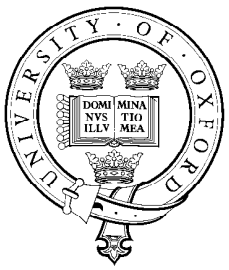


Forensic Risk Assessment: A Metareview, Novel Meta-analysis, and Empirical Study Developing a Violence Screening Tool for Schizophrenia

Jay P Singh
Department of Psychiatry and University College
University of Oxford

Submitted for the degree:
DPhil in Psychiatry
(Approximately 50,000 words)

Trinity Term 2011

GENERAL ABSTRACT

Title: Forensic Risk Assessment: A Metareview, Novel Meta-analysis, and Empirical Study Developing a Violence Screening Tool for Schizophrenia

Name and Society: Jay P Singh, University College

Degree: Doctor of Philosophy in Psychiatry, Trinity Term 2011

Abstract: Mental health professionals are routinely called upon to assess the violence risk of their patients. An increasingly common method for conducting such assessments is the use of structured risk assessment tools. The aim of this thesis was to investigate the utility of such instruments: to identify and explore current uncertainties concerning their applicability and to design a novel measure that could be used as part of a stepped strategy to risk assessment. Though a number of risk assessment tools have been developed and there is a considerable literature concerning their psychometric properties, uncertainty remains regarding their effective use. In order to identify key contemporary uncertainties, a metareview of the forensic assessment literature was conducted. The metareview found that previous systematic reviews and meta-analyses of the risk assessment literature have come to conflicting conclusions on a number of issues, including the comparative predictive validity of risk assessment tools, the efficacy of actuarial tools versus clinical judgement, and the influence of demographic factors and study design characteristics on predictive accuracy. These uncertainties were then investigated in a comprehensive meta-analysis of nine commonly used risk measures. The meta-analysis concluded that there were significant differences between the predictive validity of the risk assessment tools, with instruments designed for more specific purposes performing better than those designed for more general use. Tools performed best when administered to samples demographically similar to their calibration sample. Actuarial instruments and structured clinical judgement were found to perform comparably. The final study presented in this thesis explored the feasibility of a stepped approach to risk assessment in which individuals at very low risk of future violence are screened out prior to resource-intensive clinically based assessment. High-quality national registers were used to construct a simple tool to identify patients with schizophrenia at very low risk of violent conviction after being discharged from hospital. The tool was found to produce high rates of sensitivity as well as high negative predictive values at 1, 2, and 5 years follow-up. In light of the findings of these three studies, risk assessment procedures and guidelines by mental health services and criminal justice systems may need review.

ACKNOWLEDGEMENTS

I would like to express my thanks to everyone who has contributed to the work reported in this thesis. Special thanks to Dr. Seena Fazel, my supervisor, for his dedicated guidance and constructive criticism. I thank Professor Martin Grann for his encouragement and advice, Professor Klaus Ebmeier for his support and assistance in the translation of several articles in German, Dr. Helen Doll for her assistance with statistical analysis, and Ms. Sophie Westwood for her assistance with the inter-rater reliability checks. The following authors are thanked for providing studies and/or tabular data for the analyses reported as part of this thesis: April Beckmann, Sarah Beggs, Susanne Bengtson Pedersen, Klaus-Peter Dahle, Rebecca Dempster, Mairead Dolan, Kevin Douglas, Reinhard Eher, Jorge Folino, Monica Gammelgård, Robert Hare, Grant Harris, Leslie Helmus, Andreas Hill, Clive Hollin, Christopher Kelly, P. Randy Kropp, Michael Lacy, Calvin Langton, Henry Lodewijjks, Karin Arbach Lucioni, Jeremy Mills, Catrin Morrissey, Thierry Pham, Martin Rettenberger, Marnie Rice, Michael Seto, David Simourd, Gabrielle Sjöstedt, Cornelis Stadtland, David Thornton, Vivienne de Vogel, Zoe Walkington, and Glenn Walters. Finally, I thank my family and friends for their love and unflagging support.

PUBLICATIONS ARISING FROM THESIS

Chapter I: Singh, J. P. (in press). The history, development, and testing of forensic risk assessment tools. In E. Grigorenko (Ed.), *Handbook of juvenile forensic psychology and psychiatry*. New York: Springer.

Chapter II: Singh, J. P., & Fazel, S. (2010). Forensic risk assessment: A metareview. *Criminal Justice & Behavior*, *37*, 965-988.

Chapter III: Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review*, *31*, 499-513.

Chapter IV: Singh, J. P., Grann, M., Lichtenstein, P., Långström, N., & Fazel, S. (under review). Towards a simple screen for violence risk in schizophrenia: A development and cross-validation study of 13,806 discharged patients.

TABLE OF CONTENTS

INDEX OF TABLES

INDEX OF FIGURES

Figure Page

LIST OF COMMON ABBREVIATIONS

Abbreviation	Definition
AUC	Area Under the Curve
CI	Confidence Interval
DOR	Diagnostic Odds Ratio
FN	False Negative
FP	False Positive
HCR-20	Historical, Clinical, Risk Management – 20
HR	Hazard Ratio
IQR	Interquartile Range
N/A	Not Available/Applicable
NND	Number Needed to Detain
NPV	Negative Predictive Value
NSS	Number Safely Screened
PCL-R	Psychopathy Checklist – Revised
PPV	Positive Predictive Value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-analyses
ROC	Receiver Operating Characteristic
SARA	Spousal Assault Risk Assessment
SAVRY	Structured Assessment of Violence Risk in Youth
SCJ	Structured Clinical Judgement
SORAG	Sex Offender Risk Appraisal Guide
STARD	Standards for Reporting of Diagnostic Accuracy Studies
SVR-20	Sexual Violence Risk – 20
TN	True Negative
TP	True Positive
UK	United Kingdom
US	United States of America
VRAG	Violence Risk Appraisal Guide

Chapter I:
The Field of Forensic Risk Assessment

The contents of this chapter have since been published by SAGE Publications as: Singh, J. P. & Fazel, S. (2010). 'Forensic risk assessment: a metareview', Criminal Justice & Behavior, 37, 965-988. Available at: http://dx.doi.org/10.1177/0093854810374274.

Chapter II:
Forensic Risk Assessment: A Metareview

2.1 ABSTRACT

A large number of systematic reviews and meta-analyses have been conducted in the field of forensic risk assessment, and their conclusions have occasionally been conflicting. To examine the quality and findings of these reviews, the first metareview of this literature was conducted. Nine systematic reviews and 31 meta-analyses from 1995 to 2009 were identified. The themes covered in these reviews included the comparison of various risk assessment schemes, the utility of the actuarial approach compared with unstructured and structured clinical judgement, and the predictive validity of instruments across demographics and study designs. This metareview found that the quality and consistency of findings in these areas varied considerably, suggesting the existence of major uncertainties concerning the utility of forensic risk assessment. Review quality was generally poor, with few reviews reporting replicable search strategies, half not investigating sources of heterogeneity or excluding overlapping samples, and only a third assessing publication bias. A standardisation of review reporting with particular emphasis on methodological consistency is suggested.

2.2 INTRODUCTION

Since the first reported study on forensic risk assessment, conducted in a population of parolees in 1928 (Burgess, 1928), research in the field has expanded considerably and has included work on many measures in varied populations and settings. Searching for all previously published literature using the term *forensic risk*

assessment on the PsycINFO search engine in 1999 would have yielded a total of 1,283 citations, whereas the same search in 2009 gave a total of 4,785 records. As the number of published studies has grown, literature reviews have assisted in summarising and synthesising this work. These reviews have influenced clinicians and policymakers by providing an empirical base for their decision-making (Gendreau, Goggin, & Smith, 2000).

As discussed in the previous chapter, there are two methods of systematically summarising the results of past studies: systematic reviews, which are descriptive in nature and do not involve quantitative synthesis, and meta-analyses, which use summary statistics to combine the results of primary studies. The quality of such systematic reviews and meta-analyses varies widely (Deville et al., 2002). These differences in quality may explain why contradictory findings are occasionally reported, even when the same literature has been reviewed. For example, in the field of treatment research in mental health, reviews of the efficacy of antidepressants and the comparative advantage of second-generation antipsychotics have resulted in conflicting findings despite having examined similar literatures (Geddes, Freemantle, Harrison, & Bebbington, 2000; Leucht et al., 2003).

In the forensic risk assessment literature, reviews concerning which risk assessment tools have the highest rates of predictive validity (Campbell, French, & Gendreau, 2007; Salekin, Rogers, & Sewell, 1996; Walters, 2003b), whether actuarial risk measures or clinicians are better at detecting risk of offending (Ægisdóttir et al., 2006; Guy, 2008; Hanson & Morton-Bourgon, 2007), and whether risk assessments are equally valid in different populations (Edens, Campbell, & Weir, 2007; Leistico, Salekin, DeCoster, & Rogers, 2008) have come to different conclusions. It is possible

that this is attributable to the research base having changed with time, although it may also be related to differences in the methodology used to review the relevant evidence.

Four important methodological differences among reviews include: (1) the inclusion of a replicable systematic search, (2) the investigation of heterogeneity, (3) the extent to which publication bias is investigated, and (4) the effect sizes used to report results (Moher et al., 2007). Given that systematic reviews and meta-analyses mitigate selection bias by conducting systematic searches, it is important that detailed descriptions of search criteria, inclusion and exclusion criteria, and review flow (i.e., how many studies were initially considered for inclusion and how many were excluded) are provided. Without including such a replicable search strategy, it is difficult to assess which amongst a series of reviews with conflicting findings has included the most representative set of studies. A second methodological difference between reviews is the extent to which heterogeneity is investigated. As the field of forensic risk assessment includes primary studies that are conducted on a variety of populations in many study settings, methodologists suggest that reviewers need to statistically investigate the presence and sources of heterogeneity (Higgins, 2008; Higgins, Thompson, Deeks, & Altman, 2003). Thirdly, reviews differ in whether and how they address publication bias. Not all literature concerning forensic risk assessment has been published in peer-reviewed journals. The findings of the "grey literature" (i.e., government reports, conference presentations, Master's theses, doctoral dissertations) may differ systematically from the results published in journal articles. Therefore, reviewers should consider formally assessing publication bias. Finally, reviews often differ in the effect estimates used to report findings. Effect sizes in the risk prediction literature have relative strengths and weaknesses (Chapter I,

Section 1.4.1.1), and review findings should be qualified in light of these (Gendreau et al., 2000).

2.2.1 Metareview Methodology

To investigate discrepancies in the methodological quality and the findings of reviews, one novel approach is a metareview. Though the term *metareview* has, at times, been used interchangeably with *meta-analysis*, the former has its own methodology, systematically searching for and descriptively summarising all available meta-analyses and systematic reviews on a given topic. (Narrative reviews are not included, as they are especially vulnerable to selection bias.) Metareviews are different from second-order meta-analyses, which collect and quantitatively summarise data from all primary studies included in a set of reviews (e.g., Adams, Fenton, Quraishi, & David, 2001).[3] Although qualitative in nature, metareviews highlight inconsistencies in meta-analyses and systematic reviews and can note areas that require further research (Cipriani, Geddes, Furukawa, & Barbui, 2007; Ruddy & House, 2005). There is no accepted method (to my knowledge) of quantitatively synthesising the findings of individual meta-analyses. To do so would be problematic, as multiple reviews often use overlapping studies or samples.

In his editorial on the applications of metareview, Delgado-Rodriguez (2006) identified four strengths of this approach. The first is that metareviews allow researchers to investigate the general quality of the review literature on a given topic. This may be accomplished using a standardised checklist such as the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) Statement

[3] Second-order meta-analyses do not allow researchers to investigate the quality or reporting characteristics of previous reviews, nor do they allow for thematic analysis to identify major uncertainties in those reviews. Therefore, metareview methodology was deemed more appropriate for the purposes of the present investigation. The specific aims of this metareview are described in Section 2.2.2.

(Moher, Liberati, Tetzlaff, & Altman, 2009), a 27-item checklist of review characteristics designed to enable a transparent and consistent reporting of search strategies and results. A second strength is that metareviews can advance researchers' understanding of heterogeneity. If consistent evidence of heterogeneity is found, a metareview can suggest that sources of heterogeneity be investigated regardless of the statistical significance of a meta-analysis' findings. A third strength of metareview methodology is that by using the review as the unit of analysis, the consistency of publication bias findings can be explored. As reviews may operationalise the term "publication" differently (e.g., published in a peer-reviewed journal versus not, or available to the public in published form [journal article or government report] versus not), metareviews can qualify findings of publication bias. Finally, metareviews can identify which outcome statistics are commonly used to summarise study findings. Thus, metareviews can be used to investigate the key benchmarks of review quality.

2.2.1.1 Examples of Metareview Methodology

PsycINFO, MEDLINE, EMBASE, and the US National Criminal Justice Reference Service Abstracts were used to identify previous metareviews of the medical and social sciences literatures. Searching for the term *metareview* resulted in 150 entries as of June 30, 2009. While this search was not sensitive enough to identify all previous metareviews (e.g., Palma & Delgado-Rodriguez, 2005; Petitti, 2001), the aim was to identify examples of the novel methodology to serve as models. Using this search strategy and excluding duplicates, narrative and systematic reviews, meta-analyses, and records that were not concerned with the fields of medicine or psychology, three records remained. The topics covered by these metareviews included: (1) the short-term effectiveness and safety of antidepressants for treating

depression (Cipriani et al., 2007), (2) interventions in key areas of liaison psychiatry (Ruddy & House, 2005), and (3) the epidemiology and reporting characteristics of systematic reviews in the field of medicine (Moher et al., 2007).

Cipriani et al., 2007

Cipriani and colleagues (2007) conducted a metareview of the short-term effectiveness and safety of antidepressants used in the acute phase treatment of major depression. The researchers systematically searched eight online databases to identify reviews of short-term pharmacologic interventions for depression that used antidepressants as part of treatment. Only reviews of randomised controlled trials (RCT) were included. The metareview identified 1 relevant systematic review and 11 meta-analyses. While the reviews provided consistent evidence that antidepressants are effective in treating major depression in primary care settings, the authors concluded that there remains considerable uncertainty concerning the health-related effects of such medication. For example, maternal selective serotonin reuptake inhibitor (SSRI) usage during breast-feeding was not found to have any negative effects on infants, though there was some evidence of a relationship between maternal SSRI usage and pregnancy-related complications. The authors discussed the limitations of previous reviews of the antidepressant literature and suggested areas in need of further research.

