. (1998) 'Treatment and Management of Psychopaths'. In Cooke, D. J. Forth, A. E. and Hare, R. D. (eds) *Psychopathy: Theory, Research and Implications for Society*. Dordrecht: Kluwer, pp. 303 – 354.

McGraw, K. O. and Wong, S. P. (1996) 'Forming Inferences About Some Intraclass Correlation Coefficients'. *Psychological Methods* 1: 30–46.

Pfohl, B., Blum, N. and Zimmerman, M. (1995) *Structured Interview for DSM-IV Personality, SIDP-IV*. Iowa: University of Iowa.

Philipse, M., de Ruiter, C., Hildebrand, M. and Bouman, Y. (2000) *HCR-20. Beoordelen van het risico van gewelddadig gedrag. Versie 2* [HCR-20. Assessing the Risk of Violence. Version 2]. Nijmegen/Utrecht: Prof. Mr. W. P. J. Pompestichting/Dr. Henri van der Hoeven Stichting.

Quinsey, V. L., Harris, G. T., Rice, M. E. and Cormier, C. A. (1998) *Violent Offenders: Appraising and Managing Risk*. Washington DC: American Psychological Association.

Strand, S., Belfrage, H., Fransson, G. and Levander, S. (1999) 'Clinical and Risk Management Factors in Risk Prediction of Mentally Disordered Offenders: More Important Than Actuarial Data?' *Legal and Criminological Psychology* 4: 67–76.

Webster, C. D., Douglas, K. S., Eaves, D. and Hart, S. D. (1997a) 'Assessing Risk of Violence to Others'. In Webster, C. D., and Jackson, A. (eds) *Impulsivity. Theory, Assessment and Treatment.* New York: The Guilford Press, pp. 251–272.

Webster, C. D., Douglas, K. S., Eaves, D. and Hart, S. D. (1997b) *HCR-20. Assessing the Risk of Violence. Version 2*. Burnaby, BC, Canada: Simon Fraser University and Forensic Psychiatric Services Commission of British Columbia.

Whyte, C. R., Constantopoulos, C. and Bevans, H. G. (1982) 'Types of Countertransference Identified by Q-Analysis'. *British Journal of Medical Psychology* 55: 187–201.