# Predicting criminal recidivism

## Empirical studies and clinical practice in forensic psychiatry

The printing of this thesis, and the research reported in it, were financially supported by Pompestichting Center for Forensic Mental Health, Nijmegen, The Netherlands

THESIS RADBOUD UNIVERSITEIT NIJMEGEN

ISBN 90-9019246-8
NUGI 757
NUR 740, 776

COVER ART
"Claw nebula". Fractal image by Roger Johnston.

COVER DESIGN AND TEXT LAYOUT
Martien Philipse

Printed in The Netherlands by Febodruk, Enschede.

Voor mijn ouders

# Predicting criminal recidivism

## Empirical studies and clinical practice in forensic psychiatry

Een wetenschappelijke proeve op het gebied van de
Sociale Wetenschappen

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de
Radboud Universiteit Nijmegen,
op gezag van de Rector Magnificus prof. dr. C.W.P.M. Blom,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op

donderdag 23 juni 2005

des namiddags om 1.30 uur precies

door

Martinus Wilhelmus Gertrudis Philipse

geboren op 1 december 1966
te Boxmeer

PROMOTORES

Prof. dr. C.P.F. van der Staak
Prof. dr. W. van den Brink, Universiteit van Amsterdam

COPROMOTOR

Dr. M.W.J. Koeter, Universiteit van Amsterdam

MANUSCRIPTCOMMISSIE

Prof. Dr. H.L.I. Nijman (voorzitter)
Prof. Dr. H.J.C. van Marle, Erasmus Universiteit Rotterdam
Prof. Dr. B.C.M. Raes, Vrije Universiteit Amsterdam

Prediction is very difficult, especially of the future.
*Niels Bohr*

"Why," said Mr. Pickwick, "I may have formed some ideas upon the subject, but, as I have never submitted them to the test of experience, I should be sorry if you were to regulate your proceedings by them."
*Charles Dickens, 'The Pickwick Papers'*

# ☙ Contents

# Outline of this thesis

This thesis is comprised of three major sections. The first of these offers a general introduction, in which the present state of the art in violence risk assessment research is described. Core concepts in this area are addressed in some detail.

The second section contains three empirical studies, describing the development, psychometric characteristics and predictive validity of a dynamic violence risk assessment tool for the Dutch forensic mental health system called *terbeschikkingstelling* (TBS). The core question of this project was, whether clinically meaningful, dynamic predictors of risk could be derived from existing TBS practice, that would add significant predictive power to static risk factors.

The first paper investigates the dimensional structure of the dynamic risk assessment tool that was developed to answer the research question. It furthermore investigates whether unstructured clinical risk ratings are significantly related to dynamic items from this tool as well as their underlying dimensions.

The second paper explores psychometric characteristics of the research instrument. Interrater and retest reliability are assessed. Also, it is studied whether the instrument is able to distinguish patients recently admitted to hospital from those who were recently discharged, under the assumption that these groups represent relatively higher risk versus relatively lower risk patients and thus tell us something about the instrument's capacity to register differences in risk.

The third paper is the prediction study proper. In this part of the research, it is investigated if an increase of predictive validity can be achieved by adding dynamic dimensions and items from the research instrument, as well as unstructured clinical risk assessments, to a basic set of static predictors. For this purpose, a prospective follow-up study was conducted among 151 TBS-dischargees, of whom reconviction data were retrieved after a 5.5 to 8.5 year follow-up. Descriptive data regarding reoffending are offered in addition to the predictive validity analyses.

The third and final section of this thesis summarizes the main issues addressed in the Introduction, as well as the main findings from the three empirical studies. Several possible interpretations of findings are offered, and their implications are assessed both from research and clinical perspectives. Suggestions for future research focus are included. Finally, some general thoughts on ethics and the limits of predictability are offered.

Though the content of these sections is in many ways interrelated, all three, as well as the three studies in section II, can be read independently of each other. For a brief overview of the main issues in each section, the overall study findings and general conclusions, the reader is referred to the summaries provided at the end of this thesis.

## Legal context - TBS

The study reported in this thesis was conducted in special hospitals for the execution of *terbeschikkingstelling* or TBS. The term *terbeschikkingstelling* is in itself meaningless, being a truncation of the original name of the measure, *terbeschikkingstelling van de Regering*, literally: 'being put at the Government's disposal'. TBS is a judicial measure provided for offenders who are not held (fully) accountable for their crime due to a mental disorder. In such cases, the judge may choose to pass a TBS sentence instead of or in addition to a prison sentence, if two conditions are met: (1) the offence warrants a prison sentence of at least 4 years, and (2) there is a high risk of reoffending.

TBS entails involuntary admission to a special hospital, initially for the duration of 2 years, though the length of stay is in effect indeterminate. The hospital's goal is to protect society from any further danger posed by the patient. In the short run, this is achieved by detaining the patient within the confines of the facility; and in the longer run, by reducing his[1] risk of reoffending through treatment, so that he can be gradually reintroduced into the community. Each case is reassessed in court at least bi-annually. The judge decides whether treatment has been sufficiently successful to terminate TBS, or whether the measure needs to remain in force for another 1 or 2 years. This decision relies heavily upon the advice and risk assessment provided by an expert witness from the hospital where the patient is treated. However, the patient and his lawyer may offer opposing views and evidence, and the judge furthermore takes into account the duration of the TBS sentence relative to the average prison sentence served for similar offences ('proportionality'). These additional factors account for the occasional TBS terminations against hospital advice.

[1] Throughout this thesis, 'he', 'him' and 'his' refer to both men and women.

# I ☙ General introduction

## *Risk assessment: core concepts and state of the art*

This section is an extended and adapted version of a chapter entitled 'Risicotaxatie: Oplossingen voor een onoplosbaar probleem' [Risk assessment: Solutions to an insoluble problem], in: Oei, T.I., & Groenhuijsen, M.S. (2003). *Actuele ontwikkelingen in de forensische psychiatrie*. [Current developments in forensic psychiatry]. Deventer: Kluwer (pp. 229-255).

# I.1 Introduction

A risk assessment is a statement on the likelihood that a certain undesirable event will take place in the future. The process leading to such an estimate is generally also referred to as risk assessment, whereas interventions designed to prevent the event are labeled risk management. These activities are a routine part of every branch of forensic mental health. Decisions on treatment interventions, freedom of movement, and the imposition or termination of court-ordered treatment measures are all guided by risk assessments.

This is certainly true in the context of the Dutch *terbeschikkingstelling* (TBS) system. This court-ordered treatment measure can be imposed on any perpetrator of an offence punishable with a prison sentence of at least 4 years, if he or she was suffering from a mental illness at the moment the offence was committed, and is deemed at risk to reoffend. The measure is of indefinite duration; in fact, the length of stay in a TBS-hospital is primarily determined by treatment progress, resulting, it is presumed, in reduced risk of reoffending. In principle, the court will only end the measure when sufficiently convinced that risk of reoffending is significantly reduced – though in practice exceptions to this rule are far from rare. The hospital is under a legal obligation to provide annual or bi-annual reports on the current likelihood that the patient will reoffend when released into the community.

Such risk assessments in judicial contexts are in many ways comparable to prognostic tasks in other fields, such as weather forecasting in meteorology (Monahan & Steadman, 1996). There is however one crucial difference. In forensic risk assessment, those who are the future victims of wrong decision-making are in no way informed about the assessment process. They cannot chose to take precautions (or leave them) the way the viewer of the weather forecast can. Moreover, if they cross the path of an unduly discharged, potentially violent forensic patient, the effect is likely to be far more serious than getting wet in an unforeseen downpour. For that reason alone it is of tremendous importance that forensic risk assessments be carried out as carefully and precisely as possible.

This however is not a simple task – predicting the future never is, as Niels Bohr once succinctly pointed out. Risk assessment in correctional contexts has been the subject of ample research during the last three decades and yet the status quo cannot be summarized more pointedly than in Hart's phrase:

> *"The state of the science simply does not allow the conclusion that a solution has been found for the problem of risk assessment"* (1999, p. 487).

Nevertheless, the accumulated research has provided some insights that have changed the practice of risk assessment in many countries, including the Netherlands. Existing procedures have been reevaluated, and empirically supported risk assessment tools have been widely adopted. This introductory chapter reviews the core concepts regarding risk assessment, describes different types of risk assessment, summarizes several ways of evaluating predictive validity, and finally sums up some recent developments and current issues.

# I.2 ❧ Core concepts

## I.2.1 From dangerousness to risk assessment

Though the use of the term risk assessment is common practice nowadays, in the past prognoses about the potential for future violence have been called by many different names. These changes in terminology reflect insights derived from accumulating empirical evidence.

Until the beginning of the 1990's the term 'dangerousness' was in common use. Unfortunately, this concept was riddled with definitional problems, of which Scott (1977) provides an example. In his view, only acts that cannot be anticipated or aborted are dangerous. Accordingly, he defines dangerousness as "an unpredictable and untreatable tendency to inflict or risk serious, irreversible injury or destruction, or to induce others to do so" (p. 128). Surprisingly, this does not deter him from dedicating a full paragraph to the prediction of dangerousness. It is obvious that his view is of little help to clinicians in forensic hospitals who have to provide treatment aiming at risk reduction.

The definition offered by Kozol *et al.* (1972) is somewhat less problematic: dangerousness is "a potential for inflicting serious bodily harm on another" (p. 372). Yet it highlights another drawback of the dangerousness concept: it tends to describe the violence potential as an attribute of the individual, and downplays the role of context factors (Monahan, 1981). Due to such conceptual and definitional problems, researchers gradually abandoned the notion of 'dangerousness'. In clinical practice and legal contexts, however, the term is still widely used (e.g., Swartz *et al.*, 1999).

In 1981 John Monahan published his groundbreaking monograph entitled 'The prediction of violent behavior'. This title encapsulates another issue that has dominated research on violent behavior for many years: the concept of prediction. All the studies that Monahan reviews in this book were true prediction studies, in which yes-or-no statements by clinicians or statistical tools regarding future patient violence were linked to subsequent actual violence. On a conceptual level this dichotomization makes sense: the eventual outcome is dichotomous (either an individual reoffends or he doesn't) so it seems logical that the prognostic statement should also divide subjects into those who are expected to reoffend and those who are not.

However, predicting the future in this literal sense with any degree of meaningful detail is simply impossible (Hart, 2001). We may be able to predict that at least one individual who at some point in time received forensic treatment will reoffend in the next 12 months, and very likely be correct; but such a 'prediction' has no practical value. What we really want to predict is which particular individual will reoffend, and preferably also when, where and how. Unfortunately, such a level of precision precludes any high degree of accuracy: the error rate rapidly increases as the definition of the outcome is made more specific.

It is therefore preferable to phrase a prognosis in a way that incorporates the fundamental uncertainty of the future, and this is what the concept of risk assessment achieves. Instead of an absolute yes-or-no prediction, a relative likelihood is formulated that a specific individual will commit a new offence in the future. This statement can for instance be expressed in a percentage, or in categories (e.g., high, medium, low risk). It is never true or false, as it keeps the possibility of either kind of outcome open. One essential characteristic that distinguishes risk from mere chance, is the fact that it is calculable (Ewald, 1981). As will be seen in the following, this fact has had far reaching consequences for forensic risk assessment practice.

## I.2.2 Relative risk and absolute outcome

The evolving insights into the nature of risk have strongly influenced the way risk assessment procedures are designed. However, the notion of prediction was not in fact abandoned. Relative risk assessments may be realistically indefinite and statistically elegant, but daily practice in forensic mental health settings continuously requires absolute, black-and-white decisions. A patient is granted leave or not; a treatment measure is extended or terminated; and of course, eventually the patient will reoffend or not. On an individual level there is no such thing as a '46% discharge' or a '67% relapse' (Berlin *et al.*, 2003; Hart, 2003). Indeed, the vast majority of risk assessment studies also continues to use dichotomous outcome criteria.

It is inevitable that risk assessment tools are eventually used to generate predictions rather than relative statements on risk. Patients are classified as either likely reoffenders or likely non-reoffenders. When such a classification is made, the four possible prediction-outcome combinations can be summarized in a 2 by 2 table (Table 1).

TABLE 1. PREDICTION-OUTCOME CONTINGENCIES

| | | PREDICTION | |
|---|---|---|---|
| | | Non reoffender | Reoffender |
| OUTCOME | No new offence | True negative | False positive |
| | New offence | False negative | True positive |

As is clear from this table, predictions can result in two types of error. First, a person can be discharged after having been assessed as low risk, and nevertheless commit a (serious) new offence. Such so-called false negative predictions often result in societal upheaval, and obviously have dramatic consequences for the people who fall victim to them. In general there tends to be less worry about the second error, false positive predictions, which lead to the unnecessary detention of patients who are assessed as high risk, but who would not have reoffended had they been released. False positive predictions result in inadequate allocation of treatment resources, and in the wrongful curtailment of the individual's civil liberties. It should be noted that false negative predictions are highly visible in the form of unfortunate re-offenses, but that false positive predictions go unnoticed simply because they cannot be detected.

These two prediction errors have a 'trade-off' relationship: if one goes up, the other goes down. This is easy to see if one considers a situation where the management of an institution seeks absolute certainty that no recidivism should occur (zero false negatives), and therefore takes the drastic measure of not discharging a single patient. Obviously this results in the detention of large numbers of patients who would not have reoffended had they been discharged (false positives).

An important implication arising out of risk assessment-based decision making and its associated errors, is that the accuracy of predictions in in-patient or prison settings can only be assessed in select samples. Studies can only include offenders who were discharged from an institution, which nearly always involves some risk assessment procedure and risk management decisions. In consequence, the validity of risk assessments is routinely evaluated in samples that over-represent true and false negatives while under-representing true and false positives, an issue rarely addressed in research reports (Litwack, 2001).

Foregoing this problem unfortunately requires experimental setups (e.g., random discharge of patients) that are ruled out by ethical considerations. Historically, there have been a few occasions where the required conditions occurred naturally, most famously in the US Supreme Court Baxstrom vs. Herold ruling of 1966 (303 US 107), which resulted in the immediate discharge into non-secure settings of 967 forensic mental health patients who were considered dangerous by their therapists. Cocozza & Steadman (1976) seized the opportunity to study the outcome, and found that clinical predictions of recidivism, unmediated by subsequent risk management strategies, were completely unrelated to actual reoffending. Unfortunately this study contained some serious methodological flaws (Greenland, 1985), and it is questionable to what extent its findings still have relevance 30 years later. Recently, Canton *et al.* (2004) studied the validity of several risk assessment procedures in an untreated sample of offenders in The Netherlands, and in contrast to the bulk of risk assessment studies found, that under these circumstances unstructured clinical risk judgments had significant predictive validity and performed equally well as a structured risk assessment tool. This finding supports the assumption that samples less influenced by selective risk management strategies may yield different results in risk assessment evaluation than the usual select samples.

## I.2.3 Follow-up time

Risk assessment evaluation is also influenced by the length of follow-up time. This is the time that elapses between the moment that risk factors are assessed, and the moment that recidivism data are retrieved. In most risk assessment studies follow-up refers, more precisely, to the time at risk: the true time the person has been in the community and at risk to commit offences. The time at risk is the follow-up time corrected for periods of incarceration, premature decease, and similar situations that interfere with the possibility of committing criminal acts. Obviously, the longer the time at risk, the greater the likelihood that the target behavior will occur. For instance, patients discharged from TBS in general have an 8% risk of being reconvicted for any serious new offence within 1 year; the probability of reconviction within 5 years, however, is 17% (Leuw, 1999). Clearly, the choice of follow-up time has a direct effect on the balance of prediction-outcome combinations shown in Table 1 (p. 20). In the TBS example just quoted, there is a 9% likelihood that a patient who after a 1 year follow-up is classified as a true negative (or a false positive in case of TBS termination contrary to hospital advice), will be classified as a false negative (or true positive) after 5 years.

Follow-up time not only influences the percentage of recidivists, but also directly affects the methods than can be applied to assess risk. In (very) short term risk assessment, behavioral models may be available that allow fairly accurate prognostications for the (very) near future. Like the meteorologist who uses data on present atmospheric pressure, winds, and cloud formations to generate a reasonably accurate forecast for the next one or two days, clinicians in forensic hospitals may for instance use an offence script to check if signs of imminent danger are present. An offence script describes the combination of behaviors, thoughts, feelings and situations typically preceding a relapse for a particular individual (Van Beek, 1999). Many daily safety decisions, such as allowing a patient to leave the hospital perimeter without escort, implicitly rely on similar short term risk assessment mechanisms. Such assessments may well be fairly accurate (McNiel & Binder, 2002). Earlier research has shown that acute symptom profiles predict short term violence (McNiel & Binder, 1994). Mossman (1994) concluded that clinical risk assessments were as good as alternative prediction models and performed moderately well when assessing the risk for violence within one year.

However, just as weather forecasts cannot tell us what the weather will be on a particular day in three years time, offence scripts or their informal clinical equivalents are not well suited to assess the risk of a new offence years in advance. They only describe short term causal mechanisms. Long term risk assessment requires other risk indicators, that are nomothetically rather than ideographically founded. Grubin & Wingate (1996) have observed that in long term risk assessment it is not 'states', the value of changeable characteristics in the here and now, but 'traits', the status of characteristics that are more stable over time, that determine the predictive validity of the assessment. To complete the analogy, the meteorologist may not be able to use his knowledge of present atmospheric conditions to tell us what the weather will be like on a particular October day in three years time; but historical data on October weather do enable him to estimate the likelihood that it will be a rainy day. Similarly, follow-up time in risk assessment has an effect on the type of risk factors that are useful (i.e., valid), and the type of estimates or predictions that can be made with them. In summary, it seems that fairly specific predictions can be made for the near future mainly using state or dynamic parameters, whereas general predictions can be made for the more distant future primarily based on trait or static parameters.

## I.2.4 Outcome measures

A crucial element in any risk assessment is the targeted outcome. Over the years, risk assessment studies have included a bewildering variety of outcome measures, differing with regard to the nature of the behavior, the context of its occurrence, and its (judicial) aftermath. Recidivism has been defined as anything from in-patient transgressions of hospital rules (e.g., Hildebrand, De Ruiter & Nijman, 2004) to serious violence in the community (e.g., Grann *et al.*, 2000). Violence has been defined as acts that cause serious physical harm to others (e.g., Ward & Eccleston, 2000), but also as any act that would induce fear in an average person and may result in legal prosecution (Webster *et al.*, 1997). Some studies included all target behaviors reported by the individual, his family, and official records (e.g., Monahan *et al.*, 2001), others focused solely on violent criminal behavior that resulted in reconviction (e.g., Philipse *et al.*, this thesis).

Several authors have stressed the need for further differentiation of the outcome criterion, and the development of tools that represent a more nuanced conception of reoffending risk. Hilterman (2002) pointed out the importance of the seriousness of relapse. Mulvey & Lidz (1995) argued that risk comprises five distinct elements: nature, likelihood, frequency, seriousness and imminence. Monahan & Steadman (1994) suggested that different types of violence may correspond to different sets of risk factors.

Studies examining the predictive validity of risk assessments as a function of their precision with regard to the outcome criterion are as yet rare, but quite informative. For example, Sjöstedt & Grann (2002) showed that risk assessment instruments that have a moderate overall performance, yield results varying from high accuracy to no better than chance when the outcome criterion is broken down into measures of frequency, imminence, nature and severity of reoffending. Kroner & Mills (2001), when comparing the predictive validity of 5 different risk measures, found that all measures were better at predicting revocations due to parole violation than reconvictions for any kind of violent or other criminal activity. Douglas & Ogloff (2003) concluded that risk assessments specified with regard to imminence or severity of reoffending did not add predictive validity to an unspecified overall risk assessment.

Studies examining the predictive validity of instruments that aim to assess the risk of a particular type of offence is less scarce, especially regarding risk of sexual violence. Thus far, this literature does not seem to support the conclusion that such a specific focus results in improved predictive power (Philipse *et al.*, 2001; Hanson & Thornton, 2000; Rice & Harris, 1997). For instance, Quinsey *et al.* (1998) found that their Sex Offender Risk Appraisal Guide (SORAG) performed only marginally better than the Violence Risk Appraisal Guide (VRAG) when assessing sexual violence risk. Nunes *et al.* (2002) found the SORAG to be equally (moderately) accurate at predicting specific sexual reoffending and non-specific general reoffending. Sjöstedt & Långström (2002) found that the Sexual Violence Risk – 20 (SVR-20, Boer *et al.*, 1995) had (limited) predictive validity with regard to non-sexual reoffending, but did not predict sexual reoffending at all. Finally, Hanson & Thornton (2000) concluded from a meta-analysis that all risk levels of their sex offender risk assessment scale, the Static-99, were consistently more predictive of violent than of sexual reoffending.

One explanation for these findings might be that these instruments incorrectly equate the type of offense for which the offender is presently detained with the expected type of offence in case of relapse. Sexual violence risk assessment tools are commonly used to assess risk in sex offenders, on the implicit assumption that if such offenders reoffend, they will do so with a sex offence. This is however not necessarily the case. For instance, Sjöstedt & Långström (2002) found that in an average 7.7 year follow-up of their sample of 51 rapists, 7 reoffended with a new sex offence, but 10 committed a non-sexual violent offence, and 3 subjects were reconvicted for both offence types. De Vogel *et al.* (2003) in an average 11.7 year follow-up of 122 sex offenders found new charges for sex offences for 39% of the sample, but new charges for non-sexual violence for 46%, a balance very similar to that in the Sjöstedt & Långström study. These findings are also in accordance with those in the sex offender recidivism meta-analysis by Hanson & Bussière (1998). In the 61 studies they included, the average rearrest or reconviction rate for sex offences among rapists during a median follow-up of 4 years was 18.9%, while non-sexual violent reoffending occurred in 22.1% of cases.

### *I.2.4.1 Dark numbers*

A discussion of outcome measures in risk assessment research cannot ignore the issue of dark numbers, which is regarded as an urgent methodological problem (Grann, 1998; Monahan & Steadman, 1994). With some exceptions, most notably the McArthur Risk Assessment study, which included peer- and self-reported crime (Monahan *et al.*, 2001), officially recorded crime is the most commonly used outcome measure in risk assessment research. The reason for this is simple: official crime records provide the most easily accessible database of offending behavior. Also, retrieval of these data does not depend on the willingness of subjects in the study to cooperate. Moreover, a renewed conviction provides the most solid indication that the individual at hand did in fact commit the new offence.

However, many violent or sexual transgressions are never brought to the attention of the police, and those that are do not always result in charges, let alone in convictions. These offences go unnoticed when official files are consulted to determine outcome, compromising criterion variables in many risk assessment studies. The extent of this so-called 'dark number' can by its very nature not be ascertained, but estimates can be derived from victimization surveys and police statistics. In The Netherlands in 2003, 64% of those who were victim of a violent assault reported this to the police (Uitvoeringsconsortium Projectbureau Politiemonitor, 2003). In that same year, 52% of violent offences were solved (source: Centraal Bureau voor de Statistiek). This would suggest that about one third of violent incidents eventually appeared in official crime records, a proportion similar to that estimated for North America by Monahan in 1981.

If two thirds of all violent crime goes unrecorded, this would certainly constitute a major impediment to risk assessment research that depends on judicial and police files. However, this figure needs some qualification. First, it is instructive to explore the reasons for not reporting offences. A survey of the quality of contacts between the public and the police conducted in the Netherlands in 2003, showed that in 46% of unreported offences, the victim thought that the offence was not serious enough to go to the police. An additional 10% found reporting the offence "too much trouble" (Uitvoeringsconsortium Projectbureau Politiemonitor, 2003). This suggests that over half of unreported offences may be relatively innocuous, or at least not serious enough to warrant inclusion in the outcome criterion of violence risk assessment research. Different types of offences may however be differently affected. Sex offences especially seem vulnerable to under-reporting (Marshall & Barbaree, 1988).

A second fact worthy of note is that offences are not evenly distributed among the population (Monahan, 1981). If only one third of offences enter official records, this does not imply that only one third of offenders are detected. Many unsolved crimes were committed by offenders who appear in judicial or police files due to arrest for another offence. Wolfgang (1978, cited in Monahan, 1981), found three violent acts per single arrest. In a 4-year follow up of a random sample of 13,000 convicts in the UK, Marshall (1994) found that 7% of offenders with previous sex offences were responsible for 31% of subsequent sex offence reconvictions. Though this study focused on solved crime (reconviction), it nevertheless clearly illustrates that many reoffenders commit more than one crime.

Furthermore, it should be noted that an unspecified number of unsolved crimes are committed by first offenders rather than recidivists. Consequently, this part of the dark number does not affect risk assessment research among persons who have already been in the correctional or forensic mental health system. A 2001 survey within the Dutch police district Haaglanden showed 63.2% of arrested offenders to be 'starters' (Versteegh *et al.*, 2003); even if this number applies only roughly to unsolved crime as well, it still greatly reduces the dark number problem in reoffending risk research.

Finally, it seems reasonable to assume that reoffenders, that is, individuals who have been in contact with the law previously, have a higher likelihood than others to be identified as possible suspects in a crime investigation: they are already known to the police.

Together, these considerations suggest that official records may not underestimate the true number of reoffenders, relevant to risk assessment research in forensic settings, to the extent that has sometimes been assumed. Even more importantly, the assumption that official crime statistics misrepresent true criminality in such a way that predictor-outcome relations are significantly altered may itself be incorrect. In the so-called Cambridge Study on criminal careers, it was shown repeatedly that self-reported and officially recorded offending each resulted in similar predictor sets (Farrington, 2001). The main effect of the dark number, depending on its extent, is therefore the reduction of the so-called base rate of (re)offending, the percentage of (re)offenders identified in the population. Smaller base rates reduce the likelihood that statistically significant links will be found between predictors and outcome. Klassen & O'Connor (1987, cited in Monahan & Steadman, 1994) for instance found that including patients' self-reported violence in their outcome measure yielded a more than 25% increase in predictive accuracy over official records alone. The mechanism explaining this effect is further elucidated in the next paragraph.

## I.2.5 Base rates

Once a research sample, a target behavior and a follow-up time have been determined, these three in turn allow the determination of the base rate of the target behavior. The base rate is the proportion of individuals in a population that show a certain behavior or characteristic within a given period of time. In terms of violence risk assessment, the base rate commonly refers to the percentage of offenders or patients who reoffend violently or sexually during a particular time period.

Base rates have far-reaching consequences for the possibilities of any risk assessment procedure to make meaningful additions to results achieved by chance alone (Quinsey *et al.*, 1998; Grubin & Wingate, 1996). In a population with a very low base rate, say, reconviction for a violent offence within 3 years for 1 out of every 100 cases, one can hardly do better than assessing a zero likelihood of reconviction for every individual. This assessment will be 99% accurate. Not only can a risk assessment method add just 1% accuracy at best, it will also need to identify the few reoffenders very consistently (high sensitivity), and without designating non-reoffenders as reoffenders (high specificity), in order to establish a statistically significant improvement over the 'zero-risk' assessment. This requires a quality of content and methodology that is not even remotely approached by any risk assessment tool now in existence.

Hence, it follows that the optimum conditions for validating a risk assessment procedure occur when the base rate of the target behavior equals 50%. At this base rate, the likelihood of generating accurate assessments based on chance alone is minimal. Consequently, risk assessment instruments do not need to meet very high performance standards to improve accuracy in a meaningful (statistically significant and clinically relevant) way.

Existing studies have been conducted in populations with very different base rates. Some authors have chosen to optimize base rates deliberately by over-representing reoffenders in their sample (e.g., Hanson & Harris, 2000;

Douglas *et al.*, in press). Depending on the statistical measures used, this procedure runs the risk of artificially inflating the accuracy of the instrument. In TBS, the setting of the empirical study reported in this thesis, the base rate of reconviction of ex-patients to a prison sentence of at least six months and/or TBS within 5 years after discharge, has been fairly constant over many years at around 20% (Leuw, 1999; Van Emmerik, 1985). For a specific subsample of sexually deviant, psychopathic sex offenders discharged from one particular TBS-hospital, De Vogel *et al.* (2003) reported a base rate of 74% for any reconviction after an average 12 year follow-up.

## I.2.6 Measuring predictive validity

A great variety of statistical procedures is available to the researcher who wants to examine the validity of risk assessments. Risk assessment results may be linked to measures of frequency or severity of reoffending through Pearson correlations (e.g. Hilterman, 2002), or to dichotomous outcome using point-biserial correlations (e.g. Hildebrand, de Ruiter & de Vogel, 2004; Douglas *et al.*, in press). Odds ratios may be used to explore if and how increased risk ratings correspond to increased reoffending (e.g., Sjöstedt & Långström, 2001). Cohen's *d* can be calculated to evaluate the distance (measured in standard deviations) between the mean risk scores of reoffenders and non-reoffenders (e.g., Belfrage *et al.*, 2000). Proportions of reoffenders in groups defined by the presence or absence of risk factors can be tested for statistically significant differences using a Chi-square test (e.g., Scalora & Garbin, 2003). Other approaches, such as measuring agreement between prediction and outcome using kappa, or comparing reoffender and non-reoffender mean scores using Student's t-test are less common but also feasible. Most studies report several such measures of association.

Each of these procedures has its merits and limitations; none of them, however, offers direct clues for decision making. A significant Pearson correlation, or the finding that reoffenders' mean scores on a risk assessment instrument are significantly higher than those of non-reoffenders, are in themselves interesting, but do not tell the user of the instrument which score warrants a patient's discharge. Furthermore, none of these measures takes the population base rate into account, so that prediction results may capitalize on chance effects in the sample. Both these problems are to some extent solved by a statistical measure now commonly used in risk assessment research, the Receiver Operating Characteristic.

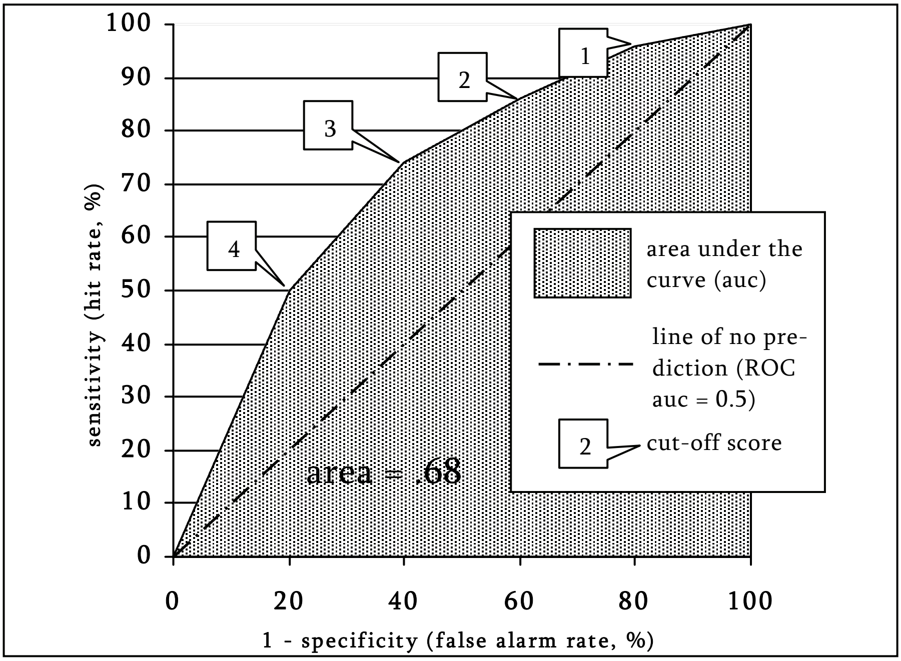
### *I.2.6.1 Receiver Operating Characteristic*

The Receiver Operating Characteristic (ROC) was introduced into the forensic risk assessment field by Mossman in 1994. This statistical technique stems from signal detection theory, and was originally used as a measure for the precision of radar systems at different system settings – hence the name. The ROC-curve is simply the visual representation of the trade-off between false positive and false negative predictions. Every value that the instrument can assume is successively used as a cut-off, and for each of these cut-offs a contingency table like Table 1 (p. 20) is produced. The percentage of 'hits' (true positives, sensitivity) is then plotted on the Y-axis against the percentage of 'false alarms' (false positives, 1- specificity) on the X-axis at every cut-point, resulting in a curve (see Figure 1 on the following page). The area under the curve (AUC) is commonly applied as a scalar measure of predictive validity (Hanley & McNiel, 1982), and can range from 1.0 (perfect prediction) through 0.5 (prediction no better than chance) to 0.0 (perfect negative prediction). A convenient rule of thumb for interpreting AUC values is to think of them as the likelihood that a randomly chosen reoffender will have a higher risk assessment score than a randomly chosen non-reoffender. If the ROC AUC equals 1.0, the lowest scoring reoffender has a higher score than the highest scoring non-reoffender, and cut-off values between their scores will distinguish recidivists from non-recidivists with 100% accuracy. Note that even in this unlikely case, only cut-off values falling between the highest non-reoffender score and the lowest reoffender score yield perfect predictions – other cut-offs will introduce prediction error.

Soon after Mossman's 1994 publication, ROC analysis was firmly established as the first choice statistical procedure in risk assessment research (e.g., Rice & Harris, 1995), owing to several advantages it has over other measures. First, by showing the balance of hits and false alarms at every cut-point in the instrument, it facilitates the identification of the cut-off corresponding to the optimum balance between false positive and false negative predictions. Several cost-benefit formulae have been described to help determine the best cut-point (e.g., Metz, 1978; Halpern *et al.*, 1996). Second, the construction of the ROC curve is straightforward and easily grasped, is available in widely used statistical software packages like SPSS, and has the added appeal of visual representation in a graph.

The most important advantage of ROC-curves is that, for a given instrument, the trade-off between false positives and false negatives has been shown to remain constant even when selection ratios (e.g., percentage of patients retained in hospital) and base rates (e.g., percentage of patients

FIGURE 1. A HYPOTHETICAL ROC-CURVE AND ASSOCIATED CONTINGENCY TABLES



Contingency table at cut-off 1

| | | Prediction | | |
|---|---|---|---|---|
| | | no | yes | |
| Outcome | no | **10** | **40** | 50 |
| | | *(20)* | *(80)* | *(100)* |
| | yes | **2** | **48** | 50 |
| | | *(4)* | *(96)* | *(100)* |
| | | 12 | 88 | 100 |

Contingency table at cut-off 2

| | | Prediction | | |
|---|---|---|---|---|
| | | no | yes | |
| Outcome | no | **20** | **30** | 50 |
| | | *(40)* | *(60)* | *(100)* |
| | yes | **7** | **43** | 50 |
| | | *(14)* | *(86)* | *(100)* |
| | | 27 | 73 | 100 |

Contingency table at cut-off 3

| | | Prediction | | |
|---|---|---|---|---|
| | | no | yes | |
| Outcome | no | **30** | **20** | 50 |
| | | *(60)* | *(40)* | *(100)* |
| | yes | **13** | **37** | 50 |
| | | *(26)* | *(74)* | *(100)* |
| | | 43 | 57 | 100 |

Contingency table at cut-off 4

| | | Prediction | | |
|---|---|---|---|---|
| | | no | yes | |
| Outcome | no | **40** | **10** | 50 |
| | | *(80)* | *(20)* | *(100)* |
| | yes | **25** | **25** | 50 |
| | | *(50)* | *(50)* | *(100)* |
| | | 65 | 35 | 100 |

░ = false alarms ▓ = 'hits'

NOTE. Population base rate of outcome = 50%. Contingency tables display overall percentages, and *(between brackets)* row percentages (proportions of the outcome predicted correctly or incorrectly).

actually violent after discharge) vary widely (Rice & Harris, 1995, 2003). Thus, the ROC area under the curve is much less strongly influenced by base rates than other measures of predictive accuracy. As a useful consequence, direct comparisons can be made between ROC's of risk assessment procedures that were evaluated in samples with different base rates. However, the relative base rate independence of ROC analysis does not change the fact that the actual performance of risk assessment procedures in practical decision making will still be influenced by the population base rate. Also, it has been noted that in clinical practice, tests will (and should) be applied with a cut-off value close to the so-called optimum operating point, and that the prediction error trade off at this particular value may be a better measure for the clinical accuracy of the test than the overall trade-off curve represented by ROC AUC values (Halpern *et al.*, 1996).

The interpretation of ROC values in risk assessment research studies tends to be rather optimistic, with values over .75 generally being regarded as highly satisfactory (Sjöstedt & Grann, 2002). Yet, at an AUC of .75 there will still be 1 false alarm for every 3 true positives at the optimum cut point, and an error rate greater than that at all other cut-points. As a more realistic framework for interpreting AUC values Sjöstedt & Grann (2002) propose to regard values below .60 as low accuracy; values between .60 - .70 as marginal accuracy; .70-.80 modest accuracy, .80-.90 moderate accuracy; and values over .90 as high accuracy.

## I.2.7 Static and dynamic risk factors

At first glance the distinction between static and dynamic risk factors seems straightforward: either the value on a predictor is fixed, a historical fact that cannot change after it has been established (e.g., age at first conviction; failure in high school); or it is a changeable, dynamic characteristic of the person, that may assume different values in successive assessments (e.g., mood; marital status). Separating the two is particularly useful when contemplating risk factors that may offer starting points for risk reduction strategies. Reducing (re)offending risk is the aim of any system of forensic mental health care; risk assessment instruments in this context lose much of their clinical appeal if they only comprise unchangeable, i.e., untreatable static predictors.