Ruddy & House, 2005

Ruddy and House (2005) conducted a metareview to investigate interventions for clinical problems likely to be treated by liaison psychiatric services. The authors systematically searched six databases for systematic reviews and meta-analyses

concerning interventions designed to treat psychological problems resulting from a physical illness, somatoform disorders, or self-harming behaviour. Using pre-specified inclusion and exclusion criteria, 51 relevant systematic reviews and 14 relevant meta-analyses were identified. The researchers appraised the quality of the collected reviews using a checklist published by Oxman and Guyatt (1988). This checklist assesses whether reviewers clearly stated their aims and methods, conducted a systematic search, specified their inclusion and exclusion criteria, assessed the validity of the included primary studies in a way that was free from bias, ensured that primary studies were quantitatively synthesised in an appropriate manner, and discussed inconsistent findings (Oxman & Guyatt, 1988). The inter-rater reliability of ratings of review quality was assessed using a random sample of the included reviews, which were rated by two of the metareview's authors.

The metareview concluded that much of the clinical practice of liaison psychiatry is based on low-quality research evidence. Further, reviews often came to conflicting conclusions about which form of treatment is most effective for different problems. The authors discussed the limitations of the identified reviews and the need for more service-oriented research focusing on common problem areas in clinical practice.

Moher et al., 2007

Moher and colleagues (2007) conducted a metareview investigating the quality of systematic reviews and meta-analyses of the medical literature. The aim of the metareview was to identify a cross-sectional sample of recently published reviews and to examine their epidemiological, descriptive, and reporting characteristics. The authors systematically searched for reviews that were published in English and

indexed on the electronic database, MEDLINE, in November of 2004. Using pre-specified inclusion and exclusion criteria, 139 systematic reviews and 161 meta-analyses were identified. Information was extracted on methodological characteristics of these reviews. In addition, the authors described the journals in which published reviews appeared (e.g., impact factors). A random sample of the reviews was chosen and coded by a second examiner to test inter-rater reliability.

The metareview concluded that the quality and consistency of systematic reviews of the medical literature vary considerably. Review findings were often conflicting, even when the same literature was reviewed. Over one-third of the included reviews did not include replicable search strategies, did not exclude duplicate studies or overlapping samples, did not investigate sources of heterogeneity, and did not assess evidence of publication bias. The authors also discussed the limitations of the included reviews and noted a series of high quality reviews that could be used as models for the field.

In summary, metareview is a promising new methodology that can help to critically appraise the quality of review literatures and the consistency of their findings. As a number of reviews have been published on forensic risk assessment, conducting a metareview of this literature may be beneficial in identifying uncertainties that merit further exploration.

2.2.2 Objectives

In the present study, a metareview of previous systematic reviews and meta-analyses in the field of forensic risk assessment was conducted. The primary goal of the metareview was to descriptively examine the review literature on risk assessment

to identify major uncertainties. Specifically, the aim was to compare previous reviews' methods of reporting results with recommended standards in the field of systematic reviewing and to descriptively analyse the methodological quality of these reviews. Finally, this study aimed to describe the broad findings of these systematic reviews and meta-analyses and to highlight areas in the field that could benefit from further research.

2.3 METHOD

2.3.1 Definitions

Articles included in this descriptive review were either systematic reviews or meta-analyses of the forensic risk assessment literature. The forensic risk assessment literature was defined as that concerned with calculating the likelihood of criminal, violent, or sexual offending (Kemshall, 1996). Articles were considered systematic or meta-analytic reviews if they descriptively or quantitatively summarised the findings of previous studies that had been located through the use of a predetermined search strategy.

2.3.2 Systematic Search

A systematic search was conducted using the following electronic databases: PsycINFO, MEDLINE, EMBASE, and the US National Criminal Justice Reference Service Abstracts. These databases were selected as they index reviews of the psychological (PsycINFO), medical (EMBASE and MEDLINE), and criminological (US National Criminal Justice Reference Service Abstracts) literatures. The following Boolean criteria (where significant phrases and keywords are combined using

standard operators) were used to search each of the databases: risk assessment AND (meta-analysis OR systematic review). The search was restricted to articles that had been published between January 1, 1995 and June 30, 2009, because the intention was for the metareview to provide a summary of the contemporary systematic review and meta-analytic literature. Additional works were located through the reference sections of previously located reviews and by communication with researchers in the field.

Reviews in all languages were considered for inclusion as were reviews not published in peer-review journals (i.e., government reports, conference presentations, Master's theses, and doctoral dissertations). Reviews were included if their titles, abstracts, and/or *Methods* sections revealed evidence that the work was a systematic review or meta-analysis of the forensic risk assessment literature. Narrative reviews (e.g., Beech, Fisher, & Thornton, 2003; Daffern, 2007; Edens, Skeem, Cruise, & Cauffman, 2001), editorials, and primary studies were excluded.

Reviews whose authors stated that their work was an attempt to update the literature used in a previous review were included, but the specified older review was not. For example, Hanson and Morton-Bourgon's (2004) meta-analysis on predictors of sexual offending was an update of two previous meta-analyses on sexual offender recidivism (Hanson & Bussière, 1996, 1998). Thus, the latter two reviews were excluded from the metareview.

Government reports (e.g., Campbell et al., 2007; Gendreau et al., 1996; Hanson & Morton-Bourgon, 2004, 2007; McCann, 2006) that had been subsequently published in a more selective form (e.g., Campbell, French, & Gendreau, 2009; Gendreau, Little, & Goggin, 1996; Hanson & Morton-Bourgon, 2005, 2009; McCann

& Lussier, 2008) were included, while their journal versions, which generally contained a narrower range of analyses, were not.[4]

The initial search identified a total of 935 records (Figure 2.1). When the records' abstracts were scrutinised to see whether they showed evidence of having reviewed the forensic risk assessment literature, the number of records was reduced to 132. When editorials, primary studies, narrative reviews, older versions of systematic reviews or meta-analyses that had been updated, and published versions of government reports were excluded, a total of 40 reviews remained: 9 systematic reviews and 31 meta-analyses.

2.3.3 Data Extraction

The format for this descriptive review was derived from Moher and colleagues (2007), who recently conducted a metareview to investigate the methodological quality of the medical review literature. Using this format, 41 epidemiological, descriptive, and reporting characteristics of the reviews were extracted. When information was unclear or seemingly conflicting, my supervisor, Dr. Seena Fazel, was consulted. When no consensus could be reached or information was missing, it was coded as such.

2.3.4 Inter-rater Reliability and Quality Assessment

As a measure of quality control, 5 (12.5%) of the included reviews were randomly selected and epidemiological, descriptive, and reporting characteristics were

[4] An exception was made for Hanson and Morton-Bourgon's (2005) journal article based on their government report concerning predictors of sexual recidivism (Hanson & Morton-Bourgon, 2004). The journal version contained a number of new analyses, including the comparison of the predictive validity estimates produced by clinically based risk assessment tools versus offence history and phallometric assessments. The journal article did not, however, contain several main findings of the government report (e.g., the risk factors that were most highly associated with sexual recidivism). Therefore, both of these reviews were included.

coded by a research assistant with an undergraduate degree in psychology working independently of the university (Ms. Sophie Westwood). The research assistant was provided with a set of coding sheets (Appendix A) and the randomly selected reviews. Inter-rater agreement was measured using Cohen's (1960) kappa coefficient. Kappa is a chance-corrected coefficient of inter-rater agreement between categorical classifications (Cohen, 1960). To calculate the kappa coefficient, each of the 41 characteristics for which data was extracted was coded into a binary category: agreements between the raters were coded as +1 and non-agreements as +0. Kappa was then calculated using the following equation:

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

where P_o was the percentage of observed agreement and P_e was the chance-expected agreement (0.50). Using this method, a high level of agreement was established (κ = 0.91; Landis & Koch, 1977). Disagreements mostly concerned the reporting of review flow and exclusion criteria. These disagreements likely arose because such information was sometimes available in figures but not in the text. After discussion with Dr. Seena Fazel, a consensus was reached on all disagreements.

The reporting quality of the included reviews was measured using the PRISMA Statement (Moher et al., 2009).[5] Also considered were the Quality of Reporting of Meta-analyses (QUOROM) Statement (Moher et al., 1999) and the Meta-analysis of Observational Studies in Epidemiology (MOOSE) Statement (Stroup et al., 2000). The PRISMA Statement was chosen as it was the most comprehensive

[5] PRISMA items relating to statistical synthesis (items 13, 14, 16, 21, and 23) were scored positively for all nine systematic reviews as they indicated that no quantitative analyses were conducted.

checklist that had published guidelines (Liberati et al., 2009) and did not presume that reviews were of randomised controlled trials (Appendix B).

2.4 RESULTS

2.4.1 Epidemiological Characteristics

Including duplicates across reviews, a total of 2,232 studies were included in the 40 reviews (Table 2.1). Articles appeared in 22 different journals, 19 of which were specialised. The journals that published the most reviews were *Criminal Justice & Behavior* (*N reviews* = 5; 12.5%) and *Law & Human Behavior* (*N* = 5; 12.5%). Reviews had often been cited between 1 and 20 times (*N* = 13; 32.5%) or more than 100 times (*N* = 5; 12.5%). The most cited reviews are shown in Table 2.2. Regarding the year of publication,[6] a trend was found such that more reviews have been published in recent years ($\chi^2_{trend}[1, N = 40] = 7.02$, $p = 0.01$). A mean of 3 (*SD* = 2) reviews were published each year between 1995 and 2009. Corresponding authors were based primarily in the United States (*N* = 19; 47.5%), Canada (*N* = 12; 30.0%), or the United Kingdom (*N* = 6; 15.0%).

2.4.2 Descriptive Characteristics

Of the included reviews, 3 (7.5%) were updates (Table 2.3). These reviews were by different authors. Reviews included a median of 25 studies (*interquartile range* [*IQR*] = 13-61). A total of 126 different risk assessment tools were investigated in the 40 reviews. The most commonly investigated tools were the Psychopathy Checklist measures: the Psychopathy Checklist – Revised (PCL-R; Hare, 1991, 2003;

[6] The year of publication refers to the year in which a review was made available (e.g., journal article printed, government report disseminated, conference presentation given, Master's thesis or doctoral dissertation submitted) as opposed to the year in which a review appeared in a published format.

N = 15; 37.5%), the Psychopathy Checklist (PCL; Hare, 1985; *N* = 9; 22.5%), the Psychopathy Checklist: Screening Version (PCL:SV; Hart, Cox, & Hare, 1995; *N* = 8; 20.0%), and the Psychopathy Checklist: Youth Version (PCL:YV; Forth, Kosson, & Hare, 2003; *N* = 6; 15.0%). Other commonly investigated instruments included the Historical, Clinical, Risk Management – 20 (HCR-20; Webster, Eaves, Douglas, & Wintrup, 1995; Webster et al., 1997; *N* = 6; 15.0%) and the Level of Service Inventory – Revised (LSI-R; Andrews & Bonta, 1995; *N* = 6; 15.0%). More than half of the reviews (*N* = 23; 57.5%) included only samples of offenders. Of the reviews that included only offenders, 5 (12.5%) included only violent offenders, 1 (2.5%) included only non-violent offenders, 13 (32.5%) included both violent and non-violent offenders, and it was unstated or unclear what category of offenders was included in 4 (10.0%) of the reviews. The majority of the reviews (*N* = 28; 70.0%) included participants from a mixture of settings, including prisons, psychiatric units, and the community.

2.4.3 Reporting Characteristics

Data were extracted on four categories of reporting characteristics: (1) eligibility criteria, (2) search characteristics, (3) results characteristics, and (4) discussion characteristics (Table 2.4).

2.4.3.1 Eligibility Criteria

Regarding inclusion criteria, 17 (42.5%) of the reviews specified that they would only include studies with a specific type of offender (e.g., violent offenders). Only prospective studies were included in 15 (37.5%) of the reviews. Five (12.5%) reviews specified that only studies published in English would be incorporated in the

analyses. The majority of the reviews (N = 34; 85.0%) did not report any language inclusion criteria. Of the included reviews, 26 (65.0%) included grey literature such as government reports, conference presentations, Master's theses, and doctoral dissertations.

2.4.3.2 Search Characteristics

A median of 3 (IQR = 2-5) databases were searched in each review. The years covered by the systematic search were reported in 19 (47.5%) reviews, whereas only the beginning or end year of the search was reported in 6 (15.0%) investigations. The other 15 (37.5%) reviews did not report the years that the search covered. Regarding the criteria used in the systematic search of databases, the majority of the reviews (N = 29; 72.5%) specified the keywords that were used. Of the included reviews, 21 (52.5%) addressed whether their included study samples overlapped.

2.4.3.3 Results Characteristics

There are three stages to a systematic search: (1) the initial search using keywords or Boolean criteria, (2) the secondary search, wherein citations are scrutinised to see whether they are relevant to the issue being investigated, and (3) the final search, in which inclusion criteria are applied and duplicate records as well as studies that could not be obtained are excluded (Moher et al., 2009). These three stages typically produce three different record counts. All three record counts were included in 7 (17.5%) reviews. Of the 40 reviews, 12 (30.0%) assessed the quality of their included studies.

Outcome measures used to report findings in the meta-analyses included correlation coefficients (k *cases* = 34), measures of heterogeneity (k = 30),

standardised distribution effect sizes (e.g., z, t, F; $k = 21$), comparison of means effect sizes (e.g., Cohen's d, Glass' Δ, Hedges' g; $k = 20$), area under the curve (AUC; $k = 6$), regression coefficients ($k = 5$), relative improvement over chance (RIOC; $k = 2$), positive predictive value (PPV; $k = 2$), negative predictive value (NPV; $k = 1$), number needed to detain (NND; $k = 1$), and sensitivity and specificity ($k = 1$) (Table 2.5).[7] Eleven (27.5%) of the meta-analyses explicitly reported having used random effects models. Six (15.0%) of these meta-analyses also calculated summary effect estimates using fixed effects models, though these results were reported in only four of the manuscripts. (For a discussion of fixed effects and random effects models, see Chapter III, Section 3.3.5.2.)

Evidence of heterogeneity was formally assessed in 27 (65.9%) of the included reviews.[8] Heterogeneity was measured using the Q statistic (Cochran, 1954) in 24 (60.0%) reviews, 4 (10.0%) of which also reported the I^2 statistic (Higgins & Thompson, 2002). The χ^2 statistic was used to report evidence of heterogeneity in an additional 2 (5.0%) reviews. One review (2.5%; Buchanan & Leese, 2001) reported that studies' sensitivities and specificities were heterogeneous at the $p = 0.04$ level but did not report which statistic was used to derive this value. Of the 27 meta-analyses that reported investigations of heterogeneity, 24 reported having found significant evidence of heterogeneity. In 8 (20.0%) reviews, heterogeneity was investigated with the explicit purpose of identifying outliers (i.e., individual effect estimates that may have skewed findings). Studies included in these reviews were classified as outliers if their removal reduced the value of Q by 50%. These outliers were removed before subsequent analyses were conducted. Reviewers investigated sources of heterogeneity by systematically testing the statistical influence of moderating variables (e.g., gender,

[7] In some cases, reviews included multiple statistics from the same family of effects sizes (e.g., both r and Φ as correlation coefficients).

ethnicity, age, type of offending, length of follow-up) on effect size in 20 (50.0%) of the reviews.

Publication bias was formally assessed in 13 (32.5%) of the meta-analyses.[8] Rosenthal's (1979) Fail-safe *N* was calculated in 5 (12.5%) reviews (Table 2.5), and subgroup analysis was used to compare the effect sizes of published and unpublished studies in 8 (20.0%). Of those reviews that investigated publication bias using subgroup analysis, 7 (17.5%) defined a published study as one which appeared in a peer-review journal, while the eighth (2.5%; Ægisdóttir et al., 2006) defined a published study as one which appeared in an APA-affiliated journal. None of the subgroup analyses found evidence of publication bias (i.e., all 95% confidence intervals [CI] overlapped).[9]

2.4.3.4 Discussion Characteristics

Main findings were summarised in the *Discussion* sections of 33 (82.5%) of the investigations. Of the included reviews, 31 (77.5%) discussed the relevance of their findings for researchers, clinicians, and policymakers. Limitations at the review, study, and outcome levels were discussed by 16 (40.0%) reviews. Finally, a general interpretation of the results in the context of other evidence and implications for future research were reported in 22 (55.0%) reviews.

2.4.4 Reporting Quality

The average review fulfilled 18 (*SD* = 3) of the 27 criteria on the PRISMA Statement. The most commonly satisfied PRISMA criteria were the incorporation of a

[8] In addition, one of the included systematic reviews (Lund, 2000) theorised reasons for heterogeneity in findings and descriptively investigated whether published studies reported different results than unpublished studies.

[9] The meta-analysis by Edens and colleagues (2007) initially found evidence of publication bias when general aggression was used as the outcome. However, when the largest study was removed from the meta-analysis, no evidence of publication bias was found using any outcome criterion.

structured abstract (*N* = 39; 97.5%), the specification of inclusion and exclusion criteria (*N* = 29; 72.5%), and the summarisation of main findings (including the strength of evidence for each main outcome) in the *Discussion* section (*N* = 33; 82.5%). The least commonly met PRISMA criteria included clarifying whether a review protocol existed and where it could be accessed (*N* = 0; 0.0%), reporting whether funding had been received for the project (*N* = 8; 20.0%), and specifying whether study quality had been investigated (*N* = 12; 30.0%).

2.4.5 Areas Covered by the Reviews

For each meta-analysis, a list was compiled of the analyses that were conducted. Analyses of the same variables/areas were grouped together into themes. As a result of this strategy, six broad themes emerged: (1) the comparison of risk assessment tool validity, (2) the analysis of individual risk assessment schemes, (3) the comparison of actuarial tools with unstructured and structured clinical judgement, (4) the comparison of tool validity in different sample demographics, (5) the comparison of predictive validity findings produced using different study designs, and (6) the comparison of relative effect sizes for individual risk factors for recidivism. The topics descriptively investigated by the included systematic reviews all conformed to one of these themes.

2.4.5.1 The Comparison of Risk Assessment Tool Validity

Of the included reviews, 10 (25.0%) compared the predictive validity of two or more risk assessment instruments (Table 2.6). All 10 reviews compared at least one of the Psychopathy Checklist measures (i.e., PCL, PCL-R, PCL:SV, and PCL:YV)

with other tools. Three of the reviews found that a Psychopathy Checklist measure was better in predicting offending than a competing measure, whereas seven did not:

- The predictive validity of the PCL, PCL-R, PCL:SV, and PCL:YV was compared to that of the Lifestyle Criminality Screening Form (LCSF; Walters, White, & Denney, 1991), and no significant differences were found (Walters, 2003b).

- When the PCL, PCL-R, and the PCL:YV were compared to the Youth Level of Service/Case Management Inventory (YLS/CMI; Hoge & Andrews, 2002), no significant differences in predictive validity were demonstrated (Edens et al., 2007).

- Olver, Stockdale, and Wormith (2009) investigated the predictive validity of the PCL:YV, the YLS/CMI, and the SAVRY and found no significant differences.

- Campbell et al. (2007) compared the predictive abilities of the PCL, PCL-R, PCL:SV, HCR-20, LSI (Andrews, 1982), LSI-R, the Violence Risk Appraisal Guide (VRAG; Quinsey et al., 1998, 2006), and the Statistical Information on Recidivism scale (SIR; Nuffield, 1982) and found no significant differences between the tools.

- Schwalbe (2007) compared the predictive validity of the PCL:SV to that of the YLS/CMI, the North Carolina Assessment of Risk (NCAR; Schwalbe, Fraser,

Day, & Arnold, 2004), and the Orange Country Risk Assessment (OCRA; Orange County Probation Department, 1988). No significant differences in predictive validity were found.

- Gendreau and colleagues (2002) found that there was no significant difference between the predictive validity of the PCL-R and the LSI-R.

- Hemphill, Hare, and Wong (1998) concluded that the predictive validity of the PCL-R was significantly higher than that of the LSI-R.

- Gendreau and colleagues (1996) compared the predictive validity of the PCL-R to that of the LSI-R, the Salient Factor Score (SFS; Hoffman, 1983), the Wisconsin Classification System (Baird, 1981), and the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1967). The PCL-R's predictive validity was found to be significantly higher than that of the MMPI. All other comparisons were non-significant.

- Walters (2006) compared the predictive validity of 7 risk assessment tools (HCR-20, LCSF, LSI, LSI-R, PCL, PCL-R, and VRAG) with that of 13 self-report measures.[10] The risk assessment tools were found to produce higher

[10] Included self-report measures were the Buss-Durkee Hostility Inventory (BDHI; Buss & Durkee, 1957), Beck Hopelessness Scale (BHS; Beck, Weissman, Lester, & Trexler, 1974), California Psychological Inventory – Socialization Scale (CPI-SO; Gough, 1957), Criminal Sentiments Scale (CSS; Andrews & Wormith, 1984), Multidimensional Anger Inventory (MAI; Siegel, 1986), Michigan Alcoholism Screening Test (MAST; Seltzer, 1971), Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1967), Novaco Anger Scale (NAS; Novaco, 1994), Neuroticism-Extroversion-Openness Personality Inventory – Revised (NEO-PI-R; Costa & McCrae, 1992), Personality Assessment Inventory (PAI; Morey, 1991), Psychological Inventory of Criminal Thinking Styles (PICTS; Walters, 1995), Peterson, Quay, and Cameron Psychopathy Scale (PQC; Peterson, Quay, & Cameron, 1959), and the Self-Appraisal Questionnaire (SAQ; Loza, Dhaliwal, Kroner, & Loza-Fanous, 2000).

rates of predictive validity than the self-report measures. The review suggested that this result may have been attributable to item content on the self-report measures which was not relevant to criminal attitudes. Subsequent analyses revealed that self-report measures whose content was restricted to items measuring criminal attitudes produced non-significantly different rates of predictive validity when compared with risk assessment tools.

- A systematic review conducted by Dolan and Doyle (2000) found evidence that the PCL-R, VRAG, HCR-20, and the Dangerous Behavior Rating Scheme (DBRS; Menzies, Webster, & Sepejak, 1985) were valid predictors of violence.

2.4.5.2 The Analysis of Individual Risk Assessment Schemes

Six (15.0%) meta-analyses investigated the predictive validity of a single risk assessment scheme. Each review reported positive significant findings for the scheme investigated:

- The relationship between of the Psychopathy Checklist measures and both general and violent recidivism was explored by two meta-analyses (Leistico et al., 2008; Salekin et al., 1996). Both reviews reported significant positive associations between scores on the PCL measures and recidivism.

- Walters (2003a) examined the ability of the Psychopathy Checklist measures to predict both offending in the community as well as institutional incidents and found the tools to be valid predictors of both.

- The Psychopathy Checklist measures' abilities to predict institutional incidents were investigated by Guy, Edens, Anthony, and Douglas (2005) as well as Edens and Campbell (2007) with positive significant results.

- Nikolova and colleagues (2006) meta-analytically investigated the utility of the HCR-20. The researchers concluded that the HCR-20 was a valid predictor of violence in both men and women in both psychiatric and correctional samples across study designs.

2.4.5.3 The Comparison of Actuarial Tools with Clinical Judgement

The question of which form of risk assessment – actuarial or clinically based – produces higher rates of predictive validity was explored by 7 (17.5%) of the reviews (Table 2.7). Two reviews concluded that actuarial instruments were more accurate than tools that employ structured clinical judgement (Hanson & Morton-Bourgon, 2004, 2007). Three reviews found actuarial measures to be superior to clinical judgement but did not dichotomise the latter into unstructured and structured assessment (Ægisdóttir et al., 2006; Buchanan & Leese, 2001; Grove, Zald, Lebow, Snitz, & Nelson, 2000). The sixth relevant review (Guy, 2008) found that no distinct advantage existed for actuarial tools as opposed to measures that employed structured clinical judgement. Finally, Hanson and Morton-Bourgon (2005) compared the predictive validity of tools that employ structured clinical judgement with phallometric assessments and offence history and found no significant differences.

2.4.5.4 The Comparison of Tool Validity in Different Sample Demographics

Gender

The comparative utility of risk assessment tools for men and women was investigated in 14 (35.0%) reviews. Two meta-analyses found that gender significantly affected rates of predictive validity. Leistico and colleagues (2008) investigated the moderating role of gender on the predictive validity of the Psychopathy Checklist measures and found that the instruments produced larger effect sizes in samples that included more female participants. In contrast, Edens et al. (2007) found that the PCL, PCL-R, and the PCL:YV predicted violent recidivism more accurately in men than in women. In addition to these meta-analyses, a systematic review conducted by Holtfreter and Cupp (2007) suggested that the predictive validity of the LSI-R differs in men and women because the items incorporated in the assessment are more sensitive to detecting risk in men. Gender was not found to moderate predictive validity in 11 meta-analyses on tools including the Psychopathy Checklist measures and the LSI-R (Gendreau et al., 2002; Guy, 2008; Guy, Edens, Anthony, & Douglas, 2005; McCann, 2006; Olver et al., 2009; Schwalbe, 2007, 2008; Skeem, Edens, Camp, & Colwell, 2004; Smith, Cullen, & Latessa, 2009; Walters, 2003a, 2006).

Ethnicity

Seven (17.5%) meta-analyses investigated whether participant ethnicity moderated effect size. Two meta-analyses concerning the Psychopathy Checklist measures found that the higher the proportion of white participants in a sample, the higher the predictive validity (Edens et al., 2007; Leistico et al., 2008), while three other meta-analyses on these instruments found no evidence of ethnic differences

(Edens & Campbell, 2007; Guy et al., 2005; Skeem et al., 2004). The remaining two reviews concerned the efficacy of risk assessment tools designed for juveniles and concluded that instruments performed comparably in white and non-white participants (Olver et al., 2009; Schwalbe, 2007). In the two reviews that specified multiple minority groups rather than using a general "non-white" category (Edens & Campbell, 2007; Guy et al., 2005), no pairwise comparisons were conducted.

Age

None of the 8 (20.0%) meta-analyses that investigated the moderating role of participant age found evidence that the variable significantly influenced effect size (Blair, Marcus, & Boccaccini, 2008; Edens & Campbell, 2007; Guy, 2008; Leistico et al., 2008; McCann, 2006; Skeem et al., 2004; Walters, 2003a, 2006). However, 3 (7.5%) meta-analyses did find that younger age was associated with higher rates of recidivism (Cottle et al., 2001; Gendreau et al., 1996; McCann, 2006), suggesting that instruments' PPVs might be higher for risk assessment tools when applied to younger participants.

Base Rate of Offending

The influence of samples' base rates of offending on their predictive validity findings was investigated by 4 (10.0%) reviews. These meta-analyses measured predictive validity using PPVs, NPVs, or point-biserial correlation coefficients, all of which are base rate dependent. In a meta-analysis of the clinical prediction literature on psychiatric patients, Bjørkly (1995) reported significantly higher PPVs for samples with higher base rates of offending and significantly higher NPVs for samples with lower base rates of offending. A review by Schwalbe (2008) on risk assessment tools

designed for use with juveniles found that higher base rates of recidivism were associated with larger correlation coefficients. However, no *p*-values or confidence intervals were reported to assess the statistical significance of this finding. The remaining two meta-analyses found that a sample's base rate of offending did not significantly influence effect size (Blair et al., 2008; Edens et al., 2007).

2.4.5.5 The Comparison of Tool Validity across Study Designs

Study Setting

The utility of risk assessment tools in different study settings was investigated in 10 (25.0%) reviews. Two meta-analyses concluded that study setting affected rates of predictive validity. Leistico and colleagues (2008) found that the Psychopathy Checklist measures' effect sizes were larger in psychiatric (forensic or general) samples than in correctional (prison, jail, detention) samples. Guy and colleagues (2005) found that the PCL, PCL-R, and PCL:SV performed significantly differently in prison and psychiatric samples. However, the direction of this effect was not reported. Three meta-analyses (including one on the Psychopathy Checklist measures) found that predictive validity did not differ between psychiatric and non-psychiatric populations (Blair et al., 2008; Guy, 2008; Skeem et al., 2004).

A meta-analysis of the prospective clinical prediction literature in psychiatric patients reported a mean PPV of 0.32 for studies with short follow-up periods (≤1 week) and 0.45 for studies with longer follow-up periods (>1 week) (Bjørkly, 1995). Studies with short-term follow-up periods produced an average NPV of 0.89, while studies with longer follow-ups produced an average NPV of 0.84.