On closer inspection, however, the distinction does not appear all that clear-cut. Risk factors such as age at the time of assessment or number of supervision failures are often labeled 'static', even though they can clearly change (be it only in the upward direction). In contrast, certain predictors that are often labeled 'dynamic', such as impulsivity or substance dependency, may be deeply ingrained in the personality or even genetically or cerebrally determined, and consequently change only very slowly, if at all. Mills *et al.* (2003) propose to blur the boundary between static and dynamic predictors by discarding the term 'dynamic' completely. In their opinion it is unrealistically suggestive of continuous flux, while change on most so-called 'dynamic' items in reality is slow. They suggest that predictors labeled 'dynamic' should more appropriately be called 'stable'.

Hanson & Harris (2000, 2001) have made a particularly compelling contribution to a more sophisticated conceptualization of the static-dynamic duality while constructing their Sex Offender Needs Assessment Rating (SONAR). First, they define as dynamic only those risk factors "that are amenable to deliberate intervention" (2001, p. 106). Only when targeted interventions can make a difference, a predictor is truly dynamic in the clinical sense. They furthermore note that a risk factor is dynamic only if changes in risk factor scores correlate consistently with a decrease or increase in actual (re)offending rates. Finally, they suggest to distinguish stable and acute dynamic risk factors. Stable dynamic risk factors only change over considerable periods of time, for instance, personality disorder (some even consider this a static factor, e.g., Webster *et al.* 1997). Acute dynamic risk factors can change from day to day (e.g., housing situation) or even hourly (e.g., intoxication). In general, the more changeable the characteristic, the less likely it is to be predictive in the long term (Hanson & Harris, 2000; Grubin & Wingate, 1996).

# I.3 ☙ Putting risk assessment into practice: three approaches

Follow-up time, time at risk, choice of recidivism criterion, and the population base rate together exert a decisive influence on the possible success of any risk assessment procedure, irrespective of its methodology. Nevertheless, the procedure used to put risk assessment into practice can itself influence the validity of assessments. In general, three main approaches are distinguished: unstructured clinical assessment, actuarial assessment, and structured clinical assessment. Many subdivisions can be made within these main categories. Doren (2001), for instance, distinguished 6 risk assessment variants (cited in Hilterman & Gresnigt, 2003). However, such exercises seem of semantic rather than practical interest, and the discussion below is therefore limited to the three main categories.

## I.3.1 Unstructured clinical risk assessment

Notwithstanding recent scientific developments, this traditional approach to risk assessment is still the one most widely used (Dernevik, 2004). The clinician is, as it were, his own risk assessment tool. He combines observations, intuition, input from other therapists, input from the patient, work experience and theoretical notions about the mechanisms of human behavior, to arrive at a conclusion about reoffending risk. The qualification 'unstructured' does not imply that this approach precludes the use of standard procedures or a fixed framework. It does imply however that the content of such a framework, i.e., the information selected, the weighing of it, and the way data are combined to reach a final conclusion, is determined subjectively (and often implicitly) by each individual clinician. Thus, a clinician adhering to the psycho-analytical school may emphasize early childhood experiences or unresolved trauma. A cognitive-behavioral therapist is more likely to focus on the chain of decisions leading up to the offence, or on cost-benefit assessments of crime, and will base his risk assessment on the extent to which the patient is aware of these mechanisms and able to control them.

Obviously, such subjectivity is detrimental to the reliability of assessments. Reliability is further compromised by the process that psychoanalysts label as 'countertransference': the assessor unwittingly projects personal emotions and associations on the patient, influencing his assessment. Extra-clinical interests may further thwart the assessment of risk either consciously or unconsciously: there may, for instance, be management pressure to increase patient flow on a ward, heightening the assessor's sensitivity to patient characteristics that coincide with this need. This tendency to favor evidence that fits a preferred hypothesis is known as 'confirmation bias' (Mynatt *et al*, 1977).

These effects, each moderating clinical risk assessments, compound each other, increasing the likelihood that two clinicians judging the same patient independently, will arrive at different risk estimates. This does not bode well for the predictive validity of unstructured clinical assessments, and empirical research has further called it into question. The aforementioned Baxstrom study (Cocozza & Steadman, 1976; see p. 21) was emblematic, but before and after it many studies were published that confirm the Baxstrom results: clinical predictions of future violence that are achieved without any standardized tool tend to have modest predictive validity at best. Comprehensive reviews of this research were published by Mossman (1994) and Grove *et al.* (2000). One finding from the latter review was that inclusion of clinical interview data in risk assessment decreases its predictive validity, which may be related to the counter transference hypothesis suggested before. Clinical interviews may also provide the clinician with an information 'overkill' which makes it harder to distinguish relevant from irrelevant information. Another finding from the Grove *et al.* review is that education and experience of the assessor are unrelated to the validity of his risk assessments. The absence of a correlation between clinical experience and the accuracy of assessments has also been reported by other authors (Slovic *et al.*, 2000; Garb, 1989).

Mossman (1994) concluded that clinicians will usually predict better than chance, but that they can claim no special expertise in this area because a simple checklist of historical risk markers performs equally well or better. This finding challenges the forensic clinician to prove the added value of his input on top of that simple checklist, which is the hallmark of the actuarial risk assessment approach.

On the opposite page, Table 2 provides a general overview of risk factors incorporated in the most prominent instruments discussed below.

## I.3.2 Actuarial risk assessment

The term actuarial refers to the world of insurances. Connecting risk assessment with insurances proves apt, for the very word 'risk' originated in the insurance business, somewhere around 1660: it derives from the old Italian word *risco*, meaning 'that which cuts', referring more specifically to reefs or rocks that may damage a vessel and incur the loss of cargo (Ewald, 1981).

Insurance premiums are calculated by an actuary, a mathematical advisor who estimates the risk that a client will actually put in a claim with the insurance company. For this purpose he uses tables in which characteristics of large numbers of previous clients are linked to their insurance claims. These tables show which combinations of risk factors have resulted in which claims. The new client is then matched with the existing data to assess which group he resembles most, and his premium is settled at the level appropriate to the average claim by that group.

TABLE 2. OVERVIEW OF RISK FACTORS INCLUDED IN SOME MAJOR RISK ASSESSMENT INSTRUMENTS

| Risk factor | VRAG | LSI-R | Static 99 | HCR-20 | SVR-20 |
|---|---|---|---|---|---|
| Historic/static | | | | | |
| Lived with both biological parents until adolescence | x | | | | |
| Childhood maladjustment (school, community) | x | x | | x | x |
| Victim of child abuse | | | | x | x |
| Employment history problems | | x | | x | x |
| Substance or alcohol abuse history | x | | | x | x |
| Relational history problems / married? | x | | x | x | x |
| Criminal history; first offender, number of previous convictions | | x | x | x | x |
| Criminal history (any details regarding previous offences) | x | x | x | | x |
| Early age at first offence/violence | | x | | x | |
| Failure on previous conditional release, institutional misconduct | x | x | | x | x |
| Age at index offence or time of assessment | x | | x | | |
| Index offence: number of offences | | x | | | |
| Index offence: victim injury | x | | | | x |
| Index offence: victim sex | x | | x | | x |
| Index offence: victim is family member | | | x | | |
| Index offence: victim is stranger | | | x | | |
| Personality disorder present | x | | | x | x |
| History of psychotic or other axis 1 disorder | | x | | x | |
| Psychopathy | x | | | x | x |

*(continued on the following page)*

*(Table 2 – continued)*

| Dynamic | VRAG | LSI-R | Static 99 | HCR-20 | SVR-20 |
|---|---|---|---|---|---|
| Sexual deviance | | | | | x |
| History of suicidal or homicidal ideation | | | | | x |
| Current psychotic disorder | x | x | | x | x |
| Current employment | | x | | | |
| Current financial situation | | x | | (x) | |
| Quality of current relationship | | x | | (x) | |
| Current accommodation | | x | | (x) | |
| Current leisure activities | | x | | | |
| Quality of current social network, criminal peers | | x | | (x) | |
| Current substance use problems | | x | | (x) | |
| Negative/procriminal attitudes | | x | | x | x |
| Denial or minimization of offence | | | | x | x |
| Treatment participation/responsiveness | | x | | x | x |
| Insight | | | | x | |
| Impulsivity | | | | x | |
| Availability of professional support | | | | x | |
| Likelihood of encountering stressful situations | | | | x | |
| Realistic plans for the future | | | | x | x |

NOTE. (x) indicates that the risk factor is not assessed directly, but implicitly pre-sent. See text for instrument references.

Actuarial risk assessment in forensic mental health was inspired by this approach, and the parallels are obvious. However, in this context the term 'actuarial' has acquired many different meanings, most of which the actuary would not recognize (Buchanan, 1999). The term is applied to designate the 'transparency' of a risk assessment procedure; to indicate the historical nature of constituent items in a tool; to indicate that any standardized procedure was used in risk assessment; or to indicate that data were processed using a mathematical procedure. Though all these interpretations highlight an aspect of actuarial models, none is complete. An actuarial risk assessment procedure is best defined as any risk assessment achieved through the use of a fixed set of risk factors, which are evaluated according to fixed rules, and are then subjected to a mathematical algorithm to produce the final risk estimate.

Actuarial risk factors are usually (but not necessarily) based on empirical research findings, and gathered in a test-like tool. However, actuarial risk assessment instruments can not be equated with psychological tests. Whereas psychological tests assess meaningful, interconnected and internally consistent constellations of symptoms or characteristics in order to establish an underlying disorder or ability, an actuarial tool merely lists a number of characteristics that have been shown to correlate with future negative events.

The original actuarial table-format is rarely encountered in risk assessment research; a rare exception are the Californian Actuarial Risk Assessment Tables (CARAT, Schiller & Marques, 1998). Relinquishing this format entails the loss of an important benefit of actuarial tables. In a table format, it is possible that the presence of, for instance, two particular risk factors results in a higher risk than the presence of five other risk factors. This relative weighing of risk factors is lacking in the linear world of actuarial checklists, where mostly the presence of more risk factors will automatically result in higher risk estimates. In this respect tables are closer to reality than checklists. That is why Monahan *et al.* (2001) recently returned to an approach closely akin to the actuarial table, the 'iterative classification tree' (ICT). Using the ICT technology, Monahan *et al.* showed, for example, that among psychopaths the impact of psychopathy on recidivism rates is strongly moderated by the presence or absence of child abuse in the individual's history, whereas for non-psychopaths childhood abuse is not a risk factor at all.

Typically an actuarial tool will comprise 10 to 20 items, to be rated on uniform scales. Some tools, like the Violence Risk Appraisal Guide (VRAG, Quinsey *et al*, 1998) attach weighted scores to risk factors, including the possibility that the absence of a risk factor results in a reduced sum score. The emphasis as regards content is mostly on the person's history, which means that a score, once established, remains fixed or can only increase (e.g., number of previous convictions). This explains why the term actuarial is often used as a synonym for 'historic' or 'static'. There is, however, no reason why an actuarial procedure couldn't also contain dynamic risk factors. But as the construction of actuarial tools is mostly empirically driven, the lack of robust and consistent empirical support for dynamic risk factors is reflected in the content of most of these instruments. A notable exception is the LSI-R (Level of Service Inventory – Revised, Andrews & Bonta, 1995) which contains many dynamic risk factors. However, given the fact that the authors allow the use of clinical discretion to modify end results when individual situations strongly call for it, it might be argued that this is in fact a structured clinical rather than an actuarial instrument.

Specific expertise is rarely required for scoring an actuarial risk scale. Anyone sufficiently acquainted with the scoring procedure and supplied with adequate files will find that the assessment amounts to little more than a clerical procedure. Quinsey *et al.* (1998) even argue that clinicians should not be involved in actuarial ratings as a matter of principle, because research shows that their input reduces the assessment's validity. Grove & Meehl (1996) go yet one step further, and state that the involvement of expensive clinicians in risk assessment procedures is an unethical waste of tax funds. The superiority of actuarial over clinical procedures, they somewhat grandiloquently claim, is the best established finding in the social sciences.

Yet, predictive validity of unstructured clinical and that of actuarial approaches do not differ quite as dramatically as such statements would lead one to expect. In their aforementioned review of 136 studies, which apart from forensic psychiatry covered areas as diverse as general medicine, general psychiatry, education and advertising, Grove *et al.* (2000) found both approaches to be equally valid in half the cases; in 40% of studies actuarial methods performed substantially better than unstructured clinical approaches, while overall actuarial methods were about 10% more accurate. It should be added that the gain of actuarial over unstructured clinical procedures was most pointed in medical and forensic assessments as opposed to the other prediction contexts under study.

In forensic psychiatric and correctional actuarial risk assessment, modest to moderate results have been reported with the VRAG, the LSI-R and (though not a risk assessment tool) the PCL-R (Psychopathy Checklist Revised, Hare, 1991). Among actuarial risk assessment tools for sex offenders the RRASOR (Rapid Risk Assessment for Sex Offender Recidivism, Hanson, 1997), subsequently incorporated in the Static-99 (Hanson & Thornton, 2000), and the MnSOST-r (Minnesota Sex Offender

Screening Tool – revised, Epperson *et al.*, 1999) have proven to be moderately accurate. Static risk prediction models generally account for some 20 to 30% of variance in reoffending (Ward & Eccleston, 2000). For illustrative purposes, a selection of recent studies is summarized in Table 3 (p. 40).

Even though actuarial prediction results are moderate at best and vary over studies, some researchers have nonetheless fallen prey to a veritable actuarial euphoria. They present their instruments as a panacea to the risk assessment problem, or voice the opinion that the results achieved are so good that any further improvement is unlikely to occur, and even more unlikely to be achieved through the addition of dynamic predictors (Harris & Rice, 2003). Such notions however ignore the fundamental limitations of our present knowledge. Actuarial risk assessment procedures typically were developed through a process aptly described by Bonta (1997) as 'dustbowl empiricism'. Risk factors were identified by retrospectively linking extant databases on offenders, mostly compiled for other than risk assessment research purposes, to reoffending data. Such a procedure of course does not yield intrinsically coherent predictor sets, and provides no insight into the dynamics governing the reoffending process. Silver & Miller (2002), for example, point to the fact that substance abuse is a risk factor present in most actuarial instruments. Yet, they argue, targeting substance abuse in treatment may have no effect on recidivism rates whatsoever, as the social and personal precursors of substance abuse are much the same as those that lead to violence and crime. Thus, there may be a spurious correlation between substance abuse and crime, while there is no causal link at all.

Bonta (1997) observes that actuarial risk assessment dissolves the individual into a random collection of risk factors. As Ewald (1991) puts it: risk is a characteristic of the collective rather than the individual. Actuarial risk assessment is the ultimate consequence of the shift from dangerousness, a propensity to commit violence that is rooted in the individual, to risk, a likelihood of violence derived from risk factors describing the collective. Thus, actuarial risk is nomothetic and deductive rather than idiographic and inductive.

As a consequence, actuarial methods are insensitive to particular characteristics of the individual that may in fact strongly influence the likelihood of a new offence. If, for instance, a patient has in the past few weeks repeatedly voiced the intent to commit a murder, which would generally be considered a salient risk factor, few actuarial tools will register it. Nor will actuarial instruments be sensitive to a physical handicap that

TABLE 3. OVERVIEW OF A SELECTION OF STUDIES INTO THE PREDICTIVE VALIDITY OF

| instrument | study | population/setting |
|---|---|---|
| VRAG<br>Violence Risk Appraisal Guide (Quinsey *et al.*, 1998) | Rice & Harris, 1997 | rapists/maximum security psychiatric facility in Canada (n=159) |
| | Cooke, Michie & Ryan, 2001 | inmates/prison, Scotland (n=250) |
| | Kroner & Mills, 2001 | inmates, non sexual offenders/prison, Canada (n=87) |
| | Loza, Villeneuve & Loza-Fanous, 2002 | federal inmates/<br>correctional services institutions, Canada (n=124) |
| | Sjöstedt & Långström, 2002 | rapists/inpatient forensic treatment or detention in Sweden (n=51) |
| | Harris & Rice, 2003 | sex offenders in treatment, Canada (n=46) |
| LSI-R<br>Level of Service Inventory-Revised (Andrews & Bonta, 1995) | Gendreau *et al.*, 1996 | meta-analysis, non-treatment settings, mainly USA & Canada (n=1,141) |
| | Kroner & Mills, 2001 | inmates, non sexual offenders/prison, Canada (n=87) |
| | Mills *et al.*, 2003 | federal prisoners, Canada (n=209) |
| Static-99<br>(Hanson & Thornton, 2000) | Hanson & Thornton, 2000 | meta-analysis, treatment and prison, Canada & UK (n=1,208) |
| | Sjöstedt & Långström, 2001 | sex offenders, prison, Sweden (n=1,400) |
| | Harris & Rice, 2003 | sex offenders in treatment, Canada (n≈37) |
| PCL-R<br>Psychopathy Checklist-Revised (Hare, 1991) | Gendreau *et al.*, 1996 | meta-analysis, non-treatment settings, mainly USA & Canada (n=1,141) |
| | Grann *et al.*, 1999 | violent crime convicts, Sweden (n=560) |
| | Cooke, Michie & Ryan, 2001 | inmates/prison, Scotland (n=250) |
| | Kroner & Mills, 2001 | inmates, non sexual offenders/prison, Canada (n=87) |

NOTE. (ns) = not significant. (?) = no significance data available. For publication details of studies

MAJOR ACTUARIAL RISK ASSESSMENT INSTRUMENTS

| Follow-up time | outcome criterion | predictive accuracy |
|---|---|---|
| average 10 years | charged with or returned to hospital due to violent offence | ROC AUC .77 |
| | charged with new sex offence | ROC AUC .62 |
| max. 3.3 years | any reconviction | ROC AUC .72 (?) |
| | any reconviction for violence | ROC AUC .67 (?) |
| average 2.2 years | reconviction for violent offence | ROC AUC .60 (?) |
| 2 years | any new charge | r = .12 (ns) |
| | violent acts | r = .05 (ns) |
| average 6.1 years | reconviction for a sex offence | ROC AUC .58 (ns) |
| | violent reoffense resulting in conviction | ROC AUC .69 |
| 5 years (fixed) | violent recidivism | ROC AUC .86 (?) |
| minimum 6 months | parole violation, rearresr, or reconviction | mean r = .35 |
| average 2.2 years | reconviction for any violent offence | ROC AUC .67 (?) |
| average time of opportunity 2.1 years | charged with new violent offence | r = .26 |
| not specified | new sexual offence, (charge, hospital readmission) | mean ROC AUC .71 |
| | new violent offence | mean ROC AUC .69 |
| average 3.7 years | reconviction for a sex offence | ROC AUC .76 |
| | reconviction for a violent offence | ROC AUC .64 |
| 2 years (fixed) | sexual recidivism | ROC AUC .84 (?) |
| 11 years (fixed) | sexual recidivism | ROC AUC .64 (?) |
| minimum 6 months | parole violation, rearresr, or reconviction | mean r = .28 |
| 2 years | reconviction for violent recidiviism | ROC AUC . 72 |
| max. 3.3 years | any reconviction | ROC AUC .70 (?) |
| | any reconviction for violence | ROC AUC .65 (?) |
| average 2.2 years | reconviction for violent offence | ROC AUC .56 (?) |

please refer to the Reference list at the end of this book.

restrains the patient's freedom of movement and renders an offence very unlikely – an exceptional situation that has supplied these individual risk-reducing variations with the epithet 'broken leg cases'. On the other hand, Grove & Meehl (1996) point out that broken leg situations are very rare in practice, and hardly compromise the validity of actuarial tools overall. In fact, they argue, if we allow the clinician to make corrections to an actuarial score to reflect individual aspects of a particular case, we know from the research into clinical risk assessment that such corrections will mostly be applied misguidedly, and would nearly always decrease predictive validity. (As will be seen below, recent studies into structured clinical risk assessment may prove this assumption to be false.)

Castel (1991) notes another important implication of the statistical approach: a patient no longer needs to display overt signs of dangerousness or deregulation. The mere presence of a number of statistical risk factors in his personal history suffices to designate him as a future risk. This results in the disconnection of risk assessment and subsequent risk management interventions. The mere conclusion that a person answers to a sufficient number of risk factors to be considered a danger, does not provide any clues as to the type of intervention this particular individual would need to prevent his actual (re)offending. Thus a strictly actuarial approach alienates the risk assessment process from the clinician and his attempt to reduce risk through treatment; hence Grubin's (1999) conclusion: "Statistical significance does not equal clinical relevance" (p. 332) - to which, however, it should be added that statistical insignificance *does* equal clinical irrelevance.

The alienating effect of actuarial predictors is exacerbated by their general triviality. Grubin & Wingate (1996) list some of the most common actuarial risk factors: young age, single marital status, psychopathy, an extensive criminal history, supervision failures, alcohol abuse and antisocial behavior in childhood. What, they then (rhetorically) ask, do such predictors tell us? They tell us that young, manipulative, egocentric, aggressive single males, who in the past have shown to care little about rules, are inclined to behave aggressively, especially when drunk. Actuarial risk assessment never quite succeeds in breaking free from this kind of circular reasoning, stating with much aplomb that aggressive people with a strongly criminal profile are likely to behave aggressively.

Silver & Miller (2002) suggest that actuarial tools should not be viewed as risk assessment instruments at all, but rather as systems describing violence base rates of population subgroups, or 'base rate dispersion'. They echo Castel (1991) in their conclusion that actuarial instruments tend to stigmatize the individual, who is designated high risk merely based on statistical group membership. Actuarial instruments, according to Silver & Miller, primarily serve to facilitate the management of institutional resources, rather than targeting individuals or social conditions in need of reform. Thus, such tools contribute to the "continued marginalization of populations already at the fringes of the economic and political mainstream" (p. 138).

To conclude, it should be emphasized that though actuarial methods may elicit ample criticism from clinical and ethical viewpoints, in the end the only predictors and prediction methods that can be clinically relevant and are ethically defensible are those that have demonstrable predictive validity. In this respect, actuarial procedures undeniably remain strong contenders in the field of risk assessment.

### *I.3.2.1 A note on psychopathy*

The selective overview in Table 3 (p. 40) includes the Psychopathy Checklist Revised (Hare, 1991), following the lead of many comparative studies of risk assessment instruments that include this influential tool (e.g., Rice & Harris, 1995; Cooke *et al.*, 2001; Kroner & Mills, 2001; Gendreau *et al.*, 2002; Sjöstedt & Långström, 2002; Douglas *et al.*, in press). Several authors have stressed the effectiveness of the PCL-R as a predictor of future violence, and this ability has commonly been considered proof of the instrument's validity (Salekin *et al.*, 1996; Hart, 1998; Grann *et al.*, 1999; Hare *et al.*, 2000).

It should, however, be noted that the PCL-R is not a risk assessment tool as such (Hemphill & Hare, 2004). Rather, it is a psychometric instrument for the assessment of psychopathy, a particular, predatory and manipulative survival strategy which has been called the most strongly predictive single risk factor for future violence (Salekin *et al*, 1996). The primary relevance of the psychopathy concept is in its implications for treatment, which are as yet contentious. Though many authors have concluded that psychopathy is associated with negative treatment response (see Hildebrand, 2004, for an overview), some even suggesting that particular treatment strategies might in fact make matters worse (Quinsey *et al.*, 1998), this view has recently been challenged in a systematic re-evaluation of 24 studies, indicating that most of them did not have appropriate designs to warrant such conclusions (D'Silva *et al.*, 2004).

The PCL-R score is included as an item in many commonly used risk assessment instruments, such as the VRAG, HCR-20, and LSI-R. Inevitably, the items used in the PCL-R to measure psychopathy overlap with items constituting the rest of the risk assessment instrument at hand, which raises questions of redundancy. Quinsey *et al.* (1998) do not only acknowledge such redundancy between PCL-R and VRAG, they argue that it is the foundation of the VRAG's predictive robustness, thus turning an apparent disadvantage into an advantage. Additionally they show that the VRAG excluding the PCL-R –score is only marginally less predictive of renewed violence than the VRAG including psychopathy. Similar results are reported by Douglas *et al.* (1999) with regard to the HCR-20. These authors find that although the HCR-20 excluding the PCL:SV (Psychopathy Checklist: Screening Version, Hart *et al.*, 1995) adds predictive accuracy to psychopathy alone, the reverse is not the case: the PCL:SV does not add to the predictive validity of HCR-20-scores excluding psychopathy.

## I.3.3 Structured clinical risk assessment

Risk assessment science on the crossroads of criminal justice and mental health has long been sharply divided between the individualistic, flexible, clinically salient but predictively dubious clinical approach, and the generalizations of group-based, rigid, clinically unpromising but statistically more effective actuarial techniques. The measure of polarization was at times unduly extreme in the light of Mossman's (1994) conclusion that ROC AUC's of actuarial methods averaged .71 and those of unstructured clinical assessments .67. The violently polemical tone nonetheless adopted by pro-actuarialists like Grove & Meehl (1996) and Quinsey *et al.* (1998) kindles a suspicion that other than scientific motives played a role in maintaining the schism.

In the mid 1990's a group of Canadian researchers concluded that the existing situation was unproductive, and developed a third approach to risk assessment that occupies the middle ground between the clinical and actuarial extremes. This was named the 'structured clinical approach'. It acknowledges the relative success of actuarial methods by adopting the standardized checklist format and including empirically well-established historical predictors. However, most structured clinical instruments also include risk factors that offer starting points for clinical intervention, even if the empirical evidence for such predictors is not very compelling as yet (Webster *et al.*, 1997). Some examples of such predictors are: insight; current psychotic symptoms; negative attitudes; and responsiveness to treatment. Moreover, the final risk assessment does not result from a mathematical operation, but from clinical reflection on the score pattern.

The assessor may attach different weights to different items according to specific case characteristics, or may even add risk factors that do not occur in the instrument. The final assessment is formulated on a simple three-point scale: low, medium or high risk; a pretence of higher precision levels seems unjustified (Webster *et al.*, 1997).

Thus, structured clinical methods, unlike actuarial ones, are not limitative and constraining, but supply a minimum framework, or as Webster *et al.* (1997) put it, an *aide-mémoire*. They guarantee that the basic structure of each risk assessment is the same, and that certain well-established predictors are in any case included in the assessment, without ruling out individual variation and clinical input.

This strategy has been operationalized in the HCR-20 (Historical / Clinical / Riskmanagement – 20, Webster *et al.*, 1997), for assessment of risk for violence; the SVR-20 (Sexual Violence Risk-20, Boer *et al.*, 1997), for assessment of risk for sexual violence; and the SARA (Spousal Assault Risk Assessment Guide, Kropp *et al.*, 1995) specifically targeting risk of domestic violence. Youth versions of the SVR-20 (Structured Assessment of Violence Risk in Youth, SAVRY, Borum *et al.*, 2002) and the HCR-20 (Early Assessment Risk List, EARL-20B for boys, Augimeri *et al.*, 2001, and EARL-21G for girls) have also been developed and are currently being tested.

A fair amount of research has meanwhile been conducted, mainly on the HCR-20; the number of studies published in peer reviewed journals however lags somewhat behind the number of comparable publications on actuarial instruments. Many of the studies are small-scale, and findings are variable; a selective overview is displayed in Table 4 (following pages). Some studies show HCR-20 and SVR-20 to have moderate predictive power, others report no or little improvement over chance. Thorough evaluation of structured clinical methods is as yet hampered by the fact that most published studies were conducted retrospectively, and merely linked file based item sum scores of the instrument to outcome. Such studies, by excluding the clinical weighing process that is the defining characteristic of structured clinical methods, in fact evaluated HCR-20 or SVR-20 as actuarial tools. Only a few studies have included at least an approximation of the structured clinical process. Findings from De Vogel *et al.* (2003) seem to suggest that a retrospective final, assessor-made categorization into low, medium or high risk is a slightly better predictor of future sexual violence than the actuarial sum-score on the SVR-20 (ROC AUCs of .82 and .77 respectively), but the difference is not statistically significant. Douglas *et al.* (in press) found structured clinical judgments to be slightly less predictive

TABLE 4. OVERVIEW OF A SELECTION OF STUDIES INTO PREDICTIVE VALIDITY OF

| Instrument | Study | Population/setting |
|---|---|---|
| HCR-20<br>Historical/Clinical/ Riskmanagement - 20 (Webster *et al.*, 1997) | Douglas *et al.*, 1999 | civil psychiatric, Canada (n=193) |
| | Belfrage *et al.*, 2000 | max. security prison inmates, Sweden (n=41) |
| | Kroner & Mills, 2001 | inmates, non-sexual offenders/prison, Canada (n=87) |
| | Coooke *et al.*, 2001 | inmates/prison Scotland (n=250) |
| | Douglas *et al.* (in press) | offenders from correctional institutions, Canada (n=188) |
| SVR-20<br>Sexual Violence Risk -20 (Boer *et al.*, 1995) | Dempster & Hart, 2002 | sex offenders from federal prisons, Canada (n=95 [violent recidivism], 71 [sexual recidivism]) |
| | Sjöstedt & Långström, 2002 | rapists, inpatient forensic treatment or detention in Sweden (n=51) |
| | De Vogel *et al.*, 2003 | sex offenders, forensic inpatients, Netherlands (n=122) |

NOTE. ns = not significant. (?) = no information on significance. $r_{pb}$= point biserial correlation.

STRUCTURED CLINICAL RISK ASSESSMENT INSTRUMENTS

| follow-up time | outcome criterion | predictive accuracy |
|---|---|---|
| average 1.7 years | criminal offence or rehospitalization records of physical violence | ROC AUC .76 |
| | criminal offence records of criminal violence | ROC AUC .80 |
| average 8 months prospective | institutional violence | HCR tot. mean difference 8.8, p < .001 (Mann-Whitney U)<br>H only: mean diff. ns |
| average 2.2 years | reconviction for any violent offence | ROC AUC .62 (?) |
| max. 3.3 years | any reconviction | HCR tot. ROC AUC .71(?)<br>H only ROC AUC .72 (?) |
| | any reconviction for violence | HCR tot. ROC AUC .69 (?)<br>H only ROC AUC .68 (?) |
| 6 to 11 years | violent recidivism (participants selected based on known outcome) | HCR tot ROC AUC = .82<br>structured clinical judgment ROC AUC = .78 |
| average 5.1 years | violent recidivism from police and correctional records | Offence history $r_{pb}$ = .47<br>Fixed psychosocial $r_{pb}$ = .41<br>Variable factors $r_{pb}$ = .37<br>No SVR-20 total score correlation reported |
| | sexual recidivism from police and correctional records | Offence history $r_{pb}$ = .52<br>Fixed psychosocial $r_{pb}$ = .43<br>Variable factors $r_{pb}$ = .32<br>No SVR-20 total score correlation reported |
| average 6.1 years | reconviction for a sex offence | ROC AUC .49 (ns) |
| | reconviction for a violent offence | ROC AUC .64 (ns) |
| average 140 months | reconviction for a sexual offence | actuarial ROC AUC .77<br>structured clinical judgment ROC AUC .82 |
| | reconviction for violent non-sexual offence | actuarial ROC AUC .66<br>structured clinical judgment ROC AUC .64 |

For publication details of studies please refer to the Reference list at the end of this book.

than the HCR-20 actuarial score (ROC AUC .78 vs. .82), although this difference, too, is not statistically significant. Interestingly, these authors also found that the final, professionally informed assessment retained a unique contribution to the prediction of reoffending when partialing out actuarial HCR-20, VRAG and PCL-R measures (partial $r_{pb} = .15$, $p < .05$). Studies like these are in need of replication, especially by researchers and clinicians outside the circle of scientists closely involved with construction or translation of these instruments.

Webster *et al.* (1997) call the HCR-20 a "work in progress", and some questions that need answering are clearly identifiable. Though the inclusion of dynamic items is an important gain to clinical users of the HCR-20, these items are operationalized very broadly, so that interpretations may vary and the link to intervention strategies remains vague. The latter problem has been countered to some extent by the publication of an HCR-20 risk management guide (Douglas *et al.*, 2001), but tauter, more internally consistent item definitions remain desirable. A related problem is the fact that dynamic items are only scored on a three-point scale (risk factor is present, is possibly or partly present, or is absent), thus putting severe constraints on the instrument's ability to measure change. Regardless of such present shortcomings, the swift assimilation of structured clinical instruments both in North American and European clinical and correctional settings shows that these tools answer a need, and at the same time provides an excellent starting point for the accumulation of data that in time may be used to improve the currently available instruments.

# I.4 ℘ Risk assessment reality

## I.4.1 From laboratory to hospital

Thus far, developing and testing new risk assessment methods has mainly been an occupation of researchers. In spite of the overwhelming amount of criticism heaped upon it, the unstructured clinical approach still dominates clinical and judicial reality (Webster *et al.*, 2002; Boothby & Clements, 2000). Gardner *et al.* (1996) surmise that clinicians decline the use of structured methods because the underlying statistical principles are too complex and badly understood, and because the amount of time needed to gather the required information in accordance with the instrument's standards is simply not feasible in busy daily practice.

Buchanan (1999) points to similar reasons, and adds that actuarial tools are too limited in scope to do justice to the great variety of situations and decisions that are part of forensic psychiatric practice. Grove & Meehl (1996) suspect some less acceptable motives for clinicians to ignore statistical tools, such as an unwillingness to relinquish blind faith in their own expertise, or even a fear of losing their job.

Whatever the reasons, the present situation is such that it remains largely unknown how risk assessment instruments perform in actual forensic practice, let alone whether they indeed help reduce levels of reoffending. It seems fairly certain though that their clinical performance will not attain the same level of quality as that reported in research studies. Nearly all risk assessment research was conducted under optimized conditions that are often impossible to replicate in a clinical setting. Raters were typically well-trained and aware of the requirements to meet good reliability standards; in many cases they were professional researchers. Often they were at leisure to perform the ratings and not burdened with clinical tasks. Mostly the ratings were performed retrospectively, using files. This means that personal impressions of the patient, including countertransference effects, were largely eradicated, as were other extraneaous motives that may influence an assessment. It also means that in reliability studies each rater worked from the same, limited set of data rather than having to make subjective selections from the plethora of information bearing down on clinicians in a clinical setting - thus reducing possible interrater variation beforehand. In summary, one must conclude that most research findings were attained under laboratory-like, idealized circumstances, that differ considerably from conditions found in clinical practice.

Bridging the gap between empirical science and clinical practice requires a shift towards a scientific approach that actively facilitates evidence based practice by providing 'practice based evidence' (Margison *et al.*, 2000). Reliability and validity tests for risk appraisal instruments should be performed under clinically realistic conditions to provide a non-inflated assessment of their practical value. Webster *et al.* (2002) observe that though the HCR-20 is used in many clinical settings, "some individual practitioners have begun and ended their study of the HCR-20 with a reading of the coding sheet." (p. 45). Slovic *et al.* (2000) conducted a study that showed clinical risk judgments to be unaffected by the presence or absence of an instruction explaining concepts such as harm and probability, suggesting that even if instructions are read, they are not certain to influence judgments. Clearly, future research needs to concentrate on implementing and evaluating risk assessment procedures in clinical settings. Instruments that are overly sophisticated or time-consuming probably stand little chance of long-term survival in this environment. Tools will need to be easily grasped and straightforward. On the other hand clinicians are under a professional and ethical obligation to incorporate scientific findings into their work-practice (Grove & Meehl, 1996): they will need to acquaint themselves with the core concepts underlying any risk assessment, and will need to structurally invest part of their time in the application of available instruments according to appropriate standards. If instruments are only used erratically (or even cosmetically) it is preferable not to use them at all (Webster *et al.*, 2002).

### I.4.2 Risk communication

A major aspect of risk assessment practice is the issue of risk communication. Risk assessment outcome may be strongly influenced by the qualities of the applied method, but is determined in the end by decision-making based on it. In correctional contexts the decision-maker will often be a judge or another judicial authority. Risk assessment findings will need to be reported in a clear, careful, and unambiguous manner that is directly comprehensible to the recipient. The complex process of 'translation' that is required (Heilbrun, 1997), has been relatively neglected until recently, but according to Monahan (1996) will be one of the dominant themes in the 'next twenty years' of risk assessment research.

In the Netherlands and abroad guidelines for good quality risk reports are in short supply (De Ruiter, 2000; Heilbrun *et al.*, 1999). In one of the most profound and insightful contributions to the risk assessment field in recent years, Hart (2001) has proposed an 'anchored narrative approach' to risk communication, in which reports are judged by their narrative quality, logical consistency and relevance of content. As a starting point, he offers 10 guiding principles, displayed in Table 5. This framework amounts to a critical but positive re-evaluation of narrative types of risk reporting, and a shift away from the rigid summing of actuarial risk factors, an approach Hart characterizes elsewhere as "ridiculously simple" (1999, p. 487). Judges and juries are not Bayesian calculators, he says. "Instead, they actively construct and evaluate possible scenarios that fit the evidence presented to them — including, but not limited to, scenarios put forward by the

Table 5. Hart's 10 criteria for an anchored narrative approach to risk reporting.

- ***Does the procedure gather information concerning multiple domains of the individual's functioning?*** If risk assessments are action plans concerning people, then their perceived utility will be enhanced by the presentation of detailed, multidimensional character sketches.
- ***Does the procedure use multiple methods to gather information?*** Narratives will be perceived as more credible or persuasive if they do not rely on information gathered using a single method.
- ***Does the procedures gather information from multiple sources?*** Narratives will be also perceived as more credible or persuasive if they do not rely on a single source of information.
- ***Does the procedure gather information concerning both static and dynamic risk factors?*** Narratives that speculate about the future based only on the past ring untrue; they assume too much stability in behavior and motivation and our social environment. To be perceived as useful, speculations about the future must be relatively robust when taking into account foreseeable changes in people and their living conditions.
- ***Does the procedure allow users to evaluate explicitly the accuracy of information gathered?*** Risk assessments should recognize that evidence supporting some narrative elements is stronger than that supporting other elements. Speculations concerning the future will be perceived as more useful if they take this into account.
- ***Does the procedure allow re-assessments to evaluate changes in risk over time?*** Risk assessments should not be too rigid or inflexible in their conclusions, so that speculations about the future can take into account changes in narrative elements over time.
- ***Is the procedure comprehensive?*** The narrative should take into account all elements that consumers consider relevant, and exclude those that consumers consider irrelevant. Importantly, a good narrative can change the opinions of consumers about what is relevant and irrelevant in a particular case.
- ***Is the procedure comprehensible to consumers?*** To be useful, narratives must be structured in way that matches the information processing style of consumers.
- ***Can mental health professionals be trained to use the procedure in a consistent manner?*** Put another way, can we train professionals to use the procedure to construct useful narratives?
- ***Does implementation of the procedure result in reduction of violence?*** If violence prevention is the primary goal of risk assessment, then risk assessment procedures should facilitate the construction of narratives that assist in the planning and delivery of services intended to prevent violence.

From: Hart, 2001

disputing parties." (2001, p. 13). Risk reports should take a form in keeping with that reality, while of course remaining solidly grounded in empirical evidence regarding predictive validity.

Fascinating research by Slovic *et al.* (2000) reinforces the awareness that recipients of risk assessment reports are not calculators, but fallible human beings, and that risk reports should take account of this. These researchers studied the influence of different numerical formats for describing risk on the decisions made by recipients of the report. One finding was that relative frequency formats ('10%') consistently yield lower risk perceptions than mathematically equivalent absolute frequency formats ('1 out of 10'). Similarly, '10 out of 100' was perceived as a higher risk than '1 out of 10'. The authors ascribe this to the fact that absolute formats tend to invoke the image of an actual offence, or in the case of '10 out of 100', of 10 actual offences, whereas '10%' remains an abstraction. Such emotionally driven misperceptions can take on extreme forms. Denes-Raj & Epstein (1994) showed that if subjects were asked to pick a 'winning' red bean either from a vase containing 1 red bean and 9 white beans, or from a vase containing 7 red beans and 93 white ones, most subjects would prefer the 100 bean vase. One in 10 was perceived as a lower chance of winning than 7 in 100, simply because 1 is less than 7. These are just a few examples of the processes at work when risk assessment reports are interpreted. They stress the need for stringent reporting guidelines, for training of both writers and recipients of such reports, and they illustrate the danger inherent in thinking about risk assessment in simplistic terms.

## I.5 ꩜ Conclusion

In summary, it can be concluded that the science of risk assessment has seen important developments in many areas over the past three decades. First of all, the conceptualization of the risk assessment question has gained both in sophistication and uniformity. Notions of dangerousness gave way to the concept of risk, that acknowledges the uncertainties inherent in any prediction, as well as the role of the individual's environment in the occurrence of new offences. Second, the number of studies and the quality of their design has increased dramatically. Findings have become more readily interpretable and comparable through the introduction of ROC-analysis. Third, a great variety of instruments and manuals have become available to provide clinicians and other practitioners with an evidence-based starting point for risk assessment and associated decision making.

In general, initial pessimism about the possibilities to predict future violence at all, has been replaced by empirically supported, though guarded, optimism that prediction is to some extent possible.

It can also be concluded that the present situation is far from perfect. Most studies still fail to address the methodological problems already outlined by Monahan & Steadman in 1994: impoverished predictor variables, weak criterion variables, constricted validation samples, and unsynchronized research efforts. In the first years of the new millennium the typical risk assessment study is, as of old, a retrospective, single site undertaking, using a relatively small population sample to link mainly historical predictors or scores on risk assessment tools to some dichotomous outcome criterion.

Furthermore, the one-sided accent on prediction success has tended to obscure the need for understanding the reoffending process in terms of causality, and has estranged the research from the needs and realities confronting the clinical practitioner and judicial decision maker. The conceptual shift towards the notion of risk has left clinicians and judiciaries uncomfortably perched on the narrow edge between group-based likelihood estimates and individual, yes-or-no decisions.

Third, notwithstanding their predictionist focus, studies rarely report predictive accuracy over ROC AUC = .75, and on average reported validity seems to tend more towards the .70 than the .80 level. This means that the tools that are now available still contain very large margins of error, even when applied under simplified and otherwise optimized conditions, as is the case in the majority of studies.

# II Three empirical studies

# II.1 ☙ Study I

## *The structural coherence of clinically derived dynamic indicators of reoffending risk*

This paper was previously published as: Philipse, M.W.G., Koeter, M.W.J., Brink, W. van den, & Staak, C.P.F. van der. (2004). The structural coherence of clinically derived dynamic indicators of reoffending risk. *Criminal Behaviour and Mental Health, 14*, 263-279. Reprinted by kind permission of Whurr Publishers, London.

## II.1.1 Introduction

Assessment of the risk of reoffending in patients is daily routine in most branches of forensic mental health care. Two main 'schools of thought' can be distinguished on this question: one that takes predictive validity as its primary criterion, and another that aims for clinical usefulness. The first school argues that predictors ought to be selected statistically. Nature, source, conceptual coherence and clinical applicability of predictors are all considered of minor importance compared to the strength of their correlation with reoffending. Following this approach, widely used risk assessment tools were constructed, such as the Violence Risk Appraisal Guide (VRAG, Quinsey *et al.* 1998), the Rapid Risk Assessment for Sex Offender Recidivism (RRASOR, Hanson, 1997) or the Iterative Classification Trees from the MacArthur risk assessment study (Monahan *et al.*, 2001). These instruments share a strong focus on historical data, with special attention to criminal history, history of substance abuse, childhood problems and characteristics of the index offence. The second school acknowledges the evidence in favour of historical factors as predictors, but views exclusive reliance on immutable factors as a major shortcoming. Dynamic risk factors are also required (Bonta, 1996), including clinically meaningful variables "that are amenable to deliberate change" (Hanson & Harris, 2001, p. 106).

This second view prompted the construction of risk assessment tools that incorporate such variables. The most widely used of these seems to be the Historical/Clinical/Risk Management - 20 (HCR-20, Webster *et al.*, 1997), that includes 10 historic and 10 dynamic risk factors. Predictive validity has repeatedly been demonstrated for the latter as well as the former (Douglas *et al.*, 1999; Douglas, 2001). The Level of Service Inventory Revised (LSI-R, Andrews & Bonta, 1995) contains over 20 dynamic predictors and has shown strong predictive validity (Gendreau, Goggin & Smith, 2002).

Yet the literature supportive of dynamic predictors of risk remains less extensive and generally more tentative than that supporting historical predictors. Some authors maintain this is due to the fact that dynamic variables have little to add to the predictive power of historical data (Quinsey *et al.*, 1998; Grove & Meehl, 1996), implying that the changeability of risk is a clinical dream rather than an empirical reality.

Dynamic predictors are however at a methodological disadvantage. Due to their complex, behavioural or psychological nature, it may be harder to achieve reliability than with historical facts. The effect of time frames on predictor-outcome relations is another obstacle in research on dynamic predictors. Predictive relations may change as the value of the variable fluctuates over time, or the change in value over time may itself be predictive rather than the discreet value at a specific moment. Indeed, only when particular changes in the indicator relate consistently to particular changes in risk can the indicator be clinically useful (Hanson & Harris, 2000).

Given these complications, the wholesale conclusion that dynamic predictors or clinical input are meaningless as estimators of reoffending risk is unwarranted (Hart, 1999). One way forward is through close study of the clinical risk assessment process, rather than its outcome (Mulvey & Lidz, 1985). Whenever risk assessments are conducted in a clinical setting, clinicians will play a crucial role and dynamic factors will be taken into account. Yet, Litwack (2002) notes, studies investigating the underpinnings of clinical risk perceptions are still virtually non-existent, which "contributes to the possibly unjustified disparagement of clinical assessment" (p. 174).

Some exceptions to this general situation deserve mention. Elbogen *et al.* (2002) asked clinicians to rate the relevance of risk factors both known from research and suggested by fellow clinicians, and found that behavioural factors suggested by clinicians may improve prediction. In general clinicians rated these clinically derived behavioural variables as significantly more relevant than the HCR-20, the VRAG, or risk cues from the historical, contextual and dispositional domains of the MacArthur study. Segal *et al.* (1988), in a study very similar to the one reported here, developed an 88-item checklist of symptoms, and found that such clinically salient characteristics as impulsivity, irritability, thought disorders and expansiveness were consistently associated with the clinical perception of higher risk.

The Dutch hospital order called *terbeschikkingstelling* (TBS), has always been dominated by an unstructured clinical approach to risk assessment. TBS can be imposed by a judge in case of diminished responsibility due to a mental disorder, if the perpetrator committed an offence that warrants a prison sentence of at least 4 years, and a significant risk of serious reoffending is present (Dutch Penal Code, art. 37a). The measure entails mandatory admission to a forensic psychiatric facility, initially for the duration of two years. The court decides on one- or two year extension periods, assisted by hospital reports on treatment progress, including a risk assessment. The treatment supervisor appears in court as expert witness. The average length of stay in TBS is over 5 years (for more details, see Van Marle, 2002).

In the past, foreign visitors have noted that while TBS operates state of the art facilities with highly qualified staff, some working procedures could only be described as out-of-date (McInerny, 2000). Risk assessment practices were further challenged under the influence of North American studies (Verhagen & Philipse, 1995; De Ruiter, 2000). Currently, there is a requirement to shift from unstructured to structured assessment procedures.

The main question for the present research was whether risk indicators from mental state and behavioural domains, as suggested by clinicians in TBS, represent coherent, higher order dynamic concepts (factors) that are clinically meaningful and amenable to change. The extent to which both single items and higher order concepts were linked to the clinical assessment of risk of reoffending was also investigated.

## II.1.2 Method

### *II.1.2.1 The measure*

The research tool, the Clinical Inventory of Dynamic Reoffending Risk Indicators (CIDRRI), is a checklist containing 47 statements describing potentially changeable patient characteristics considered by clinicians in forensic in-patient settings as pertinent to the assessment of reoffending risk. Items were derived from interviews with 12 expert witnesses in 4 hospitals, and review of treatment-related documents. The item-pool was edited by the research-team, and redundancies eliminated; this process was monitored by clinicians in all the hospitals, to ensure that the resulting checklist remained an adequate reflection of clinical considerations, and that terminology was transferable.

This exercise was designed to maximise clinician participation in the project in settings then entirely unaccustomed to checklists. Gardner *et al.* (1996) argued that clinicians often ignore statistical tools for mundane reasons: they are too time-consuming, and the statistical principles underlying them are overly complex. The editorial goals were therefore to ensure that our tool should meet the following requirements: straightforward item formulations in an idiom directly recognizable to clinicians; no requirement of training other than the usual professional education and experience; no extensive and time-consuming scoring instructions; straightforward, uniform rating scales.

The final list of items in the CIDRRI is shown in Appendix I (p. 183). The respondent is asked to rate each item on a six-point scale ranging from 'not at all characteristic' to 'very characteristic' of the patient. The six-point format was chosen to prevent ‘flight’ to the middle value, and to facilitate dichotomization if necessary. We also wanted a scoring range wide enough to register change.

Finally, a 48th item was added asking the respondent for a direct clinical assessment of reoffending risk, also to be rated on a six-point scale. This rating served as the dependent variable in the assessment of predictive validity of CIDRRI items and scales with regard to clinical assessment of risk.

### *II.1.2.2 Dataset*

The four sets of CIDRRI-data merged for this study are described in Table 6. The present report is therefore based on 370 checklists on 370 patients with the earliest rating made in January 1996 and the most recent in May 2002. Ratings covered all stages of treatment, though the categories of recently admitted patients (set 2) and patients about to be discharged (set 1) are over-represented. It was assumed that if robust dimensions underlie the items, these will be independent of treatment stage; merging datasets would therefore cause no problems. This assumption was tested post hoc, as reported in the results section. Possible changes in predictive relations between items or dimensions and clinical risk assessments over the course of treatment were also taken into consideration.

Of CIDRRI's, 82% had been completed by the preferred rater: a clinical psychologist or psychiatrist responsible for treatment planning and evaluation and for advising the court. Where circumstances necessitated this, the checklist was instead rated by a psychotherapist (13.4%) or a head nurse or social worker (4.6%). The data represent ratings by a total of 84 respondents, in all cases having known the patient assessed for at least six months. On average, completing the checklist took 10 to 15 minutes.

TABLE 6. CHARACTERISTICS OF THE FOUR DATASETS COMPRISED IN THE PRESENT STUDY

| Dataset | N | Treatment stage | Data collected | Purpose |
|---|---|---|---|---|
| 1 | 151 | Moment of discharge or start of probationary leave | 1996-1998 | Testing predictive validity of dynamic risk and historic risk factors with regard to actual 3 offending during a 3-5 year follow-up |
| 2 | 115 | Admitted to the hospital less than 12 but more than 6 months ago | 1997 | Testing whether items on the dynamic checklist showed the expected differences between patients at the beginning and at the end of treatment |
| 3 | 75 | Unspecified, at least one year in the hospital | 2001 | Measuring interrater reliability (Only treatment supervisor rating used in present set) |
| 4 | 29 | Unspecified, at least one year in the hospital | 2001-2002 | Measuring test-retest reliability<br>(Only first of both ratings used in present set) |
| Present study | 370 | Unspecified, at least 6 months in the hospital | 1996-2002 | Examining the structure of dynamic risk factors in relation to clinical risk assessment |

Patients were men (93%) and women (7%) who stayed in any of the seven hospitals at any time between January 1996 and May 2002. Their average age at the time of rating was approximately 32 years. Half of the group had been convicted for a violent offence without any clear external motive; 25% for a violent offence with property motive. About 15% had committed a sex offence against an adult, 5% a sex offence against a minor. The remaining 5% were arsonists. Over 80% of patients had at least one diagnosis of personality disorder according to DSM-III-R (American Psychiatric Association, 1987), mostly of the ‘not otherwise specified’ or antisocial type. Also, around 70% had at least one code on DSM-III-R axis 1, most frequently for psychotic or substance use disorders. Over half of the group showed comorbidity on axes 1 and 2. The discharge group had an average length of stay of 4.8 years in the TBS-hospital from which they were discharged . All these characteristics are similar to overall statistics for the TBS population as published by the Dutch Ministry of Justice (Van Emmerik & Diks, 1999).

## II.1.3 Analysis

The CIDRRI contains both positive and negative item formulations. For ease of interpretation, items were recoded so that higher scores (towards the 'very characteristic' end of the scale) always corresponded with higher problem levels.

With n=370, the minimum subjects-to-variables ratio of 5, required for factor analysis, was met (Bryant and Yarnold, 1995). The items were entered into an exploratory factor analysis. The principal axis (or common) factor analysis method was applied, because its exclusive focus on common variance in items puts the dimensional structure of the data in strong relief (cf. Widaman, 1993). We expected factors to be correlated, as psychological constructs usually are (Zillmer & Vuz, 1995). This would suggest oblique rotation, that allows correlated factors. Orthogonal (varimax) rotation was preferred, however, because of its easier interpretability, and more specifically its ability to show the relative weight of factors in terms of variance explained. To establish correlations between constructs, nonetheless, scale scores were computed based on binary weighted raw scores rather than factor coefficients, which in orthogonal rotation will by definition yield non-correlated factor scores.

All scores on items loading on a factor were assigned a weight of 1, others were given a weight of 0. Resulting weighted raw scores were then summed and averaged. If items loaded on more than one factor, only the highest loading was considered. Missing values were replaced by estimates generated through the estimation maximization procedure in the SPSS Missing Values Analysis module. Factor analysis was repeated for admittees and dischargees separately, to check whether the model was similar for these particular subgroups of patients.

Finally, the relations between items and scales on the one hand, and clinician estimate of risk of reoffending on the other, were assessed through Receiver Operating Characteristic (ROC) curves (Mossman, 1994). In an ROC graph, the balance of true positive predictions (hit rate) is plotted as a function of the number of false positives (false alarm rate) for every cut off point on the scale. The area under the resulting curve (AUC), expressed in a decimal between 0 and 1, indicates predictive validity, with 0.5 equalling non-prediction and values below 0.5 indicating negative prediction.

## II.1.4 Ethics

The project was approved by the Medical Ethics Board of one of the participating hospitals.

## II.1.5 Results

### *II.1.5.1 Factor analysis*

The correlation matrix showed a great number of significant inter-item correlations, which is an essential prerequisite for factor analysis. In default principal axis factoring, nine factors were extracted. This solution, however, did not yield interpretable dimensions, and resulted in very large numbers of high secondary loadings. Several other solutions were therefore tested. Of these, the six factor model proved the most satisfactory: it is shown in Appendix II (p. 187). This more parsimonious model explained 46% of variance in the correlation matrix, not much less than the 50.6% explained by the initial nine factor solution. Though a higher percentage of explained variance would be desirable, achieving it would involve the introduction of a great number of meaningless or single-item factors, which was not preferred.

The appropriateness of factor names was checked by asking three independent judges to place every item under the factor label where they thought it fitted best. Of these classifications, 67% were in accordance with the model. Of items with loadings > .50, 78% were classified correctly. It was therefore concluded that factor names were sufficiently well-chosen.

The first factor concerns the extent to which the patient has a true, empathic understanding of his offence and takes responsibility for it, and was named 'Empathic acceptance of responsibility for the offence'. The second factor, called 'Lack of self-reliance', represents the level of daily life skills, possibly related to psychotic illness and substance abuse. Factor 3 unites a set of symptoms directly related to DSM narcissistic and anti-social personality disorders, and was therefore named 'Anti-social narcissism'. The fourth factor, 'Treatment compliance', focuses on the patient’s participation in the treatment process. Factor 5 comprises themes that can be seen as markers of treatment goals, and was named 'Attainment of treatment goals'. The final factor consists of only two items ('avoids contact' and 'does not tolerate intimacy'), but as these occurred together as a separate factor in all solutions that were computed, it seems they indicate a robust dimension for which items are underrepresented in the CIDRRI. It refers to the patient's tendency to seclude himself and avoid interaction, and was named 'Avoidance'. Correlations of the constituting items with other items suggest that this tendency is not related to anxiety, but more likely to a passive, schizoid-like personality structure.

Table 7 shows that five out of six scales had excellent internal consistency. The lower value of alpha for scale 6 is attributable to the fact that it comprises only two items. Interrater reliability, shown in the same table, was moderate, but similar to results found for dynamic scales in tools like the HCR-20 (Philipse *et al.*, in press). Evaluation of these coefficients should take into account that they were achieved without any special training or instruction of respondents; thus, they provide a clinically realistic reflection of the CIDRRI's performance. Finally, scale intercorrelations are shown, which were strong as expected, ranging from .34 to .81.

The six-factor solutions computed for admittees and dischargees separately could only be tentative, due to the reduced subject-to-variables ratio. They sufficed however for a general comparison with the overall solution. The model for the admittees yielded the same six dimensions as the over-all model, with only minor variations. For dischargees no well-fitting six-factor solution was found; instead, a five-factor model resulted that showed great similarity to the overall-model, but with the 'compliance' and 'goal attainment' factors merged. This suggests that as over the course of treatment (non)compliance results in the (non)realisation of treatment goals, the dimensions measured by both factors converge. However, retaining them as separate dimensions has added value for patients who are not in the final stages of treatment.

TABLE 7. SCALE STATISTICS FOR THE SIX-FACTOR MODEL

| | | internal consistency | interrater reliability | scale intercorrelations (Pearson r) | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $ICC_1$ | scale number | | | | |
| scale | n its | Cronbach α | $/ICC_{av}$ | 1 | 2 | 3 | 4 | 5 |
| 1 Responsi-bility | 11 | .92 | .52/.68 | - | | | | |
| 2 Self reliance | 12 | .87 | .46/.63 | .67 | - | | | |
| 3 Antisocial narcissicm. | 10 | .88 | .67/.81 | .76 | .54 | - | | |
| 4 Compliance | 5 | .85 | .57/.73 | .78 | .64 | .61 | - | |
| 5 Goal attainment | 7 | .85 | .53/.70 | .81 | .59 | .60 | .73 | - |
| 6 Avoidance | 2 | .60 | .59/.74 | .62 | .34 | .51 | .61 | .50 |

NOTE. N = 370, all p < .001, two-tailed. n its. = number of items in the scale. $ICC_1$ = single measure intraclass correlation; $ICC_{av}$ = average measure ICC.

### *II.1.5.2 Items and scales as predictors of clinical assessment of risk*

The concluding step in this study was to test if dynamic patient characteristics that clinicians say they find important for their risk judgement, were in fact predictive of that judgement. This was examined using ROC-curves, which required dichotomization of our dependent variable, the six-point rating of risk. The cut-point was chosen so as to yield a percentage of dischargees with high perceived risk corresponding as closely as possible to the 30% actual serious reoffending reported after TBS (Leuw, 1999; serious reoffending defined as any offence resulting in an unconditional prison sentence or treatment order). This was achieved by a split between values 3 and 4, resulting in the designation of 29.1% of dischargees and 60.3% of all cases as ‘high risk'. Individual item scores, the average score on the entire checklist, and scale scores were all plotted in ROC-graphs. A summary of the findings is displayed in Table 8.

Of individual items, seven had AUC's below .60; three of these values were nonsignificant at the .05 level. Thirty-one items were significantly but weakly predictive of clinical risk assessments (AUC’s between .60 and .70). The remaining nine items had ROC-values > .70, though only in three cases did the lower bound of the 95% confidence interval exceed .70 and could it

TABLE 8. ITEMS, CHECKLIST TOTAL SCORE AND SCALES AS PREDICTORS OF CLINICAL ASSESSMENT OF RISK

| Predictor variable[a] | CLINICAL ASSESSMENT OF RISK | | |
|---|---|---|---|
| | 6-point scale | dichotomous criterion | |
| | Pearson r | ROC area under the curve | 95% CI of area |
| 5. Little control over impulses | .50 | .73 | .68-.79 |
| 19. Understands his pathology | .45 | .70 | .65-.76 |
| 20. Pathology mitigated | .65 | .83 | .78-.87 |
| 28. Knows offence script | .50 | .77 | .72-.82 |
| 32. External situation improved | .56 | .78 | .74-.83 |
| 35. Lacks social skills | .41 | .70 | .65-.76 |
| 39. Guided by impulses | .45 | .72 | .68-.78 |
| 41. Responsibility | .47 | .73 | .68-.79 |
| 45. Allows insight into leave | .48 | .74 | .69-.79 |
| Checklist total score | .64 | .82 | .77-.87 |
| Scale 1: Responsibility offence | .59 | .77 | .74-.83 |
| Scale 2: Self-reliance | .57 | .79 | .74-.84 |
| Scale 3: Cluster B | .52 | .74 | .69-.79 |
| Scale 4: Compliance | .42 | .71 | .65-.76 |
| Scale 5: Treatment goals | .68 | .85 | .81-.89 |
| Scale 6: Avoidance | .30 | .64 | .58-.69 |

NOTE. All p<.001 [a]Of individual items, only those with ROC area's >.70 are displayed

be safely assumed that the item truly had moderate to fairly strong predictive value for the clinical risk assessment. These were item 28, 'Patient knows his offence script'; item 32, 'External circumstances have changed for the better', and item 20, 'Pathology that was at the root of the offence has vanished or mitigated'. This final item seemed the strongest predictive single item with ROC AUC = .83 (95% CI .78-.87). This means that a randomly selected patient who is assessed by the clinician as being at high risk of reoffending has an 83% chance of having a more unfavourable score on this item than a randomly selected patient who was assessed as low risk. It should be noted though that the confidence intervals of most items overlapped, so that a strongest predictor could not be identified with certainty.

In summary, most of the items in the CIDRRI were significantly related to clinicians' estimates of reoffending risk. The checklist total score did not perform better than the strongest predictive individual items. This suggests that when assessing risk, specific characteristics of the patient may have an equally strong influence on clinical risk assessment as has the more general picture of the patient's condition; on the other hand, it may also simply indicate that the CIDRRI contains more items than are necessary.

The 6 scales were all significantly related to clinical risk assessment. All except the 'Avoidance' scale yielded AUC values over .70, and can be considered at least moderately predictive of clinical assessment of risk. The strongest predictor of clinical assessment that could be extracted from the present data overall was scale 5, 'Achievement of treatment goals'. The AUC value was .85 (95% CI .81-.89), indicating a strong relation to the clinical assessment of long-term risk, and, not surprisingly, stressing the importance clinicians attach to treatment success as an indicator of risk of reoffending.

Possibly, the significance of predictors for the clinical risk estimate changes depending on the moment of assessment. The findings show this to be the case. Interestingly, scale scores are uniformly stronger predictors of clinical risk perceptions for newly admitted patients than for those about to be discharged. One explanation may be that for the latter group, external factors not accounted for in the checklist, such as availability of aftercare, or having lodgings and a job, will help to determine the perceived risk levels. In accordance with this interpretation, 'Lack of self-reliance' is the strongest predictor of perceived risk for admittees, with an ROC area under the curve of . 87 (95% CI .78-.96), as opposed to an area of .77 for dischargees (95% CI .62-.80). For this second group, empathic acceptance of the offence and attainment of treatment goals emerge as the strongest predictors of the clinical risk estimate (area's of .79 both). For both groups, 'Avoidance' has the least impact on clinical risk perceptions.

To what extent differences reflect differences in samples rather than treatment stages could not be established in detail, though available descriptive data on both patient groups and the general homogeneity of the TBS-population provide no reason to assume a strong distortion of this kind.

## II.1.6 Discussion

This study was concerned with exploring dynamic estimators of risk, which clinicians had previously identified as important. The main questions were: first, whether a meaningful dimensional structure underlied the patient characteristics quoted by clinicians as pertinent to their assessment of risk; second, whether these characteristics and the underlying dimensions were in fact significantly related to the clinical assessment of risk.

The answer to both questions was affirmative. Checklist items could be grouped into six well interpretable factors. Five of six scales based on these factors, and nine individual items, showed substantial predictive power with regard to clinical risk assessment. Additionally, it was found that both factor structure and patterns of association between items or scales and clinical risk assessment somewhat depended on the time of assessment (start or end of treatment). These variations were however not at odds with the interpretation of the overall findings.

It is reassuring to note that the most salient characteristics underpinning clinical risk assessment were similar to risk factors found in commonly used risk assessment instruments (e.g., HCR-20 or LSI-R): cluster B personality disorder, compliance, reduction of pathology, openness about activities, a functional and supportive social network, and impulse control. More specifically, our checklist and its underlying dimensions show a marked similarity to the dynamic scales empirically constructed by Quinsey *et al.* (1997). Of the 28 items listed by these authors, 16 have a direct equivalent in the CIDRRI. These include: lack of remorse; lack of consideration of others; being suggestible and easily led; problems with money management, housekeeping and personal hygiene; as well as pathology similar to that at the time of offence, and impulsivity. These similarities are all the more interesting as our checklist was compiled at a point in time when neither the HCR-20 nor the Quinsey study had yet been published. The advantage of our checklist is that, in contrast to these other scales, it can be completed to an acceptable level of reliability without specific score related training.

One obvious shortcoming of the present study is the fact that dynamic risk factors and clinical risk assessments were not linked to actual reoffending - as will be done in the concluding stages of this project. One might argue that the links we found between ratings of dynamic patient characteristics and ratings of perceived risk of reoffending, both provided by the same clinician, are hardly surprising, if not redundant. However, as Monahan &

Steadman (1994) have noted, "[t]o understand better how clinicians reach clinical judgments is of intrinsic scholarly interest and may yield valuable information about the factors clinicians believe to be predictive of violence" (p. 7). The continued neglect of this theme in published research, with its strong focus on predictive validity and methodologically sound prediction procedures, has arguably widened the gap between scientific ideal and clinical reality rather than bridged it, as exemplified by an increasingly acrimonious clinical-versus-actuarial debate (Webster *et al.*, 1997). We think studies like ours may help identify points of contact between clinical and scientific thinking on risk assessment, and bring the two closer together. Apart from that, the conclusion that the underpinnings of clinical risk assessments are coherent and orderly is hardly a trivial one given the amount of literature claiming otherwise (e.g., Grove & Meehl, 1996; Quinsey *et al.*, 1998; Meehl, 1954).

To conclude, the findings in this study show that the CIDRRI reflects to a significant extent the true dynamic underpinning of risk assessments by clinicians. Clinical risk assessment relies heavily on treatment related characteristics of the patient: have treatment goals been achieved, has pathology been mitigated, and does patient have sufficient skills to guarantee an acceptable level of functioning in daily life? It is also considered important that the patient takes responsibility for his offence(s) and shows remorse.

It remains to be seen whether these characteristics also predict actual reoffending rather than only the clinical expectation of it. This test will also clarify whether the dimensions we found are merely the result of clinical stereotyping (e.g., 'the psychotic', 'the remorseless criminal'), or may be viewed as actual patient attributes related to risk of reoffending.

# II.2 ☙ Study II

## *Reliability and discriminant validity of dynamic reoffending risk indicators in forensic clinical practice*

This chapter is a reprint of: Philipse, M.W.G., Koeter, M.W.J., Staak. C.P.F. van der, & Brink, W. van den. (in press). Reliability and discriminant validity of dynamic reoffending risk indicators in forensic clinical practice. *Criminal Justice and Behavior*. Reprinted by kind permission of Sage Publishers, Thousand Oaks, CA.

## II.2.1 Introduction

Over the years many tools for assessing risk of criminal reoffending have been developed. Doren (1999) noted that in the United States 25 such tools were available for sex offenders alone. Yet at the same time Hart (1999) observed that "the state of the science simply does not allow the conclusion that a solution has been found for the problem of risk assessment" (p. 487). The profusion of instruments itself demonstrates the lack of consensus on the risk assessment question. Moreover, variations in method of construction, item content and rating procedure rarely seem to yield variations in predictive performance consistent enough to support a preference for a particular tool (e.g., Gardner, Lidz, Mulvey & Shaw, 1996; Kroner & Mills, 2001; Sjöstedt & Långström, 2002).

The need for reliable risk assessment tools is strongly felt in forensic psychiatry, where reoffending risk is arguably the single most important criterion for treatment evaluation. From this viewpoint, existing instruments and research have two major shortcomings. First, well-established tools such as the Violence Risk Appraisal Guide (VRAG; Quinsey, Harris, Rice & Cormier, 1998), the Iterative Classification Trees (Monahan *et al.*, 2001) or the Static-99 (Hanson & Thornton, 1999) are mainly comprised of historical factors, and thus resulting risk levels are unlikely to change. Therefore, in risk management, these tools can do little more than suggest desirable levels of external control, such as community supervision or incarceration.

Only a small number of risk assessment instruments include dynamic items in order to enhance clinical relevance. One of these, the Level of Service Inventory-revised (LSI-R; Andrews & Bonta, 1995) has been shown to be a good measure of risk (Gendreau, Little & Goggin, 1996; Gendreau, Goggin & Smith, 2002). Findings concerning other notable dynamic instruments, the Historical/Clinical/Riskmanagement-20 (HCR-20; Webster, Douglas, Eaves & Hart, 1997) and the Sex Offenders Needs Assessment Rating (SONAR; Hanson & Harris, 1998) may be more contentious, but seem to hold some promise (Douglas, Ogloff, Nicholls & Grant, 1999; Belfrage, Fransson & Strand, 2000; Hanson & Harris, 2001). Nevertheless, studies examining dynamic risk predictors are far outnumbered by those focusing on static characteristics. Apparently the long-running actuarial-versus-clinical debate, which resulted in the general conclusion that simple actuarial methods outperform unstructured clinical risk assessment (Grove & Meehl, 1996; Mossman, 1994), has prompted researchers to turn away from the needs, let alone the input of clinicians. Yet several studies have shown that clinically relevant, dynamic indicators can yield equally accurate predictions as static factors (Gendreau *et al.*, 1996). Quinsey, Coleman, Jones & Altrows (1997) identified seven dynamic factors with short term predictive validity for eloping when controlling for historical factors.

A second shortcoming of risk assessment research from the forensic mental health viewpoint is the preponderance of clinically unrealistic, retrospective instrument validation. This approach is optimized to achieve good reliability in ways that are hard, if not impossible to replicate in quotidian clinical practice. The plethora of information available in a clinical environment is often replaced by files only (or even mere case vignettes), while extensive and time consuming training and instruction procedures are put into place. Often raters are not clinicians but researchers, familiar with the prerequisites for attaining good reliability, at leisure to perform ratings free from personal sympathies or daily clinical hassles, and thus likely to produce better results than clinicians who lack these advantages. Accordingly, De Vogel & De Ruiter (2002) found between-clinician interrater reliability on dynamic risk items in a forensic in-patient setting to be substantially lower than that between researchers.

Unfortunately, the alternative prospective research design is generally unpopular in risk assessment research, due to the expansive timeframe required. Available prospective work, most notably the MacArthur study (Monahan *et al.*, 2001), tends to focus on short term risk in the community or within the hospital context itself.

The present study set out to meet the shortcomings discussed above by evaluating a clinically derived, dynamic risk assessment tool in a long-term, prospective follow-up design spanning the years 1996 through 2004. More specifically, we examine whether a clinically realistic data gathering procedure with a focus on dynamic predictors is compatible with acceptable reliability and discriminant validity. Findings are discussed taking into account the strengths and limitations of the checklist in comparison to similar instruments.

## II.2.2 Method

This multi-center study was set in The Netherlands, within the system of *terbeschikkingstelling* (TBS), a court-ordered treatment measure that can be imposed on perpetrators of severe offences who are not fully accountable due to a mental disorder. TBS entails an involuntary stay of indefinite length in an in-patient forensic mental health facility, aiming at an eventual safe reintegration into society. The in-hospital treatment phase on average takes about 5 years. Release decisions are court-based, whereas permission for probationary leave is granted by the Ministry of Justice. Both decisions depend strongly on the assessment of reoffending risk.

Separate sets of checklist data were compiled to address each of the three research issues: interrater reliability (Study A), retest reliability (Study B) and discriminant validity (Study C).

### *II.2.2.1 Participants*

Staff from 8 TBS-hospitals participated in one or more of the studies. An overview of participants per study is shown in Table 9. The majority of respondents were treatment supervisors: academically educated, mostly experienced psychologists or psychiatrists who are sworn expert witnesses. Head nurses, participating as co-raters in Study A, received higher vocational education as psychiatric nurse or social worker and are responsible for day-to-day patient management on the ward as well as directing the team of social therapists. Proportions of male and female respondents within each group did not differ significantly.

TABLE 9. NUMBERS OF PARTICIPANTS, PARTICIPATING HOSPITALS AND RATERS IN THE STUDIES

| study | patient n | hospital n | raters: n, function |
|---|---|---|---|
| A. Interrater reliability | 75 | 5 | 26 ts[a] 34 hn[b] → 36 pairs |
| B. Retest reliability | 29 | 3 | 13 ts |
| C. Discriminant validity | | | |
| admittees | 114 | 8 | 38 ts (108 ratings),<br>8 no[c] (6 ratings) |
| dischargees | 118 | 7 | 34 ts (109 ratings),<br>9 no (9 ratings) |

NOTE. [a]ts = treatment supervisor. [b]hn = head nurse. [c]no = nurse or other functionary.

Patients in all three studies were male. Mean age was 33.7 years (SD = 8.1) in Study A, and 33.9 years (SD = 8.7) in Study B. Patients in Study A had been convicted to TBS for sex offences in 38.4% of cases and for violent offences in 49.3% of cases. Of all offences, 25.1% resulted in victim death. In Study B, the corresponding figures were 41.3%, 48.2% and 27.6%

respectively. At the time of rating, patients had been in the caseload of respondents for at least 6 months.

In Study C, the average age of admittees (n=114) was 31.5 years (SD = 8.2), of dischargees (n = 118) 34.6 years (SD = 8.6). This significant age difference ($p < .05$) is explained by the duration of TBS treatment. As for index offenses, both groups were similar with regard to proportions of sex offenses (21% of admittees, 17% of dischargees, ns), violent offenses (64% vs. 66%, ns) or offenses resulting in victim death (32% vs. 27%, ns).

### *II.2.2.2 Materials*

The Clinical Inventory of Dynamic Reoffending Risk Indicators (CIDRRI) is a checklist containing 47 statements on dynamic patient characteristics (listed in Appendix I, p. 183). Each item is to be rated on a six-point scale, with extremes defined as 1 = 'not at all characteristic' and 6 = 'very characteristic'.

CIDRRI items resulted from interviews with 12 treatment supervisors in 4 Dutch forensic in-patient settings on the question which patient characteristics they considered pertinent to the assessment of reoffending risk. The item-pool was edited by a research-team of representatives from the hospitals involved in the present study. Clinicians monitored this process to ensure that the resulting checklist would remain an adequate reflection of clinical considerations in all hospitals. Finally, a 48th item was included that asked for a direct rating of the perceived long-term risk of reoffending by the patient after discharge. The CIDRRI is designed to be rated by the treatment supervisor and requires no additional training. Scoring instructions briefly explain the 6-point rating scale and encourage the respondent not to skip any questions. Judgments should reflect the clinician's current view of the patient. Completing the checklist typically takes 10 to 15 minutes.

An earlier investigation showed clinical item ratings to be significantly associated with clinical assessment of risk according to the 48th item. Also, principal axis factoring with orthogonal rotation yielded six clinically meaningful scales, as shown in Table 10 (Philipse, Koeter, Van den Brink & Van der Staak, 2004).