The utility of risk measures in psychiatric settings was further investigated in four systematic reviews (Blank, 2001; Kumar & Simpson, 2005; Turgut, Lagace,

Izmir, & Dursun, 2006; Woods & Ashley, 2007). Although none of these reviews quantitatively synthesised the literature, they all concluded that risk assessment in psychiatric populations was a topic of considerable importance to the field.

Temporal Design

Four (10.0%) meta-analyses investigated whether a study was prospectively or retrospectively designed influenced effect size. In a meta-analysis comparing the predictive validity of actuarial and SCJ tools, Guy (2008) reported that retrospective studies produced higher predictive validity estimates than prospective studies. In support, Leistico and colleagues (2008) found that retrospective investigations of the predictive validity of the Psychopathy Checklist measures produced larger effect sizes than prospective ones. Evidence that temporal design did not influence effect size was provided by two meta-analyses that investigated the performance of actuarial measures (Blair et al., 2008; Guy et al., 2005).

Length of Follow-up

Seven (17.5%) reviews investigated the moderating role of length of follow-up on predictive validity. In their review of the Psychopathy Checklist measures, Leistico and colleagues (2008) found that samples with longer follow-up periods produced significantly larger effect sizes than samples with shorter follow-up periods. Two other meta-analyses of the Psychopathy Checklist measures found that length of follow-up did not influence predictive validity estimates (Edens & Campbell, 2007; Edens et al., 2007). In a meta-analysis on the predictive validity of risk assessments designed for use with juveniles, Schwalbe (2008) reported that shorter lengths of follow-up were associated with larger effect sizes, but no *p*-values or confidence

intervals were reported to assess the statistical significance of this finding. A second review on the juvenile risk assessment literature by Schwalbe (2007) found that length of follow-up did not moderate effect size. In their review of the LSI-R literature, Smith and colleagues (2009) found that samples with shorter periods of follow-up produced significantly larger effect sizes than samples with longer periods of follow-up. Finally, Blair and colleagues (2008) found that length of follow-up did not moderate the effect sizes produced by actuarial instruments.

Sample Size

None of the 5 (12.5%) meta-analyses that investigated the moderating role of sample size found that the variable significantly influenced predictive validity (Blair et al., 2008; Grove et al., 2000; Hanson & Morton-Bourgon, 2007; Schwalbe, 2007, 2008).[11]

Country of Origin

Four (10.0%) meta-analyses found that the country in which a study was conducted influenced its predictive validity findings. In a meta-analysis on the performance of the Psychopathy Checklist measures, Leistico and colleagues (2008) reported that studies from Canada and Europe produced larger effect sizes studies conducted in the United States. Olver and colleagues (2009) reported that the PCL:YV predicted general recidivism with higher accuracy in Canadian studies than in non-Canadian studies. Guy and colleagues (2005) found that US studies of the PCL, PCL-R, and PCL:SV produced larger effect sizes than non-US studies investigating the same tools. In their review on the accuracy of risk assessments for

[11] While Schwalbe (2007, 2008) reported some evidence that larger effect sizes were produced by juvenile risk assessment tools when used with smaller samples, no *p*-values or confidence intervals were reported to assess the statistical significance of these findings.

sexual offenders, Hanson and Morton-Bourgon (2007) found that studies conducted in the UK produced larger effect sizes than studies from the US or Canada. Three meta-analyses compared the effect sizes produced by risk assessment tools in North American studies versus European studies and found no significant differences (Blair et al., 2008; Edens et al., 2007; Guy, 2008).

Type of Offending

Whether risk assessment tools predicted the likelihood of certain types of offending better than others was explored by 7 (17.5%) reviews. Gendreau and colleagues (2002) found that the LSI-R produced higher rates of predictive validity when used to assess the risk of general offending as opposed to violent offending. Six meta-analyses found that the type of offending being predicted by a risk assessment tool did not affect rates of predictive validity. Two of these reviews dichotomised type of offending into violent versus non-violent (Leistico et al., 2008; Walters, 2003a), two as general versus violent (Hemphill et al., 1998; Olver et al., 2009), and two reviews divided types of offending into five subcategories: sexually violent offending, non-sexually violent offending, non-sexually non-violent offending, general offending, and institutional misconduct (Blair et al., 2008; Walters, 2006).

Type of Outcome

Five (12.5%) reviews investigated whether risk assessment tools were able to predict certain outcomes better than others. In a meta-analysis of the predictive validity of risk assessment tools designed for use with juveniles, Schwalbe (2008) reported that samples that used arrest or referral as their outcomes produced significantly larger effect estimates than samples that used incarceration as their

outcome. Campbell and colleagues (2007) found that the PCL and the PCL-R performed better in samples that defined recidivism as institutional violence as opposed to violence in the community. In their review of the literature on the PCL-R and the LSI-R, Gendreau et al. (2002) reported that using either conviction or incarceration as the definition of recidivism resulted in different rates of predictive validity. However, no *p*-values or confidence intervals were reported to assess the statistical significance or direction of this finding. Two meta-analyses on Psychopathy Checklist measures found that using institutional misconduct as opposed to offending in the community as the outcome of interest did not influence effect size (Leistico et al., 2008; Walters, 2003a).

Source of Information Used to Score Tool

The moderating influence of the source of information used to score risk assessment tools on effect size was explored in 7 (17.5%) reviews. Three meta-analyses reported that risk assessments based on information from file review alone produced larger effect sizes than assessments based on information from self-report, interview, or a combination of file review and interview (Campbell et al., 2007; Grove et al., 2000; Leistico et al., 2008). In contrast, four meta-analyses found that using different sources of information to score instruments did not moderate effect size (Edens et al., 2007; Gendreau et al., 2002; Guy, 2008; Schwalbe, 2007).

2.4.5.6 The Comparison of Individual Risk Factors for Recidivism

Nine (22.5%) reviews synthesised the evidence on risk factors for recidivism, though the studies were on different categories of offender. Six reviews measured the association of three or more risk factors with recidivism. Three meta-analyses

investigated the strongest risk factors for recidivism among juvenile offenders (Cottle et al., 2001; McCann, 2006; Redlak, 2003). Hanson and Morton-Bourgon (2004) conducted a meta-analysis on predictors of sexual recidivism among adult sex offenders. Gendreau and colleagues (1996) conducted a meta-analysis to broadly investigate risk factors for adult offender recidivism. Finally, a review by Bonta, Law, and Hanson (1998) examined the predictors of general and violent recidivism in mentally disordered offenders. The risk factors reported to be most strongly associated with recidivism in each of these reviews is reported in Table 2.8.

In addition to these six reviews, Lund (2000) conducted a systematic review of a subsample of seven studies measuring the relationship between denial and sexual recidivism that were selected by Hanson and Bussière (1998)[12] as part of their meta-analysis investigating risk factors for sexual offending. The review concluded that heterogeneity in the definition of denial, as well as in populations of interest and study settings, made it difficult to state decisively whether denial predicted sexual recidivism. Finally, two systematic reviews examined the risk factor literature for adolescent sexual offenders and concluded that there are several promising approaches (e.g., the combination of evidence-based risk factors with case-specific clinical considerations) for assessing recidivism risk in this population (Gerhold, Browne, & Beckett, 2007; Worling & Långström, 2003).

2.4.5.7 Other Areas Covered

Two additional topics covered by the included reviews were authorship effects and preventative detention. Blair and colleagues (2008) investigated the existence of an "allegiance effect" (p. 346) such that larger effect sizes are found when an author

[12] The review by Hanson and Bussière (1998) is a predecessor to the meta-analysis of Hanson and Morton-Bourgon (2004).

of a tool is also an author of a study investigating that instrument's predictive validity. Literature on three actuarial tools – the VRAG, the Sex Offender Risk Appraisal Guide (SORAG; Quinsey et al., 1998, 2006), and the Static-99 (Harris, Phenix, Hanson, & Thornton, 2003; Hanson & Thornton, 1999) – was examined, and evidence of a significant authorship effect was found. In contrast, Guy (2008) explored the existence of this authorship bias in both actuarial measures (including the VRAG, SORAG, and Static-99) and tools that employ structured clinical judgement and found no evidence that effect estimates produced by studies where a tool author or translator was also a study author were different from those reported in studies conducted by independent researchers.

Buchanan and Leese (2001) conducted a meta-analysis of the literature concerning the prediction of violent behaviour in adults residing in the community to investigate the feasibility of the Dangerous Severe Personality Disorder (DSPD) Programme, a UK policy initiative that proposed to preventatively detain individuals diagnosed with DSPD who were judged to be at risk of harming others. Using the NND, the review quantitatively synthesised 21 studies in which risk of future violence was assessed. The meta-analysis concluded that an average of six individuals judged to be at high risk of future offending would need to be detained to prevent a single act of violence from occurring in the community.

2.5 DISCUSSION

The aim of this metareview was to identify current uncertainties in the field of forensic risk assessment. Forty systematic reviews and meta-analyses comprising 2,232 studies were identified and the quality and consistency of their findings were descriptively examined. The included reviews came to conflicting conclusions on a

number of issues, including: (1) the comparative predictive validity of risk assessment tools, (2) the efficacy of actuarial tools versus unstructured and structured clinical judgement, (3) the influence of demographic factors on predictive validity, (4) the influence of study design characteristics on predictive validity, and (5) the relative strength of association of individual risk factors for recidivism.

2.5.1 Uncertainties in the Forensic Risk Assessment Literature

2.5.1.1 Uncertainty Concerning the Comparative Predictive Validity of Risk Assessment Tools

Despite the fact that 126 risk assessment tools were investigated in the included reviews, no one measure was consistently found to be more valid than any other. Of the meta-analyses that compared the predictive utility of the Psychopathy Checklist measures with at least one other instrument, almost all found that assessments based on the Psychopathy Checklist (i.e., PCL, PCL-R, PCL:SV, PCL:YV) produced non-significantly different rates of predictive validity than the other tool(s). This suggests that the view of some experts who have, in the past, argued that the Psychopathy Checklist measures are unparalleled in their ability to predict future offending (Hart, 1998; Salekin et al., 1996) should now be reconsidered. However, meta-analyses did indicate that the measures were valid predictors of both institutional incidents and offending in the community. There was a notable lack of meta-analyses on the predictive literature of tools other than the Psychopathy Checklist measures.

Meta-analyses frequently investigated the predictive validity of both widely implemented risk measures (e.g., HCR-20, PCL-R, VRAG) and less commonly used alternatives (e.g., NCAR, OCRA, Wisconsin Classification System). As meta-analyses are often cited to support statements about the general utility of forensic risk

assessment tools, it may be that calculating overall summary effect estimates and investigating sources of heterogeneity using such an inclusive approach has provided researchers, clinicians, and policymakers with a biased view of how the most commonly used risk measures (i.e., those with the largest clinical impact) perform.

2.5.1.2 Uncertainty Concerning the Efficacy of Actuarial versus Clinically Based Risk Assessment

There was conflicting evidence as to whether actuarial instruments predict offending more accurately than clinically based assessments. Five meta-analyses that compared actuarial measures with clinical judgement found that the former produced higher rates of predictive validity. A sixth meta-analysis, however, found no difference in efficacy between actuarial tools and those that employ structured clinical judgement. Although the majority of the reviews favoured actuarial instruments over clinically based tools, this finding should be interpreted in light of the methodological quality of the included reviews.

2.5.1.3 Uncertainty Concerning the Influence of Demographic Factors on Predictive Validity

There was mixed evidence of risk assessment tools' validity in individuals of both genders. One meta-analysis of the PCL measures reported that the more female participants were in a sample, the higher the effect estimate, whereas a second reported that the instruments were more predictively valid in men. Other reviews concluded that the risk assessment tools performed comparably in men and women. As individual risk factors are hypothesised to be different for men and women (Belknap & Holsinger, 2006; Holtfreter & Cupp, 2007), future reviews could assess

gender validity in a wider range of tools, thus broadening the frame of risk factors measured.

Evidence of predictive validity was also inconsistent with regard to participant ethnicity. There was some meta-analytic evidence that the higher the proportion of white participants in a sample, the larger the effect size. However, other reviews concluded that instruments' predictive validity did not vary depending on participants' ethnic backgrounds. Given that many risk assessment tools were calibrated on predominantly white samples but are administered to individuals of a variety of ethnic backgrounds, future reviews may wish to further investigate the moderating role of this demographic variable.

The metareview also identified uncertainty as to whether participant age or the base rate of offending affects the predictive validity of risk assessment instruments. As risk assessment tools are currently used to assess dangerousness in both juvenile justice and adult correctional settings as well as in psychiatric subgroups with markedly different base rates of offending (Stuart & Arboleda-Florez, 2001), the influence of these variables on predictive validity should continue to be measured in future reviews.

2.5.1.4 Uncertainty Concerning the Influence of Study Design Characteristics on Predictive Validity

Previous reviews of the forensic risk assessment literature came to conflicting conclusions as to whether existing risk measures are more or less valid in different settings. As risk instruments are now being increasingly used in psychiatric (Higgins et al., 2005; Khiroya et al., 2009), correctional (Archer et al., 2006; Viljoen et al., 2010), and court settings (DeMatteo & Edens, 2006; Young, 2009), it is important that future reviews investigate the utility of risk assessment tools across contexts.

There was some evidence that length of follow-up moderates effect size. Two reviews concluded that longer periods of follow-up were associated with larger effect sizes, while a third review reported the opposite. Risk assessment tools are being used with increasing regularity in criminal cases when making decisions regarding the length of institutionalisation and length of community supervision (Lyon, Hart, & Webster, 2001; Sreenivasan et al., 2000; Vess, 2008). Therefore, it is important that future meta-analyses investigate this variable to determine whether using risk instruments in such decisions is appropriate.

Whether the country in which a study was conducted influences its predictive validity findings was investigated by seven meta-analyses with mixed findings. Two reviews reported that studies conducted outside of the United States produced higher rates of accuracy than those conducted therein. A third review concluded the opposite. Other meta-analyses compared the effect sizes produced by North American studies with those from European studies and found no significant differences. These findings are complicated by the fact that some tools use different cut-off scores to classify individuals as being at high risk for offending in North America and Europe (Grann, Långström, Tengström, & Kullgren, 1999; Hare, 1991, 2003). When cut-off dependent outcome statistics such as PPV, NPV, NND, and DOR are used, it is important that reviewers quantitatively summarise the results of studies that use the same score threshold, regardless of the country in which the data were collected. This may require future reviewers to request new tabular data from study authors. Taking such precautions is important, because rates of true positives, false positives, true negatives, and false negatives vary depending on which cut-off score is used in a study.