TABLE 10. SIX SCALES OF THE CLINICAL INVENTORY OF DYNAMIC REOFFENDING RISK INDICATORS

| scale name *(and brief description)* | n of items | items[a] | α[b] |
|---|---|---|---|
| 1. Empathic acceptance of responsi-bility for the offence *(patient takes responsibility for his actions, does not minimize the seriousness of the offence, and is capable of empathy towards victims and others)* | 11 | 2, 3, 8, 9, 11, 12, 19, 24, 30, 33, 41 | .92 |
| 2. Lack of self-reliance *(patient has limited daily life skills regarding finances, hygiene, day structure etc.)* | 12 | 1, 10, 16, 17, 18, 31, 35, 36, 37, 38, 39, 46 | .87 |
| 3. Anti-social narcissism *(patient is self-centered, aggressive, impulsive and grandiose)* | 10 | 5, 6, 22, 23, 25, 27, 29, 40, 43, 44 | .88 |
| 4. Treatment compliance *(patient participates actively in treatment to the best of his capabilities)* | 5 | 13, 14, 15, 34, 42 | .85 |
| 5. Attainment of treatment goals *(general TBS treatment goals have been achieved, e.g. patient knows offence script, asks for help if necessary)* | 7 | 4, 20, 21, 28, 32, 45, 47 | .85 |
| 6. Avoidance *(patient does not tolerate intimacy and stays away from others)* | 2 | 7, 26 | .60 |

NOTE. [a]See Appendix I (p. 183) for item descriptions. [b]Cronbach's alpha for internal consistency

### *II.2.2.3 Procedure*

*Study A*. Because each patient has only one treatment supervisor, no two truly equivalent raters were available for the interrater reliability study. Instead, as the closest approximate the head nurse was asked to perform the second rating. Fifteen cases were obtained from each of 5 participating hospitals; nearly all treatment supervisors and approximately half of all head nurses in these hospitals contributed to the study. For practical reasons, the choice of patients was left to respondents. Raters were instructed to perform paired ratings of each case within the same week, and not to discuss the checklist together. Inevitably, they will have been in contact for regular treatment consultation, but we felt this would not compromise the validity of the study as a reflection of the CIDRRI's reliability under normal clinical circumstances.

*Study B.* Nearly all treatment supervisors in 3 hospitals provided 2 or 3 double ratings each, for patients of their own choice. Respondents were instructed to complete the second rating one month after the first. The purpose of this repeated rating was explicitly stated in the instruction, and raters were encouraged not to make any effort to remember their scores. The checklist for the second rating contained some additional questions regarding any unusual circumstances between ratings that might have influenced item-scores.

*Study C.* As yet in the absence of recidivism data, we chose to compare ratings for all patients newly admitted to any of 8 hospitals during 1996, to those of all patients from 7 hospitals about to re-enter the community on probationary leave between January 1 1996 and December 31 1998. Forensic mental health treatment presupposes that untreated patients will generally pose a higher risk than those who have finished treatment. Some support for this assumption is found in research. Greeven (1997) showed that treatment in a TBS-hospital resulted in significant reduction of acting-out symptoms and improvements in interpersonal functioning, as well as reduction in overall DSM-III personality pathology, characteristics known to correlate with future violence.

Probationary leave is the preferred way of reintroducing the patient into the community, usually signaling satisfactory treatment completion. Hanson & Bussière's (1998) finding that treatment completion is related to reduced reoffending risk in sex offenders, is of interest in this respect. Wormith & Olver (2002) found treatment completion to be related to lower recidivism risk in violent offenders. During probationary leave, the TBS-measure remains in force and some level of supervision is maintained. Leuw (1999) showed that patients who had probationary leave prior to the end of their TBS-measure had a significantly lower likelihood of reoffending than those discharged directly into the community. Finally, it should be noted that one hospital only supplied admittee subjects: the high security Mesdag Hospital, which specializes in managing high-risk TBS-cases. This coincidental factor (the hospital withdrew from the rest of the study due to lack of research staff to co-ordinate data-collection) helped to increase the 'risk-contrast' between admittees and dischargees. Taking all this information into account, it appeared probable that risk levels would in fact generally be higher for newly admitted patients than for those starting their probationary leave.

## II.2.3 Results

*Study A*. Mean differences within rater-pairs for items, scales, and total score were examined through *t* tests. Item mean scores did not differ significantly except on 3 items. Item 14, 'Patient perseveres in resistance against treatment', was systematically judged more unfavorably by treatment supervisors than head nurses (mean difference = .40, t(74) = 2.33, p < .05), as was item 38, 'Patient has insufficient internal structure to stand on his own' (mean difference = .53, t(74) = 2.36, p < .05). Reversely, item 47, 'Patient alternates between idealization and devaluation of people close to him', was rated more unfavorably by head nurses than treatment supervisors (mean difference = -.62, t(74) = -3.51, p < .01). The same held true for Scale 5, 'Attainment of treatment goals' (mean difference = -.16, t(74) = -2.37, p < .05). No differences were found for the means on the six-point clinical assessment of risk (item 48) and the CIDRRI total score.

Next, intraclass correlations (ICC's; Bartko, 1966; Shrout & Fleiss, 1979) were computed for items, scales and list total. The two-way random effects, single measure, absolute agreement ICC model was appropriate, according to the decision rules set out by Shrout & Fleiss and McGraw & Wong (1996). The two–way model reflects the fact that in the present data both rows (items, scales) and columns (raters) were sources of variance. We did not intend to investigate reliability between two specific rater-categories, but assumed raters in a pair to be equivalent, requiring a model with random rather than fixed rater effects. Absolute agreement ICC takes full account of between-rater variance, thus testing exact agreement of scores rather than mere consistency in score patterns. Single measure ICCs best reflect the most likely clinical use of the CIDRRI, as they indicate the level of reliability attained with only one judge per case. Average measure ICCs, the higher reliability levels resulting from averaged ratings of the same case by several judges, were computed to gain insight into the number of judges required to attain improved ICC-levels. Fleiss (1981) and Cicchetti & Sparrow (1981) have proposed cut-off levels for the evaluation of ICCs as follows: values below .40 are low; .40 through .59 are fair; .60 through. 74 are good; and values over .75 are excellent. The Spearman-Brown prophecy formula, nj = ICC*(1 - rl)/rl(1 - ICC*), was applied to average measure ICC's to establish the number of raters required to attain the .75 level of reliability (where nj is the required number of judges, ICC* is the desired minimum ICC-value, and rl is the lower bound of the average measure 95% confidence interval).

As shown in Appendix I (p. 183), all single measure ICC's for items were at or above the .40 minimum except for items 26, 27, 32, 35, 38, and 40. ICC single and average measure values with associated confidence intervals for the six scales and total score, as well as some additional descriptive data, are listed in Table 11. All scales had at least fair interrater reliability. A minimum of 5 averaged ratings would be required to guarantee a .75 level of interrater reliability on all scales, though 3 raters would suffice for most.

*Study* B. The strategy for data analysis was identical to Study A, with the omission of average measure ICC's, as these lack relevance to the test-retest measurement. The average time between first and second rating was 39 days (range 22 – 92, SD = 18). In 5 cases some special circumstances were reported to have occurred between both ratings. These pertained to reduced (n = 1) or increased (n = 3) liberty of movement, or to the successful start of psychopharmacological treatment (n = 1). None of these circumstances could be clearly related to lack of stability in any items.

TABLE 11. INTERRRATER RELIABILITY OF SCALES AND CIDRRI TOTAL SCORE.

| scale | range | | | SD[b] | agree- | s.m. ICC[d] | a.m. ICC | n judges |
|---|---|---|---|---|---|---|---|---|
| | rater[a] | min. | max. | | ment[c] | (95% C.I[e].) | (95% C.I.) | .75[f] |
| 1 | a | 2.27 | 4.45 | .495 | 69% | .52 (.33-.67) | .68 (.50-.80) | 3 |
| | b | 2.18 | 4.55 | .529 | | | | |
| 2 | a | 2.83 | 4.58 | .426 | 68% | .43 (.23-.60) | .60 (.38-.75) | 5 |
| | b | 2.83 | 4.58 | .370 | | | | |
| 3 | a | 1.20 | 5.40 | .890 | 71% | .67 (.52-.78) | .80 (.68-87) | 2 |
| | b | 1.50 | 5.30 | .941 | | | | |
| 4 | a | 2.40 | 5.00 | .585 | 63% | .57 (.40-.70) | .73 (.57-.83) | 3 |
| | b | 1.80 | 5.10 | .669 | | | | |
| 5 | a | 1.86 | 4.86 | .664 | 77% | .56 (.38-.70) | .72 (.55-.82) | 3 |
| | b | 1.71 | 4.57 | .641 | | | | |
| 6 | a | 1.00 | 6.00 | 1.120 | 77% | .59 (.42-.72) | .74 (.60-.84) | 2 |
| | b | 1.00 | 6.00 | 1.200 | | | | |
| Tot. | a | 126 | 200 | 15.14 | 73% | .55 (.37-.69) | .71 (.54-.82) | 3 |
| | b | 127 | 195 | 15.99 | | | | |

NOTE. N=75. All p < .0001. Scales: 1. Empathic acceptance of responsibility for the offence; 2. Lack of self-reliance; 3. Antisocial narcissism; 4. Treatment compliance; 5. Attainment of treatment goals; 6. Avoidance. Tot. = CIDRRI total score. [a]Rater a = treatment co-ordinator, rater b = head nurse. [b]SD = standard deviation. [c]Agreement = percentage of rater pairs with scores less than 1 SD from each other, using lowest of both SD's. [d]ICC = intraclass correlation; s.m. = single measure, a.m. = average measure. [e]C.I. = confidence interval. [f]Number of judges required to obtain an a.m. ICC of at least .75.

A significant mean difference was only found for item 20, 'The pathology underlying the offence has been lifted or mitigated', which, surprisingly, was rated less favorably the second time (mean difference = -.3448, t(28) = -2.167, p < .05).

Item single measure ICC's are reported in Appendix I (p. 183). Of these, 37 (78%) were good to excellent (ICC > .60); only item 43 fell below the .40 threshold. Retest ICC's for scale and total scores are displayed in Table 12. These were all good to excellent.

*Study C.* CIDRRI-means for admittees and dischargees were compared (t tests). Next, a k-means cluster analysis was performed to see if two contrasting groups of cases clustered on the basis of checklist ratings would correspond to a significant degree with the low and high risk patient groups. The cluster procedure identifies n reference cases, with n being the required number of groups, that have maximally contrasting scores on the variables determining the clusters. Other cases are then assigned to the reference case they resemble most.

TABLE 12. TEST-RETEST RELIABILITY OF SCALES AND TOTAL SCORE

| Scale | | score range | | SD[b] | agree | s.m.[d] | 95% |
|---|---|---|---|---|---|---|---|
| | t[a] | min. | max. | | ment[c] | ICC[e] | C.I.[f] |
| 1. Empathic | 1 | 2.45 | 3.82 | .398 | 86% | .72 | .49-.86 |
| responsibility | 2 | 2.36 | 3.91 | .435 | | | |
| 2. Lack of self- | 1 | 2.83 | 4.42 | .406 | 86% | .71 | .47-.85 |
| reliance | 2 | 2.58 | 4.42 | .451 | | | |
| 3. Anti-social | 1 | 1.40 | 5.00 | .764 | 93% | .82 | .65-.91 |
| narcissism | 2 | 1.60 | 5.00 | .824 | | | |
| 4. Treatment | 1 | 1.80 | 5.00 | .766 | 76% | .73 | .50-.86 |
| compliance | 2 | 2.00 | 4.60 | .708 | | | |
| 5. Attainm. of | 1 | 1.14 | 4.14 | .829 | 86% | .75 | .53-.87 |
| treatm. goals | 2 | 1.57 | 4.29 | .792 | | | |
| 6. Avoidance | 1 | 1.00 | 5.50 | 1.316 | 90% | .79 | .60-.89 |
| | 2 | 1.00 | 5.50 | 1.300 | | | |
| CIDRRI total | 1 | 122.20 | 187.06 | 11.801 | 86% | .74 | .51-.87 |
| score | 2 | 125.02 | 192.23 | 14.518 | | | |

NOTE. N = 29. All p < .0001. [a]t = test (1) or retest (2); [b]SD = standard deviation; [c]agreement = percentage of t=2 scores differing less than 1 SD from t=1 score, using the lowest of both SD's; [d]s.m. = single measure; [e]ICC = intraclass correlation; [f]C.I. = confidence interval.

Items 13, 'Patient faithfully takes prescribed medication' and 45, 'He allows insight into the way he spends his leave', had large numbers of missing values in the admittee group and were deleted from the analysis. This also affected the computation of scales 4 and 5. Thirty five out of the 45 remaining dynamic characteristics as well as the risk assessment item showed significant differences in mean scores between admittees and dischargees ($p < .05$, see Appendix I for details), as did the checklist total score and 5 of 6 scales. All differences were in the expected direction, with newly admitted patients receiving the more unfavorable rating. Contrast was particularly high for the 6-point rating of perceived long-term risk on the risk assessment item (mean difference = -2.65, t(216) = -16.20, $p < .01$). The 10 items with non-significant mean differences pertained to being easily influenced; contact avoidance; treatment participation; circadian rhythm; self-care; active use of leisure time; fear of abandonment; inflated self-image; psychotic symptoms; and active use of learning opportunities. The non-significant scale was scale 4, 'Treatment compliance'.

Sensitivity (i.e, the percentage of admittees clustered together in the group containing their majority) and specificity (i.e., the percentage of dischargees clustered in the group containing their majority) of the cluster groupings are displayed in Table 13. Scores on items and scales discriminated to an appreciable extent between recently admitted, high risk patients and lower risk patients starting their probationary leave. The list total score however did not. The extremely accurate clustering based on the clinical risk assessment may be somewhat redundant, as a clinician is not likely to rate a newly admitted patient as low risk, nor a patient who has been granted probationary leave as high risk.

TABLE 13. RISK GROUP CLASSIFICATION RESULTS FOR SEVERAL *k*-MEANS CLUSTER SOLUTIONS BASED ON CHECKLIST DATA

| cluster variable(s) | percentage of risk-group clustered together | | | kappa[a] |
|---|---|---|---|---|
| | % low risk (specificity) | % high risk (sensitivity) | % low + high (overall accuracy) | |
| 47 dynamic items | 77.1** | 60.5 | 68.5** | .37** |
| 6 scales | 88.6** | 71.9** | 79.9** | .61** |
| Long term risk assessment (it. 48) | 74.3** | 92.9** | 83.9** | .67** |
| CIDRRI totalscore | 59.4 | 57.0 | 58.0* | .16* |

NOTE. *$p < .05$. **$p < .01$. For percentages, p indicates significance of the difference with chance classification. [a]Kappa is the measure of agreement between risk-group membership and cluster membership.

## II.2.4 Discussion

This study investigated reliability and discriminant validity of a 47-item (6 scale) dynamic risk assessment tool under naturalistic hospital conditions in Dutch forensic in-patient facilities. Aiming for clinically feasible procedures, rater training and a coding manual were not utilized.

No significant mean differences between paired raters were found except on 3 items and 1 scale. Single measure intraclass correlations were fair to good for 41 items and all scales. For items that had differing means (items 14, 38, 47), more explicit definitions for the rating scale anchor points may draw the mean score levels of raters closer together. For items with insufficient ICC's (items 26, 27, 32, 35, 38, 40), the item content itself may need elaboration or further specification. Item 38 yielded both a significant mean difference and low ICC, indicating significant disagreement between raters both in absolute and relative score levels. It is interesting to note that the assessment of patients' ability to function self-reliantly in daily life is the greatest source of discrepancy between raters.

CIDRRI interrater ICC's were grossly comparable to, if somewhat lower than those of the dynamic Clinical and Riskmanagement domains from the widely used HCR-20 (Webster *et al.*, 1997). For these, Cooke, Michie & Ryan (2001) reported single measure ICCs of .74 and .70 respectively; Vincent *et al.* (2001) found .70 and .58 respectively; Coté (2001) reported $ICC_1$ = .71 for the C-scale, and Müller-Isberner & Jöckel (1997) kappa = .49. For a structured clinical assessment of risk based on HCR-20 total scores, Douglas, Ogloff & Hart (2003) found $ICC_1$ = .61 (cited in Douglas & Ogloff, 2003).

Schene *et al.* (2000) note that such values do not meet reliability standards current for psychological tests (kappa > .70, ICC > .90). They add, however, that these may be unrealistic for instruments that, like the CIDRRI, are not 'tests' for measuring a particular well defined construct, but tools for gauging more diffuse concepts (i.e., that have blurred boundaries with other constructs), and involve diverse item-pools. In such cases, these authors suggest, values from .50 to .70 can be considered 'moderate'.

Yet, for dynamic risk assessment instruments of this kind like the LSI-R or the SONAR high reliability has been reported. Andrews & Bonta (1995) found LSI-R interrater reliability to be in the range of .80 to .99 (Pearson r). However, the Pearson r is a less critical measure than the ICC, and moreover these findings only pertain to the final risk/needs assessment on a 5-point scale. For the SONAR Hanson & Harris (2001) reported levels of agreement between 94 and 97%. These impressive rates of concordance were however achieved at the cost of employing field researchers, provided with a week of training, an extensive coding manual, full supervision at the start of data collection, and regular teleconferences to reduce rater drift. Coding of interviews and files took from 4 to 6 hours per case (Hanson & Harris, 1998). Though from a research viewpoint these efforts to increase reliability are exemplary, they will not easily be replicated on a day-to-day basis in an average forensic hospital, and resulting reliability findings may thus have little practical meaning.

In contrast, the paired raters in our study were clinicians who received no specific training, performed ratings on 6-point scales (as opposed to the more common dichotomous or 3-point scales), operated in the everyday clinical context, and completed the checklist in 10 to 15 minutes. It should be noted, too, that paired raters were from different clinical disciplines, which may have resulted in lower interrater reliability than truly equivalent raters would have achieved. Taking all these circumstances into account it seems reasonable to assume that in clinical practice our instrument will perform at least equally well as comparable risk assessment tools with regard to reliability. Test-retest reliability was satisfactory: within-rater consistency was good to excellent for ratings repeated within a timespan of one to two months.

The true test of the CIDRRI's validity is its predictive power with regard to actual reoffending, which will be the focus in the final stage of the present project. Theoretically the CIDRRI, as an example of structured clinical assessment, may be assumed to outperform pure clinical judgment, as any form of structure improves predictive performance over subjective, unstructured approaches (Webster *et al.*, 1997). In the absence of final recidivism data, predictive validity was approximated by comparing ratings for admittees with those for dischargees. Mean scores differed significantly and consistently in the expected direction, with newly admitted patients receiving more unfavorable ratings than patients starting probationary leave on 80% of items, all but one of the scales, and the checklist total. Two scale-based clusters corresponded to a large extent with these groups. In all, the CIDRRI therefore appears to be fairly well able to discriminate newly admitted patients, assumed to be high risk from, assumedly lower risk patients about to be sent on probationary leave.

A certain amount of biased clinical perception cannot be ruled out as a possible explanation of this finding. Clinicians who know that a patient is new to the hospital may be inclined to perceive him as more dangerous, disturbed, and difficult to manage than others, irrespective of his actual condition. Unfortunately the nature of the checklist and the required raters precluded blind ratings. However, ratings for the two patient groups were performed at different points in time by nonselect raters, and clinicians rating newly admitted patients were unaware of the purpose of this rating. These circumstances, as well as the similarity of both patient groups on some key characteristics, provide reason to assume that score differences are to an appreciable extent related to actual differences in patient characteristics.

In sum then, our findings warrant the conclusion that the instrument we constructed is sufficiently reliable as compared to available alternatives, and there are positive indications for its validity. Future research will assess predictive validity of the CIDRRI in detail using 5-year follow-up data on criminal recidivism. If findings give occasion for it, the checklist will then be revised to yield a concise clinical risk assessment tool for TBS.

*Clinical implications*. The moderate levels of interrater reliability raise some questions with regard to clinical practice. Both content and wording of our items were largely determined by clinicians themselves, reflecting terminology that is deeply engrained in daily clinical practice and routinely found in treatment progress reports, applications for leave, and court advise on TBS-extension. Yet, two clinical practitioners will apparently quite often judge these items differently for a case with which they are both intimately familiar. Different professional perspectives as well as ambiguities in terminology itself may underlie these discrepancies. Both need to be remedied to increase the reliability of any communication about the patient's condition. This implies some form of interdisciplinary consensus on definitions, formalized in a set of instructions. However, it is questionable to what extent such instructions would prove effective. As Slovic, Monahan & MacGregor (2000) have compellingly demonstrated, clinicians may be inclined to ignore available instructions. Webster, Müller-Isberner & Fransson (2003) noted this with regard to the HCR-20, complaining that "some individual practitioners have begun and ended their study of the HCR-20 with a reading of the coding sheet" (p. 45).

The alternative solution, as our findings suggest in accordance with conclusions by Cooke *et al.* (2001) and McNiel, Lam & Binder (2000), would be to always base assessments of dynamic patient characteristics on averaged multiple independent judgments. Such an investment would be advisable especially in the context of risk assessment: the more far-reaching instrument-based decisions are, the stricter reliability requirements should be (Nunnally & Bernstein, 1994). We can argue that for research purposes reliability levels from .40 upwards are acceptable, but they are not acceptable as a basis for decisions on freedom of movement (i.e., that entail the possibility of new offences and new victims). Given the levels of reliability found in our study, a minimum of 3 raters would generally suffice to guarantee the .75 threshold suggested by Fleiss (1981) as the lower bound of excellence.

If any risk assessment tool is to be used successfully in clinical practice, either some aspects of that practice need to change drastically, or the instrument needs to be simple, easy-to-use, and not encumbered by extensive training requirements or 50-page instructions (cf. Gardner *et al.*, 1996). Aiming at the second, less Utopian of both goals, the CIDRRI appears a viable option. As a consequence, psychometric results are less spectacular than those reported for some other instruments, but they are to all likelihood rather more realistic. Three to 5 independent clinical raters who subsequently establish one consensus rating will produce excellent reliability and increased validity. For this tool and indeed any form of risk assessment such a consensus procedure seems highly recommendable in clinical practice.

# II.3 ಆ Study III

## *Static and dynamic patient characteristics as predictors of criminal recidivism. A prospective study in a Dutch forensic psychiatric sample.*

This chapter is a reprint of: Philipse, M.W.G., Koeter, M.W.J., Staak, C.P.F. van der, & Brink, W. van den. (submitted). Static and dynamic patient characteristics as predictors of criminal recidivism. A prospective study in a Dutch forensic psychiatric sample.

## II.3.1 Introduction

Assessment of risk of criminal recidivism in patients is one of the core activitities in forensic mental health, and has been the subject of many studies. Although confidence in predictive possibilities has increased over time (Bonta, Law, & Hanson, 1998), several issues remain unsolved. The preponderance in risk assessment research of static risk factors, characteristics that cannot be altered through clinical intervention, has created a gap between research and clinical practice, where risk indicators are needed that offer handles for treatment (Dolan & Doyle, 2000). Researchers continue to differ as to whether a search for alternative, clinically useful 'dynamic' predictors is necessary and whether such predictors can be identified with any certainty (Dempster & Hart, 2002). Research on dynamic risk assessment tools that is thus far available, largely leaves unclear whether dynamic predictors have any incremental value given a basic set of static predictors.

The actual efficacy of established predictors in (prospective) clinical reality, including their effect on recidivism rates, also remains unknown. Research is dominated by retrospective designs, optimized to yield good reliability in ways that are often not clinically feasible; they provide an idealised image of instrument performance (Dolan & Doyle, 2000; Gardner *et al.*, 1996). Practical implementation of such instruments is likely to result in shrinkage of their reliabilty and predictive power

The present study set out to identify clinically relevant dynamic risk factors. Predictive validity of these factors was evaluated while controlling for static risk markers. This was done in a prospective, naturalistic setup.

## II.3.2 Method

### *II.3.2.1 Setting*

The study was set in The Netherlands, in forensic inpatient settings for the execution of the measure of so-called *terbeschikkingstelling* (TBS). TBS is a court-ordered treatment measure that can be imposed on perpetrators of serious violent or sex offences who were not fully accountable for their criminal acts due to a mental disorder at the time of the offence, and who are considered to be at risk to reoffend. TBS is imposed and extended or terminated by a judge, whotakes into account the advise of a psychologist or psychiatrist expert witness.

### II.3.2.2 Participants

Data on static and dynamic patient characteristics were collected in seven (of the, then, nine) forensic psychiatric hospitals in The Netherlands between January 1, 1996 and December 31, 1998. Two hospitals did not participate in the study, in one case due to the fact that the hospital was newly established and lacked dischargees, and in the other case due to lack of a co-ordinator to support the research effort on site.

*Patients.* Assessments of dynamic and static risk factors were completed for patients discharged from any of the seven hospitals, either due to termination of the TBS-measure, or at the start of probationary leave. Patients who were transferred to other forensic hsopitals or to prison were excluded from the study because they would not be at risk in the community. Probationary leave is the preferred mode of discharge in TBS. It means the patient lives and works outside the hospital with unrestricted liberties and minimum supervision, usually by a probation officer. The TBS-measure, however, remains in force and if circumstances require re-hospitalisation the patient can be readmitted.

*Raters.* Forty-three different raters participated in the study. Of these, 34 were treatment supervisors, clinical psychologists or psychiatrists responsible for treatment planning as well as advising the court in TBS extension hearings. They provided over 90% of the ratings. The remaining 9 raters were head nurses or psychotherapists, standing in for treatment supervisors who, for practical reasons, were unable to complete a particular rating.

### II.3.2.3 Materials

Data on dynamic patient characteristics were collected using the Clinical Inventory of Dynamic Reoffending Risk Indicators (CIDRRI, Philipse *et al*, 2004). This rating scale was developed specifically for this study. It contains 47 statements on patient behaviour, affect, and clinical symptoms that were identified by clinicians from participating hospitals as pivotal to their risk management decision making. A supervising clinician (psychiatrist or clinical psychologist) rates the items on a 6 point scale, ranging from 'not at all characteristic of the patient' to 'very characteristic'. At the end of the list the clinician is asked to rate the patient's risk of reoffending directly, again on a 6-point scale. No specific training or instruction is required: any clinician familiar with the patient can use the instrument.

The construction process of the CIDRRI, as well as its factor structure are described in detail in Philipse *et al.* (2004). CIDRRI items were condensed into 6 scales, shown in Table 14. An earlier study showed the instrument to have acceptable interrater reliability. Also, it was shown to discriminate to an appreciable extent between patients just starting their treatment and those recently discharged (Philipse *et al.*, in press).

Data on static risk factors were collected from hospital files by clerical workers, using an additional structured inventory called the File Checklist. It contained 41 items, most of which were directly derived from the contents of a national patient data management system that was developed in the early 1990's for use in forensic hospitals. The File Checklist covered demographic characteristics; DSM axes 1 and 2 diagnoses at admittance; data on intelligence, education and employment; details of the TBS index offence; characteristics of TBS offence victims; criminal history; mental health history; irregularities during TBS; and personal circumstances at discharge.

TABLE 14. DYNAMIC SCALES OF THE CLINICAL INVENTORY OF DYNAMIC REOFFENDING RISK INDICATORS

1. **Empathic acceptance of responsibility for the offence**. *Patient acknowledges his responsibility for the offence and does so while truly aware of the impact of his actions on the victim(s)*

2. **Self-reliance.** *Patient has sufficient skills to function acceptably in daily life without professional help*

3. **Anti-social narcissism**. *Patient has traits from narcissistic and anti-social personality disorders*

4. **Treatment compliance.** *Patient has engaged in the treatment process to the best of his abilities*

5. **Attainment of treatment goals.** *Treatment goals that are generally considered important, such as establishing an offence script or improving network conditions, have been achieved*

6. **Avoidance.** *Patient is inclined to stay away from others and to dislike intimacy*

### *II.3.2.4 Procedure*

Data collection was coordinated by a research team, with a member in each participating hospital. These representatives kept track of discharges meeting inclusion criteria, and asked the responsible treatment supervisor to complete the CIDRRI. At the same time a clerical worker completed the File Checklist.

### *II.3.2.5 Outcome measure*

Reconviction data were coded dichotomously, as the presence or absence of any judicial sanction confirming the patient as the perpetrator of a new offence that involved sexual or other violence, including attempts at or threats of such violence. Data on post-discharge conviction were retrieved from the *Centrale Justitiële Documentatie* (Central Criminal Justice Files of the Ministry of Justice) on June 16, 2004.

### *II.3.2.6 Statistical analysis*

Entering all 41 available static predictors into the analysis would compromise the meaningfulness of results due to chance capitalisation. Therefore, static predictors were selected for inclusion in the final analyses in three successive steps. First, we dropped all items from the File Checklist for which there was insufficient support in the risk assessment literature. Fifteen variables similar to items in established risk assessment tools, specifically the Historical/Clinical/Riskmanagement-20 (HCR-20, Webster *et al*, 1997a) and the Violence Risk Appraisal Guide (VRAG, Quinsey *et al.*, 1998), were retained; these are shown in Table 15. In the second step, univariate predictive validity of each variable was assessed in Cox regression analysis. Only items with univariate predictive validity at p < .05 were included in the final multivariate analysis.

In the final analysis, Cox regression (survival) analysis was used to develop the prediction model. Like other forms of regression analysis, it eliminates redundant predictors from the model and retains only items that each add significant predictive validity to the other predictors in the set. It has the advantage of also taking into account the variable times at risk among subjects.

Multivariate analysis was performed using forward stepwise entry of variables, with likelihood ratio significance testing. Significance levels for entering as well as remaining in the model were set at p = .05.

Predictors were entered in three successive blocks. First, the preselected static predictors were entered in Block 1, to establish the baseline multivariate static prediction model. The six CIDRRI-scales were then entered in Block 2. All scales were included irrespective of univariate significance, as the emphasis in this study was on finding dynamic predictors of risk. CIDRRI scales that did not significantly add to the predictive value of the static predictor set or to other CDDRI scales, were

## Table 15. Overview of static risk factors from the File Checklist that were included in the final Cox regression analyses

| |
|---|
| DSM axis 1: Psychosis at time of admission (yes/no) |
| DSM axis 1: substance use disorder at time of admission (yes/no) |
| DSM axis 2: B-cluster personality disorder (yes/no) |
| Comorbidity of personality disorder and substance use disorder at time of admission (yes/no) |
| Highest level of employment (7 categories) |
| Only female victims in TBS index offence(s) (yes/no) |
| Execution of the TBS measure was preceded by a period of incarceration due to a concurrent prison sentence (yes/no)[a] |
| Patient had been criminally violent before TBS-offence (yes/no) |
| TBS-offence was first offence (yes/no) |
| Age at first conviction (years) |
| Number of institutional homes where patient lived before 18th year (3 categories) |
| Number of times absent without leave during TBS (n) |
| Arrangements for therapeutic aftercare after discharge have been made (yes/no) |
| There is a regular source of income after discharge (yes/no) |

NOTE. [a]TBS sentences are in some cases imposed in combination with a prison sentence. This option may be applied by judges in case the offence has "severely upset the legal order", which means the offence resulted in serious damage to the physical integrity of others. The length of a prison sentence preceding TBS may thus be regarded as an indication of the type and seriousness of the offence.

removed from the model. Finally, separate constituent CIDRRI items from discarded scales were entered in Block 3, but only if the item odds ratio was statistically significant at p < .05 in univariate Cox regression analysis. This limitation was again set to prevent the introduction of an unfeasibly large number of variables into the equation.

To evaluate the predictive power of the ensuing final prediction model while avoiding effects of overfitting, an unweighted prediction score was computed by summing the rough scores on predictor variables. A Receiver Operating Characteristic (ROC) curve of sum scores was then plotted, and the area under the curve computed to establish predictive power. An ROC-curve shows the balance between false positive and false negative predictions at every cut-point in the assessment scale. The area under the curve (AUC) represents general predictive power, with 0.5 equalling non-prediction, 1.0 equalling perfect positive prediction, and 0.0 equalling perfect negative prediction.

## II.3.3 Results

### *II.3.3.1 Participants*

Nation-wide, 180 patients met our discharge criteria during the data collection period (data provided by Dienst Justitiële Inrichtingen, Ministry of Justice), of whom 151 entered our study. This means that the study covered 83.9% of our target population. After screening, 19 cases were discarded for the following reasons: (1) For 2 non-Dutch nationals who were repatriated directly after discharge, and 1 person who deceased shortly after discharge, it would be either unfeasible or meaningless to retrieve reconviction data. (2) CIDRRI ratings were sometimes returned with considerable time delay. It was decided to discard ratings conducted more than eight months after discharge (n=16). Although this cut-off was to some extent arbitrary, it provided an optimum balance between inclusion of possibly less valid data on the one hand, and loss of reoffenders on the other. Of the remaining CIDRRI's, 70% were rated within 3 months after discharge.

The final sample thus comprised 132 cases, or 73.3% of our target population. Of these, 92.4% were male, 7.6% female. This is in accordance with the general TBS-population, which is over 90% male. TBS had been imposed for (mostly violent) property offences in 12.9% of cases; for violent offences in 50.7% of cases; for sex offences in 19.7% of cases; and for arson in 16.7% of cases.

At the beginning of treatment, 75.7% of patients were diagnosed with at least one personality disorder according to DSM-III or DSM-III-R (American Psychiatric Association 1980, 1987), mostly falling into the B-cluster (33.3%) or the Not Otherwise Specified category (32.6%). Also, 69.7% of patients had at least one Axis 1 disorder, with psychotic disorders (29.6%) and substance-related disorders (18.9%) being the most common primary diagnoses.

Length of stay in the hospital from which the patient was discharged varied from 1.2 to 11.4 years, with an average of 4.7 years. The vast majority of patients, 83.3% (n=110), left the hospital on probationary leave. The remaining cases (n=22) were discharged into the community because the court terminated the TBS-measure. This ruling was in accordance with hospital advice in only 4 cases, and went against hospital advise in the remaining 18 cases.

### II.3.3.2 Follow up and reconviction

At the point in time when reconviction data were retrieved, the last patient to enter the study had been discharged exactly 5.5 years ago, whereas follow-up for the first patient entering the study amounted to 8.5 years. Time at risk either to first reconviction or to the end of the study varied from 71 days (2.4 months) to 3,088 days (8.5 years), with a mean of 2,272 days (6.2 years) and a median of 2,493 days (6.8 years). The quickest relapse occurred 71 days after discharge, whereas the longest time-span between discharge and relapse was 2,177 days (5.9 years). The Kaplan-Meier Survival curve is shown in Figure 2 (p. 98).

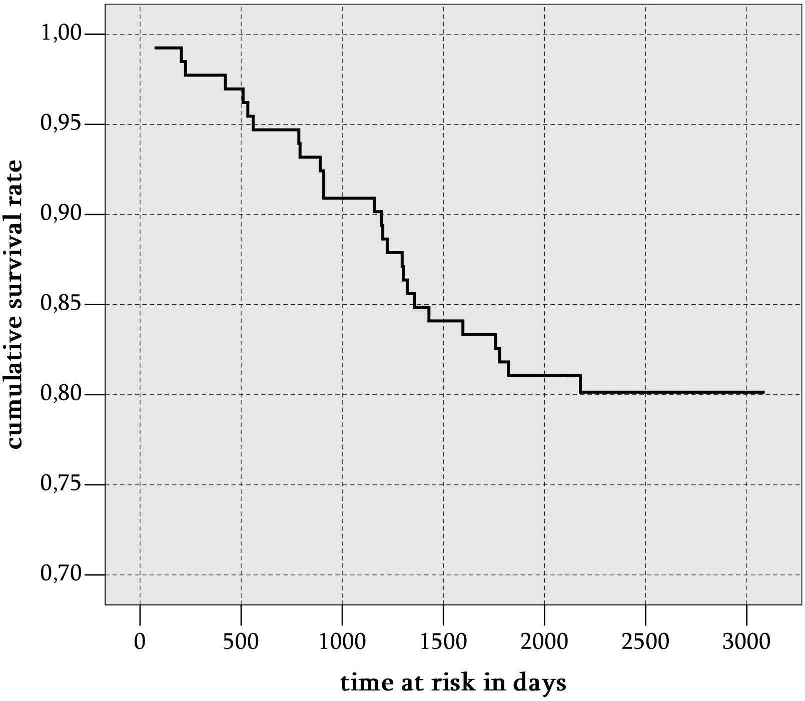
A total of 26 dischargees (19.7%) were again convicted for a violent (n=21) or sexual (n=5) offence during follow-up. In 11 cases the new offence was similar to the offence for which TBS had been imposed.

### II.3.3.3 Prediction of reconviction

The direct risk assessment given by raters on a 6-point scale at the end of the CIDRRI did not have any predictive power with regard to actual reconviction for a violent or sexual offence (ROC AUC = .44; 95% CI .31-.56; ns). Of the 15 static variables selected, only 7 proved significant predictors in univariate Cox regression analysis ($p < .05$), as shown in Table 16. Of these, 4 remained in the initial multivariate static prediction model.

None of the 6 dynamic scales added to the predictive power of the static model. Therefore, single CIDRRI items were assessed for their univariate

FIGURE 2. KAPLAN-MEIER SURVIVAL CURVE FOR RENEWED CONVICTION OF A VIOLENT OR SEXUAL OFFENCE AFTER DISCHARGE.



significance as a predictor (p < .05), to establish a selection for inclusion in the multivariate analysis. Only four items remained, these are again shown in Table 16. Interestingly, for all items the direction of the relation between item scores and reconviction ran counter to expectations: it appeared that higher levels of observed dysfunction were associated with lower risk of reconviction.

When these 4 dynamic items were entered stepwise into the analysis in addition to the already established static predictor set, none of them added significantly to the prediction of reconviction. Thus, the final model, displayed in Table 17, was identical to the initial static prediction model. It shows that being absent without leave during TBS, comorbidity of axis 2 and substance use disorder at admission, and presence of cluster B personality disorder each increase the risk of reconviction, whereas presence of psychosis at admission reduces reconviction risk.