Inconsistent evidence was also found regarding the moderating role of the following variables on predictive validity: temporal design (i.e., prospective or retrospective study design), type of offending (e.g., general versus violent), type of outcome predicted (e.g., arrest, charge, conviction, incarceration, or institutional incident), source of information used to score risk assessment instruments (e.g., file review only versus other), and tool authorship (i.e., variations in effect size attributable to the presence or absence of a tool author as a study author).

2.5.1.5 Uncertainty Concerning the Relative Strength of Risk Factors for Recidivism

Risk factors for recidivism were examined in 10 reviews, and it was found that both static and dynamic factors were linked with future antisocial behaviour. However, it was notable that different risk factors were reported as having the strongest associations with recidivism in each review. Although this may be attributable to heterogeneous populations and outcomes (e.g., general, violent, or sexual recidivism) having been investigated in these reviews, different risk factors were reported even when similar populations and outcomes were examined.

2.5.2 Review Quality

With the average review meeting two-thirds of PRISMA criteria, the present metareview found that many reviews of the field of forensic risk assessment share significant methodological weaknesses. First, only six reviews contained replicable descriptions of their systematic searches (i.e., specified which study characteristics were used for eligibility, described all information sources used in the search, presented the electronic search strategy for at least one database, and reported

numbers for review flow). This suggests that future reviews may wish to include more comprehensive and transparent descriptions of their systematic searches.

Second, in approximately half of the reviews, authors did not specify whether they had ensured that duplicate studies were not included and that the samples from the primary studies making up their review did not overlap. In such cases, the reviews are likely to have overestimated effect sizes.

Third, although significant evidence of heterogeneity was found in over 90% of reviews in which it was investigated, sources of heterogeneity were not assessed in half of the reviews. Broadly, heterogeneity as a construct can be divided into three forms: clinical heterogeneity, methodological heterogeneity, and statistical heterogeneity (Gagnier, Moher, Boon, Beyene, & Bombardier, in press; West et al., 2010). Clinical heterogeneity refers to between-study differences in the type of risk assessment tool being administered (e.g., actuarial versus SCJ), sample demographics (e.g., gender, ethnicity, age, or psychiatric diagnosis), and outcome characteristics (e.g., type of offending being predicted or length of follow-up). Methodological heterogeneity refers to differences in study design (e.g., prospective versus retrospective orientation) and quality (e.g., standardised reporting checklist score). Finally, statistical heterogeneity refers to variability in the effect estimates produced by different studies beyond that expected by chance. These three forms of heterogeneity are interrelated: Clinical and methodological heterogeneity can and do co-occur, and either can result in statistical heterogeneity. None of the reviews included in the metareview investigated all three sources of variation. Given the nature of the field, which investigates tool validity in individuals in different settings from different socio-demographic and offending backgrounds and includes a variety of study methodologies, it could be argued that sources of between-study

heterogeneity should be investigated regardless of the statistical significance of findings.

Fourth, the detection of publication bias was also limited, with only a third of the reviews assessing this. This may have biased results in favour of positive significant findings; however, little evidence of publication bias was found in those reviews that did include such investigations.

Fifth, the statistics used to report on predictive validity in the included reviews had limitations. The effect sizes most frequently used to describe the association between a risk assessment tool and offending were correlation coefficients (commonly the product-moment or point-biserial *r*). Caution is warranted when interpreting the strength of associations using correlation coefficients, as these statistics are base rate dependent (Rice & Harris, 1995). That is, guidelines for the interpretation of these effect sizes only hold at certain base rates of outcome prevalence. For example, an *r* of 0.30 may be large at a base rate of 15% but small at a base rate of 50% (Cohen, 1992). Such issues of interpretation were rarely mentioned, and the influence of base rate on effect size was only explored in four of the included meta-analyses. Thus, conclusions that instrument scores are "highly" correlated with offending should be interpreted with caution.

The AUC is another commonly used outcome statistic that has its limitations. For the purposes of meta-analysis, the usefulness of pooling AUCs is limited in that it does not allow researchers to explore sources of heterogeneity by metaregression, an increasingly important statistical approach that allows for the systematic exploration of the influence of continuous variables on effect size and adjustment for the influence of one moderating variable on another (Thompson & Higgins, 2002). Another single effect indicator, the diagnostic odds ratio (DOR), allows for the systematic

investigation of heterogeneity using metaregression and may be a useful alternative to correlation coefficients and the AUC.

2.5.3 Development of the MARQ Checklist

The main benefit of metareview methodology appears to be its usefulness in conducting thematic analyses to identify major uncertainties that warrant further exploration. By comparing reviews' reporting characteristics against standardised quality checklists such as the PRISMA Statement, metareviews also provide fields with high standards that future systematic reviews and meta-analyses can follow. According to a recently developed checklist designed by expert systematic reviewers (Shea et al., 2007) as well as guidelines published by the Cochrane Collaboration (Becker & Oxman, 2008), metareviews should assess, at a minimum, whether included reviews state their objectives *a priori*, report a reproducible search strategy, include unpublished studies, provide a list of included (and, if applicable, excluded) studies, summarise the sample and design characteristics of included studies, conduct an inter-rater reliability check to assess the consistency of the data extraction process, investigate within-study and between-study heterogeneity, assess evidence of publication bias, and disclose conflicts of interest. As part of this process, metareviews should attempt to develop a standard list of clinical and methodological covariates that subsequent reviews can investigate as potential sources of between-study heterogeneity (West et al., 2010).

To my knowledge, the metareview presented in this chapter was the first use of this novel methodology to investigate the reporting characteristics and findings of a diagnostic accuracy literature. Therefore, to provide a template for future metareviews, I have revised the PRISMA Statement into a modified inventory, the

Metareview Assessment of Reporting Quality (MARQ) Checklist (Appendix C). The purpose of this 21-item checklist is to encourage a transparent and consistent reporting of metareview methodology. Future research could be conducted to assess the face and construct validity as well as the inter-rater reliability of this instrument.

2.5.4 Implications

Given the increasing number of primary studies and uncertainties in the field of forensic risk assessment, the evidence base needs updating on a regular basis. The most commonly cited reviews of the field are between 8 to 14 years old, suggesting that clinicians and policymakers' views of risk assessment may be based on outdated literature. This finding highlights the responsibility of those who publish influential systematic reviews and meta-analyses to publish updates.

The present metareview also suggests that some topics may benefit from new systematic reviews and meta-analyses. For example, meta-analyses comparing the predictive validity of commonly used risk measures (both actuarial and clinically based) other than the Psychopathy Checklist measures are needed. Annotated bibliographies of replication studies conducted on tools such as the Static-99 (Helmus, 2008), the SORAG (Mental Health Centre Penetanguishene, 2009), and the VRAG (Mental Health Centre Penetanguishene, 2009) assessment schemes reveal that there is substantial literature on non-PCL measures that warrants meta-analytic investigation.

Finally, meta-analyses on forensic risk assessment could consider including the number needed to detain as an outcome statistic. This promising effect size enables more informed consideration of the ethical dilemma of unnecessarily detaining individuals whom risk assessment tools predict will become violent.

However, the NND should not be reported alone, as it only measures a tool's ability to identify individuals who will go on to offend (i.e., the ability to make accurate "rule in" decisions). A novel outcome statistic developed as part of my postgraduate work, the *number safely screened* (NSS), may be reported alongside the NND to provide a complementary "rule out" index of tool performance. The NSS calculates the number of individuals who a risk assessment tool judges to be at low risk of offending who can be screened out (e.g., released from correctional institution or discharged from hospital) before a single criminal incident occurs (Table 1.1).

2.5.5 Limitations

The present metareview is limited by being a primarily descriptive review. To date, no accepted statistical methods have been developed to deal with combining the results of individual meta-analyses, which is partly attributable to the potential complexity of overlapping studies and/or samples. Thus, this metareview has some of the same limitations as systematic reviews, in that summary effect estimates could not be calculated and sources of between-review heterogeneity could not be statistically investigated.

A second limitation of the present study is that it was difficult to directly compare the findings of the various reviews, as they compared the predictive validity literatures of different instruments using studies obtained through systematic searches with different inclusion and exclusion criteria. In addition, reviews explored tool utility in different populations and used different outcome measures. It may be that these differences contributed more to disparate findings than review quality. Therefore, this metareview may best be viewed as an investigation into the overall

quality of the review literature on forensic risk assessment and a thematic analysis identifying key uncertainties regarding the use of risk assessment tools.

To address the limitations of the metareview, a potential solution would be to conduct a high-quality meta-analysis of the most commonly used risk assessment tools, in which the uncertainties identified in the metareview are investigated using a variety of outcome measures. This would potentially clarify the general utility of forensic risk assessment tools and the conditions under which the risk measures with the largest clinical impact perform best. In addition, the meta-analysis could serve as a model if the review were to be designed using a standardised checklist to address the issues of low quality identified in the metareview. (Such a comprehensive meta-analysis will be conducted in the next chapter.)

2.5.6 Conclusion

Systematic reviews and meta-analyses of the forensic risk assessment literature have come to conflicting conclusions on key issues and have a number of potentially important limitations, suggesting that their findings should be considered as provisional. Limitations of previous reviews include not providing replicable search strategies, not excluding overlapping samples, not investigating sources of heterogeneity, not assessing the presence of publication bias, and using a narrow range of summary effect sizes. There are some areas that would benefit from systematic review and meta-analysis, in particular, the validity of commonly used risk assessment instruments other than the Psychopathy Checklist measures. Future reviews should attempt to follow standardised guidelines such as the PRISMA Statement in order to address the lack of consistency in review methodology and statistical reporting. In the following chapter, the uncertainties identified and

methodological limitations noted in the metareview will be addressed in a novel meta-analysis.

Table 2.1 Epidemiological Characteristics of Reviews Included in a Metareview of the Forensic Risk Assessment Literature

Category	Subcategory	Number of *N* = 40 (%)
Total number of journals		22
Number of journals publishing	One review	18
	Two reviews	2
	Three reviews	0
	Four reviews or more	2
Journal type	General	3
	Specialty	19
Journal impact factor by review (2 year)	0.0-5.0	12
	5.1-10.0	1
	10.1-15.0	1
	>15	1
	N/A	7
Number of citations	1-20	13 (32.5)
	21-40	3 (7.5)
	41-60	1 (2.5)
	61-80	2 (5.0)
	81-100	1 (2.5)
	>100	5 (12.5)
	N/A	15 (37.5)

Note. N = number of reviews; N/A = not available.

Table 2.2 Five Most Cited Reviews Concerning Forensic Risk Assessment (as of June 2009)

Citation	Citation Count	Mean Annual Citation Count
Grove, Zald, Lebow, Snitz, & Nelson (2000)	179	20
Salekin, Rogers, & Sewell (1996)	238	18
Bonta, Law, & Hanson (1998)	194	18
Hemphill, Hare, & Wong (1998)	194	18
Gendreau, Goggin, & Smith (2002)	84	12

Table 2.3 Descriptive Characteristics of Reviews Included in a Metareview of the Forensic Risk Assessment Literature

Category	Subcategory	Number of *N* = 40 (%)
Use of terms "systematic review" or "meta-analysis" in title or abstract		28 (70.0)
Update of a previous review		3 (7.5)
Number of included studies	Median (*IQR*)	25 (13-61)
Quantitative synthesis performed		31 (77.5)
Tools used	PCL-R	17 (42.5)
	PCL	9 (22.5)
	PCL:SV	8 (20.0)
	PCL:YV	8 (20.0)
	HCR-20	6 (15.0)
	LSI-R	6 (15.0)
Offender status of participants in included samples	Offenders only	23 (57.5)
	Non-offenders only	1 (2.5)
	Both	9 (22.5)
	Unstated/Unclear	7 (17.5)
Population	Prisoners only	0 (0.0)
	Patients only	2 (5.0)
	Community only	1 (2.5)
	Mixed	28 (70.0)
	Unstated/Unclear	9 (22.5)

Note. N = number of reviews; IQR = interquartile range.

Table 2.4 Reporting Characteristics of Reviews Included in a Metareview of the Forensic Risk Assessment Literature

Domain	Category	Subcategory	Number of *N* = 40 (%)
Abstract	Structured summary of review		39 (97.5)
Introduction	Rationale for review provided		35 (87.5)
	Explicit statement of questions addressed by review		29 (72.5)
Eligibility criteria	Eligibility criteria based on study population	Specific offender type	17 (42.5)
		Gender	3 (7.5)
		Other	4 (10.0)
		N/A	16 (40.0)
	Eligibility criteria based on study design	Prospective	15 (37.5)
		Retrospective	0 (0.0)
		Mixed/Unstated	25 (62.5)
	Eligibility criteria based on language of report	English only	5 (12.5)
		All languages considered	1 (2.5)
		Language criteria not reported	34 (85.0)
	Grey literature investigated	No	14 (35.0)
		Yes	26 (65.0)
Search	Number of databases searched	Median (*IQR*)	3 (2-5)
	Number of other sources searched	Median (*IQR*)	3 (2-3)
	All information sources listed	No	7 (17.5)
		Yes	33 (82.5)
	Years of coverage were reported	No	15 (37.5)
		Partially	6 (15.0)
		Yes	19 (47.5)
	Search terms reported for one or more electronic databases	No search terms	7 (17.5)
		Keywords	29 (72.5)
		Full Boolean	3 (7.5)
		Readers referred elsewhere	1 (2.5)
	Duplicates considered	No	19 (47.5)
		Yes	21 (52.5)
Results	Review flow reported	No	25 (62.5)
		Partial	8 (20.0)
		Yes	7 (17.5)
	Publication bias assessed	No	27 (67.5)
		Yes	13 (32.5)
	Quality of studies assessed	No	28 (70.0)
		Yes	12 (30.0)
Discussion	Main findings summarised		33 (82.5)
	Relevance of review for policymakers, clinicians, and researchers discussed		31 (77.5)
	Limitations discussed		16 (40.0)
	General conclusion provided		22 (55.0)

Note. N = number of reviews; N/A = not available; IQR = interquartile range.

Table 2.5 Outcome Measures Reported in Reviews of the Forensic Risk Assessment Literature

Outcome Measure	*k*
Correlation coefficient	34
Heterogeneity	30
Standardised distribution effect size	21
Comparison of means effect size	20
Area under the curve	6
Regression coefficient	5
Fail-safe *N*	5
Relative improvement over chance	2
Positive predictive value	2
Negative predictive value	1
Number needed to detain	1
Sensitivity	1
Specificity	1

Note. k = number of cases.

Table 2.6 Metareview Results: The Comparison of Risk Assessment Tool Validity

Tools	Significant	Non-significant	N/A
PCL, PCL-R, PCL:SV, PCL:YV vs. LCSF		●	
PCL, PCL-R, PCL:YV vs. YLS/CMI		●	
PCL:YV vs. YLS/CMI, SAVRY		●	
PCL, PCL-R, PCL:SV vs. HCR-20, LSI, LSI-R, VRAG, SIR		●	
PCL:YV vs. YLS/CMI, NCAR, OCRA		●	
PCL-R vs. LSI-R		●	
PCL-R vs. LSI-R	● (PCL-R > LSI-R)		
PCL-R vs. LSI-R, SFS, Wisconsin Classification System, MMPI	● (PCL-R > MMPI)		
Risk assessment tools (PCL, PCL-R, HCR-20, LCSF, LSI, LSI-R, VRAG) vs. Self-report measures (BDHI, BHS, CPI-SO, CSS, MAI, MAST, MMPI, NAS, NEO-PI-R, PAI, PICTS, PQC, SAQ)	● (RAT > SRM)		
PCL-R, DBRS, VRAG, HCR-20			● [a]

Note. ● = One review; N/A = not applicable; RAT = risk assessment tools; SRM = self-report measures.

[a] The work of Dolan and Doyle (2000) was a systematic review in which no summary statistics were calculated.

Table 2.7 Metareview Results: Actuarial versus Clinically Based Risk Assessment

Topic	Significant	Non-significant
Actuarial vs. Structured clinical judgement vs. Unstructured clinical judgement	●● (Actuarial > SCJ > UCJ)	
Actuarial vs. Clinical judgement [a]	●●● (Actuarial > Clinical)	
Actuarial vs. Structured clinical judgement		●

Note. ● = One review; SCJ = structured clinical judgement; UCJ = unstructured clinical judgement. The meta-analysis by Hanson and Morton-Bourgon (2005) did not conform to this method of description and was, therefore, omitted from this table.

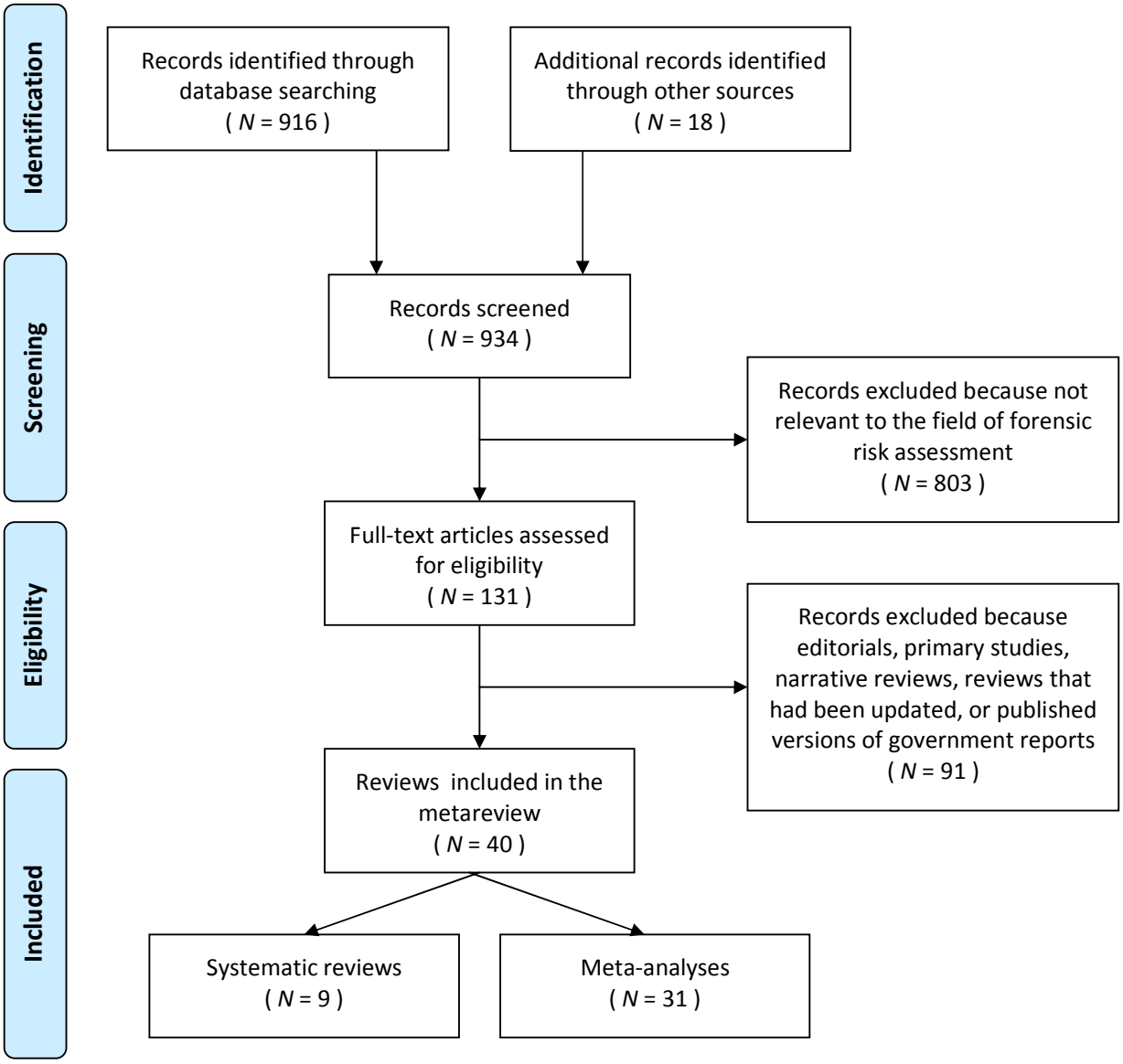
[a] In these reviews (Ægisdóttir et al., 2006; Buchanan & Leese, 2001; Grove et al., 2000), the authors did not dichotomise clinical prediction into unstructured and structured clinical judgement.

Table 2.8 Metareview Results: The Three Risk Factors Most Highly Correlated with Recidivism across Reviews of the Forensic Risk Assessment Literature

Citation	Risk Factors	Population
Gendreau, Goggin, & Little (1996)	1. Score on risk scale 2. Identification/socialization with other offenders 3a. Antisocial personality 3b. Criminogenic needs	Adult offenders
Hanson & Morton-Bourgon (2004)	1. Score on actuarial risk scale 2. General self-regulation problems 3. Violation of conditional release	Sex offenders
McCann (2006)	1. Non-Caucasian 2. Psychopathy 3. Victim of sexual abuse	Juvenile sex offenders
Redlak (2003)	1. Deviant sexuality 2. Having more than one victim 3. Out-of-home placement	Juvenile sex offenders
Cottle, Lee, & Heilbrun (2001)	1. Age at first commitment 2. Age at first contact with the law 3. Non-severe pathology	Juvenile offenders
Bonta, Law, & Hanson (1998)	1. Objective risk assessment 2. Adult criminal history 3. Juvenile delinquency	Mentally disordered offenders

Note. For studies in which the correlation between risk factors and multiple forms of recidivism (e.g., general, violent, sexual) was measured, the most conservative category of offending (i.e., the definition that would produce the highest sensitivity) was used to determine the risk factors appropriate for this table. General offending was considered to be the most conservative definition followed by (in order) violent (including sexual) offending, violent (non-sexual) offending, and sexual offending. Three systematic reviews (Gerhold, Browne, & Beckett, 2007; Lund, 2000; Worling & Långström, 2003) did not conform to this method of description and were, therefore, omitted from this table.

Figure 2.1 Results of a Systematic Search Conducted to Identify Reviews of the Forensic Risk Assessment Literature



***

Chapter III:
A Comparative Meta-analysis of Commonly Used Risk Assessment Tools

3.1 ABSTRACT

There are a large number of structured instruments that assist in the assessment of offending risk, and their use appears to be increasing in mental health and criminal justice settings. However, there remains uncertainty about which commonly used risk assessment tools produce the highest rates of predictive validity, whether actuarial instruments perform better than clinically based measures, and whether overall rates of predictive validity differ by gender, ethnicity, length of follow-up, and other sample- and study-level characteristics. A meta-analysis of nine commonly used risk assessment instruments was conducted following PRISMA guidelines. Data was collected from 68 independent studies based on 25,980 participants in 88 samples. For 54 (61.4%) of the samples, new tabular data was provided directly by authors. Rates of predictive validity were assessed using six outcome statistics. Clinical and methodological between-study heterogeneity were analysed using subgroup analysis and metaregression, and statistical heterogeneity was measured using the I^2 statistic. The review found that forensic risk assessment tools produce similar rates of predictive validity to other prognostic behavioural measures but poor predictive validity compared to medical diagnostic instruments. When risk assessment tools were compared with each other, the Structured Assessment of Violence Risk in Youth (SAVRY), a tool designed to detect violence risk in juveniles, was found to produce the highest rates of predictive validity, while an instrument used to identify adults at risk for general offending, the Level of Service Inventory – Revised (LSI-R), and a personality scale commonly used for the purposes of risk assessment, the Psychopathy

Checklist – Revised (PCL-R), produced the lowest. Instruments produced higher rates of accuracy in older, predominantly white samples and when predicting violent outcomes or institutional incidents. Some evidence of an authorship effect was also found such that studies on which a tool author was also a study author produced higher rates of predictive validity than studies conducted by independent investigators. Risk assessment procedures and guidelines by mental health services and criminal justice systems may need review in light of these findings.

3.2 INTRODUCTION

Many risk assessment tools are currently used to assist in the identification and management of individuals at risk of harmful behaviour. In the previous chapter, it was found that over 120 different risk assessment instruments are currently implemented in correctional and psychiatric settings. These measures range from internationally utilised tools such as the Historical, Clinical, Risk Management – 20 (HCR-20; Webster et al., 1995; Webster et al., 1997) to locally developed and implemented risk measures such as the North Carolina Assessment of Risk (NCAR; Schwalbe et al., 2004). Given the large selection of tools available to general and secure hospitals, prisons, the courts, and other criminal justice settings, a central question is which measures have the highest rates of predictive accuracy. As the metareview concluded, however, no single risk assessment tool has been consistently shown to have superior ability to predict offending, and major uncertainties remain regarding the populations and settings in which structured risk instruments may be accurately used.

Such uncertainties are important given that risk assessment tools have been increasingly used to influence decisions regarding accessibility of inpatient and

outpatient resources, preventative detention, parole and probation, and length of community supervision in many Western countries. Recent work has suggested that the influence of risk assessment tools appears to be growing in both general and forensic settings. For example, risk assessment schemes are now used in more than 70% of general and secure psychiatric hospitals in England (Higgins et al., 2005; Khiroya et al., 2009) and in the majority of the 20 states that have involuntary commitment laws for sex offenders in the United States (Craig & Beech, 2010). Risk measures are also being used with increasing regularity in both criminal and civil court cases in the UK and the US (DeMatteo & Edens, 2006; Young, 2009). The widespread, often legally required use of risk measures necessitates the regular and high-quality review of the evidence base (Seto, 2005).

3.2.1 Uncertainties in Risk Assessment

The research base on the predictive validity of risk assessment tools has expanded considerably in recent decades; however, as concluded in the metareview, policymakers and clinicians continue to be faced with conflicting findings of primary and review literature on a number of central issues (Gendreau et al., 2000). In addition to whether there are differences between the predictive validity of risk assessment instruments, key uncertainties include:

(1) Do actuarial instruments or tools that employ structured clinical judgement produce higher rates of predictive validity?

(2) Do risk assessment tools predict the likelihood of offending with similar validity across demographic backgrounds?

(3) Do aspects of study design influence the predictive validity of risk assessment tools?

3.2.1.1 Actuarial Instruments versus Structured Clinical Judgement

As discussed in Chapter I (Section 1.2.1), there are currently two dominant approaches to forensic risk assessment: the actuarial approach and the structured clinical approach. Actuarial risk assessment tools estimate the likelihood of offending by assigning arithmetic values to factors associated with offending and then combining these values using a statistical algorithm to translate an individual's total score into a probabilistic estimate of the likelihood of antisocial behaviour. Instruments that employ structured clinical judgement are composed of empirically-based risk and protective factors for offending. These factors, commonly organised into scales, are used to guide clinicians' predictions of antisocial behaviour. As found in the previous chapter, while past reviews have provided evidence that actuarial tools produce higher rates of predictive validity than instruments that rely on structured clinical judgement, other researchers have presented evidence that both forms of risk assessment produce equally accurate predictions.

3.2.1.2 Demographic Factors

There is contrasting evidence whether risk assessment tools are equally valid in men and women. Several recent reviews have found no difference in tool performance between the genders (e.g., Schwalbe, 2008; Smith et al., 2009). Schwalbe (2008) conducted a meta-analysis on the validity literature of risk assessment instruments adapted for use in juvenile justice systems and found no differences in predictive validity based on gender. This finding was supported by a

meta-analysis conducted by Smith, Cullen, and Latessa (2009), who found that the Level of Service Inventory – Revised (LSI-R) produced non-significantly different rates of predictive validity in men and women. In contrast, recent meta-analyses have found that the predictive validity of certain risk assessment tools is higher in samples of juvenile men (Edens et al., 2007) or in women (Leistico et al., 2008).

Another uncertainty is whether risk measures' predictive validity differs across ethnic backgrounds. There is evidence from primary studies and meta-analyses that risk assessment tools provide more accurate risk predictions for white participants than for individuals of other ethnic backgrounds (Bhui, 1999; Edens et al., 2007; Långström, 2004; Leistico et al., 2008). This variation may be due to differences in the base rate of offending among individuals of different ethnicities (Federal Bureau of Investigation, 2002). These differences are seen in inpatient settings (Fujii, Tokioka, Lichton, & Hishinuma, 2005; Hoptman et al., 1999; Lawson, Yesavage, & Werner, 1984; McNiel & Binder, 1995; Wang & Diamon, 1999) and upon discharge into the community (Lidz, Mulvey, & Gardner, 1993). Contrary evidence has been provided by reviews that have assessed the moderating influence of ethnicity on predictive validity rates in white, black, Hispanic, Asian, and Aboriginal participants and have found no differences (Edens & Campbell, 2007; Guy et al., 2005).