In order to assess the predictive power of this model, a simple sum score was computed. One point was added for cluster B personality disorder at

TABLE 16. OVERVIEW OF RISK FACTORS WITH UNIVARIATE PREDICTIVE VALIDITY IN COX REGRESSION ANALYSIS

| static factors | β | p |
|---|---|---|
| DSM axis 1: Psychosis at time of admission | -1.530 | .005 |
| DSM axis 1: substance use disorder at time of admission | 0.930 | .021 |
| DSM axis 2: B-cluster personality disorder | 0.843 | .036 |
| Comorbidity of personality disorder and substance use disorder at time of admission | 1.149 | .004 |
| Execution of the TBS measure was preceded by a period of incarceration due to a concurrent prison sentence | 1.273 | .038 |
| TBS-offence was first offence | -1.111 | .017 |
| Number of times absent without leave during TBS | 0.467 | .000 |
| dynamic items (CIDRRI) | | |
| 7. Avoids contact | -0.392 | .023 |
| 8. Is unable to empathise with the victim's suffering | -0.271 | .046 |
| 11. Completely denies his offence | -0.744 | .018 |
| 36. Is unable to live on his own and take care of himself | -0.290 | .023 |

admission; comorbidity of personality disorder with substance related disorder at admission; and having been absent without leave at least once. One point was subtracted for presence of a psychotic disorder at admission. This unweighted score produced an ROC area under the curve for reconviction for renewed violent or sexual offending of .79 (95% CI .69-.89, p < .001), showing that the model, however succinct, possessed significant and substantial predictive power.

TABLE 17. PREDICTION MODEL FOR RISK OF RECONVICTION FOR VIOLENT OFFENDING (MULTIVARIATE COX REGRESSION ANALYSIS)

| Step | Predictor | β | p | odds ratio[a] |
|---|---|---|---|---|
| 1 | Number of times absent without leave during TBS (count) | .422 | .000 | 1.525 |
| 2 | Comorbidity of any personality disorder with substance use disorder at time of admission (dichotomous score) | .981 | .016 | 2.667 |
| 3 | DSM axis 1: psychosis, at time of admission (dichotomous score) | -1.153 | .037 | .316 |
| 4 | DSM axis 2: any cluster B personality disorder at time of admission (dichotomous score) | .818 | .048 | 2.267 |

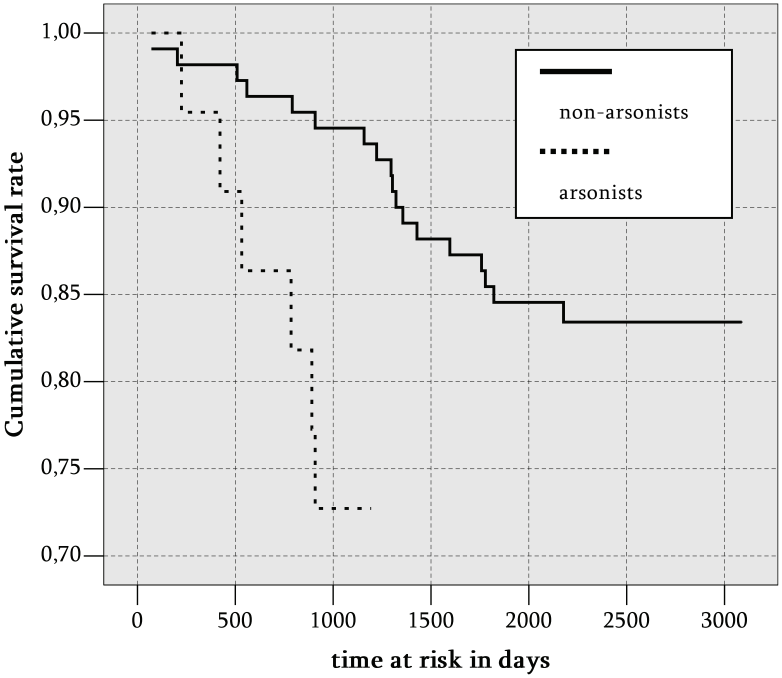
NOTE. Model $\chi^2(4)=43.261$, p <.001. [a]Odds ratio is the factor by which the likelihood of reconviction multiplies with every incremental step on the predictor variable.

## *II.3.3.4 Post hoc analyses*

Because of the limited scope of the resulting prediction model, it was decided to conduct post hoc analyses to establish (1) whether findings might have been influenced by particular subgroups of patients, and a more extensive model might result if such a subgroup were deleted from the analysis; and (2) whether the variability of time at risk might have influenced findings.

*(1) Influence of subgroups*. To identify subgroups, reconviction rates were compared for patients with different TBS-offence types, different types of diagnoses, different modes of discharge and different modes of aftercare. It was found that patients with psychotic disorders at admission were reconvicted significantly less often than others ($\chi^2(1)=9.201$, p = .002); patients with substance use disorders were reconvicted significantly more than others ($\chi^2(1)=5.700$, p=.017), as were patients with cluster B personality disorders ($\chi^2(1)=4.447$, p=.035). None of these findings provided new insights, as these characteristics are all part of the prediction model.

FIGURE 3. KAPLAN-MEIER SURVIVAL CURVES FOR RENEWED CONVICTION OF A VIOLENT OR SEXUAL OFFENCE AFTER DISCHARGE FOR ARSONISTS AND NON-ARSONISTS SEPARATELY.



However, it was also found that arsonists were reconvicted substantially more often than perpetrators of any other type of offence: 8 out of 22 arsonists were reconvicted (36.4%) as opposed to 18 out of 110 non arsonists (16.4%) ($\chi^2(1)=4.636$, p=.03). It is interesting to note that new offences by arsonists were mostly of a violent or sexual nature rather than involving new arson. Kaplan Meier survival curves of arsonists versus non-arsonists are shown in figure 3.

On the basis of this finding, the Cox regression analysis was repeated for the sample excluding arsonists. This yielded a three factor prediction model that again lacked clinical salience. In this model the DSM-III disorders from the main model (psychosis and Cluster B personality disorder) were replaced by a single CIDRRI item, 35: 'Lacks essential social skills'. Like other dynamic variables in univariate analysis this one too had a 'reversed' relation with reoffending: patients seen as having deficient social skills were reconvicted less often than those judged to have good social skills. Absence without leave remained in the model with an odds ratio similar to that in the main model. Comorbidity of personality disorder with substance use disorder also remained, and had increased predictive power as compared to the main model.

An unweighted sum score was again computed by adding or subtracting the dichotomous scores on the three predictors. The 6-point rating on CIDRRI item 35 was dichotomized for this purpose by splitting scores between ratings 3 and 4. This new unweighted score, based on only three predictors, again had considerable predictive validity: the ROC AUC curve was .75 (95% CI .62-.88, p < .01). Nevertheless, the inclusion of a counterintuitive predictor (item 35) as well as the reduced ROC AUC left the main model for the full sample of patients unchallenged as the best prediction scheme derivable from the data.

*(2) Influence of variable follow-up duration*. Harris & Rice (2003) have argued that variable follow-up duration is an important cause of underestimation of predictive validity. Though Cox regression takes this variability into account, it was nevertheless decided to perform an additional logistic regression analysis with a fixed follow-up of 2,011 days (5.5 years), which is the time span between last discharge and date of retrieval of reconviction data. Thus every patient in the analysis had been outside the hospital for at least 2,011 days (taking into account time-outs in the hospital for some patients on probationary leave). A new recidivism score was created, designating patients who were reconvicted after more than 2,011 days as non-reconvicted.

Logistic regression analysis was conducted following the same procedure as described for Cox regression. The resulting model again contained only 4 variables, 3 of which were the same as in the main model. Cluster B personality disorder however was replaced by CIDRRI item 8, 'Is unable to empathise with the victim's suffering'. Its relation with outcome again ran counter to intuition: patients rated as more empathic were reconvicted more often. An unweighted sum score of these predictors yielded an ROC area under the curve for predictive validity of .76 (95% CI .66-.87, p < .001). This result is comparable to that of the original model, but the inclusion of a predictor with a counterintuitive predictive direction renders it less attractive. With regard to model content the similarities between both models are such that it can be safely concluded that variability of time at risk did not unduly influence the main outcome of the study.

## II.3.4 Discussion

The present study investigated whether clinically relevant, dynamic patient characteristics that are routinely regarded as reconviction risk indicators in forensic psychiatric settings, added significantly to the predictive power of static risk factors. The research was conducted under naturalistic conditions in a prospective setup.

A four-predictor risk assessment model was found, comprising: absence without leave during TBS; comorbidity of axis 2 and substance use disorder at admission; Cluster B personality disorder; and psychosis at admission. This set predicted future reconvictions for sexual or violent offences with considerable accuracy, comparable to that of the VRAG, HCR-20 and PCL-R (Dolan & Doyle, 2000).

Regrettably however, this model is fully static. Some dynamic factors were shown to have predictive power when considered separately, but proved redundant in a multivariate model including the static predictors. Interestingly, univariate Beta coefficients of dynamic risk factors in all cases pointed in the opposite direction of that expected in clinical practice, suggesting substantial discrepancies between behaviour observed on the outside by clinicians, and possible underlying drives and intentions of the patient. In accordance with this, direct clinical assessments of risk at the moment of discharge were completely unrelated to subsequent reconviction. This last finding is unsurprising in the light of international research (Grove & Meehl, 1996), but had not been previously reproduced in the context of Dutch forensic psychiatry.

### *II.3.4.1 Implications for clinical risk assessment*

The dynamic variables included in this study were all directly derived from clinical practice in TBS (Philipse *et al.*, 2004). They represented patient characteristics that were, and are, routinely used by clinicians to assess the risk of future reoffending, and to identify focus points for treatment. Though the data were gathered 6 to 9 years ago, the items that were tested continue to be typical of risk factors cited by TBS clincians in risk assessment reports. Moreover, many of them are, in one form or another, also part of the definition of dynamic items in the HCR-20 or the Level of Service Inventory – revised (LSI-R, Andrews & Bonta, 1995). The finding that none of these variables added any predictive power to a handful of static predictors is therefore worrying. It suggests that clinicians need to be very careful when considering unstructured impressions of dynamic patient characteristics as part of a risk assessment. As Webster *et al.* (1997b) have noted, "historical variables deserve a position of primacy in any scheme used in attempts to assess violence potential in persons with psychiatric disorders" (p. 256), and, "Risk assessments should conform to some generally known scheme or device" (Webster *et al.*, 1997a, p. 7).

Earlier findings with regard to CIDRRI interrater reliability were sufficiently satisfactory to render it unlikely that predictive validity of dynamic factors was grossly underestimated due to reliability problems (Philipse *et al.*, in press). Nevertheless dynamic risk factors are more complex than static ones, and more difficult to assess (Quinsey *et al*, 1998). Even when items like those in the CIDRRI are scored reliably, a scoring procedure designed to allow easy use in the busy everyday practice may incur the risk of only skimming the surface.

Here lies a possible explanation for the counterintuitive direction of predictive relationships between dynamic items and outcome in univariate analysis (for instance, patients denying their offence are reconvicted less often). It may indicate that adequate risk management strategies are put into place for patients with obvious areas of dysfunction, while risk management for patients who project an image of adequate functioning relies too heavily on the patient's apparent abilities. In this context, psychopathy may be of crucial influence: for instance, a patient may seem very contrite with regard to his offence, but may in fact be faking this sentiment because he knows it will enhance his chances of discharge. Unfortunately, psychopathy could not be included as a variable in the present study because no Dutch version of the Psychopathy Checklist - Revised (PCL-R; Hare, 1991) was available at the time data collection started. Though we can therefore not be sure that such mechanisms are at work in the present study, the CIDRRI relies exclusively on the clinician's individual perceptions of the patient, and is thus vulnerable to 'impression management'. In this respect CIDRRI performance might be improved by basing judgments on systematic observation and formal psychological assessments, and by involving multiple raters in the process.

### II.3.4.2 Research considerations and study limitations

The benefit of a hospital-based, prospective research procedure is that it yields a realistic impression of an instrument's validity under the circumstances of use for which it was intended. In this context only a very small set of predictors survived in the final model, and even several retrospectively well-established historic predictors were discarded (for instance, age at first conviction and number of previous convictions). This would suggest that much used risk assessment tools like the VRAG and HCR-20 are in need of systematic testing under everyday clinical circumstances before any definite conclusions are drawn about their ecological validity and usefulness. As Margison *et al.* (2000) have noted, evidence based practice cannot exist without such "practice based evidence".

Two methodological issues regarding the present study warrant brief consideration. First, 30% of checklists were only returned to the researchers over 3 months after the patient had actually left the hospital. It is, of course, possible that these checklists were rated based on imperfect recollections of the patient, and thus negatively affected predictive validity. Yet, it can also be argued that the delay worked in favor of dynamic predictors, as the clinician may have benefited from current information about the patient's functioning in the community context.

Second, like nearly every risk assessment study the present study is hampered by the reoffending 'dark number'. There can be no doubt that the reconvictions for violent and sexual crimes, used as prediction criterion in this study, represent only part of the true amount of reoffending. Many crimes are never brought to the attention of the police, and others remain unsolved. Estimates of the extent of this dark number vary greatly. Research setups that solve this problem through intensive follow-up with self-report and collateral interviews are rarely feasible, if only because they require willingness on the patient's behalf to keep in close contact with the forensic system after discharge. This will inevitably result in a select sample. The problem of dark numbers appears in essence to be insoluble, and needs to be taken into account when evaluating any risk assessment research findings.

To conclude, it should be noted that in recent years important changes have occurred that may have improved risk assessment practice in TBS already. This is one drawback of prospective research: due to the time needed, developments in the field can easily overtake it. Notably, from 2000 onwards Dutch forensic hospitals have seen the introduction of standardized risk assessment tools as well as the PCL-R, by now mandatory input in any risk assessment. However, it should not be forgotten that long-term prospective, hospital based validation of these instruments is still largely absent even internationally, and non-existent in the Netherlands. Whether they have added value as compared to a compact static prediction model like the one presented in this article, remains to be seen.

The findings from our study suggest that observational and clinically interpreted input may not be a good basis for risk assessment, and that other sources of information need to be considered. Future research into dynamic predictors of reoffending may need to focus on new measures that are less susceptible to manipulation by the patient or to clinical observation and evaluation bias. Experimental performance tasks and psycho-physiological measures are alternatives that warrant closer investigation and could well provide a way forward.

# III ☙ General conclusions and discussion

This thesis started with an overview of core concepts and the current state of the art in violence risk assessment. In the second part, three empirical studies were reported that centered on psychometric characteristics and predictive validity of a checklist containing hypothesized dynamic risk factors for TBS. This final section contains general conclusions from the preceding parts; addresses some limitations of the present research; discusses the meaning of these findings both in clinical and research terms; and looks at ways forward, taking into account current issues and developments in the risk assessment field.

# III.1 ꞓ Strengths and weaknesses of the current risk assessment knowledgebase

The science of violence risk assessment has made important strides forward over the last thirty years. Crucial developments were:

- A conceptual shift from dangerousness as a personality characteristic, towards risk of reoffending as the result of an interaction between particular characteristics of the individual and his environment.
- A conceptual shift from predicting absolute outcome towards assessing the relative likelihood of that outcome.
- A general increase in the number of scientific studies on risk assessment, resulting in a shift from early pessimism about predictive possibilities towards (mostly) guarded optimism.
- A general increase in the methodological quality of studies, and the introduction of the receiver operating characteristic (ROC) as a uniform and less base rate sensitive, decision-maker friendly measure of predictive validity.
- The introduction of risk assessment instruments grounded in empirical research as viable alternatives to unstructured clinical risk assessment procedures.
- The introduction of structured clinical risk assessment tools as a way out of the clinical-versus-actuarial deadlock.
- Nascent interest in issues of risk communication and risk management as corollaries of risk assessment strategies.

These improvements have been of great importance to the field, yet are far from providing all the answers needed. It comes as no surprise that current state of the art risk assessment is by no means capable of satisfying the unrealistic demands for safety and security rife among politicians and the public today. Unfortunately, political pressure to adopt empirically valid state of the art methods may promote token use of risk assessment tools: there is a danger that they remain mere hasty and superficial additions to an unchanged underlying practice, while they should be agents of fundamental change to that practice. Such change can only be effected from within, and not through external obligations. The user of a risk assessment instrument needs himself to be convinced of its worth, and thoroughly aware of its probable superiority over his own clinical judgment. In this respect it is a worrying fact that empirically derived risk assessment instruments still hold only limited appeal to the clinical practitioner in forensic mental health (Elbogen *et al.*, 2002).

Several specific factors can be identified that have estranged available risk assessment techniques from the practitioner for whose use they are intended. First, many available tools are fully or to a large extent comprised of risk factors that cannot be changed by treatment interventions. Static risk factors help to make the clinician aware of important base rate determinants, but do not help him plan treatment. Instruments that do include dynamic predictors define them vaguely and their rating scales allow little room for actually measuring and registering change. Sturidsson *et al.* (2004) have noted that in clinical reality, predictor-outcome relations tend to be viewed as more complex than is reflected in the one-dimensional operationalisations preferred in risk assessment instruments. Global, simplified assessments of current functioning furthermore lack clear handles for intervention. For instance, relational problems may have many different causes, and targeted intervention requires an individual understanding of the underlying processes and the specific way in which they are linked to reoffending risk, while a risk assessment tool provides insight into neither of these. Finally, after completing a structured risk assessment procedure, the clinician is left with a relative risk statement that offers few clues when it comes to dichotomous decision making – he sees himself confronted with the question that Harris (2003) quoted in a commentary title: "Men in his category have a 50% likelihood, but which half is he in?" (p. 389).

In addition, it is worth noting that dynamic risk factors such as are present in several instruments probably offer little that is new to the clinician; if, for instance, a patient displays overt antisocial attitudes, any clinician would most certainly take these into account when assessing offending risk, no matter which risk assessment procedure he is adhering to. For example, the CIDRRI, the research instrument used in the empirical study reported in the preceding pages, contains items directly suggested by clinicians who were unfamiliar with either the HCR-20 or the LSI-R, but it nevertheless covers nearly all dynamic topics in both these instruments.

Webster *et al.* (1997) have said that the main value of the HCR-20 "at this point may lie in the general principles it espouses rather than in its detail" (p. 5). Indeed, the main strength of risk assessment tools seems to lie in the structure they offer rather than in their specific content, an assumption supported by the fact that instruments with varying item pools yield very similar (and similarly varying) prediction results (as can be concluded from a comparison of findings listed in Tables 3 and 4 in the General Introduction, see pp. 40 and 46). Unfortunately, however, optimum use of the structuring abilities of risk assessment tools seems more feasible in research than in clinical contexts. In forensic mental health settings the time requirements attached to instruments like the HCR-20 or the LSI-R may easily be overruled by other priorities, incurring the risk of erratic use as signaled by Webster *et al.* (2002). This situation in turn has caused studies into risk assessment validity to be preponderantly academic - as yet they tell us little about the validity of instruments in clinical practice, and nothing at all about the effect of their application on reoffending rates.

## III.2 ꕤ The empirical studies

The empirical studies reported in this thesis aimed to address some of the issues mentioned above. The main question was whether, within the in-patient forensic mental health context of TBS, dynamic predictors of risk could be found that would be meaningful to clinicians, while at the same time adding predictive validity to known static predictors. To answer this question, a multi-site, longitudinal study was conducted with a checklist constructed specifically for this purpose, the Clinical Inventory of Dynamic Reoffending Risk Indicators (CIDRRI). Seven TBS-hospitals participated in rating the CIDRRI and the accompanying checklist of static risk factors during three years for every patient leaving the hospital on probationary leave or due to termination of the TBS measure (n=151). Reoffending data were retrieved from judicial files after a minimum follow-up of 5.5 years. Additional datasets were compiled to investigate interrater (n=75) and test-retest (n=29) reliability; a further set of checklists was rated for patients who had recently been admitted to any of the participating hospitals (n=114).

Somewhat as a 'side-product', the study design also allowed us to test the predictive validity of clinical risk assessment by practitioners in TBS, which had never been investigated before

### III.2.1 Study I. Structural coherence of clinically derived dynamic indicators of reoffending risk.

This study was concerned with two questions:

1. The implicit structure of the CIDRRI: did the items in the checklist represent a limited set of underlying clinically relevant dimensions?
2. The predictive relations between CIDRRI items, CIDRRI dimensions, and the clinical risk estimate.

To investigate these questions, the datasets compiled for different parts of the study were merged to yield an overall dataset comprising 370 cases. Factor analysis of these data showed that six underlying dimensions could be meaningfully distinguished. These were:

- Empathic acceptance of responsibility for the offence
- Lack of self-reliance
- Anti-social narcissism
- Treatment compliance
- Attainment of treatment goals
- Avoidance

The overall model was used to compute 6 scales corresponding to these factors. It was then investigated to what extent CIDRRI items and CIDRRI scales were related to the clinical risk estimate, as rated on a six-point scale at the end of the checklist. It was found that 44 out of the 47 items had significant predictive validity with regard to this clinical risk estimate. The strongest bivariate predictor among the items was item 20, reduction in psychopathology assumed to underlie the index offence (ROC AUC =.83). All six scales, too, were clearly related to the clinical risk estimate, with 'Attainment of treatment goals' showing the strongest relation (ROC AUC =

.85). Finally, the CIDRRI total score was also strongly related to the clinical risk estimate (ROC AUC = .82).

These findings confirmed the CIDRRI as an adequate representation of dynamic patient characteristics on which clinicians base their risk estimate. They further showed that CIDRRI-items represented clinically meaningful underlying concepts. Finally, the results showed that clinicians set particular store by reduction of psychopathology and the attainment of treatment goals when assessing reoffending risk. The nine items most strongly related to the clinical risk estimate were very similar to the "top 10" of risk factors perceived as relevant by Swedish clinicians in a study by Sturidsson *et al.* (2004), notably with regard to insight into pathology, treatment motivation, and instability.

## III.2.2 Study II. Reliability and discriminant validity of dynamic reoffending risk indicators in forensic clinical practice.

The second study investigated psychometric properties of the CIDRRI, by answering the following questions:

1. What are the levels of interrater reliability of CIDRRI items and scales?
2. What are the levels of test-retest reliability of CIDRRI items and scales?
3. To what extent is the CIDRRI able to distinguish recently admitted patients, assumed to represent higher risk levels, from patients about to start probationary leave, assuming that the latter group of patients represent a lower risk level?

Though there were marked differences in score patterns among paired raters, overall levels of interrater reliability for CIDRRI-scales as well as most items were fair. In contrast to current risk assessment tools, whose reliability was more often than not tested under optimized circumstances, the reliability findings in our study may be regarded as an ecologically realistic estimate of CIDRRI's true reliability in clinical use.

Test-retest reliability was good for all scales, and satisfying for all individual items save one. Finally, the comparative study of admittees ('high risk' group) and dischargees ('low risk' group) showed that the 6 scale-scores together clustered these groups with 72% sensitivity and 89%

specificity, thus achieving a high level of accuracy (80%). The clinical rating of the risk of reoffending showed even better classification results, with 93% sensitivity and 74% specificity (84% overall accuracy). When scores on the 47 items were used as a basis for clustering, it appeared that these were able to identify most low-risk patients correctly, but did so at the cost of a large proportion of false negatives. The CIDRRI total sumscore had no significant discriminating power.

### III.2.3 Study III. Static and dynamic patient characteristics as predictors of criminal recidivism.

The final study addressed the question of predictive validity:

1. Do items and scales contained in the CIDRRI have predictive validity with regard to reconviction after discharge?
2. Does the clinical assessment of risk provided by raters at the end of the CIDRRI checklist have predictive validity with regard to reconviction after discharge?
3. If CIDRRI items, scales, or clinical assessment have predictive validity, do they also add predictive power to models already containing static risk factors?

It was found that 26 of the 132 patients in this study (19.7%) were reconvicted for a violent or sexual offence during a 5.5 to 8.5 year follow-up. Univariate Cox regression analysis showed 6 static factors, none of the scales, and 4 individual CIDRRI items to have predictive validity, while clinical risk estimates were unrelated to outcome. The CIDRRI items predicted negatively.

These findings answer the first two questions. With regard to the third question it was found that when static predictors alone were entered into multivariate regression analysis, a four-factor prediction model resulted. When this was retained and dynamic items were added in a stepwise procedure, no significant predictive power was added to the model. This means that the final prediction model was fully static, containing only 4 variables: the number of times patient had been absent without leave during TBS; comorbidity of personality disorder with substance use disorder at the time of admission; psychosis at the time of admission; and any cluster B personality disorder at the time of admission. It should be noted that presence of psychosis at admission reduced rather than increased the risk level.

### III.2.4 Overall summary of findings

The most salient overall findings of these studies may be summarized as follows:

- Clinicians in TBS are strongly concerned with treatment goals, treatment participation and mitigation of pathology when assessing risk of reoffending.
- Two clinicians in TBS who work closely together on the same case, may have considerable differences of opinion regarding patient characteristics that are routine points of reference in treatment and clinical risk assessments. Tacit assumptions of consensus may often be overly optimistic.
- Achieving high interrater reliability on a low-threshold instrument like the CIDRRI, when rated by practitioners in a forensic mental health setting, would require the averaging of scores by at least 3, and preferably 5 independent raters.
- Ratings on dynamic factors viewed by clinicians in TBS as important indicators of reoffending risk, are able to distinguish patients in early treatment stages from those about to be discharged with a considerable level of accuracy.
- Risk of reoffending of patients about to leave the hospital, as assessed by treatment supervisors without a structured aid, bears no relation to actual reconviction after discharge.
- Some dynamic risk indicators suggested by clinicians in TBS show an actual but negative relation to reconviction after discharge in univariate analysis.
- Once a multivariate, static risk assessment model is established, dynamic risk factors as are commonly used in unstructured risk assessments by clinicians in TBS, do not have any added predictive value.
- Risk of reconviction after TBS can be estimated with considerable accuracy utilizing a highly compact static predictor set.

# III.3 ☙ Discussion

Hope with regard to useful dynamic predictors of reoffending risk for TBS is not encouraged by the present study. The most salient finding is the large discrepancy between the clinical assumptions about risk factors, and empirical reality. Though clinicians, when asked, are able to generate an extensive and detailed list of patient characteristics considered to be pivotal in risk assessment, none of these characteristics, either individually or clustered to represent underlying dimensions, appear to add any predictive power to a handful of basic historical data. Given this fact, it is not surprising (but still rather disturbing) that clinical risk assessments themselves, which were found to be strongly related to the dynamic risk factors, also lack predictive validity.

These findings need to be evaluated from two angles at least: first, the question needs to be considered whether they are adequate reflections of TBS reality rather than artifacts resulting from deficiencies in the study design; secondly, assuming these findings to some extent reflect TBS reality, clinical and judicial implications demand consideration.

## III.3.1 Strengths and limitations of the study design

### *III.3.1.1 Three major strengths of the study*

The present study design had three outstanding strengths. First, it was a multi-site project in which all but 1 eligible TBS-hospital participated. Apart from narrowing down the definition of discharge, to make sure all dischargees would in fact be at risk in the community, the sample was non-select and represented over 80% of the national target population.

Second, data collection procedures had high ecological validity: the data used in this study, both static and dynamic, were the same data that would be available to clinicians in everyday practice. By including extensive sets of both predictor types, the poverty of prediction variables noted by Monahan & Steadman (1994) was avoided. Moreover, these data were collected by the functionaries who are also responsible for them in daily life: treatment supervisors provided treatment related dynamic data, whereas clerical workers collected historical information from files.

Third, the follow-up was prospective, so that dynamic data could be based on current knowledge of the patient rather than retrospective file inspection. Prospective study designs are generally regarded as the most desirable in follow-up research (Grann, 1998); with regard to research in clinical psychiatry, Andreasen (2000) calls them "the coin of the realm" if one wants to make valid predictions (p. 1374).

Thus, though the present research project did not aspire to address all the methodological issues pointed out by Monahan & Steadman (1994) (in fact could not do so as it germinated at a time when their admonishments were not yet published) it nonetheless meets the requirements set out by these authors on several major points.

### *III.3.1.2 Limitations of the study*

Ecological validity may however also constitute a weakness. The present study set out to identify useful dynamic predictors of reoffending risk in TBS. It could be argued that by favoring clinical realism and feasibility, requirements for measuring complex dynamic phenomena were insufficiently met, and the emphasis in the study subtly shifted towards an assessment of the validity of existing clinical practice rather than of science based dynamic risk factors. This and several other limitations of the study are discussed in more detail below.

#### *III.3.1.2.a Reliability issues*

From the very beginning, the present project was planned as field research. The study would be implemented among practitioners who, in most of the institutions, had no previous experience with systematic data collection for research, and had no time allotted in their schedules for such activities. Of necessity, our instrument therefore needed to be straightforward and simple to use. We decided to dispense with time-consuming training sessions and detailed instructions altogether, guided by the fact that the CIDRRI contained only items directly derived from clinical practice, and therefore, we assumed, readily comprehensible to raters.

Obviously, this choice left room for individual differences in item-interpretation, as well as other subjective influences such as the quality of rater-patient contact or countertransference. This is reflected in mostly moderate interrater reliabilities. However, it is not necessarily true that this moderate reliability explains the lack of predictive validity. This is most clearly visible when comparing univariate predictive validity of items and scales in Study III to the reliability findings reported in Study II. Scales have generally higher reliability than items, yet none of them has significant univariate predictive validity; reversely, the four items that do have some predictive power do not necessarily have high reliability. Moderate reliability is likely to dampen the predictive power of risk factors, but is less likely to be the sole factor rendering them completely invalid. If it did, and in such large numbers at that, this would indicate that the predictive validity of the dynamic factors, if any, is not very robust.

### *III.3.1.2.b Construct validity of dynamic items*

The second issue we need to face when considering the research findings, is the construct validity of dynamic predictors. How likely is it that these risk factors actually measured the characteristics they describe? The most obvious study limitation in this respect is the fact that characteristics were not measured directly, but were assessed indirectly by the treatment supervisor. It may be argued that CIDRRI items did not in fact measure dynamic patient characteristics, but rather the clinical perception of them. Thus, an element of clinical judgment is introduced that may be liable to suffer from several of the same drawbacks from which clinical risk assessments suffer. Indeed, the data offer some clues indicating that clinical ratings of the CIDRRI may have represented what the clinician wanted to see or what the patient wanted the clinician to see, rather than the actual condition of the patient.

A first indication of this is the fact that the univariate predictive relationships to outcome of the four CIDRRI items in Table 16 (p. 99) are the reverse of what would be expected. Patients assessed as functioning more adequately with regard to those characteristics were at a consistently higher risk of reconviction than patients judged to function less adequately. This effect is fairly strong, as becomes apparent when the scores on the four items are summed, and the 20% highest scorers (high level of perceived dysfunction) on the sum-variable are compared to the 80% lower scorers. Reconvictions occurred for only 6% of the high scorers, as opposed to 23% of the low scorers. Put differently, 92% of reconviction occurred among low scorers. What this suggests is that the perception and judgment of adequate functioning is based on a superficial appearance of normality which disguises either underlying dysfunctions or unchanged, static risk.

A second, related indication that CIDRRI items measure clinical perceptions of patient characteristics rather than the characteristics themselves follows from the admittee-dischargee comparison of Study II. This study showed that as long as observations are restricted to the situation inside the hospital, relations between CIDRRI-predictors, clinical risk estimates and patient treatment phase are fully consistent and all point in the expected direction: higher item scores correspond to higher risk estimates, and risk estimates for newly admitted patients are consistently higher than for patients who were recently discharged. It is telling that these relations are then either reversed or eradicated once scores and risk estimates are linked to actual outcome. To put it strongly, it seems to suggest that an intramural verisimilitude is established that can maintain itself within these limited confines, but crumbles when confronted with external reality. This edifice is based on clinicians' observations and patients' behaviors that may both be governed by ulterior motives. Clinicians will feel a need to see improvement in their patients, or at least to maintain patient flow on their wards, and may well unwittingly modify their evaluations of particular patient behaviors as treatment progresses, even without any true change occurring. Patients in turn may learn to adapt to the demands made by the clinical environment, and shape their behavior so as to enhance access to freedom of movement and to increase the likelihood of discharge, possibly without this change reflecting any true change in underlying motivations and pathology. Though these scenarios cannot be proved or disproved from the present data, they offer themselves as viable explanations for the curious discrepancies between Studies II and III.

Implicit in both these indications with regard to the validity of CIDRRI items is the concept of psychopathy. Recent research suggests that among TBS-patients, up to 35% may be psychopaths when a PCL-R cut-off of 26 is applied (Hildebrand, 2004). Those within this group who have heightened scores on the affective and manipulative factor 1 may be very apt at projecting a 'mask of sanity' that fits clinical wishes and expectations, while underlying antisocial and predatory tendencies remain unchanged, thus increasing the risk of clinical misjudgments. It is a clear limitation of the present study that PCL-R scores could not be included, as a Dutch version of this instrument was not available at the time the project started. Had such scores been available, they could have provided insight into any modifying effects of psychopathy on CIDRRI item ratings.

A final consideration with regard to construct validity follows from Hanson & Harris's (2000) conceptualization of dynamic risk factors. CIDRRI items were mostly of the 'stable' rather than the 'acute' dynamic type, as test-retest findings in Study II confirm. However, in the prediction study the indicators were only measured at one point in time. It would have been preferable had changes in item ratings over longer periods of time been correlated to outcome. Possibly, such a procedure would have yielded different results with regard to predictive validity of dynamic risk factors. If this is considered a limitation of the current study, it should be noted that it is one it shares with nearly all studies presently available regarding any dynamic risk assessment tool.

Nevertheless, Gagliardi *et al.*'s (2004) recent argument needs to be taken into account, that dynamic factors that are assessed only once, are in fact static, and that truly dynamic risk assessment requires frequent 'local readings', similar to the way weather forecasts are constantly updated. Though this viewpoint is persuasive, it is not necessarily valid from every perspective. It seems that most forensic treatment settings aim to effect changes in behavior and pathology that will durably persist after treatment is ended. This is certainly the case in the setting where the present study was conducted. Dynamic risk factors are 'dynamic' as defined by Hanson & Harris (2001): they are susceptible to treatment intervention. But as soon as the patient is discharged the changes that were effected are hoped to endure, and as such indeed to be static. In that respect, a single measurement is commensurate with clinical practice

Furthermore, Gagliardi *et al.* may take the meteorological analogy, originally introduced by Monahan & Steadman (1996), one step too far. The policy statement of the American Meteorological Society (1998) holds that "The predictability of the day-to-day weather for periods beyond day 7 is usually small", and that "no verifiable skill exists or is likely to exist for forecasting day-to-day weather changes beyond two weeks" (p. 2162). Surely, it is not proposed to reassess every former forensic patient or former inmate on a weekly basis for an indefinite number of years after discharge? Clearly, this is not a realistic option. Indeed, often it will simply be judicially impossible to enforce such a regime on persons no longer serving a sentence. Thus, validation of a single risk assessment at discharge, or at best of repeated assessments during treatment, seems an adequate reflection of clinical reality, both as it now exists and as is likely to exist in the future.

### *III.3.1.2.c Validity of predictor-outcome relations*

The general introduction contains an exposé on dark numbers which explains why it was assumed that retrieval of reconviction data from official files only would not result in a misrepresentation of predictor-outcome relations, nor would be likely to greatly underestimate the number of reoffenders. Nonetheless, inclusion of outcome sources other than official files might have increased the statistical power of the study, and helped predictors that did not emerge from the present study to attain statistical significance. Klassen & O'Connor (1987, cited in Monahan & Steadman,

1994) found that including patients' self-reported violence in their outcome measure yielded a more than 25% increase in predictive accuracy over official records alone.

Furthermore, our study, like nearly all follow-up studies conducted in institutional settings, may have suffered from that other dark number: we had no way of validating CIDRRI-assessments with regard to high risk patients remaining in the hospital. Arguably, had these patients been discharged and followed up as well, predictor-outcome relations might have turned out very differently. Possibly, reoffenders in our sample represented a minority of erroneously discharged patients, who were either perceived as low risk or whose treatment was ended by the court against hospital advice, while the majority of high risk patients rightly remained inside the hospitals on the basis of accurate risk assessments. As has been argued in the general introduction to this thesis, this effect cannot be ascertained directly. The data, however, do offer some possibilities to investigate this issue more closely; specifically, there are two pointers that reduce the likelihood that our findings were skewed because we could not study true positives.

First, if risk assessment and risk management strategies were accurately targeting high risk patients and adequately preventing their discharge, one would not expect to find many patients assessed as high risk among those discharged. One would furthermore expect the high risk group that was discharged to be largely equivalent to the group discharged against hospital advice. Neither of these expectations were confirmed by the data. Of the discharged patients, more than 25% was rated 4 or higher by treatment supervisors on a 6-point risk scale, whereas the reoffending risk of nearly 60% of patients for whom TBS-extension was requested by the hospital was in fact rated 3 or lower on the CIDRRI risk assessment item. Of the 17 patients discharged against hospital advice, only 7 (41%) fell into the group rated to be at high risk by clinicians. These findings together indicate that there is no straightforward equivalence between clinical risk judgments and risk management strategies, even when accounting for court decisions that are not supported by the hospital.

Second, due to contrary discharges we can study reoffending rates of a small group of patients who would have been retained in hospital had clinical risk assessment been the only deciding factor. If clinical risk assessments are usually highly adequate, higher rates of reoffending are to be expected among patients discharged against hospital advice than among other dischargees. Reconviction rate was indeed higher among contrary dischargees than among others (29% vs. 18%), but the difference was unspectacular and not statistically significant.

All in all these findings do not offer compelling support for the assumption that risk assessments for patients not released from hospital are much more accurate than for those who are discharged. Given the strong correlations between clinical risk assessments and item scores, as demonstrated in Study I, we may assume that if relations between clinical risk assessments and outcome were not significantly misrepresented in our study, consequently the same holds true with regard to predictive power of items (and scales).

### *III.3.1.2.d Moderating effects among dynamic items*

It has been noted in the General Introduction that originally, actuarial tables take into account the fact that particular item combinations rather than linear accumulations of risk factors may be predictive, an effect illustrated with reference to the Iterative Classification Trees used by Monahan *et al.* (2001) (see p. 37). In the present research, only linear combinations of items were tested. The research sample was too small to allow meaningful division into subgroups using classification trees. Thus, the possibility remains open that particular subsets of CIDRRI items and/or scales may be predictive for particular offender subgroups.

### *III.3.1.2.e Follow-up time*

As has been shown in the general introduction, longer follow-ups favor stable risk factors. The relatively long follow-up in the present study (5.5 to 8.5 years), which served the need of identifying sufficient reoffenders for meaningful analysis, may have worked against the establishment of valid dynamic risk factors. However, as was shown in Study III, if the follow-up period is limited to a fixed 5.5 years, this does not significantly change the predictor set. Additional post hoc analyses not reported in the study revealed, that if the outcome criterion was changed to reconviction within the first year at risk, the 6 dynamic scales of the CIDRRI remained non-predictive, irrespective of inclusion or exclusion of the static prediction model in the analysis. Only when low threshold 'reoffending' indicators such as repeal of probationary leave were included in the outcome criterion, did Scale 1 (empathic responsibility for the offence) become a statistically significant predictor (Cox regression, forward entry of scales (Block 2) after entering the static predictor set (Block 1), with Likelihood Ratio significance test: odds ratio .676, p=.034) . Thus, predictive power of the CIDRRI dynamic risk factors seems more likely to depend on the type of outcome that is predicted rather than the duration of follow-up. Note that, in accordance with previous findings, the predictor-outcome relation for Scale 1 again was the reverse of what would be expected.

### *III.3.1.2.f External validity*

The scope of our sample would suggest high external validity of the study reported in this thesis. However, external validity is somewhat compromised by the fact that data reflect the situation in TBS during the years 1996 through 1998. Considerable advances have been made in the area of risk assessment in the years since. Notable events in TBS in this regard were the introduction of structured clinical risk assessment procedures (e.g. HCR-20) and the PCL-R, in the year 2000. In general, it may be assumed that clinicians in 2004 are more aware of empirically established risk predictors, more familiar with requirements regarding reliable item rating, and less naïve with regard to manipulations and superficial adaptation by patients than they or their colleagues were 6 to 8 years ago. On the other hand, as Webster *et al.* (2002) have noted, the availability of instruments in itself does not fully safeguard clinicians against such pitfalls – this depends on correct use, and as these and other authors have pointed out, much is often left to be desired in this regard (Grubin & Wingate, 1996; Buchanan, 1999).

## III.3.2 Implications for research and clinical practice

This thesis started out on an optimistic note regarding the possibilities of dynamic risk assessment – without optimism, a study of the scope and duration of the present one would not have been undertaken, let alone finished. In the general introduction, authors who did not share this optimism were mildly castigated for their fatalism, and methods overly reliant on static predictors were also questioned from an ethical viewpoint. It is both interesting and worrying that the findings of the present study now seem to compel us to back these authors and their methods rather than criticize them. Implications of our findings for future research and clinical practice deserve careful consideration.

### ***III.3.2.3 Research implications***

The findings reported in this study imply several possible conclusions with regard to dynamic risk assessment that have implications for future research.

- Risk is static. This depressing conclusion would be the most far-reaching to draw. It would simply mean that both researchers and clinicians must join Rice & Harris (2003) in their view that it is a wasted effort to try and find dynamic predictors that add predictive validity to a basic set of static risk factors. However, even though the evidence from the present study may seem to support such a stance, we feel it should not be adopted before the alternative possibilities below have been thoroughly addressed.
- Dynamic risk is so individually determined that nomothetic methods do not apply. It is very well possible that present conceptualizations of dynamic risk assessment, including those in the study reported here, have been too quick to assume that dynamic factors, like their static counterparts, will generally be the same for every offender. In fact, however, dynamic risk may be so individually determined that it cannot be described from the top down (nomothetic), but only from the bottom up (ideographic). If this is the case, dynamic risk research will need to focus on causal structures underlying dynamic risk on case level, rather than trying to isolate the content of such structures by analyzing data from large samples. In other words, a theory of reoffending needs to be developed. This would require thorough and extensive comparative case studies of recidivists and non-recidivists, and sophisticated techniques of content analysis. If, through such methods, a theory can indeed be generated, predictions based on such a theory can then be empirically validated.
- Dynamic risk factors need to be measured in a different way. Another possible consequence of the complexity of dynamic risk factors is that the methods applied for their measurement need reconsideration. Strategies based on clinical observation, such as found in the CIDRRI or the HCR-20, may not sufficiently succeed in assessing the patient's true condition; they are susceptible to wishful thinking by the clinician as well as 'impression management' by the patient. Several alternatives should be considered.
    - One option would be to reduce the risk of subjective distortions by providing very detailed item instructions. This has, for instance, been attempted by Reed *et al.* (1997), in their Behavioural Status Index (BEST-Index or BSI). Every item provides detailed examples of the type of behavior that is targeted. Also, explicit definitions are given for terms like 'mostly', 'rarely', 'frequently', et cetera. The down side to this approach is that it is highly labor-intensive and works best if the entire treatment planning and reporting cycle is organized along

BSI-lines. Moreover, this kind of approach needs to be supplemented by ongoing and extensive rater training.

- A related approach would include the use of psychometric tools specifically designed to measure particular psychological concepts and disorders. The application of the PCL-R is only the most obvious (as well as most successful) example of this. Though tentative instruments for measuring hostile aggression, impulsivity and anger have become available (Buss & Perry, 1992; Barratt, 1994; and Novaco, 1994, respectively), these need further study, as do concepts like empathy or insight, which remain in need of detailed operationalisation. In The Netherlands, several out-patient forensic psychiatric facilities have joined forces to construct and validate an instrument for measuring treatment motivation (Drieschner, 2002).
- The direct measurement of dynamic conditions may be extended to include the application of objective measures, such as physiological, neurocognitive, and neuroimaging parameters. Phallometry (for assessing deviant sexual preferences) or polygraphy ('lie detector test'), though controversial still, are alternatives gradually becoming open to discussion in the Netherlands (Rassin *et al.*, 2002). MRI-scanning has already been successfully applied in psychopathy research (e.g., Kiehl *et al.*, 2001). Such measures, though costly, may provide a far safer means of circumventing manipulative answering strategies and rater bias than any other available. At the least, they deserve serious consideration for risk assessment applications, though it is likely that if they can be used for this purpose, this use will probably be limited to specific subgroups of offenders (e.g., pedophiles).

Given the changeability of dynamic items, any type of measurement ought to be repeated over time during treatment, and the changes in scores rather than discrete scores at a particular point in time should be applied as the independent variable with regard to reoffending.

### *III.3.2.4 Clinical implications*

First of all, the present research like many previous studies, helps to remind clinicians of the crucial importance of historic risk factors. The model presented in the final study fortunately shows that such factors need not necessarily be trivial from a clinical point of view: three out of the four predictors in fact concern psychiatric diagnoses. As the research design was geared to reflect clinical practice in TBS, findings also hold some specific clues for practitioners in that area.

- Dynamic patient characteristics that are routinely used in treatment evaluation, communications about patients, and risk assessment in TBS, when assessed on the basis of clinical perceptions and without reference to structured methods, appear not to be predictive at all of true reoffending. Clinicians should be very careful when considering such charac-teristics, and be critical of their own and others' use of them.
- To the extent that dynamic items in this study were valid measures of true dynamic patient characteristics, the findings call into question the changeability of reoffending risk. The assumption that risk of reoffending can be significantly and durably reduced by therapeutic interventions targeting symptoms of mental disorders, skill deficiencies, and behavioral problems in the here and now, may be overly optimistic. Intervention may generally need to focus on control rather than cure. They may also need to target the patient's context at least as intensively as the patient himself.
- Practitioners should not be too quick to assume that they agree among each other about the meaning of routine clinical terminology. They should be aware that interpretations of such terminology may in fact differ considerably even among colleagues who work closely together. It is advisable to discuss the meaning of particular words and labels directly. In risk assessment procedures, it is furthermore advisable to neutralize subjective differences by involving multiple independent raters in the process. Agreement between raters strengthens the significance of clinical observations, whereas disagreements should lead to discussion and reevaluation of the observations (Groen & Van den Brink, 1992).
- Clinicians in TBS should be wary of individual risk intuitions based on general, unstructured clinical observations or subjective hunches. The unstructured assessment of risk by TBS-clinicians was shown to be non-predictive in the present study. Given the superior performance of any of the available structured risk assessment tools as compared to the unstructured clinical assessment in TBS, standardized use of one or more of such tools is strongly recommended. Moreover, these tools should not be used as a mere addition to clinical risk assessment - in fact, the reverse should be the case: professional discretion should be regarded as an optional addition to the results from a risk assessment instrument. Structured risk assessment results should only be clinically modified if there are compelling reasons to do so, and modifications should be clearly argued.

A final recommendation regards clinicians and researchers alike. Rather than withdrawing each into their own specialism and ignoring each other's priorities, or worse, attacking each other, these disciplines need to find common ground and work from there.

By far the greatest challenge they face together remains the development of a causal theory of risk. It needs to become clear *why* known risk factors are predictive of violence. Clinicians are a prime source of knowledge and ideas when it comes to this, and researchers should make use of this knowledge. It is likely that such a theory will show some predictors not to be predictors at all, but mere correlates of the predicted violent behavior, that share with it an underlying explanatory mechanism (e.g., Silver & Miller, 2002). If, any time in the future, risk of reoffending can be operationalized in causal instead of correlational terms, this will also allow a return to the clinically preferred assessment of cases on an individualized level. However, even if we should succeed in formulating such a theory, on the individual level its predictions will remain dependent on complex constellations of parameters that cannot all be determined with certainty. That is, even a theory-driven, causal risk assessment model will only be able to produce probabilistic risk estimates.

In general, if researchers want their work to have meaning in clinical practice, they will have to address the needs and realities of that practice rather than remain in the 'laboratory'. This involves finding the optimum balance between reliable and valid methodology on the one hand, and practical feasibility on the other. It seems that risk assessment methods will stand the best chance of finding their way into clinicians' hands and staying there, if they have a 'plug-and-play' character. The iterative classification trees derived from their MacArthur risk assessment study by Monahan *et al.* (2001), are a good example of this. In appearance, they are similar to the diagnostic decision trees in DSM-IV (APA, 1994) and as such readily comprehensible to clinicians.

Researchers, clinicians, the judiciary and policy makers should work together to agree on acceptable false positive/false negative balances in risk assessment decision making, incorporated in uniform decision-making guidelines for commonly used risk assessment procedures. The question which relative risk level corresponds to which decision cannot be left unanswered. In formulating guidelines, the relative weight of actuarial findings, professional discretion and contextual influences, as well as prerequisites for their consideration, should be clearly outlined. Similarly, researchers, clinicians and the judiciary need to look critically at present risk reports, and develop improved and standardized formats that incorporate up-to-date risk assessment knowledge, yet speak the language of the court-room. As others have noted (De Ruiter, 2000), the introduction of specialized training with regard to this would be highly expedient.

### *III.3.2.5 Ethical considerations with regard to follow-up*

The choice of follow-up time is not merely a technical matter in research planning: it carries ethical implications as well. From a research point of view very long follow-ups are enticing, as larger numbers of reoffenders tend to increase the statistical power of a study. Rates of reoffending have been described over extremely long follow-ups. Prentky *et al.* (1997), for instance, followed a sample of child molesters over a 25 year period. Quinsey *et al.* (1998) reported an astonishing 100% probability of reoffending within 10 years for offenders in the highest scoring 'bin' of their Violence Risk Appraisal Guide (VRAG). However, the question arises if these are realistic time frames for a forecast of human behavior. Harris & Rice (2003) found that predictive accuracy of four actuarial instruments was consistently highest after a 2 year follow-up, and subsequently decreased as the follow-up was extended up to 11 years.

At the very least, the question needs to be raised whether we actually expect a risk assessment procedure to offer risk estimates not for the next few years, but for the next decade, several decades or even the rest of the offender's life. This question is all the more pressing when instruments for long-term risk assessment mainly consist of static risk factors and contain few dynamic risk indicators, or none at all. A high risk outcome on a static risk assessment tool that has been validated over a long follow-up period, leaves few risk management options other than long-term incarceration. Given the imperfections of the existing risk assessment procedures, such an intervention could lead to serious injustice. As was shown in Study III, and was demonstrated by Leuw (1999), the bulk of reoffending after TBS occurs in the first 5 years after discharge. In this context, it seems advisable not to project risk estimates beyond that timeframe. In general, it would seem that the forensic expert has at least as much reason as the meteorologist to be wary of long term predictions

Obviously, a limited risk assessment timeframe is only truly meaningful in case the risk assessment procedure includes changeable characteristics; static assessment tools will merely return the same result when reassessed after a 5-year interval. In agreement with Silver & Miller (2002), it is proposed here that static instruments should be regarded as describing base rate dispersion. Whether the protection of society should prevail over individual freedom to such an extent that certain people can be incarcerated for very long times simply because they answer certain descriptive characteristics of a particular population subgroup, is a complex ethical issue beyond the scope of the present thesis, that needs to be addressed by scientists as well as policy-makers and the law.

## III.4 ☙ Towards the future – some final remarks

After the 1990's produced many studies aglow with predictionist optimism, the tide at present seems to be turning somewhat. New validation studies with existing tools confirm earlier results, but do not add new insights. Different authors note similar bottlenecks, but few offer solutions for them. With only a few exceptions, and taking an optimistic view, risk assessment validity seems firmly 'stuck' somewhere around ROC AUC = .80.

One of the few new impulses the field has seen in recent years is the search for so-called protective factors, in an attempt to break through the one-sidedly negative approach typical of risk assessment instruments (a rare exception being the SAVRY, by Bartel *et al.*, 1999, which includes six protective factors). Protective factors indicate characteristics and circumstances that reduce reoffending risk rather than increase it. They promote non-deviant behavior, resilience, and desistance from criminal activity (Lösel & Bender, 2003). Howells (1998) suggests that any needs assessment of violent offenders should include such 'buffer' factors.

A challenging conceptual question inherent in this approach is, whether or not a protective factor is more than merely the absence or the negative of a risk factor (Lösel & Bender, 2003; De Vogel *et al.*, 2004). It has been argued that protective factors can be independent characteristics or circumstances that buffer or mediate the effect of present risk factors (Fitzpatrick, 1997). In The Netherlands, De Vogel *et al.* (2004) created a research instrument based on this assumption, called the Structured Assessment of Protective Factors (SAPROF). It comprises 16 possible buffering or mediating protective factors, to be rated in conjunction with either HCR-20 or SVR-20. Simultaneously a group of Canadian researchers developed the Short Term Assessment of Risk and Treatability (START, Martin *et al.*, 2004), which lists 20 factors that can be assessed on a continuum as either a risk or a strength. The START thus adheres to the view that risk and protective factors are two sides of the same medal. The validity and additive value of both these new instruments awaits demonstration. As yet, research support for protective factors as valid indicators of reduced violence risk is wanting (De Vogel *et al.*, 2004). Research involving START and SAPROF offers opportunities to change this, and may in time add a valuable and clinically salient new dimension to risk assessment.

A second notable development in the last few years has been the increased accent on risk management. This shift of focus is motivated by an awareness that risk assessment by itself has little meaning – only the decisions and interventions based on it lends meaning to the activity. Some authors have argued that the forensic field should let go of 'predictionism' and start to concentrate more exclusively on risk management issues. However, though this approach helps to further the integration of research and practice, it also in a way seems to beg the question: as long as it remains unclear which dynamic characteristics increase (or reduce) risk, how do we know what to manage? The interest in risk management only further stresses the need to identify valid dynamic risk factors – if, that is, risk management strategies want to extend anywhere beyond mere incarceration.

## III.4.1 Limits of predictability

Present-day western society is obsessed with issues of safety and security (Castel, 1991; Van Swaaningen, 1996: Bouttelier, 2002). Governments are pressed for fail-safe guarantees against any kind of undesirable event. Van Swaaningen (1996) characterizes this attitude as infantile, and accuses politicians of cultivating a 'populist ecology of fear'. Meanwhile, and contrary to these demands, the risk assessment literature has recently shown an increased awareness of the fundamental uncertainties ingrained in reality, and the impossibility of predicting the future with anything approaching perfect accuracy. Several authors have illustrated this by referring to chaos-theory (Hart, 2001; Williams & Arriogo, 2002). Chaos-theory's main tenet is that very small causes can have very great effects. To predict anything with high accuracy, the initial situation needs to be measured with a level of precision that is, literally, humanly unattainable (Ford, 1983). This holds true even in simple, deterministic physical systems; evidently, it will therefore also hold true in the far more complex world of human behavior. Though we are not likely to see the implementation of esoteric chaos-theory mathematics in risk assessment tools any time soon, the analogy is instructive and certainly exceeds the level of mere metaphor. Recognition of the complex, chaotic, and non-linear nature of reality helps researchers, clinicians and the public alike to a better understanding of the limits on predictive aspirations. In time, it may also help to find a 'fractal' order underlying the apparently chaotic process of recidivism, analagous to the underlying mathematical order that describes such diverse chaotic phenomena as the shape of clouds (and of broccoli), the coast line of the United Kingdom, or the graphic on the cover of this book

Nevertheless, the possibility cannot be ruled out that prediction results will never be much better than those presently achieved. In saying that, it should be noted that effect sizes attained by forensic risk assessment methods are at least as good and often (far) better than those of many medical interventions, psychotherapeutic methods or educational decisions (Lipsey & Wilson, 1993; Andreasen, 2000). The unbridgeable part of the gap dividing that level of accuracy from absolute certainty will always remain the domain of moral, judicial and political judgments.

# Summaries

# Samenvatting

Dit proefschrift bestaat uit drie onderdelen. Het eerste deel is een algemene inleiding waarin kernbegrippen en de stand van kennis op het terrein van delictrisicotaxatie uiteen worden gezet. Het tweede deel bevat een drietal empirische studies naar risicotaxatie in de terbeschikkingstelling (TBS), die tussen 1996 en 2004 werden uitgevoerd in 8 TBS-klinieken. In het derde deel worden de belangrijkste bevindingen uit de studies samengevat, en wordt ingegaan op de mogelijke betekenis ervan.

*TBS*. Het in dit proefschrift beschreven onderzoek speelde zich af in de context van de maatregel van terbeschikkingstelling (TBS). TBS kan worden opgelegd aan de pleger van een ernstig delict, wanneer zijn misdrijf hem door de aanwezigheid van een geesteziekte niet (volledig) kan worden toegerekend, en wanneer de kans op delictherhaling groot wordt geacht. De TBS-gestelde wordt, in eerste instantie voor 2 jaar, opgenomen in een TBS-kliniek, waar hij verpleegd en behandeld wordt tot de mate van delictrisico voldoende is verminderd. De kliniek rapporteert tenminste eenmaal per 2 jaar aan de rechter over de behandelvoortgang en het delictgevaar. Mede op basis daarvan beslist de rechter of de maatregel wordt verlengd dan wel beëindigd.

## I Algemene inleiding

### *Risicotaxatie: kernbegrippen en huidige stand van kennis*

### I.1 Inleiding

De inschatting van de kans dat een patiënt in de toekomst opnieuw een delict zal plegen, ook wel 'risicotaxatie' genoemd, is een even alledaags als essentieel onderdeel van de klinische praktijk in de forensische psychiatrie. De laatste drie decennia is veel wetenschappelijk onderzoek gedaan naar dergelijke inschattingen. Tot op heden blijven verschillen van mening bestaan over wat de beste benaderingswijze is en hoe accuraat zulke inschattingen in het beste geval kunnen zijn.

### I.2 Kernbegrippen

Aanvankelijk werd het wetenschappelijk denken over delictrisico beheerst door het begrip 'gevaarlijkheid', dat primair werd gezien als een eigenschap van de persoon. Mettertijd groeide evenwel het inzicht dat de kans op een delict de uitkomst is van een interactie tussen persoon, omgeving en situatie. Tegelijk verschoof het accent van voorspelling van criminele recidive naar inschatting van de relatieve kans daarop, analoog aan de manier waarop weersverwachtingen worden opgesteld. Dit doet recht aan de onvermijdelijke onzekerheidsmarge die in elke prognose aanwezig is.

In de klinische praktijk moeten niettemin op zulke relatieve kansinschattingen absolute beslissingen worden gebaseerd. Er kunnen dan twee soorten fouten worden gemaakt: een patiënt kan onterecht als een hoog risico worden gezien en langer van zijn vrijheid beroofd blijven dan nodig is (vals-positieve voorspelling); of een patiënt kan onterecht als een laag risico worden beschouwd, waardoor hij voortijdig uit behandeling ontslagen wordt en een nieuw delict begaat (vals-negatieve voorspelling).

Metingen van de omvang van deze fouten worden door diverse factoren beïnvloed. Ten eerste speelt de tijd dat patiënten na hun ontslag worden gevolgd een belangrijke rol: hoe langer de volgtijd, hoe meer recidivisten. Verder is van invloed welke uitkomstmaat wordt gehanteerd om recidive vast te stellen. Wanneer, bijvoorbeeld, iedere vorm van gewelddadig gedrag als recidive wordt opgevat, levert dat meer recidivisten op dan wanneer alleen hernieuwde veroordelingen als recidive worden gezien. Ten derde leveren de meest gebruikte bronnen van recidivegegevens, officiële justitiële registers, minder recidivisten op dan zelfrapportage door betrokkenen, omdat niet alle gepleegde criminaliteit bij politie of justitie bekend wordt. In dit verband wordt wel gesproken van het *dark number*: een groep recidivisten die door onderzoekers niet wordt waargenomen. Dit 'donkere getal' wordt vaak beschouwd als een essentiële tekortkoming in risicotaxatiestudies. Recent onderzoek liet echter zien dat officiële bronnen en zelfrapportage nagenoeg identieke statistische voorspellingsmodellen opleveren. Mogelijk is het *dark number* dus niet zo'n grote hindernis voor onderzoek als wel eens wordt verondersteld.

Een vierde factor die het beeld van de accuraatheid van risicotaxaties beïnvloedt is de *base rate* van terugval. Dit is het percentage personen in de hele populatie dat binnen een bepaalde tijdsspanne recidiveert. Bij zeer lage of hoge *base rates* zal de inschatting dat terugval bij niemand respectievelijk bij iedereen optreedt al zeer accuraat zijn, en hebben specifieke risicotaxatiemethodes weinig meer toe te voegen. Bij een *base rate* van 50% is de potentiële toegevoegde waarde van zulke methodes daarentegen het grootst.

Twee laatste kernbegrippen zijn statische en dynamische risicofactoren. Factoren die samenhangen met delictrisico worden vaak onderscheiden in onveranderbare (historische) gegevenheden (bijvoorbeeld: leeftijd waarop betrokkene voor het eerst werd veroordeeld), en veranderbare, en mogelijk behandelbare, persoons- of omgevingsfactoren (bijvoorbeeld: betrokkene's stemming). De afgrenzing tussen beide categorieën is niet scherp. Veranderlijke en veranderbare kenmerken zijn klinisch interessant, maar kenmerken die zo veranderlijk zijn dat ze sterk fluctueren per dag of week, hebben minder kans op de langere termijn voorspellend te zijn dan meer stabiele kenmerken.

## I.3 Risicotaxatie in de praktijk: drie benaderingen

In de praktijk kan een risicotaxatie grofweg op drie verschillende manieren worden uitgevoerd. De oudste daarvan is de ongestructureerd klinische benadering. Kenmerkend voor deze aanpak is dat zij sterk bepaald wordt door de subjectieve inbreng van de beoordelaar. Er wordt geen gebruik gemaakt van een inhoudelijk sturend, gestandaardiseerd instrument. Een groot aantal empirische studies heeft duidelijk gemaakt dat deze werkwijze in het slechtste geval leidt tot onbetrouwbare, inaccurate inschattingen, en in het beste geval tot inschattingen waarvan de kwaliteit langs andere, veel eenvoudigere weg kan worden geëvenaard.

Dat eenvoudiger alternatief is de actuariële risicotaxatie. Bij deze werkwijze wordt het subjectieve element volledig losgelaten. De inschatting gebeurt aan de hand van een vooraf gegeven lijst kenmerken waarvan uit onderzoek gebleken is dat zij daadwerkelijk met recidivegevaar samenhangen. Ook de wijze van beoordeling en de rekenformule om tot een eindconclusie te komen zijn strikt voorgeschreven. Voor gebruik van zulke instrumenten is specialistische kennis meestal niet nodig en volgens sommigen zelfs onwenselijk. Actuariële risicotaxatie instrumenten laten in onderzoek redelijke tot goede resultaten zien. Een belangrijk nadeel is dat hun inhoud vaak statisch is, en geen aanknopingspunten biedt voor behandelinterventie: een eenmaal vastgesteld risico wordt nooit meer kleiner.

De sterk gepolariseerde impasse tussen voorstanders van een klinische en die van een actuariële benadering werd pas doorbroken toen het derde alternatief ten tonele verscheen: de gestructureerd klinische benadering. Deze aanpak neemt de gestandaardiseerde lijst en empirische onderbouwing uit de actuariële benadering over. Zij voegt daar echter klinisch relevante risicofactoren aan toe, en geeft de beoordelaar ook de ruimte in het eindoordeel af te wijken van de instrumentscore, wanneer daar goede redenen voor zijn. Voorspellingsonderzoek met deze categorie instrumenten laat wisselende, gemiddeld genomen echter redelijke resultaten zien. Verder onderzoek is nodig.

## I.4 De werkelijkheid van risicotaxatie

De evaluatie van de kwaliteit van risicotaxatie instrumenten is tot nu toe vooral een laboratoriumactiviteit geweest. Over de feitelijke effectiviteit van instrumenten in de klinische praktijk, en hun effect op (vermindering van) recidive, is vooralsnog vrijwel niets bekend. Instrumentgebruik zal in klinische settings zelden zo fraai aan alle 'regelen der kunst' voldoen als in onderzoek, en het effect ervan zal daardoor krimpen.

Het effect van risicotaxatie op recidive hangt ook af van de manier waarop over risico wordt gerapporteerd. Zo is uit onderzoek bekend dat ontvangers van risicorapportages (bijvoorbeeld rechters) in hun besluitvorming worden beïnvloed door het gebruik van absolute ('1 op 10') of relatieve ('10%') termen in de verwoording van de terugvalkans. Absolute getallen leiden tot een hoger waargenomen risico dan hun procentuele equivalenten. Dit soort bevindingen onderstrepen het belang van uniforme richtlijnen voor risicorapportage, die vrijwel ontbreken.

## I.5 Conclusie

Het wetenschapsgebied van risicotaxatie heeft in drie decennia een grote ontwikkeling doorgemaakt, resulterend in gematigd optimisme over prognosemogelijkheden. Veel vragen blijven echter open, en de vervreemding van het risicotaxatie onderzoek van de klinische praktijk is een belangrijk probleem.

# II Drie empirische studies

## II.1 Studie I. De structurele samenhang van klinisch afgeleide dynamische indicatoren voor terugvalrisico

In deze studie werd onderzocht of psychische en gedragskenmerken van patiënten, die volgens clinici in de TBS bepalend zijn voor recidiverisico, samenhangende concepten van hogere orde vertegenwoordigen. Tevens werd onderzocht in hoeverre de risicotaxatie door TBS-clinici inderdaad met hun beoordeling van deze kenmerken samenhing.

***Methode***. Clinici in 4 TBS-klinieken werd gevraagd aan te geven welke gedragskenmerken van patiënten volgens hen bepalend zijn voor terugvalrisico. Na bewerking en toetsing van de bevindingen in 8 klinieken, resulteerde een instrument dat 47 van zulke kenmerken bevat, de 'Clinical Inventory of Reoffending Risk Indicators' (CIDRRI). De hoofdbehandelaar geeft op een 6-puntsschaal aan in welke mate elk kenmerk op een bepaalde patiënt van toepassing is. Voorts geeft hij op een 48$^{ste}$ item aan hoe groot hij op langere termijn de kans acht op een nieuw delict.

Op deze itemlijst werd een exploratieve principale factoranalyse uitgevoerd met orthogonale (varimax) rotatie. Deze methode identificeert groepen van items die met elkaar samenhangen en samen een overkoepelend concept vertegenwoordigen. Ook werden gegevens over kort geleden opgenomen en kort geleden ontslagen patiënten apart geanalyseerd, om te toetsen of het factormodel van de behandelfase afhankelijk was. Tot slot werd de mate waarin CIDRRI items en factoren de klinische risicotaxatie voorspelden getoetst middels zogenaamde ROC-analyse.

***Resultaten***. Er waren 370 CIDRRI's beschikbaar voor analyse. Er werden 6 factoren gevonden. Dit waren:

- Empathische acceptatie van verantwoordelijkheid voor het delict (11 items)
- Gebrek aan zelfstandigheid (12 items)
- Antisociaal narcisme (10 items)
- Medewerking aan de behandeling (5 items)
- Bereiken van behandeldoelen (7 items)
- Vermijding (2 items)

Door optelling van ruwe scores werden deze factoren omgezet in schalen. Het deelmodel voor net opgenomen patiënten was gelijk aan het totaalmodel. Het deelmodel voor net vertrokken patiënten omvatte 5 in plaats van 6 factoren, doordat de factoren 'medewerking aan behandeling' en 'bereiken van behandeldoelen' tot één factor samenvielen.

De meeste CIDRRI items hielden significant verband met de klinische inschatting van de terugvalkans. De schalen hielden duidelijk verband met de klinische risico-inschatting, behalve de schaal 'Vermijding'. Verbanden tussen items en schalen enerzijds en klinische risico-inschatting anderzijds waren voor net opgenomen patiënten anders dan voor net vertrokken patiënten. Bij laatstgenoemde groep lag het accent op delictverantwoordelijkheid en bereiken van behandeldoelen, terwijl bij 'starters' gebrek aan zelfstandigheid het sterkst voorspellend was voor de klinische risico-inschatting.

***Discussie***. Dynamische patiëntkenmerken die volgens clinici in de TBS belangrijk zijn bij het inschatten van delictrisico, blijken een zestal betekenisvolle hogere orde concepten te vertegenwoordigen. Zowel deze concepten als de oorspronkelijke items hangen inderdaad samen met de klinische inschatting van terugvalrisico. Hoewel dit niets zegt over verbanden met feitelijke terugval, is zulke kennis over klinische denkwijzen aangaande terugvalrisico ook op zichzelf van belang.

## II.2 Studie II. Betrouwbaarheid en discriminante validiteit van dynamische delictrisico-indicatoren in de forensische klinische praktijk

In deze studie werd onderzocht in hoeverre de CIDRRI, een op klinisch gebruiksgemak toegesneden instrument, toegepast onder alledaagse klinische omstandigheden acceptabele maten van betrouwbaarheid en discriminante validiteit opleverde.

***Methode***. Er werd een interbeoordelaar betrouwbaarheidsstudie (A), een test-hertest betrouwbaarheidsstudie (B) en een discriminante validiteitsstudie (C) uitgevoerd. In deelstudie A werden voor 75 patiënten twee CIDRRI's tegelijk gescoord door twee beoordelaars onafhankelijk van elkaar. In deelstudie B werden voor 29 patiënten twee CIDRRI's na elkaar, met een maand tussenpoos, gescoord door dezelfde behandelcoördinator. In deelstudie C werd onderzocht of een groep van 115 net opgenomen patiënten in een *k*-means clusteranalyse kon worden onderscheiden van een groep van 118 net vertrokken patiënten op basis van hun CIDRRI scores. Achterliggende veronderstelling was dat net opgenomen patiënten in het algemeen een hoger delictrisico kennen dan net vertrokken patiënten.

***Resultaten***. Uit deelstudie A bleek dat de meeste items aan minimumeisen voor acceptabele interbeoordelaar betrouwbaarheid voldeden. Dat betekent dat twee onafhankelijke beoordelaars voor dezelfde patiënt meestal tot redelijk vergelijkbare beoordelingen kwamen. Uit deelstudie B bleek een goede mate van test-hertest betrouwbaarheid voor de meeste items en alle schalen. Deelstudie C liet zien dat 'startende' en 'vertrekkende' patiënten vooral op basis van CIDRRI schaalscores, maar ook op basis van de klinische risico-inschatting, goed van elkaar te onderscheiden waren.

***Discussie***. Betrouwbaarheidsgegevens over de CIDRRI wijken niet erg veel af van die van veelgebruikte andere risicotaxatie instrumenten; dit ondanks het feit dat, in tegenstelling tot het meeste betrouwbaarheidsonderzoek, geen bijzondere inspanningen (training, instructie, handleiding) werden verricht om betrouwbaarheid te verhogen. Bovendien bevatte de CIDRRI 6-

punts- in plaats van de gebruikelijker 3-puntsscoreschalen en was er dus meer ruimte voor discrepanties in beoordelingen.

De CIDRRI blijkt net opgenomen patiënten goed te kunnen onderscheiden van net vertrokken patiënten. Dit onderscheidend vermogen is echter niet noodzakelijk gerelateerd aan feitelijke risicoverschillen; het kan ook wijzen op klinische vooroordelen met betrekking tot deze groepen zoals die in CIDRRI beoordelingen tot uiting komen. Toetsing van de CIDRRI aan feitelijke recidive zal moeten uitwijzen of dit het geval is.

## II.3 Studie III. Statische en dynamische patiëntkenmerken als voorspellers van criminele recidive. Een prospectieve studie in een Nederlandse forensisch psychiatrische steekproef

In deze studie werden de dynamische risico-indicatoren uit de CIDRRI in een prospectieve opzet getoetst op hun voorspellende waarde voor feitelijke recidive, rekening houdend met vooraf gegeven statische voorspellers.

***Methode***. Gedurende de jaren 1996 tot en met 1998 werden voor alle patiënten die uit een van 7 TBS-klinieken vertrokken met proefverlof of einde maatregel, gegevens verzameld met de CIDRRI en met een checklist voor (statische) achtergrondgegevens. Scores werden na een volgperiode van minimaal 5,5 en maximaal 8,5 jaar gekoppeld aan recidivegegevens uit de Centrale Justitiële Documentatie. Recidivecriterium was hernieuwde veroordeling voor een delict met een gewelddadige of seksuele component (inclusief pogingsdelicten). Middels survivalanalyse, een vorm van regressieanalyse die rekening houdt met de uiteenlopende tijdsduren dat ex-patiënten in de samenleving hebben verbleven, werd een multivariaat voorspellingsmodel opgebouwd. Een regressieanalyse berekent stap voor stap welke kenmerken het sterkst voorspellend zijn, en vervolgens welke andere kenmerken nog voorspellingskracht toevoegen aan de voorspellers die al in het model zijn opgenomen. Alvorens items in de multivariate analyse te brengen werden enkele voorselecties toegepast om te voorkomen dat een resulterend model teveel op toevalsinvloeden zou berusten (kanskapitalisatie).

***Resultaten***. De dataset bevatte 132 personen: bijna driekwart van alle TBS-gestelden die gedurende de jaren van dataverzameling aan het uitstroomcriterium voldeden. De meeste van hen kregen TBS opgelegd wegens een gewelds- of zedendelict. Bij driekwart was bij opname een persoonlijkheidsstoornis vastgesteld, voornamelijk van het antisociale, narcistische of ongespecificeerde type. Ook stoornissen op As 1 van DSM-III-r kwamen veelvuldig voor, met name psychose en verslavingsproblematiek. De mediane volgperiode na ontslag was 6,8 jaar. Van betrokkenen werd 19,7% opnieuw veroordeeld voor een delict dat aan het terugvalcriterium voldeed.

De klinische inschatting van het terugvalrisico aan het einde van de CIDRRI had geen enkele voorspellende waarde voor feitelijke nieuwe veroordeling. Wel hadden enkele dynamische kenmerken op zichzelf beschouwd voorspellende waarde, maar hun verband met recidive was in alle gevallen tegengesteld aan de verwachting (bijvoorbeeld: patiënten die als *meer* empathisch werden gezien recidiveerden *vaker*). Multivariate analyse van statische voorspellers leverde een model met vier predictoren op. CIDRRI-schalen en losse CIDRRI-items voegden hieraan niets meer toe. Het uiteindelijke predictie-model was dus geheel statisch, en bevatte de volgende voorspellers:

- Aantal malen ongeoorloofd afwezig gedurende TBS (meer = hoger risico)
- Comorbiditeit van enige persoonlijkheidsstoornis met middelenproblematiek, bij opname (ja = hoger risico)
- Psychose bij opname (ja = lager risico)
- Enige persoonlijkheidsstoornis uit het DSM-III-r B-cluster, bij opname (ja = hoger risico).

Dit model bleek een tamelijk goede recidiveprognose op te leveren, met een ROC-waarde van 0,79 (waarbij 0,5 gelijk staat aan voorspelling niet beter dan toeval, en 1,0 aan perfecte voorspelling). Aanvullende analyses voor subgroepen van patiënten qua stoornissen en delicten, en met gelijke volgtijden voor alle patiënten, voegden geen nieuwe inzichten toe.