As reported in the metareview, previous meta-analyses (e.g., Blair et al., 2008; Guy, 2008; Leistico et al., 2008) have found that participant age does not affect the predictive validity of risk assessment tools. However, epidemiological investigations and reviews (e.g., Gendreau et al., 1996) suggest that younger age is a significant risk factor for offending. As risk assessment instruments are routinely used in samples of both males and females, whites and non-whites, and both young and older individuals, it is important to further investigate the influence of gender, ethnicity, and age on predictive validity, particularly for commonly used risk assessment tools.

3.2.1.3 Study Design Characteristics

As identified by the metareview, additional uncertainties in the field of risk assessment include whether study design characteristics such as temporal design, length of follow-up, study setting, country of origin, type of offending, type of outcome, the source of information used to score an instrument, or having a tool author as a study author influence predictive validity.

Temporal Design

Whether a study has a prospective or retrospective design may influence predictive validity findings. Being that the primary goal of risk assessment is to predict future offending, some researchers have stated that prospective research is not just appropriate, but necessary to establish a tool's predictive validity (Caldwell, Bogat, & Davison, 1988). However, a strength of retrospective study designs is that researchers do not have to wait for time to elapse before they investigate whether the studied individuals offended or not. This methodology is particularly useful with low base rate outcomes such as violent crime (Maden, 2001). Both actuarial and clinically based instruments can be used retrospectively. Tools that employ structured clinical judgement (SCJ) can be scored using file information from sources such as psychological reports, institutional records, and/or court reports (de Vogel, de Ruiter, Hildebrand, Bos, & van de Ven, 2004).

Length of Follow-up

The evidence regarding the comparative efficacy of risk assessment tools in predicting offending in the short-term versus the longer-term is mixed. While meta-analytic evidence suggests that risk assessment tools produce higher rates of

predictive validity in the longer-term (Leistico et al., 2008), other reviews (Schwalbe, 2008; Smith et al., 2009) and primary studies (Bauer, Rosca, Khawalled, Gruzniewski, & Grinshpoon, 2003; Sreenivasan et al., 2000) have found the opposite. In addition, several recent meta-analyses have concluded that length of follow-up does not moderate effect size at all (Blair et al., 2008; Edens & Campbell, 2007; Edens et al., 2007; Schwalbe, 2007). Given that studies often follow participants for different lengths of time and given that effect size may vary with time at risk, the potentially moderating role of this variable needs to be examined (Cottle et al., 2001).

Study Setting

Prisons, psychiatric hospitals, and the community are typical settings in most research in the forensic risk assessment literature (Bauer et al., 2003; Bjørkly, 1995; DeMatteo & Edens, 2006; Edens, 2001). Meta-analytic evidence regarding the moderating role of study setting on effect size varies (Leistico et al., 2008; Skeem et al., 2004). Some experts (e.g., Edens et al., 2001) suggest that differences in the accuracy of risk assessments may be attributed to contextual differences in study setting. Due to these differences, measures may predict offending exceptionally well in one setting but poorly in others (Hanson & Morton-Bourgon, 2007). To explore whether risk assessment tools perform differently in correctional, psychiatric, and community settings, this variable will be explored as part of the present meta-analysis.

Country of Origin

As legal systems and rates of resolved criminality differ between countries, the predictive validity estimates produced by risk assessment tools may differ between nations. Previous meta-analyses have found mixed evidence as to whether predictive validity estimates vary depending on the country in which a study is conducted (Blair

et al., 2008; Edens et al., 2007; Guy, 2008; Guy et al., 2005; Hanson & Morton-Bourgon, 2007; Leistico et al., 2008; Olver et al., 2009). Given that these instruments are now routinely used to influence decisions regarding public protection and individual liberty in a number of Western countries, it may be important to further investigate the role of study location as a moderating variable.

Type of Offending

The most commonly used risk assessment tools have been designed to predict general or violent (including sexual) offending (Archer et al., 2006; Khiroya et al., 2009; Viljoen et al., 2010). Risk measures may, however, be used to predict a form of offending for which they were not designed (e.g., the Violence Risk Appraisal Guide [VRAG; Quinsey et al., 1998, 2006], a tool designed to predict violence in mentally disordered offenders, has been used to predict general offending in patients discharged from forensic units [Kroner, Stadtland, Eidt, & Nedopil, 2007]). Though there is some evidence that instruments produce commensurate rates of predictive validity when predicting the likelihood of more versus less severe antisocial behaviour (Blair et al., 2008; Hemphill et al., 1998; Leistico et al., 2008; Olver et al., 2009; Walters, 2003b; Walters, 2006), there is also evidence to suggest that instruments may produce higher rates of predictive validity when predicting different forms of offending (Gendreau et al., 2002).

Type of Outcome

The outcomes that risk assessment tools are most commonly used to predict include arrest, charge, conviction, incarceration, and institutional infractions (Cottle et al., 2001; Schwalbe, 2007). Previous meta-analyses have come to conflicting conclusions as to whether risk assessment tools are better at predicting certain of these

events (Campbell et al., 2007; Leistico et al., 2008). It may be important to investigate this uncertainty as different parties (e.g., mental health professionals, parole boards, policymakers, judges) may be interested in predicting the likelihood of different outcomes.

Source of Information

Sources of information commonly used to score risk assessment instruments include file review (i.e., criminal registers, medical reports, and case notes), interviews, and self-report. While there is evidence that instruments produce commensurate rates of predictive validity regardless of which source(s) is/are used to score an assessment tool (Edens et al., 2007; Gendreau et al., 2002; Guy, 2008; Schwalbe, 2007), there is evidence from reviews on the Psychopathy Checklist measures which suggests that studies in which file information alone is used produce larger effect sizes than those that use self-report, interviews, or some combination of these sources (Campbell et al., 2007; Leistico et al., 2008).

Tool Authorship

Evidence from the pharmaceutical and, more recently, psychiatric literatures that industry sponsorship results in more positive significant research findings has led to a growing demand for transparency regarding conflicts of interest (Bekelman, Li, & Gross, 2003; Lexchin, Bero, Djulbegovic, & Clark, 2003; Perlis et al., 2005). If investigator objectivity may have been compromised due to financial or non-financial interests, the credibility of research findings may be questioned in the absence of clear disclosure (Fava, 2009; Maj, 2008).

A conflict of interest may result when the authors of a risk assessment tool investigate the predictive validity of that instrument (Blair et al., 2008). Tool authors

may have a vested interest in their measure performing well, as such empirical support can lead to both financial benefits (e.g., selling tool manuals and coding sheets, offering training sessions, being hired as an expert witness, attracting funding) as well as non-financial benefits (e.g., increased recognition in the field and more opportunities for career advancement). This may result in an *authorship effect* whereby the authors of a risk assessment tool find more positive significant results when investigating their measure's predictive validity than do independent researchers.

The meta-analytic evidence concerning the existence of an authorship effect is mixed (Blair et al., 2008; Guy, 2008; Harris, Rice, & Quinsey, 2010). Those reviews that have found evidence of an authorship effect have not used multivariate analyses to clarify whether the effect may be explained by other sample- or study-level variables (e.g., differences in the length of follow-up in studies by tool authors versus studies by independent researchers). Such analyses have likely not been conducted due to reliance on outcome statistics such as correlation coefficients and the AUC that do not allow researchers to use metaregression methodology.

3.2.2 Objectives

Despite the increasing use and potential importance of risk assessment instruments, it is unclear which instruments have the highest rates of predictive validity and whether these rates differ by important demographic and study design characteristics. To further investigate these uncertainties using methodology that addresses issues of quality and reporting identified in the metareview, a novel meta-analysis was conducted. The aim of this meta-analysis was to explore the general utility of forensic risk assessment instruments, to examine rates of predictive validity

in commonly used tools and to assess the potential sources of heterogeneity outlined above.

3.3 METHOD

3.3.1 Review Protocol

The Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) Statement (Moher et al., 2009), a 27-item checklist of review characteristics designed to enable a transparent and consistent reporting of results (Appendix B), was followed.

3.3.2 Tool Selection

The goal was to analyse the predictive validity of the most commonly used risk assessment schemes in the field today. Based on surveys of clinicians working in forensic settings (e.g., Archer et al., 2006; Higgins et al., 2005; Khiroya et al., 2009; Viljoen et al., 2010) and reviews of the literature (e.g., Bonta, 2002; Doren, 2002; Kemshall, 2001), the following nine instruments were identified as those most often used in the context of forensic risk assessment (Table 3.1): the *Level of Service Inventory – Revised* (LSI-R; Andrews & Bonta, 1995), the *Psychopathy Checklist – Revised* (PCL-R; Hare, 1991, 2003), the *Violence Risk Appraisal Guide* (VRAG; Quinsey et al., 1998, 2006), the *Sex Offender Risk Appraisal Guide* (SORAG; Quinsey et al., 1998, 2006), the *Static-99* (Harris et al., 2003; Hanson & Thornton, 1999), the *Historical, Clinical, Risk Management – 20* (HCR-20; Webster et al., 1995; Webster et al., 1997), the *Sexual Violence Risk – 20* (SVR-20; Boer et al., 1997), the *Spousal Assault Risk Assessment* (SARA; Kropp, Hart, Webster, & Eaves, 1994, 1995, 1999), and the *Structured Assessment of Violence Risk in Youth* (SAVRY;

Borum, Bartel, & Forth, 2002, 2003). These instruments are a sample of both actuarial and SCJ tools (Appendix D).

Though also found to be commonly used in clinical practice (Archer et al., 2006; Khiroya et al., 2009; Viljoen et al., 2010), adaptations of the PCL-R (e.g., PCL:SV and PCL:YV) were excluded to increase the diversity of the included measures. Meta-analytic evidence suggests that PCL-R findings may be generalisable to the other Psychopathy Checklist measures, as the tools have been found to produce similar rates of predictive validity (Campbell et al., 2007). Though frequently used as part of the risk assessment process (Viljoen et al., 2010), clinical inventories such as the Minnesota Multiphasic Personality Inventory (MMPI; Hathaway & McKinley, 1967) and intelligence scales such as the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1997) were also excluded, as they lack guidance for use in predicting antisocial behaviour.

3.3.2.1 Actuarial Tools

Level of Service Inventory – Revised

The LSI-R was designed to use psychosocial status to predict the likelihood of general recidivism in adult offenders. The tool provides professionals with information that can be used to help make decisions regarding level of supervision and treatment. Information acquired through a semi-structured interview is used to score individuals on 54 items measuring 10 domains: criminal history (10 items), leisure/recreation (2 items), education/employment (10 items), companions (5 items), financial (2 items), alcohol/drug problems (9 items), family/marital (4 items), emotional/personal (5 items), accommodation (3 items), and attitude/orientation (4 items). Each item is scored in a present (+1 point) versus absent (+0 points) format. Total LSI-R score is then used to place individuals into one of five risk bins: low

(scores +0 to +13), low-moderate (scores +14 to +23), moderate (scores +24 to +33), moderate-high (scores +34 to +40), or high (scores +41 to +54). The LSI-R has been shown to be effective in predicting parole outcome, success in halfway houses, institutional misconduct, and recidivism in the community (Andrews & Bonta, 1998).

Psychopathy Checklist – Revised

The PCL-R is a personality assessment designed to diagnose psychopathy. The item content on the PCL-R is based on Cleckley's (1941) operational definition of the psychopathic personality outlined in his work, *The Mask of Sanity*. The tool is composed of 20 items which load onto two factors: selfish, callous, and remorseless use of others (10 items) and chronically unstable and antisocial lifestyle (10 items). The PCL-R is scored using data collected in a file review and a semi-structured interview. Each item is scored out of two points: +0 = item does not apply to the individual, +1 = item applies to a certain extent, or +2 = item does apply. Total PCL-R score is used to place individuals into one of two classifications: non-psychopathic (scores +0 to +29) or psychopathic (scores +30 to +40). Although the PCL-R was not originally designed as a risk assessment tool, meta-analytic reviews have reported that the tool is able to predict future offending accurately in a number of settings and populations (Leistico et al., 2008; Salekin et al., 1996; Walters, 2003). It is also notable that PCL-R scores have been incorporated into other risk assessment tools such as the VRAG, SORAG, HCR-20, and SVR-20.

Violence Risk Appraisal Guide

The VRAG was designed to predict the likelihood of violence in previously violent, mentally disordered offenders. Scoring the VRAG's 12 historical items requires the administrator to have information about the offender's childhood conduct,

family background, previous criminal behaviour, psychological issues, and index offence. Information is collected using file review, self-report, and third party sources (e.g., interviews with family members). An individual's total score places him or her into one of three risk bins: low (scores –24 to –8), moderate (scores –7 to +13), or high risk (scores +14 to +32). Replication studies have established the VRAG's ability to accurately predict violent recidivism in a variety of settings and in a number of additional populations, including child molesters, rapists, and general offenders (Mental Health Centre Penetanguishene, 2009).

Sex Offender Risk Appraisal Guide

The SORAG was designed to assess the likelihood of violent and sexual recidivism in previously convicted sex offenders. Information acquired from file review, self-report, and third party sources is used to score participants on 14 items, 10 of which are identical to VRAG items. The four unique items include: history of violent offences, number of previous convictions for sexual offences, history of sex offences only against girls under 14, and phallometric test results. An individual's total SORAG score places him or her into one of three risk bins: low (scores –17 to +2), moderate (scores +3 to +19), or high risk (scores +20 to +34). Studies conducted in a variety of study settings on a number of populations have reported the SORAG to be a valid predictor of violent and sexual offending (Mental Health Centre Penetanguishene, 2009).

Static-99

The Static-99, a derivation of the Rapid Risk Assessment of Sexual Offence Recidivism (RRASOR; Hanson, 1997) and the Structured Anchored Clinical Judgement (SACJ-Min; Grubin, 1998), was designed to predict the long-term

probability of sexual recidivism in adult male offenders who have previously committed a sexual offence. The manual for the Static-99 (Harris et al., 2003; Hanson & Thornton, 1999) states that the tool should not be used with female offenders, male offenders under age 18, those who have been detained due to illegal consensual activity (e.g., prostitution), or those who have been detained for a non-sexual offence. Information acquired via file review is used to score individuals on 10 historical items. Scored out of 12 points, the Static-99 has four risk bins: low (scores +0 to +1), moderate-low (scores +2 to +3), moderate-high (scores +4 to +5), and high risk (scores +6 to +12). A number of replication studies have found the Static-99 to be a valid predictor of sexual recidivism in correctional and psychiatric samples (Helmus, 2008).