***Discussie.*** Deze studie leverde geen aanknopingspunten voor dynamische risicotaxatie in de TBS. De beoordelingscriteria en beoordelingswijze voor recidivegevaar zoals in dit onderzoek vormgegeven, bleken geen voorspellende waarde te hebben voor feitelijke terugval. Mogelijk spelen discrepanties tussen percepties van de beoordelaar en de feitelijke toestand van de patiënt hierin een rol. Om dit soort vertekeningen te voorkomen verdienen gestandaardiseerde of psychofysiologische methoden aanbeveling bij het operationaliseren van potentiële dynamische risicofactoren.

# III Algemene conclusies en discussie

## III.1 Sterke en zwakke punten van het huidige kennisbestand over delictrisicotaxatie

Het onderzoeksveld betreffende delictrisicotaxatie heeft veel vooruitgang geboekt in de laatste decennia, met name door toegenomen aandacht voor contextinvloeden, een explosieve toename in het aantal onderzoeksstudies, en verbeteringen in onderzoeksmethodiek. Daar staat tegenover dat empirisch onderzoek zich over het algemeen weinig rekenschap heeft gegeven van de klinische praktijk. De winst van instrumentgebruik lijkt tot nu toe vooral te zitten in aandacht voor historische voorspellers, en het bieden van een uniforme structuur.

## III.2 De empirische studies

In de empirische studies, uitgevoerd in 8 TBS-klinieken, bleek dat dynamische risico-indicatoren zoals door clinici toegepast een duidelijke onderliggende structuur kennen. Ook bleken deze indicatoren met een acceptabele mate van betrouwbaarheid gescoord te kunnen worden, en te differentiëren tussen net opgenomen en net ontslagen patiënten. De dynamische risico-indicatoren bleken echter geen toegevoegde voorspellende waarde te hebben met betrekking tot feitelijke hernieuwde veroordeling, gegeven een model met vier statische voorspellers. Het (statische) eindmodel dat uit de verzamelde data werd afgeleid had een tamelijk sterke voorspellende kracht.

## III.3 Discussie

De resultaten zijn niet hoopgevend wat betreft dynamische voorspellers van delictrisico in de TBS. Klinische aannamen over zulke voorspellers blijken empirisch niet bevestigd te worden. Het ontwikkelde instrument heeft weliswaar enige betrouwbaarheid en hangt samen met de klinische inschatting van terugvalrisico, maar houdt geen enkel verband met feitelijke recidive. In de eerste plaats is de vraag of deze uitkomst mede een gevolg kan zijn van gebreken in de onderzoeksopzet. Hoewel het landelijke bereik, de nauwe aansluiting bij de klinische werkelijkheid, en de prospectieve opzet het onderzoek een sterke basis gaven, zijn ook zwakke punten aanwijsbaar. Ten tweede moet worden stilgestaan bij inhoudelijke implicaties van de uitkomsten.

***Zwakke punten in de onderzoeksmethode***. De niet optimale betrouwbaarheid van het gebruikte instrument kan hebben geleid tot een onderschatting van de kracht van dynamische voorspellers. Verder is de vraag of de CIDRRI-items werkelijk de kenmerken gemeten hebben die ze beschrijven, of alleen maar de klinische perceptie daarvan. De omgekeerde verbanden die gevonden werden tussen dynamische voorspellers en terugval wekken het vermoeden dat bepaalde CIDRRI-scores vertekende waarnemingen van de werkelijkheid weerspiegelen. Mogelijk heeft de patiënt zich beter voorgedaan dan hij was; en mogelijk werd de clinicus teveel geleid door wat hij wilde zien. Psychopathie zou hierin een rol kunnen spelen: deze stoornis kenmerkt zich door moedwillig manipulerend en oppervlakkig functioneel gedrag van patiënten die in feite zeer recidivegevaarlijk zijn.

In dit onderzoek is verder alleen gekeken naar patiënten die uit de inrichting werden ontslagen. Juist-positieve voorspellingen (gevaarlijk geachte patiënten die terecht niet werden ontslagen) kunnen daardoor onderbelicht zijn. Zo'n vertekening is echter niet erg waarschijnlijk gezien het feit dat ook onder vertrokken patiënten een ruim aantal als 'hoog risico' werd ingeschat. Beëindigingen tegen advies van de kliniek leidden bovendien niet significant vaker tot recidive dan andere vormen van beëindiging.

De beperkte omvang van de dataset liet geen onderzoek toe naar het effect van specifieke combinaties van dynamische items bij specifieke subgroepen. De uniforme, lineaire benadering is een tekortkoming van het huidige onderzoek. Ook de lange volgtijd kan in het nadeel van dynamische voorspellers hebben gewerkt. Aanvullende analyses waarin de volgtijd aanzienlijk verkort werd leverden hiervoor echter geen aanwijzingen op. Andere varianten van het terugvalcriterium (bijvoorbeeld, aantal vermeldingen op het strafblad, of intrekking proefverlof) leverden evenmin wezenlijk andere resultaten op.

Tot slot is de vraag in hoeverre gegevens verzameld tussen 1996 en 1998 in 2005 nog geldig zijn. In de dagelijkse rapportage in TBS-klinieken zijn termen zoals die in de CIDRRI werden gebruikt echter nog onverminderd aanwezig; bovendien vormen deze termen vaak onderdeel van items in gestandaardiseerde internationale risicotaxatie instrumenten, zoals de HCR-20. Daarom lijken de uitkomsten van dit onderzoek onverminderd relevant voor de hedendaagse TBS-praktijk.

***Betekenis van resultaten vanuit wetenschappelijk perspectief.*** In relatie met wetenschappelijk onderzoek hebben de bevindingen de volgende mogelijke betekenissen:

- Risico is statisch. Dit is de meest vèrgaande conclusie, die pas getrokken mag worden nadat onderstaande mogelijkheden afdoende zijn onderzocht.
- Dynamisch risico is zo individueel bepaald dat generaliserende onderzoeksmethodes niet geschikt zijn om het aan te tonen. In dat geval zijn casestudie methodes en verdere theorievorming noodzakelijk.
- Dynamische risicofactoren moeten op andere wijze gemeten worden dan in dit onderzoek is gebeurd. 'Impressionistische' beoordelingen zoals in het huidige onderzoek zijn te vatbaar voor vertekening. Uitvoerigere instructies en training, multipele beoordelaars, meer gebruik van standaard meetinstrumenten voor dynamische concepten, of gebruik van fysiologische maten (bijvoorbeeld, 'leugendetector', penisplethysmograaf, MRI scans) zijn denkbare alternatieven.
- Dynamische risicofactoren moeten op een andere wijze gecombineerd worden dan in dit (en veel ander) onderzoek is gebeurd. Een lineaire benadering moet wellicht plaatsmaken voor een benadering waarbij specifieke combinaties van risicofactoren worden toegepast voor specifieke categorieën pati;enten.

***Betekenis van resultaten vanuit klinisch perspectief.*** Vanuit klinisch perspectief leidt het onderzoek tot de volgende inzichten:

- Kinische indrukken van dynamische patiëntkenmerken, zoals die in de TBS dagelijks worden gebruikt bij communicatie over delictrisico, zijn niet voorspellend voor feitelijke recidive.
- De resultaten wijzen erop dat zelfs ervaren clinici, op basis van klinische indrukken alleen, vermoedelijk niet in staat zijn feitelijke vooruitgang bij de patiënt van schijnaanapassing te onderscheiden, Het is daarom raadzaam voor clinici om bij risicotaxatie hun subjectieve inschattingen en intuïties ondergeschikt te maken aan bevindingen uit gestructureerde, gestandaardiseerde risicotaxatiemethodes en instrumenten als de PCL-r.
- De bevindingen roepen vragen op over de veranderbaarheid van risico. Interventies moeten zich mogelijk veel meer richten op beheersing van de context (*control* in plaats van *cure*).
- Clinici zouden niet te snel moeten veronderstellen dat zij het onderling eens zijn over de betekenis van dagelijks gebruikte klinische

termen: de bevindingen laten zien dat de interpretatie daarvan tamelijk sterk uiteen kan lopen.

Er is beter maatwerk nodig van onderzoekers om instrumenten in de klinische werkelijkheid in te passen, en clinici dienen in toenemende mate bereid te zijn om gestandaardiseerde procedures op correcte wijze in hun dagelijks werk te gebruiken.

Vanuit ethische optiek moet men zich tot slot afvragen of inschattingen over zeer lange termijn wel te verantwoorden zijn, gezien de gebreken in bestaande risicotaxatiemethodes. Een maximale voorspellingstermijn van 5 jaar lijkt redelijk, al heeft zo'n afgrenzing alleen zin wanneer het instrument dat gebruikt wordt ook dynamische voorspellers bevat (een statisch instrument levert immers ook 5 jaar later weer dezelfde inschatting op).

## III.4 Naar de toekomst – enkele slotopmerkingen

De vooruitgang in risicotaxatie-onderzoek lijkt de laatste jaren wat te stagneren. Nieuw is wel de aandacht voor beschermende factoren, kenmerken die de kans op recidive verkleinen. Ook is het toenemende accent op risicobeheersing interessant: risicotaxatie heeft immers alleen zin wanneer daaraan risicobeperkende interventies kunnen worden verbonden. Zonder kennis over dynamische risicofactoren is betekenisvolle risicobeheersing echter niet mogelijk (afgezien van langdurige opsluiting).

De hedendaagse westerse maatschappij is geobsedeerd door veiligheid en stelt onrealistisch hoge eisen aan risicotaxatieprocedures. In het veld van delictrisicotaxatie daagt echter juist het inzicht dat de werkelijkheid complex en niet-lineair is: de toekomst is inherent onvoorspelbaar. Er zijn grenzen aan voorspellingsmogelijkheden, en het is niet uitgesloten dat die grenzen wat betreft de forensisch psychiatrische delictrisicotaxatie al bereikt zijn. Hoe om te gaan met de dan overblijvende onzekerheidsmarge is geen wetenschappelijke, maar een morele, juridische, en politieke vraag.

# Summary

This thesis is comprised of three parts. The first part is a general introduction that addresses core concepts and the state of the art in the area of crime risk assessment. The second part contains three empirical studies into risk assessment within the judicial context of *terbeschikkingstelling* (TBS), conducted in 8 TBS-hospitals between 1996 and 2004. The main findings and their possible interpretations are set forth in part three.

***TBS***. The judicial measure of *terbeschikkingstelling* (TBS) can be imposed on perpetrators of serious offences, if they are not held fully accountable for their acts due to a mental disorder. Additionally, there must be a prognosis of high reoffending risk. The TBS-patient is admitted to a special hospital, initially for the duration of two years. He is nursed and treated until the risk of reoffending is sufficiently reduced. At least once every two years the hospital reports to the court on treatment progress and reoffending risk. On the basis of this input, among other things, the judge decides whether the TBS-measure should be extended or terminated.

## I General introduction
### *Risk assessment: core concepts and state of the art*

### I.1 Introduction
Risk assessment, making prognoses about the likelihood that patients will commit new offences in the future, is both a routine and essential part of clinical practice in forensic psychiatry. The last three decades have yielded ample research regarding this subject. Opinions continue to differ as to which approach to risk assessment is preferable, and what levels of accuracy are at best attainable.

### I.2 Core concepts
Initially, risk assessment research was dominated by the notion of 'dangerousness', which was primarily viewed as an attribute of the person. Eventually, the likelihood of an offence came to be seen as the result of an interaction between the person, his context, and situation. Simultaneously, there was a shift in accent away from predicting criminal recidivism, towards establishing the relative risk of reoffending, similar to the way

weather forecasts are made. This approach reflects the inevitable margin of error that is part of any prognosis.

In clinical practice, however, relative risk statements are the basis for absolute decisions. This can entail two types of error: either a patient can be seen as high risk while he isn't, resulting in unnecessarily long detention (false-positive prediction); or a patient can be incorrectly designated low risk, resulting in premature discharge and renewed offending (false-negative prediction).

Several factors influence the measurements of the extent of these errors. First, the time that patients are monitored after discharge affects the error rates: the longer the follow-up, the more reoffenders there will be. Next, the choice of outcome measure has an effect as well. If, for instance, any form of violent behavior is counted as relapse, this will result in a higher number of recidivists than an outcome measure that only includes reconvictions. Furthermore, the most commonly used sources of reoffending data, official judicial and police files, will yield less reoffenders than self-report data from the individuals under study, because not all crime comes to the attention of the police. This problem is referred to as the 'dark number', a group of (re)offenders not perceived by researchers. It is often considered to be one of the most serious impediments to risk assessment research. However, a recent study has demonstrated that official files and self-report result in nearly identical statistical predictor sets. Therefore, the dark number problem may not be quite as crippling to research as is sometime assumed.

A fourth influence on the accuracy of risk assessments is the 'base rate' of reoffending. This is the proportion of persons in a given population who recidivate within a given time span. At very high or very low base rates, the prognosis that, respectively, everybody or nobody will recidivate will be already highly accurate, and leave little room for improvement through the use of special risk assessment procedures. At a base rate of 50%, the potential added value of such procedures is maximal.

Two final core concepts are static and dynamic risk factors. Factors that relate to reoffending risk are often distinguished in unchangeable, historic givens (for instance, age at first conviction), and changeable characteristics of the person and his environment (for instance, mood). The boundary between these two categories is indistinct. Changeable characteristics are clinically interesting, but characteristics that fluctuate strongly over short time spans have less chance to be predictive in the long run than more stable features.

## I.3 Putting risk assessment into practice: three approaches

In practice, risk assessment can be conducted, roughly speaking, in three different ways. The most traditional of these is the unstructured clinical approach, which is characterized by the strong subjective influence of the assessor. The content of such assessments is not guided by any standardized instrument or procedure. A large number of empirical studies have shown this approach to be unreliable and inaccurate at worst, and at best to yield assessments of which the quality can be equaled by different, far simpler means.

This simpler alternative is the so-called actuarial risk assessment, which completely discards the subjective element. Prognoses are established using a fixed set of predictors that have been shown to correlate with subsequent violence in empirical studies. The method for assessing these predictors, and the algorithm converting predictor scores into a final risk judgment, are also strictly defined. Completing such instruments does usually not require specialist knowledge, which some authors even consider undesirable.

Actuarial risk assessment yields moderate to good results in research. An important disadvantage of these instruments is their predominantly static content, that does not offer clues for treatment: a risk level, once established, will never decrease afterwards.

The strongly polarized stalemate between proponents of the clinical and those of the actuarial approach was breached by the appearance of the third alternative, the structured clinical approach. This method adopts the standardized itemlist and empirical base of actuarial tools. It adds, however, a number of clinically relevant risk factors, and allows the assessor to formulate a final risk judgment that diverges from the numerical score, if there is good reason for this. Prediction research with this category of instruments shows variable, but generally fair, results, More research is needed.

## I.4 Risk assessesement reality

By and large, the evaluation of the quality of risk assessment tools has thus far been a laboratory activity. As yet very little is known about the actual effectiveness of such instruments in clinical practice, let alone of their effect on reoffending rates. The use of instruments in clinical contexts will rarely answer to the high standards common in research studies, and shrinkage of their reliability and predictive validity is to be expected.

The effect of risk assessment on reoffending rates also depends on the way risk is reported. The importance of reporting standards is, for instance, underlined by research demonstrating that decision-making by recipients of risk reports (such as judges) is influenced by the use of either absolute ('1 in 10') or relative ('10%' ) formats to describe the likelihood of relapse. Absolute formats results in higher perceived risk than their relative equivalents. Uniform guidelines for reporting risk are therefore much needed, yet generally lacking.

## I.5 Conclusion

The science of risk assessment has seen considerable progress during the last three decades, resulting in guarded optimism with regard to prediction possibilities. Yet, many questions remain unanswered, and the estrangement of risk assessment research from clinical practice is an important problem.

# II Three empirical studies

## II.1 Study I. Structural coherence of clinically derived dynamic indicators of reoffending risk

This study investigated whether mental or behavioral characteristics of patients, considered tob e pivotal in risk assessment by clinicians in TBS, represent meaningful higher order concepts. Also, it was investigated to what extent risk assessments by TBS clinicians were indeed related to these characteristics.

***Method***. Clinicians in 4 TBS-hospitals were asked which dynamic patient characteristics they considered to be essential for risk assessment. The results were reviewed in 8 hospitals and subsequently edited, after which a 47-item instrument resulted: the Clinical Inventory of Reoffending Risk Indicators (CIDRRI). A treatment supervisor is asked to rate on a 6-point scale to what extent each characteristic applies to a particular patient. Furthermore, he rates the general perceived risk of reoffending on a 48th item.

An explorative principal axis factor analysis with Varimax rotation was conducted on these items. Subgroups of recently discharged and recently admitted patients were also analyzed separately to establish whether the factor solution depended on treatment stage. Finally, the extent to which CIDRRI itemscores and scales predicted the clinical risk assessment was tested using Receiver Operating Characteristic (ROC) analysis.

***Results***. 370 CIDRRI's were available for analysis. Six factors were established. They were:

- Empathic acceptance of responsibility for the offence (11 items)
- Lack of self-reliance (12 items)
- Antisocial narcissism (10 items)
- Treatment compliance (5 items)
- Attainment of treatment goals (7 items)
- Avoidance (2 items)

Factors were converted to scales by summing rough item scores. The partial factor model for recently admitted patients was similar to the main model. The solution for recently discharged patients contained only 5 factors, due to convergence of the factors 'treatment compliance' and 'attainment of treatment goals'.

Most CIDRRI items were significantly related to clinical risk assessment. as were all scales, with the exception of the 'avoidance' scale. Links between items and scales on the one hand and the clinical risk assessment on the other, were different for recently admitted and recently discharged patients. In the latter group the accent was on offence responsibility, and attainment of treatment goals, while among recently admitted patients lack of self-reliance was the strongest predictor of the clinical risk assessment.

***Discussion.*** Dynamic patient characteristics that clinicians in TBS consider to be important for risk assessment, appeared to represent 6 meaningful higher order concepts. Both these concepts and underlying items were indeed demonstrably related to clinical risk assessments. Though this does not imply anything about their links to actual reoffending, such insights into clinical ways of thinking are in themselves useful.

## II.2 Study II. Reliability and discriminant validity of dynamic reoffending risk indicators in forensic clinical practice

This study investigated whether an easy to use instrument, the CIDRRI, applied under routine clinical circumstances, yielded acceptable measures of reliability and discriminant validity.

***Method***. An interrater reliability study (A), a retest reliability study (B) and a discriminant validity study (C) were conducted. In Study A, pairs of CIDRRI's were scored simultaneously for the same patients by 2 raters independently of each other, for a total of 75 patients. In Study B, 2 CIDRRI's were rated twice by the same rater over a one-month period for a total of 29 patients. Study C tested whether 118 newly admitted patients could be distinguished from 115 recently discharged patients through *k*-means cluster analysis of CIDRRI ratings, on the assumption that admittees would generally represent higher risks than dischargees.

***Results***. Study A showed most items to meet minimum interrater reliability requirements. Interrater reliability of scales was fair. Study B showed retest reliability to be good for most items and all scales. Study C demonstrated that dischargees and admittees could be distinguished from each other with high accuracy using CIDRRI scale scores as well as the clinical assessment of risk.

***Discussion***. CIDRRI reliability findings are similar to those reported for other, generally accepted risk assessment tools. This result is satisfying considering the fact that, in contrast to most risk assessment research, no special measures (such as training, instructions, manual) were included to enhance reliability; and also considering the fact that the CIDRRI utilises 6-point rating scales rather than the more common 3-point scales, thus allowing more room for interrater divergence.

The instrument was shown to have substantial power to discriminate newly admitted from recently discharged patients. This ability does, however, not necessarily indicate discriminatory power regarding reoffending risk: it may also merely reflect biased clinical perceptions with regard to these patient groups. CIDRRI validation using actual relapse data will show which of these scenarios is more likely the case.

## II.3 Study III. Static and dynamic patient characteristics as predictors of criminal recidivism

In this study, predictive validity of dynamic risk indicators from the CIDRRI was tested with regard to reconviction after discharge, while taking into account an initial set of static predictors.

***Method.*** In the years 1996 through 1998, CIDRRI and static file data were collected for all patients discharged from any of 7 hospitals either on probationary leave or due to termination of the TBS measure. After a follow-up of at least 5.5 and at most 8.5 years, data were linked to reoffending data from national criminal justice files. Recidivism was defined as any reconviction for a violent or sexual offence (including attempts). A prediction model was derived using survival analysis, which takes into account the variable times at risk among subjects. Several preselections were applied to the independent variables to reduce chance capitalization effects.

***Results***. The dataset comprised 132 subjects, or nearly three quarters of all TBS-patients who met our discharge criteria during the years of datacollection. Most of them were committed to TBS due to a violent or sex offence. Over 75% of subjects was diagnosed with at least one personality disorder at admittance, mostly of the antisocial, narcissistic and unspecified type. Diagnoses on axis 1 of DSM-III-r were also common, notably of psychosis and substance related disorders. The median follow-up was 6.8 years. Of dischargees, 19.7% was again convicted for a violent or sex offence.

The clinical risk assessment at the end of the CIDRRI was unrelated to actual reconviction. Some dynamic items however did have univariate predictive validity. In all cases, the direction of theit relation to reconviction ran counter to expectations (for instance, patients rated as more empathic were reconvicted more often). Multivariate analysis of static risk factors yielded a 4-item prediction model. CIDRRI-scales and items did not add any further predictive validity to this, so that the final prediction model was fully static, and comprised the following predictors:

- Number of times absent without leave during TBS (more = higher risk)
- Comorbiity of any personality disorder with a substance use disorder, diagnosed at admittance (yes = higher risk)
- Psychosis diagnosed at admittance (yes = lower risk)
- Any personality disorder from the DSM-III-r B-cluster, diagnosed at admittance (yes = higher risk).

This model yielded a fair prognosis of actual reconviction, with an ROC area under the curve of .79. Additional analyses regarding patient subgroups representing different offence types or disorders, and with equal follow-up times for all subjects, did not add new insights.

***Discussion***. This study did not yield any starting points for dynamic risk assessment in TBS. Dynamic risk factors as comonly used in TBS, did not predict actual reconviction. This may to some extent be explained by discrepancies between rater perceptions of the patient's condition and the patient's actual condition. To prevent this kind of distortion, standardized tools, multiple rater procedures and psychofysiological measures are recommended when operationalizing potential dynamic risk factors.

# III General conclusions and discussion

## III.1 Strengths and weaknesses in the current risk assessment knowledge base

Risk assessment research has made important strides in the last few decades, notably through increased attention to context influences, a dramatic increase in the number of research studies, and improvements in research methods. On the other hand, empirical research has been little concerned with practical clinical needs. Thus far, the benefits of structured instruments seem to reside mainly in their attention to historical predictors, and the use of a uniform structure.

## III.2 The empirical studies

The empirical studies, conducted in 8 TBS-hospitals, showed that dynamic risk indicators as used by clinicians in TBS, represent a meaningful higher order structure. These indicators could be rated with an acceptable level of reliability, and were able to distinguish newly admitted from recently discharged patients. Dynamic risk indicators did however not appear to have any added value on top of a four-item static prediction model. The static final prediction model that was derived from the data had fairly strong predictive power.

## III.3 Discussion

The findings are not encouraging with regard to dynamic predictors of reoffending risk in TBS. Clinical assumptions about such predictors could not be empirically confirmed. The instrument that was developed had moderate reliability and was clearly related to clinical assessments of risk, but proved to be fully unrelated to actual reconviction. First of all, it needs to be asked to what extent these findings may be explained by shortcomings in the research design. Though the national scope, the close adherence to clinical practice, and the prospective design provided a strong basis for the research, some weaknesses are also in evidence. Secondly, clinical and research implications of the findings need to be considered.

***Weaknesses of the research design***. The less than ideal reliability of the CIDRRI may have resulted in an underestimation of the predictive power of dynamic risk indicators. Furthermore, it is unclear whether CIDRRI items actually measured patient characteristics, or merely the clinical perception thereof. The reversed relations between some dynamic variables and outcome give rise to the suspicion that CIDRRI ratings may have represented distorted views of reality. Possibly the patient was putting up a superficial show of good behavior; and possibly the clinician was guided too much by what he wanted to see. Psychopathy may have been a factor in this process: this disorder is typified by wantonly manipulating and superficially conformist behavior in persons who in fact pose a high risk.

Furthermore, the follow-up study only included patients who were discharged from the hospital. As a consequence, true positive predictions (patients who were rightly detained) may have been underrepresented. This kind of bias is however not very likely given the fact that there was a considerable number of persons rated 'high risk' among dischargees. Also, termination of TBS against hospital advice did not result in a significantly larger proportion of reoffenders than other modes of discharge.

The limited size of the dataset did not allow study of subsets of variables as predictors within specific subgroups of patients. The uniform, linear approach is a shortcoming of the present research. Also, the long follow-up may have worked against dynamic predictors, though additional analyses with considerably shorter follow-ups did not yield support for this assumption. Other variants of the relapse criterion (for instance, number of recorded incidents, or repeal of probationary leave) did not lead to different overall findings either.

Finally, one may wonder whether data collected between 1996 and 1998 have any bearing on TBS in 2005. However, phrases and concepts such as laid down in the CIDRRI are as much part and parcel of daily reports in TBS now as they were then. Moreover, these same concepts are often part of item definitions in standardized international risk assessment tools like the HCR-20. Therefore, findings seem of undiminished relevance to present-day TBS practice.

***Implications of results from a research perspective***. With regard to research, findings carry the following possible implications:

- Risk is static. This far-reaching conclusion should only be drawn when the alternative conclusions listed below have been thoroughly investigated without yielding satisfactory results.
- Dynamic risk is individually determined to such an extent that nomothetic research methods are unsuitable to explain it. In this event case studies and the development of a theory of reoffending are needed.
- Dynamic risk factors should be measured by other means than those used in the present research. 'Impressionistic' methods like the

CIDRRI may be overly susceptible to distortions. Extensive training and instruction, the use of standardized diagnostic tools for measuring dynamic concepts, multiple rater procedures, and the use of physiological measures (such as 'lie detectors' or phallometry) are possible alternatives.

- Dynamic risk factors need to be combined differently than was done in this (and much other) research. A linear approach may need to be discarded in favour of a methods that apply specific sets of predictors for specific groups of patients.

***Implications of findings from a clinical perspective***. From a clinical viewpoint, the present studies provide the following insights:

- Clinical impressions of dynamic patient characteristics, such as are used daily in communications about patients in TBS, do not have any predictive validity with regard to actual reoffending.
- Results suggest that even experienced clinicians, when guided by clinical impressions alone, are unable to distinguish actual progress in the patient from simulated adaptation. Clinicians should therefore be advised to favour results from structured, standardized risk assessment tools and instruments like the PCL-r, over their own subjective assessments and intuitions.
- As far as dynamic risk factors in the present research were truly representative of patient characteristics, findings cast some doubt on the changeability of risk. Possibly, interventions need to focus on control rather than cure.
- Clinicians should be careful to assume too easily that they are in agreement regarding the meaning of clinical jargon: findings show that commonly used terms may be interpreted differently even by clinicians who work closely together.

Improvement of prognostic results requires improved tailoring of instruments by researchers to fit clinical reality, and increased preparedness of clinicians to correctly apply standardized methods in their daily work.

From an ethical perspective, finally, one must wonder whether very long term risk assessments are justifiable, given the obvious shortcomings of extant risk assessment procedures. A maximum prediction 'horizon' of 5 years seems reasonable, though such a delineation is only useful if an instrument contains at least some dynamic predictors (for, a static instrument would merely return the same risk level when reassessed 5 years later).

## III.4 Towards the future – some final remarks

Developments in risk assessment research seem to stall somewhat in recent years. Yet, a new impulse is provided by the increased attention to protective factors, that reduce rather than increase offending risk. Also of interest is the increased stress on risk management: for, risk assessment is only useful when it is linked to risk reducing interventions. However, without knowledge of dynamic risk factors meaningful risk management is impossible (apart from long term incarceration).

Contemporary Western society is obsessed with safety and puts unrealistically high demands on risk assessment procedures. Meanwhile, the risk assessment research field is starting to open up to the fact that reality is complex and non-linear and therefore inherently unpredictable. Prediction possibilities are limited, and we cannot be sure whether these limits haven't already been reached as far as reoffending risk assessment is concerned. How to deal with the remaining margin of uncertainty is a moral, judicial and political rather than a scientific question.

# ℭ References

American Psychiatric Association (1980) *Diagnostic and Statistical Manual of Mental Disorders. Third edition. (DSM-III).* Washington DC: APA.

American Psychiatric Association. (1987). *Diagnostic and Statistical Manual of Mental Disorders Third edition – revised (DSM-III-r).* Washington DC: APA.

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders. Fourth Edition. (DSM-IV).* Washington DC: APA.

Andreasen, N.C. (2000). Prediction in clinical psychiatry: What does the future hold? *American Journal of Psychiatry, 157* (9), 1373-1374.

Augimeri, L. K., Webster, C.D., Koegl, C.J., & Levene, K. (2001). *Early Assessment Risk List for Boys: Earl-20B, Version 2.* Toronto: Earlscourt Child and Family Centre.

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

Barratt, E.S. (1994). Impulsiveness and aggression. In: J. Monahan & H.J. Steadman (Eds.). *Violence and mental disorder. Developments in risk assessment* (pp. 61-79). Chicago: The University of Chicago Press.

Bartko, J.J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, 19, 3-11.

Beek, D. van. (1999). *De delictscenarioprocedure bij seksueel agressieve delinquenten.* [The offence script procedure in sexually violent offenders]. Doctoral thesis, University of Amsterdam. Arnhem: Gouda Quint.

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.

Berlin, F.S., Galbreath, N.W., Geary, B., & McGlone, G. (2003). The use of actuarials at civil commitment hearings to predict the likelihood of future sexual violence. *Sexual Abuse: A Journal of Research and Treatment, 15* (4), 377-382.

Boer, D.P., Hart, S.D., Kropp, P.R., & Webster, C.D. (1995). *Manual for the Sexual Violence Risk-20: Professional guidelines for assessing risk of sexual violence.* Vancouver: Simon Fraser University.

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

Bonta, J. (1997). Do we need theory for offender risk assessment? *Forum on Corrections Research, 9* (1), 42-45.

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

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

Borum, R., Bartel, P.A.. & Forth, A. (2002). *Structured Assessment of Violence Risk in Youth (SAVRY).* San Diego: Specialized Training Services.

Boutellier, H. (2002). *De veiligheidsutopie. Hedendaags onbehagen en verlangen rond misdaad en straf.* [Safety Utopia. Contemporary discomfort and desire around crime and punishment]. Den Haag: Boom.

Bryant, F.B. & Yarnold, P.R. (1995). Principal components analysis and exploratory and confirmatory factor analysis. In: Grimm, L. G., & P.R. Yarnold (Eds.), *Reading and understanding multivariate analysis* (pp. 99-136). Washington, DC: American Psychological Association.

Buchanan, A. (1999). Risk and dangerousness. *Psychological Medicine, 29*, 465-473.

Buss, A.H., & Perry, M. (1992). The Agression Questionnaire. *Journal of Personality and Social Psychology, 63*, 452-459.

Canton, W., Veer, T.S. van der, Panhuis, P. van, Verheul, R., & Brink, W. van den. (2004). De voorspellende waarde van risicotaxatie in de pro justitia rapportage: onderzoek naar de HKT-30 en de klinische inschatting. [The predictive validity of risk assessments in 'pro justitia' reports: a study of the HKT-30 and clinical assessment.] *Tijdschrift voor Psychiatrie, 8*, 525-535.

Castel, R. (1991). From dangerousness to risk. In: G. Burchell, C. Gordon & P. Miller (eds.). *The Foucault effect. Studies in governmentality* (pp. 281-298). Hemel Hempstead: Harvester Wheatsheaf.

Cicchetti, D.V. & Sparrow, S.S. (1981). Developing criteria for establishing the interrater reliability of specific items in a given inventory. *American Journal of Mental Deficiency, 86*, 127-137.

Cocozza, J.J., & Steadman, H.J. (1976). The failure of psychiatric prediction of dangerousness: Clear and convincing evidence. *Rutgers Law Review, 29*, 1048-1101.

Cooke, D., Michie, C. and Ryan, J. (2001). *Evaluating risk for violence: A preliminary study of the HCR-20, PCL-R and VRAG in a Scottish prison sample (Occasional Paper 5/2001).* Edinburgh: Scottish Prison Service.

Côté, G. (2001, April). *Violent behaviour, PCL-R and HCR-20 among involuntary inpatients, forensic patients and severely mentally disordered inmates.* Paper presented at the First Annual Meeting of the International Association of Forensic Mental Health Services, Vancouver, British Columbia, Canada.

Coulson, G., Ilacqua, G., Nutbrown, V., Giulekas, D., & Cudjoe, F. (1996). Predictive utility of the LSI for incarcerated female offenders. *Criminal Justice and Behavior, 23* (3), 427-439.

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

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology, 66*, 819-829.

Dernevik, M. (2004). *Structured clinical assessments and management of risk of violent recidivism in mentally disordered offenders.* Doctoral thesis. Stockholm: Karolinska Insitutet.

Dolan, M., & Doyle, M. (2000) Violence risk prediction. Clinical and actuarial measures and the role of the Psychopathy Checklist. *The British Journal of Psychiatry, 177*, 303-311.

Doren, D.M. (1999, September). *Comprehensive view of sex offender risk assessment instruments.* Paper presented at the 18th Annual Research and Treatment Conference of the Association for the Treatment of Sexual Abusers, Orlando, Florida.

Douglas, K.S. (2001, November). *The reliability and validity of structured professional judgments using the HCR-20 violence risk assessment scheme.* Paper presented at the International Conference 'Violence, Risk Asssessment and management: Bringing Science and Practice closer together'. Sundsvall, Sweden.

Douglas, K.S., & Ogloff, J.R.P. (2003). Multiple facets of risk for violence: The impact of judgmental specificity on structured decisions about violence risk. *International Journal of Forensic Mental Health, 2* (1), 19-34.

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

Douglas, K.S., Webster, C.D., Hart, S.D., Eaves, D., & Ogloff, J.R.P. (2001). *HCR-20: Violence risk management companion guide.* Vancouver: Simon Fraser University, Mental Health, Law & Policy Institute.

Douglas, K. S., Boer, D. P., & Yeomans, M. (in press). Comparative validity analysis of multiple measures of violence risk in a general population sample of criminal offenders. *Criminal Justice and Behavior.*

Driesschner, K. (2002, July). *Presentation of a new questionnaire for the treatment motivation of patients in court-mandated outpatient treatment.* Paper presented at the 27th Annual Conference of the International Academy of Law and Mental Health, Amsterdam, The Netherlands.

D'Silva, K., Duggan, C., & McCarthy, L. (2004). Does treatment really make psychopaths worse? A review of the evidence. *Journal of Personality Disorders, 18* (2), 163-177.

Elbogen, E.B., Calkins Mercado, C., Scalora, M.J., & Tomkins, A.J. (2002). Perceived relevance of factors for violence risk assessment: A survey of clinicians. International *Journal of Forensic Mental Health Services, 1* (1), 37-46.

Emmerik, J.L. van. (1985). *TBS en recidive. Een beschrijving van ter beschikking gestelden van wie de maatregel is beëindigd in de periode 1974-1979.* [TBS and recidivism. A description of TBS patients whose treatmen measure was terminated in the years 1974 through 1979.] Den Haag: Staatsuitgeverij.

Emmerik, J.L. van, & Diks, G.J.M. (1999). *De terbeschikkingstelling in maat en getal.* [TBS in facts and figures]. Utrecht: Dr. F.S. Meijers Instituut, Afdeling Monitoring & Research.

Ewald, F. (1981). Insurance and risk. In: G. Burchell, C. Gordon & P. Miller (eds.). *The Foucault effect. Studies in governmentality* (pp. 197-210). Hemel Hempstead: Harvester Wheatsheaf.

Farrington, D.P. (2001). *What has been learned from self-reports about criminal careers and the causes of offending? (Report for the Home Office).* London: Home Office.

Fitzpatrick, K.M. (1997). Fighting among America's youth: A risk and protective factors approach. *Journal of Health and Social Behavior, 38*, 131-148.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions (2nd ed.).* New York: John Wiley & Sons.

Ford, J. (1989). What is chaos, that we should be mindful of it? In: P.C.W. Davies, (Ed.), *The new physics* (pp. 348-372). Cambridge: Cambridge University Press.

Gagliardi, G.J., Lovell, D., Peterson, P.D., & Jemelka, R. (2004). Forecasting recidivism in mentally ill offenders released from prison. *Law and Human Behavior, 28* (2), 133-155.

Garb, H.N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin, 105* (3), 387-396.

Gardner, W., Lidz, C.D., Mulvey, E.P., & Shaw, E.C. (1996). A comparison of actuarial methods for identifying repetitively violent patients with mental illness. *Law and Human Behavior, 20* (1), 35-48.

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

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

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

Grann, M., Långström, N., Tengström, A., & Kullgren, G. (1999). Psychopathy (PCL-R) predicts violent recidivism among criminal offenders with personality disorders in Sweden. *Law and Human Behavior, 23* (2), 205-217.

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.

Greenland, C. (1985). Dangerousness, mental disorder, and politics. In: C.D. Webster, M.H. Ben-Aron, & S.J. Hucker (Eds.), *Dangerousness. Probability & prediction, psychiatry & public policy* (pp. 25-40). Cambridge: Cambridge University Press.

Greeven, P.G.J. (1997). *De intramurale behandeling van forensische patienten met een persoonlijkheidsstoornis. Een empirische studie*. [The in-patient treatment of forensic patients with personality disorder. An empirical study]. Doctoral thesis, University of Utrecht. Deventer: Gouda Quint.

Groen, H., & Brink, W. van den. (1992). De klinische diagnostiek van persoonlijkheidspathologie. Een multi-conceptuele, multi-instrumentele benadering. [Clinical assessment of personality pathology. A multi-conceptual, multi-instrumental approach]. *Tijdschrift voor Psychiatrie, 34* (3), 170-184.

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

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

Grubin, D. (1999). Actuarial and clinical assessment of risk in sex offenders. *Journal of Interpersonal Violence, 14* (3), 331-343.

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

Halpern, E.J., Alpert, M., Krieger, A.M., Metz, C.E., & Maidment, A.D. (1996). Comparisons of ROC curves on the basis of optimal operating points. *Academic Radiology, 3*, 245-253.

Hanley, J.A., & McNiel, B.J. (1982). The meaning and use of the area under the Receiver Operating Characteristic (ROC) curve. *Radiology, 143* (1), 29-36.

Hanson, R.K. (1997). *The development of a brief actuarial scale for sexual offense recidivism. (User report 1997-4).* Ottawa: Department of the Solicitor General of Canada.

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

Hanson, R.K., & Harris, A.J.R. (1998). *Dynamic predictors of sexual recidivism (Report 1998-1).* Ottawa, Ontario: Department of the Solicitor General of Canada.

Hanson, R.K., & Harris. A.J.R. (2000). Where should we intervene? Dynamic predictors of sexual offense recidivism. *Criminal Justice and Behavior, 27* (1), 6-35.

Hanson, R.K., & Harris, A.J.R. (2001). A structured approach to evaluating change among sexual offenders. *Sexual Abuse. A Journal of Research and Treatment, 13* (2), 105-122.

Hanson, R.K., & Thornton, D. (1999). *Static-99: Improving actuarial risk assessments for sex offenders. (Report 1999-2).* Ottawa, Ontario: Department of the Solicitor General of Canada.

Hanson, R.K., & Thornton, D. (2000). Improving risk assessment for sex offenders: A comparison of three actuarial scales. *Law and Human Behavior, 24* (1), 119-136.

Hare, R.D. (1991). *The Psychopathy Checklist-Revised. (PCL-R).* Toronto: Multi-Health systems.

Harris, G. (2003). Men in his category have a 50% likelihood, but which half is he in? Comments on Berlin, Galbreath, Geary and McGlone. *Sexual Abuse: A Journal of Research and Treatment, 15* (4), 389-392.

Harris, G.T., & Rice, M. (2003). Actuarial assessment of risk among sex offenders. *Annals of the New York Academy of Sciences, 989*, 198-210.

Hart, S.D. (1998). The role of psychopathy in assessing risk for violence: Conceptual and methodological issues. *Legal and Criminological Psychology, 3*, 121-137.

Hart, S.D. (1999). Assessing violence risk: thoughts and second thoughts. *Contemporary Psychology, 44* (6), 486-487.

Hart, S.D. (2001, June). *Complexity, uncertainty, and the reconceptualization of violence risk.* Paper presented at the 11th Annual Conference of the European Association of Psychology and Law, Lisbon, Portugal.

Hart, S.D. (2003). Actuarial risk assessment: Commentary on Berlin et al. *Sexual Abuse: A Journal of Research and Treatment, 15* (4), 383-388.

Hart, S.D., Cox, D.N., & Hare, R.D. (1995). *The Hare PCL:SV. Psychopathy Checklist: Screening version.* Toronto: Multi-Health Systems.

Heilbrun, K. (1997). Prediction versus management models relevant to risk assessment: The importance of legel decision-making context. *Law and Human Behavior, 21* (4), 347-359.

Heilbrun, K., Dvoskin, J., Hart, S.D., & McNiel, D. (1999). Violence risk communication: Implications for research policy and practice. *Health, Risk and Society, 1* (1), 91-106.

Hemphill, J.F., & Hare, R.D. (2004). Some misconceptions about the Hare PCL-R and risk assessment: A reply to Gendreau, Goggin and Smith. *Criminal Justice and Behavior, 31* (2), 203-243.

Hildebrand, M. (2004). *Psychopathy in the treatment of forensic psychiatric patients. Assessment, prevalence, predictive validity and clinical implications.* Doctoral thesis. Amsterdam: University of Amsterdam.

Hildebrand, M., Ruiter, C. de, & Nijman, H. (2004). PCL-R psychopathy predicts disruptive behavior among male offenders in a Dutch forensic psychiatric hospital. *Journal of Interpersonal Violence, 19*, 13-29.

Hildebrand, M., Ruiter, C. de, & Vogel, V. de (2004). Psychopathy and sexual deviance in treated rapists: Association with sexual and non-sexual recidivism. *Sexual Abuse. A Journal of Research and Treatment, 16*, 1-24.

Hilterman, E.L.B. (2002). Van kwaad tot erger: De ernst van recidive door ter beschikking gestelden tijdens verlof nader bekeken. [From bad to worse: A closer examination of seriousness of reoffending during leave by forensic psychiatric in-patients]. *Tijdschrift voor Criminologie, 44* (1), 61-80.

Hilterman, E.L.B., & Gresnigt, J. (2003). Het onderbuikgevoel en risicotaxatie in de forensische psychiatrie: van tweewieler naar multi-purpose vehicle. [Intuition and risk assessment in forensic psychiatry: From bicycle to multi-purpose vehicle]. In: H. Groen & M. Drost (Eds.), *Handboek forensische geestelijke gezondheidszorg*, (pp. 319-332). Utrecht: De Tijdstroom.

Howells, K. (1998). Cognitive behavioural therapy for anger, aggression and violence. In N. Tarrier (Ed.), *Cognitive Behavioural Therapy for Complex Cases*. Chichester: Wiley.

Kiehl, K.A., Smith, A.M., Hare, R.D., Mendrek, A., Forster, B.B., Brink, J., & Liddle, P.F. (2001). Limbic abnormalities in affective processing by criminal psychopaths as revealed by functional magnetic resonance imaging. *Biological Psychiatry, 50*, 677-684.

Kozol, H.L., Boucher, R.J., & Garofalo, R.F. (1972). The diagnosis and treatment of dangerousness. *Crime and Delinquency, 18* (4), 371-392.

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.

Kropp, R.D., Hart, S.D., Webster, C.D., & Eaves, D. (1995). *Spousal Assault Risk Assessment Guide (SARA).* Vancouver: British Columbia Institute Against Family Violence

Leuw, E. (1999). Recidive na de TBS. *Patronen, trends en processen en de inschatting van gevaar.* [Reoffending after TBS. Patterns, trends and process and the assessment of danger]. Den Haag: Ministerie van Justitie, WODC.

Lipsey, M.W. & Wilson, D.B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48* (12), 1181-1209.

Litwack, T.R. (2001). Actuarial versus clinical assessments of dangerousness. *Psychology, Public Policy, and Law, 7* (2), 409-443.

Litwack, T.R. (2002). Some questions for the field of violence risk assessment and forensic mental health: Or, "Back to basics" revisited. *International Journal of Forensic Mental Health, 1* (2), 171-178.

Lösel, F., & Bender, D. (2003). Protective factors and resilience. In: D.P. Farrington & J.W. Coid (Eds.), *Early prevention of adult antisocial behaviour* (pp. 130-204). Cambridge: Cambridge University Press.

Loza, W., Villeneuve, D.B., & Loza-Fanous, A. (2002). Predictive validity of the Violence Risk Appraisal Guide: A tool for assessing violent offender's recidivism. International *Journal of Law and Psychiatry, 25* (1), 85-92.

Margison, F.R., Barkham, M., Evans, C., McGrath, G., Mellor Clark, J., Audin, K., & Connell, J. (2000) Measurement and psychotherapy. Evidence-based practice and practice-based evidence. *The British Journal of Psychiatry, 177*, 123-130.

Marle, H.J.C. van. (2002). The Dutch Entrustment Act (TBS): Its principles and innovations. *International Journal of Forensic Mental Health, 1* (1), 83-92.

Marshall, P. (1994). *Reconviction of imprisoned sexual offenders (Research Bulletin 36).* London: Home Office.

Martin, M.-L., Middleton, C., Webster, C.D., Nicholls, T., & Brink, J. (2004, June). *Changing practice: Assessing individuals' strengths to attenuate risk of harm to self and others.* Paper presented at the 4th Annual Conference of the International Association of Forensic Mental Health Services, Stockholm, Sweden.

McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30-46.

McInerney, T. (2000). Dutch TBS forensic services: A personal view. *Criminal Behaviour and Mental Health, 10*, 213-228.

McNiel, D.E., & Binder, R.L. (1994). The relationship between acute psychiatric symptoms, diagnosis, and short-term risk of violence. *Hospital and Community Psychiatry, 45* (2), 133-137.

McNiel, D.E., & Binder, R.L. (2002). Clinical assessment of the risk of violence among psychiatric inpatients. *American Journal of Psychiatry, 148* (10), 1317-1321.

McNiel, D.E., Lam, J.N., & Binder, R.L. (2000). Relevance of interrater agreement to violence risk assessment. *Journal of Consulting and Clinical Psychology, 68*, 1111-1115.

Meehl, P.E. (1954). *Clinical versus statistical prediction*. Minneapolis: University of Minnesota Press.

Metz, C.E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8*, 283-298.

Mills, J.F., Kroner, D.G., & Hemmati, T. (2003). Predicting violent behavior through a static-stable variable lens. *Journal of Interpersonal Violence, 18* (8), 891-904.

Monahan, J. (1981). *Predicting violent behavior. An assessment of clinical techniques*. Beverly Hills: Sage.

Monahan, J. (1996). Violence prediction: The past twenty and the next twenty years. *Criminal Justice and Behavior, 23* (1), 107-120.

Monahan, J., & Steadman, H.J. (1994). *Violence and mental disorder: Developments in risk assessment*. Chicago: University of Chicago Press.

Monahan, J. & Steadman, H.J. (1996). Violent storms and violent people. How meteorology can inform risk communication in mental health law. *American Psychologist, 51* (9), 931-938.

Monahan, J., Steadman, H.J., Silver, E., Appelbaum, P.S., Clark Robbins, P., Mulvey, P.E., Roth, L., Grisso, T., & Banks, S. (2001). *Rethinking risk assessment. The MacArthur study of mental disorder and violence.* Oxford: Oxford University Press.

Mossman, D. (1994). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology, 62* (4), 783-792.

Müller-Isberner, R., & Jöckel, D. (1997, September). *The implementation of the HCR-20 in a German hospital order institution.* Paper presented at the Seventh European Conference on Psychology and Law, Solna, Sweden.

Mulvey, E. & Lidz, C. (1985). Back to basics: A critical analysis of dangerousness research in a new legal environment. *Law and Human Behavior, 9*, 209-218.

Mulvey, E.P., & Lidz, C.W. (1995). Conditional prediction: A model for research on dangerousness to others in a new era. *International Journal of Law and Psychiatry, 18*, 129-143.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: an experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 29*, 85-95.

Norman, G.R. & Streiner, D.L. (1994). *Biostatistics: The bare essentials*. St. Louis, MO: Mosby.

Novaco, R.W. (1994). Anger as a risk factor for violence among the mentally disordered. In: J. Monahan & H.J. Steadman (Eds.). *Violence and mental disorder. Developments in risk assessment* (pp. 21-59). Chicago: The University of Chicago Press.

Nunes, K.L., Firestone, P., Bradford, J., Greenberg, D., & Broom, D. (2002). A comparison of modified versions of the Static-99 and the Sex Offender Risk Appraisal Guide. *Sexual Abuse: A Journal of Research and Treatment, 14*(3), 253-269.

Nunnally, J.C., & Bernstein, I.H. (1994). *Psychometric theory (3rd ed.).* New York: McGraw-Hill.

Oei, T.I., & Groenhuijsen, M.S. (2003). *Actuele ontwikkelingen in de forensische psychiatrie*. [Current developments in forensic psychiatry]. Deventer: Kluwer.

Philipse, M.W.G., Hilterman, E.L.B., & Doren, D. (2001). Tussen mogelijkheid en illusie: een review van elf risicotaxatie-instrumenten voor seksuele delinquenten. [Between possibility and illusion: a review of eleven risk assessment instruments for sex offenders]. *Tijdschrift voor Criminologie, 43* (1), 2-24.

Philipse, M.W.G., Koeter, M.W.J , Van den Brink, W., & Van der Staak, C.P.F. (2004). The structural coherence of clinically derived dynamic indicators of reoffending risk. *Criminal Behaviour and Mental Health, 14*, 263-279.

Philipse, M.W.G., Koeter, M., Staak, C.F.P. van der, en Brink, W. van den. (in press). Reliability and discriminant validity of dynamic reoffending risk indicators in forensic clinical practice. *Criminal Justice and Behavior.*

Prentky, R.A., Lee, A.F.S., Knight, R.A., & Cerce, D.D. (1997). Recidivism rates among child molesters and rapists: A methodological analysis. *Law and Human Behavior, 6*, 635-659.

Quinsey, V.L., Coleman, G., Jones, B., & Altrows, I. (1997). Proximal antecedents of eloping and reoffending among mentally disordered offenders. *Journal of Interpersonal Violence, 12*, 794-813.

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

Rassin, E., Koppen, P.J. van, & Vrij, A. (2002). Van Othello tot Pinokkio. Over leugendetectie en haar achterliggende ratio. [From Othello to Pinocchio. On lie detection and its underlying rationale]. *Nederlands Juristenblad, 77* (43), 2130-2136.

Reed, V., Robinson, D., Woods, P., Henderson, S., & Erven, T. van. (1997) *Behavioural Status Index. Nederlandse versie. (BSI-D).* Eindhoven: GGzE.

Rice, M.E., & Harris, G.T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology, 63* (5), 737-748.

Ruiter, C. de. (2000) *Voor verbetering vatbaar.* [Susceptible for improvement]. Inaugural lecture. Amsterdam: University of Amsterdam.

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

Scalora, M.J., & Garbin, C. (2003). A multivariate analysis of sex offender recidivism. International *Journal of Offender Therapy and Comparative Criminology, 47* (3), 309-323.

Schene, A.H., Koeter, M.W.J. ,Van Wijngaarden, B., Knudsen, H.C., Leese, M., Ruggeri, M., *et al.* (2000). Methodology of a multi-site reliability study. *The British Journal of Psychiatry, 177*, s15-s20.

Schiller, G., & Marques J. (1998). *The California Actuarial Risk Assessment Tables (CARAT) for Rapists and Child Molesters.* Sacramento: California Department of Mental Health.

Scott, P.D. (1977). Assessing dangerousness in criminals. *The British Journal of Pychiatry, 131*, 127-142.

Segal, S., Watson, M., Goldfinger, S., & Averbuck, D. (1988). Civil commitment in the psychiatric emergency room I: The assessment of dangerousness by emergency room clinicians. *Archives of General Psychiatry, 45*, 748-752.

Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlation: uses in assessing rater reliability. *Psychological Bulletin, 86,* 420-428.

Silver, E. & Miller, L.L. (2002). A cautionary note on the use of actuarial risk assessment tools for social control. *Crime & Delinquency, 48* (1), 138-161.

Sjöstedt, G., & Grann, M. (2002). Risk assessment: What is being predicted by actuarial prediction instruments? *International Journal of Forensic Mental Health, 1* (2), 179-183.

Sjöstedt, G., & Långström, N. (2001). Actuarial assessment of sex offender recidivism risk: A cross-validation of the RRASOR and the Static-99 in Sweden. *Law and Human Behavior, 25* (6), 629-645.

Sjöstedt, G., & Långström, N. (2002). Assessment of risk for criminal recidivism among rapists: A comparison of four different measures. *Psychology, Crime and Law, 8*, 25-40.

Slovic, P., Monahan, J. & MacGregor, D.G.. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats. *Law and Human Behavior, 24* (3), 271-296.

Sturidsson, K., Haggård-Grann, U., Lotterber, M., Dernevik, M., & Grann, M. (2004). Clinicians' perceptions of which factors increase or decrease the risk of violence among forensic out-patients. *International Journal of Forensic Mental Health, 3* (1), 23-36.

Swaaningen, R. van. (1996). Justitie als verzekeringsmaatschappij: 'Actuarial justice' in Nederland. [The judiciary as insurance company: 'Actuarial justice' in the Netherlands]. *Justitiële Verkenningen, 5*, 80-97.

Swartz, M., Swanson, J., Wagner, H., Burns, B., Hiday, V. & Borum, R. (1999). Can involuntary outpatient commitment reduce hospital recidivism? Findings from a randomized trial with severely mentally disordered patients. *American Journal of Psychiatry, 156*, 1968-1975.

Uitvoeringsconsortium Projectbureau Politiemonitor (2003). *Politiemonitor Bevolking 2003. Beleidsrapportage.* [Population Police Monitor 2003. Executive report]. Den Haag: Ministeries van Binnenlandse Zaken en Justitie.

Verhagen, M.F.M., & Philipse, M.W.G. (1995). Het voorspellen van het risico van delictherhaling in de TBS. [Predicting the risk of reoffending in TBS]. *Tijdschrift voor Psychiatrie, 37* (7), 537-552.

Versteegh, P., Janssen, J., & Bernasco, W. (2003). Beginners, doorstromers en veelplegers. Carrièrecriminaliteit in de politieregio Haaglanden. [From starters to persistent offenders. Criminal careers in the Haaglanden police district]. *Tijdschrift voor Criminologie, 45* (2), 127-139.

Vincent, G. M., Ross, D. J., Whittemore, K., Eaves, D., Hart, S. D., Ogloff, J. R. P., *et al.* (2001, April). *Using the HCR-20: File-based researcher ratings vs. file + interview-based clinician ratings.* Paper presented at the Founding conference of the International Association of Forensic Mental Health Services, Vancouver, British Columbia, Canada.

Vogel, V. de, & Ruiter, C. de (2002). Verschillen tussen onderzoekers en behandelaars in het inschatten van het risico van gewelddadig gedrag. [Differences between researchers and clinicians in the assessment of risk of violent behavior]. *Directieve Therapie, 23,* 43-62.

Vogel, V. de, Ruiter,C. de, & Bouman, Y.M.H. (2004). *Handleiding bij de SAPROF. Structured Assessment of Protective Factors: Research Version. (SAPROF Manual).* Utrecht: Trimbos Instituut.

Vogel, V. de, Ruiter, C. de, Beek, D. van, Mead, G. (2003). De waarde van gestructureerde risicotaxatie: een retrospectief empirisch onderzoek bij behandelde seksuele delinquenten. [The value of structured risk assessment: a retrospective empirical study among treated sex offenders]. *Maandblad Geestelijke volksgezondheid, 58*, 9-29.

Ward, T., & Eccleston, L. (2000). The assessment of dangerous behaviour: Researh and clinical issues. *Behaviour Change, 17* (2), 53-68.

Webster, C.D., Douglas, K.S., Eaves, D., & Hart, S.D. (1997a). *HCR-20. Assessing risk for violence. Version 2.* Burnaby, British Columbia: Simon Fraser University, Mental Health, Law and Policy Institute.

Webster, C.D., Douglas, K.S., Eaves, D., & Hart, S.D. (1997b). Assessing risk of violence to others. In: C.D. Webster & M.A. Jackson (Eds.), *Impulsivity: Theory, assessment and treatment* (pp. 251-277). New York: Guilford Press.

Webster, C.D., Müller-Isberner, R., & Fransson, G. (2002). Violence risk assessment: Using structured clinical guidelines professionally. *International Journal of Forensic Mental Health, 1* (2), 43-51.

Widaman, K.F. (1993). Common factor analysis versus principal components analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research, 28* (3), 263-311.

Williams, C.R., & Arrigo, B.A. (2002). Law, psychology and the 'New Sciences': Rethinking mental illness and dangerousness. International *Journal of Offender Therapy and Comparative Criminology, 46* (1), 6-29.

Wormith, J.S., & Olver, M.E. (2002). Offender treatment attrition and its relationship with risk, responsivity, and recidivism. *Criminal Justice and Behavior, 29*, 447-471.

Zillmer, E.A., & Vuz, J. (1995). Factor analysis with Rorschach data. In: J. Exner, Jr. (Ed.), *Methods and issues in Rorschach research* (pp. 251-306). Hillsdale, NJ: Lawrence Erlbaum Associates.

# ☙ Dankwoord

Het is in dankwoorden van proefschriften goed gebruik om uitvoerig te belijden wat een zware klus het is geweest, en het boetekleed aan te trekken voor de enorme wissel die dat op het sociale netwerk heeft getrokken. Graag breek ik hier met die traditie. Ik ben gezegend met een werkgever die niet alleen het schrijven van deze dissertatie heeft gestimuleerd, maar binnen mijn werk ook zodanig gefaciliteerd dat avonden, weekenden en vakanties er weinig onder geleden hebben, en mijn naasten al evenmin (althans, ik heb geen klachten vernomen). Mijn dank aan bestuur en directie van de Pompestichting, die dit werkstuk mogelijk hebben gemaakt.

Evenveel dank ben ik verschuldigd aan Thieu Verhagen. Als eenzame pionier in een TBS-veld waar geen mens nog van risicotaxatie had gehoord, bedacht hij ergens eind jaren '80 al dat het misschien wel een goed idee zou zijn eens te kijken hoe wij dat eigenlijk doen, inschatten of onze patiënten nog gevaarlijk zijn. "De bonbon van het TBS-onderzoek," noemde een collega het met minzame afgunst. Het project dat Thieu opstartte, en de checklist die hij ervoor ontwikkelde, waren in feite wereldprimeurs in een tijd dat de HCR-20 nog amper een gedachte in het hoofd van Webster was. Ik prijs mij gelukkig dat hij dat geesteskind aan mijn zorgen heeft toevertrouwd - en ik hoop dat het eindresultaat zijn goedkeuring kan wegdragen, zelfs al ziet het er niet helemaal uit zoals we toen in ons optimisme verwachtten.

Onmisbaar waren natuurlijk de leden van de Projectgroep en hun achterban en management in de 8 klinieken die aan dit onderzoek meewerkten. Dank aan de directies voor de bereidwillig geboden 'kijk in de keuken', aan de behandelcoördinatoren/supervisoren ter plaatse voor het invullen van de lijsten; en aan Anke ten Wolde, Tom van Erven, Jan Niemantsverdriet, Alex Hooischuur, Rob Ziel, Eddy Brand, Sylvia Lammers, Jos Peters en anderen voor de onvermoeibare, soms esoterische maar altijd inspirerende discussies in onze bijeenkomsten, en het geduld waarmee mijn ad hoc toevoegingen aan de geplande dataverzameling werden getolereerd (èn gerealiseerd!).

Cees, Wim en Maarten, *last but not least*: jullie inbreng tilde dit onderzoek naar een hoger wetenschappelijk plan. Onze gesprekken waren steeds een leerzaam genoegen, en ik ben blij door jullie te zijn behoed voor overmatige bescheidenheid over de bevindingen, en ook voor een loopbaan in de visserij. Heel veel dank.

Tijdens een masterclass voor promovendi merkte een van de docenten ooit op, dat de hedendaagse trend van promoveren op artikelen de promovendus onder het juk van de tijdschriftredactie plaatst, waardoor hij gewoonlijk wordt beroofd van elke mogelijkheid om een persoonlijke noot in zijn product te brengen. Gevolg: almaar uitdijende dankwoorden waarin dit gemis wordt gecompenseerd in een stortvloed van soms pijnlijk persoonlijke ontboezemingen. Van mijn ter plekke genomen besluit mijn dankwoord summier te houden komt, zoals u ziet, niks terecht. Erger nog, behalve de mensen die een directe bijdrage leverden aan dit proefschrift, voel ik mij, ook al ben ik al op bladzijde twee, toch geroepen nog enkele anderen te danken.

Margot en Terrence, bijvoorbeeld, voor het proeflezen en corrigeren van mijn Engels, desnoods tot bloedens toe. Yvonne en Mieke voor collegiale en sociale steun, leven in de brouwerij, en gezamenlijke sponsoring van het Nijmeegse restaurantwezen. Wilma, gewoon, voor alles. Ed, voor de onnavolgbare telefoontirades en vele smakelijke congresherinneringen. Marjon, voor het onvermoeibaar aanleveren van artikelen en boeken en het gedogen van mijn privébibliotheek op de Zeedijk. Marian, voor de assistentie bij de layoutperikelen – jammer dat het nog steeds niet botert tussen jou en mijn draakje. Dank aan Piet, die als doorgewinterd filosoof wist aan te wijzen wat het allemaal betekende zelfs op momenten dat ik even dacht dat het helemaal nergens over ging (en met wie ik nog veel bierglazen hoop te legen bij discussies over de PRECIEZE plaats van punten en komma's). Mijn redactiewerk voor PS bleek ook een ervaring van grote waarde: zonder de mantra "kill your darlings" (dankjewel Janneke) was ik mijn artikelen aan de straatstenen niet kwijt geraakt, en was dit proefschrift minstens tweemaal zo dik geweest.

Tenslotte. Als de TBS je iets leert, is het wel hoe wezenlijk vrienden en familie zijn. Het stemt me dankbaar dat ik ze niet alleen heb, maar dat ze in hun verscheidenheid zoveel warmte, humor, steun en relativering bieden als ik dagelijks mag ervaren. De donderdagse eetclub (waar de inwendige mens op zoveel meer manieren dan alleen culinair aan zijn trekken komt), het strijkkwartet (op naar de wederopstanding!), de avondjes V&F met Theo en Henny, de gezellige zondagen in Sambeek; – zulke dingen maken het allemaal de moeite waard. Er is ook geen omgeving zo geschikt als de TBS om je te leren dat een warm ouderlijk nest geen vanzelfsprekendheid is, en een onvervangbaar goed. Zonder de niet aflatende liefde en goede zorgen van mijn ouders zou ik nu niet zijn wie ik ben; aan hen draag ik dit proefschrift in dankbaarheid op.

# Curriculum vitae

Martien Philipse was born in 1966 in Boxmeer, The Netherlands, where he finished secondary school education in 1985. He went on to study Psychology at the (now) Radboud University of Nijmegen, with Psychology of Culture and Religion as his primary subject. He graduated in 1991 on a doctoral essay on valuation systems in cancer patients.

That same year he began work as a researcher in the Pompekliniek, a forensic mental health institution in Nijmegen, and has remained there since. Risk assessment has been his core subject from the beginning, a topic on which he frequently lectures and provides training, has published repeatedly, and on which he has presented papers at numerous conferences. He was one of the instigators and the first author of the Dutch translation of the HCR-20, and is a founding member of the European Network for Structured Risk Assessment (ENSRA). For several years, he was also a board member of the Dutch Society for Forensic Sexology.

In his spare time Martien Philipse takes an active interest in architecture, music and graphic design, and plays the cello.

# Appendices

## Appendix I. Items of the CIDRRI, with reliability measures and admittee-discharge mean score comparisons.

| Item | s.m.[a] interrater ICC[b] (n=75) | s.m. retest ICC (n=29) | Admittee minus dischargee mean score (n=114, 118) |
|---|---|---|---|
| 1. Patient is easily influenced by others | .50** | .95** | .39; t(230) = 1.94 |
| 2. Patient acknowledges that he has committed a serious offence | .45** | .80** | .53; t(230) = 2.54* |
| 3. His conscience is impaired: internalization of generally accepted norms and values is largely or completely absent | .41** | .63** | .67; t(230) = 3.68** |
| 4. His social network offers actual support | .56** | .83** | 1.13; t(230) = 5.66** |
| 5. Patient has little control over his aggressive and/or sexual impulses | .61** | .66** | 1.07; t(230) = 6.08** |
| 6. His behavior transgresses other people's boundaries | .51** | .81** | .57; t(223.46) = 2.81** |
| 7. Patient avoids contact with others | .59** | .63** | .33; t(219.50) = 1.69 |
| 8. Patient does not feel what he has done to his victims | .45** | .62** | .82; t(230) = 4.44** |
| 9. He's unable to imagine how other people are feeling | .41** | .51** | .48; t(230) = 2.84** |
| 10. Patient has a craving for drugs and/or alcohol | .66** | .86** | .62; t(230) = 2.66** |
| 11. Patient denies his offence completely | .40** | .70** | .55; t(214.41) = 2.60* |
| 12. Patient is compulsively preoccupied with sex | .62** | .55** | .37; t(230) = 2.09* |
| 13. Patient faithfully takes prescribed medication | .49** | .78** | *not available* |
| 14. Patient perseveres in resistance against treatment | .46** | .68** | .54; t(223.38) = 2.84** |

*Appendix I - continued*

| Item | s.m. interrater ICC (n = 75) | s.m. retest ICC (n = 29) | Admittee minus dischargee mean score (n=114, 118) |
|---|---|---|---|
| 15. Given his capacities, patient has shown optimum participation in the treatment program | .51** | .49** | .16; t(230) = .86 |
| 16. His sleep-wake pattern is disturbed | .52** | .71** | .27; t(230) = 1.82 |
| 17. Patient has good personal hygiene | .50** | .86** | .05; t(230) = .31 |
| 18. Patient spends his leisure time in an active way | .45** | .82** | .18; t(230) = .88 |
| 19. Patient understands the nature of his psychopathology and its influence on his behaviour | .51** | .71** | 1.01; t(230) = 5.71** |
| 20. The pathology underlying the offence has been lifted or mitigated | .52** | .74** | 1.83; t(209.95) = 12.69** |
| 21. If he encounters trouble, he calls in help | .46** | .75** | .81; t(230) = 5.15** |
| 22. Patient compels others to adjust to his needs and wishes | .44** | .75** | .64; t(217.34) = 3.43** |
| 23. Patient has extreme fear of abandonment | .46** | .56** | .25; t(230) = 1.31 |
| 24. Does not give room to personal or emotional needs of others | .49** | .66** | .69; t(230) = 3.89** |
| 25. Staff feel they have to be on their guard when dealing with patient | .65** | .76** | .94; t(222.53) = 4.68** |
| 26. Patient does not tolerate intimacy | .26* | .64** | .59; t(230) = 3.35** |
| 27. Relational patterns at the time of the offence seem to repeat themselves | .31** | .68** | .89; t(230) = 4.75** |
| 28. Patient knows the chain of events leading up to his offences | .52** | .84** | 1.74; t(230) = 9.64** |

*Appendix I - continued*

| Item | s.m. interrater ICC (n = 75) | s.m. retest ICC (n = 29) | Admittee minus dischargee mean score (n=114, 118) |
|---|---|---|---|
| 29. Patient's inflated self image easily makes him feel hurt | .55** | .76** | .39; t(223.28) = 1.85 |
| 30. Patient has unrealistic ideas about his future | .56** | .48** | .85; t(230) = 4.26** |
| 31. Patient has spells during which his actions and thoughts are no longer adjusted to reality | .46** | .76** | .20; t(230) = .86 |
| 32. External circumstances in which patient committed earlier offences have changed for the better | .32** | .74** | 2.59; t(206.51) = 16.00** |
| 33. With regard to his offence, patient feels he is a victim rather than a perpetrator | .47** | .58** | .49; t(223.39) = 2.45* |
| 34. Patient actively uses opportunities to learn new things | .60** | .59** | .09; t(230) = .48 |
| 35. Patient lacks essential social skills | .33** | .55** | .79; t(230) = 4.30** |
| 36. Patient has sufficient skills to live on his own and take care of himself | .64** | .72** | .65; t(230) = 2.91** |
| 37. Patient is sufficiently capable of controlling his finances | .68** | .82** | .62; t(222.83) = 3.20** |
| 38. Patient has insufficient internal structure to stand on his own | .19* | .63** | .84; t(230) = 4.28** |
| 39. Patient allows himself to be guided by current needs and impulses | .58** | .62** | .93; t(230) = 5.17** |
| 40. Patient plays people off against each other | .37** | .90** | .76; t(220.05) = 4.00** |
| 41. Patient accepts responsibility for his own share in his problems | .54** | .58** | .75; t(230) = 4.37** |

*Appendix I - continued*

| Item | s.m. interrater ICC (n = 75) | s.m. retest ICC (n = 29) | Admittee minus dischargee mean score (n=114, 118) |
|---|---|---|---|
| 42. He keeps to his appointments | .42** | .68** | .62; t(230) = 4.10** |
| 43. Patient usually discusses others in negative terms | .56** | .24 | .63; t(210.72) = 3.67** |
| 44. Patient is easily provoked | .40** | .70** | .84; t(230) = 4.63** |
| 45. He allows insight into the way he spends  his leave | .47** [c] | .70** | *not available* |
| 46. Patient"s daily functioning strongly depends on the hospital | .54** | .82** | .77; t(230) = 3.70** |
| 47 Patient alternates between idealisation and devaluation of people close to him | .47** | .85** | .52; t(223.51) = 2.80** |

NOTE. Items are rated: "highly uncharacteristic 0 0 0 0 0 0 highly characteristic". For the purpose of analyses and ease of interpretation items were recoded if necessary, so that higher scores always corresponded to higher levels of dysfunction. [a]s.m. = single measure. [b] ICC = intraclass correlation. [c]n=46. *p<.05; **p<.01.

## APPENDIX II. FACTOR STRUCTURE OF THE CIDRRI[a]

| Factor | 1 | 2 | 3 | 4 | 5 | 6 | |
|---|---|---|---|---|---|---|---|
| Name | Responsibility | Self reliance | Antisocial narcissism | Compliance | Goal attainment | Avoidance | Total $R^2$ |
| Eigenvalue | 12.59 | 3.50 | 2.92 | 2.22 | 1.84 | 1.78 | |
| % of variance explained | 10.6 | 10.0 | 9.8 | 6.6 | 5.6 | 3.5 | 26 |
| Items | Factor loadings[b] | | | | | | $h^2$ [c] |
| 2. Faces offence | **.80** | | | | | | .74 |
| 8. Lack of victim empathy | **.74** | | | | | | .63 |
| 9. Unable to empathize | **.64** | | | | | | .55 |
| 11. Denies offenc | **.61** | | | | | | .40 |
| 41. Responsibility | **.53** | | | .39 | .29 | | .64 |
| 19. Understands his pathology | **.52** | | | | .44 | | .57 |
| 3. Impaired conscience | **.51** | | .33 | | | | .49 |
| 33. Feels victim, not offender | **.49** | | | | | | .31 |
| 24. No room for needs others | **.45** | | .38 | | | .31 | .50 |
| 30. Unrealistic view of future | **.35** | | | .28 | .29 | | .43 |
| 12. Sexually obsessive | **.34** | | | | | | .22 |
| 46. Depends on hospital | | **.79** | | | | | .67 |
| 36. Able to live by himself | | **.76** | | | | | .68 |
| 38. Unable to stand on his own | | **.58** | | | | | .41 |
| 37. Able to manage finances | | **.58** | | | | -.27 | .51 |
| 39. Guided by impulses | | **.56** | .43 | | | | .64 |
| 17. Good hygiene and self-care | | **.54** | | | | | .37 |
| 31. Psychotic episodes | | **.53** | | .31 | | | .47 |
| 18. Active use of leisure time | | **.46** | | .42 | | | .47 |
| 16. Disturbed circadian rhythm | | **.44** | | .41 | | | .38 |
| 1. Easily influenced | | **.44** | | | | -.32 | .34 |
| 35. Lacks social skills | .32 | **.43** | | | | .28 | .47 |
| 10. Craves alcohol/drugs | | **.33** | | | | | .28 |

*Appendix II - continued*

| Item | factor 1 | factor 2 | factor 3 | factor 4 | factor 5 | factor 6 | $h^2$ |
|---|---|---|---|---|---|---|---|
| 22. Forces others to adjust | | | .70 | | | | .56 |
| 40. Plays people off against e.o | | | .66 | | | | .52 |
| 44. Irritable | | | .63 | | | | .46 |
| 6. Oversteps limits of others | .31 | | .63 | | | | .57 |
| 43. Talks negatively about others | | | .62 | | | | .49 |
| 25. Staff are on their guard | | | .56 | | | | .45 |
| 5. Little control over impulses | .31 | .36 | .50 | | | | .55 |
| 23. Extreme fear of abandonment | | | .44 | | | | .23 |
| 27. Crime related patterns repeat | | | .43 | | | | .28 |
| 29. Self-image easily injured | | | .43 | | | | .31 |
| 15. Optimum treatment particip. | | | | .67 | | | .55 |
| 14. Persists in resistance | .29 | | .30 | .62 | | | .61 |
| 34. Uses learning opportunities | .34 | | | .51 | | .29 | .53 |
| 13. Takes prescribed medication | | | | .50 | .35 | | .42 |
| 42. Keeps to agreements | | | | .48 | | | .47 |
| 32. External situation improved | | | | | .61 | | .47 |
| 28. Knows offence script | .49 | | | | .56 | | .64 |
| 20. Pathology mitigated | .31 | .30 | | | .49 | | .51 |
| 47. Praises and reviles people | | | | | .48 | | .24 |
| 21. Asks help when in trouble | .33 | | | .33 | .40 | | .41 |
| 4. Supportive social network | | | | | .36 | | .21 |
| 45. Allows insight in use of leave | | | | | .35 | | .28 |
| 7. Avoids contact | | | | | | .58 | .38 |
| 26. Doesn't tolerate intimacy | | | .29 | | | .42 | .36 |

NOTE. [a]Principal axis factoring with Varimax rotation. Kaiser Meyer Olkin (KMO) = .90, Bartlett's test of sphericity approximate $\chi^2$=2,137.53, df=1,081, p<.001. [b]Factor loadings < .27 have been suppressed (cf. Norman & Streiner, 1994: loadings should be $\geq$ 5.152/√(N-2)). [c]Communalities after factor rotation.