3.3.2.2 Structured Clinical Judgement Tools

Historical, Clinical, Risk Management – 20

The HCR-20 was designed to assess the risk of violent behaviour in forensic and general psychiatric patients. The instrument is composed of 20 items organised into three scales: historical factors (10 items), clinical factors (5 items), and risk management factors (5 items). Each item is scored out of two points: +0 = item not present, +1 = item possibly present, +2 = item definitely present. The total score out of a possible 40 points is not directly used to calculate an individual's likelihood of offending. Rather, the administering clinician's judgement is used to interpret the findings and place the individual into one of three risk bins: low, moderate, or high risk. The HCR-20 has been validated in general and forensic psychiatric as well as correctional settings with both adults and juveniles (Douglas, Blanchard, Guy, Reeves, & Weir, 2010). Due to the dynamic nature of their item content, the clinical

and risk management scales of the HCR-20 may be particularly useful in assisting professionals in making treatment decisions (Heilbrun, 2003).

Sexual Violence Risk – 20

The SVR-20 was designed to predict future violence (including sexual violence) in sex offenders. Information acquired via criminal records, psychological reports, and collateral interviews is used to score the tool's 20 items, which are organised into three scales: psychosocial adjustment (11 items), sexual offences (7 items), and future plans (3 items). Items are scored out of two points: +0 = item does not apply to the individual, +1 = item applies to a certain extent, or +2 = item does apply. Taking into consideration total SVR-20 score as well as additional risk and protective factors not included on the instrument, the administering clinician places individuals into one of three risk bins: low, moderate, or high risk. A recent narrative review concluded that, in light of the findings of replication studies, the SVR-20 should be considered a valid and reliable predictor of violent recidivism in sexual offenders (Rettenberger, Hucker, Boer, & Eher, 2009).

Spousal Assault Risk Assessment

The SARA was designed to predict future violence in men arrested for spousal assault. Composed of risk factors relating to domestic abuse, the SARA consists of 20 items organised into four scales: criminal history (3 items), psychosocial adjustment (7 items), spousal assault history (7 items), and alleged/current offence (3 items). Items are scored out of two points: +0 = item not present, +1 = item possibly present, or +2 = item definitely present. The rater also has the ability to mark items as "critical". Such a scoring method draws a clinician's attention to items that may be particularly influential when deciding into which risk bin to place a subject. As with

other SCJ tools, the SARA does not use score cut-offs to determine an individual's risk for future violence. Rather, the total SARA score out of 40 is taken into account by the administering clinician when judging whether an individual is at low, moderate, or high risk of violence.

Structured Assessment of Violence Risk in Youth

The SAVRY was designed to assess the risk of future violence in adolescents. As such, the item content of the SAVRY focuses on factors relevant to this age group. The instrument is composed of 24 risk factors organised into three scales: historical factors (10 items), social/contextual factors (6 items), and individual/clinical factors (8 items). Items are scored out of two possible points: +0 = item presents a low risk of reoffending, +1 = item presents a moderate risk of reoffending, or +2 = item presents a high risk of reoffending. In addition to scoring individuals on this item content, the SAVRY also includes a scale of protective factors that is composed of six items. These items are scored as being either present or absent. Similar to the SARA, the SAVRY allows raters to mark items as "critical". As with other SCJ instruments, the SAVRY manual does not advocate using numerical indices or cut-off scores (Borum et al., 2002, 2003). Rather, a clinician's judgement is used to interpret the risk total and presence or absence of protective factors. The administering clinician uses his or her discretion to place an individual into one of three risk bins: low, moderate, or high risk. As it includes dynamic risk and protective factors, the SAVRY may be a useful aid to intervention and management planning in addition to risk assessment (Lodewijks et al., 2010).

3.3.3 Search Strategy

A systematic search was conducted using the following electronic search databases to identify studies in which the predictive validity of at least one of the instruments of interest was measured: PsycINFO, EMBASE, MEDLINE, and the US National Criminal Justice Reference Service Abstracts. These databases were selected as they index studies in the psychological (PsycINFO), medical (EMBASE and MEDLINE), and criminological (National Criminal Justice Reference Service Abstracts) literatures. The search criteria consisted of the acronyms and full names of the nine risk assessment tools. The search was restricted to articles that had been published between January 1, 1995 and November 30, 2008, because the objective of the review was to summarise the contemporary literature and most of the instruments were not published before 1995. Additional articles were located through the reference lists of previous reviews, through annotated bibliographies of the included risk assessment tools, and through discussion with researchers in the field.

Studies in all languages from all countries were considered for inclusion as were studies not published in peer-reviewed journals (i.e., government reports, conference presentations, Master's theses, and doctoral dissertations). Studies were included in the meta-analysis if their titles, abstracts, and/or *Methods* sections revealed evidence of the work having measured the predictive validity of one of the nine risk assessment tools of interest. Articles were excluded if they only included select scales of a tool (e.g., Douglas & Webster, 1999). Such studies were excluded so that the predictive validity of complete tools could be compared rather than individual scales. In addition, the original calibration studies of the tools (e.g., Hanson & Thornton, 1999) were excluded to control for the potential inflation of effect size found in development samples (Blair et al., 2008). In cases where the same

participants were used to investigate the predictive validity of a tool in several studies, the study with the most participants was included to avoid double-counting.

The initial search identified a total of 1,743 records (Figure 3.1). When the records' abstracts and *Methods* sections were scrutinised to see whether they showed evidence of having investigated one of the nine tools of interest, the number of records was reduced to 401. Due to the use of select scales of one of the tools, being a calibration sample, not having assessed predictive validity, or the use of overlapping samples, an additional 198 studies were excluded, leaving a total of 203 studies of interest. For a study sample to be included in the meta-analysis, outcome data needed to be available for a 2 x 2 contingency table (Figure 1.1), which contains the data used to calculate relevant effect sizes.

3.3.3.1 Binning Strategies

Tools for prognostic clinical decision-making are used differently depending on the context: Either for specific case identification, whereby high specificity is strived for, or for screening purposes, whereby high sensitivity is required. In the literature these situations are referred to as "rule in" decisions and "rule out" decisions, respectively (Ransohoff & Feinstein, 1978).

Therefore, two sets of analyses of the predictive accuracy of the included risk assessment instruments were conducted: The first grouped participants who were classified as low or moderate risk and compared them with those classified by the instrument as high risk. This "rule in" approach will be referred to as the high risk vs. low/moderate risk binning strategy. The second set of analyses grouped participants who were classified as moderate or high risk and compared their scores with participants who were low risk. This "rule out" approach will be referred to as the moderate/high risk versus low risk binning strategy.

Six of the nine risk assessment tools covered by this review allow for placing scores into one of three classifications: low, moderate, or high risk of offending. The LSI-R uses five risk bins: low, low-moderate, moderate, moderate-high, and high risk. For the purposes of this study, the low and low-moderate risk bins were combined and considered the low risk bin. The moderate-high and high risk bins were also combined and considered the high risk bin.[13] The Static-99 uses four risk bins: low, moderate-low, moderate-high, and high risk. For the purposes of analysis, the moderate-low risk and moderate-high risk bins were combined and considered the moderate risk bin.[13] The PCL-R dichotomously classifies individuals as being either non-psychopathic or psychopathic, and for the purposes of this study, psychopathic individuals (i.e., with scores of +30 and above) were classified as high risk and non-psychopathic individuals (i.e., with scores of +29 and below) as low risk. As there is no moderate risk bin for this tool, both binning strategies produced the same results.

3.3.3.2 Data Collection

Preliminary inspection of the 203 studies of interest revealed that different score thresholds (i.e., cut-off scores) on the risk assessment tools had been used to place participants into risk bins (i.e., low, moderate, or high risk of offending). In these cases, study authors were contacted and asked to complete a standardised form into which outcome data could be entered at the manual-suggested thresholds scores for these tools (Appendix E). Study authors were asked to provide outcome information for the overall sample as well as for male and female participants,

[13] As this review is the first in the field to have analysed predictive validity using tabular data, there was no precedent for how to combine risk bins for these instruments. Therefore, which risk categories to combine to form the low, moderate, and high risk bins was decided upon after discussion with my supervisor, Dr. Seena Fazel, and risk assessment expert, Professor Martin Grann.

separately.[14,15] When multiple datasets were available for a sample because different tools were administered to the same participants, all datasets were included and counted as different samples.[16] In cases where multiple indices of offending (e.g., general offending, violent offending, sexual offending) had been used, authors were asked for outcome data using the most conservative definition of offending (i.e., the definition that would produce the highest sensitivity). General offending was considered to be the most conservative outcome followed by (in order) violent (including sexual) offending, violent (non-sexual) offending, and sexual offending.[17]

Data for 2 x 2 tables were extracted from studies that either placed participants in low, moderate, or high risk bins according to the SCJ approach or, for actuarial instruments, using manual-suggested cut-off scores.[18] Such outcome data was available in the manuscripts of 27 eligible studies (*k samples* = 34). Additional data was requested from the authors of 133 studies (*k* = 268) and obtained for 41 studies (*k* = 54). For 8 studies (*k* = 10), outcome data was available in the manuscript for the high risk versus low/moderate risk binning strategy but not the moderate/high risk versus low risk binning strategy. Efforts to contact the authors of these investigations to obtain data for the second binning strategy were unsuccessful. Thus, data on 88 samples from 68 independent studies were included in the meta-analysis for the "rule

[14] Tabular data on male participants was available either in study manuscripts or from authors for 71 samples and on female participants for 7 samples.
[15] Study authors were also asked to provide sample AUCs if such information was not available in the manuscript.
[16] Studies in which multiple risk assessment tools were administered would be expected to produce intercorrelated effect sizes for those instruments. As this strategy is not expected to introduce a systematic source of bias (Grove et al., 2000; Guy, 2008; Hanson & Bussiere, 1998), I followed the precedent in this field of ignoring the possibility of such intercorrelations. This strategy would have reduced the overall rate of within-study variability.
[17] Sexual offences were considered violence offences for the purposes of this review.
[18] When zero counts in the 2 x 2 table of a sample were found, a constant of +1.0 was added to each cell. Zero counts can result in the inability to extract odds ratio data due to division by zero. Adding a constant allowed those samples' data to be included in the meta-analysis (Higgins, Deeks, & Altman, 2008).

in" binning strategy and data on 78 samples from 60 independent studies were included in the "rule out" binning strategy (Appendix F).

Whether there were differences between the effect sizes of the included studies and the studies from which tabular data could not be collected was tested using Cohen's *d*, a single effect indicator commonly used to investigate global effect size (Cohen, 1988; Ferguson, 2009).[19] Cohen's *d* values were able to be calculated for 195 of the 214 samples from which data was not available. Effect sizes were converted to *d* using formulae published by Cohen (1988), Rosenthal (1994), and Ruscio (2008).[20]

As variance parameters were commonly not reported alongside the effect sizes used to calculate *d*, pooling was deemed inappropriate. The median *d* value produced by the included samples (0.75; *IQR* = 0.54-0.94) and those samples that were eligible but not included (0.70; *IQR* = 0.47-0.88) were similar. Using the "cendif" command in STATA/IC 10.1 for Windows (StataCorp, 2007), the Hodges-Lehmann median percentile difference between the effect sizes was calculated to be 0.01 (95% CI = 0.00-0.07).[21] This evidence suggests that the studies included in the present meta-analysis were not different (in terms of effect size) to studies that were eligible but not included.

3.3.3.3 Data Extraction

Data was extracted on 39 descriptive, demographic, and design characteristics of the predictive validity studies. When information was unclear or seemingly

---

[19] In the context of forensic risk assessment, Cohen's *d* would refer to the standardised mean difference between the risk scores of offenders and non-offenders.

[20] Formulae were applied using DeCoster's (2009) effect size conversion calculator (available at http://web.me.com/rsbalkin/Site/Research_Methods_and_Statistics_files/Converting%20effect%20sizes--calculator.xls).

[21] Given the two sets of samples (those for which data was able to be obtained and those for which data was not able to be obtained), the Hodges-Lehmann median difference is the median value of $X_1 - X_2$, where X_1 is a randomly-selected *d* value from the samples for which data was able to be obtained and X_2 is a randomly-selected *d* value from the samples for which data was not able to be obtained (Campbell & Gardner, 1988; Hodges & Lehmann, 1963).

conflicting, Dr. Seena Fazel was consulted. When information was missing or no consensus could be reached, it was coded as such.

3.3.4 Inter-rater Reliability and Quality Assessment

As a measure of quality control, 10 (14.7%) of the included studies were randomly selected and coded by an independent research assistant with an undergraduate degree in psychology (Ms. Sophie Westwood). Studies' descriptive and demographic characteristics were extracted and true positive, false positive, true negative, and false negative counts for both binning strategies were calculated. The research assistant was provided with a set of coding sheets (Appendix G), the 10 study manuscripts, and, where appropriate, the data sheets provided by the study's authors. A high level of inter-rater agreement was established using the kappa coefficient (κ = 0.93; Landis & Koch, 1977). Kappa was calculated using the same method as in the metareview (Chapter II, Section 2.3.4). Disagreements mostly concerned the percentage of male and white participants as well as the mean age of participants in samples. These disagreements likely arose because demographic information was often commonly reported for both full samples as well as subgroups. After discussion with my supervisor, Dr. Seena Fazel, a consensus was reached on all disagreements.

The reporting quality of the included studies was measured using the Standards for Reporting of Diagnostic Accuracy Studies (STARD) Statement (Bossuyt et al., 2003), a 25-item checklist of reporting characteristics (Appendix H). Also considered were the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) Statement (Whitting, Rutjes, Reitsma, Bossuyt, & Kleijnen, 2003), the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement (Vandenbroucke et al., 2007), the Transparent Reporting of Evaluations

with Nonrandomized Designs (TREND) Statement (Des Jarlais, Lyles, Crepaz, & the TREND Group, 2004), and the Consolidated Standards of Reporting Trials (CONSORT) Statement (Begg et al., 1996). The STARD Statement was selected as it was the most comprehensive checklist that did not presume that studies were randomised or controlled.

3.3.5 Data Analysis

3.3.5.1 Choice of Outcome Measures

Six effect sizes that measure different aspects of predictive validity were chosen as outcome measures for the present meta-analysis. These effect sizes included the area under the curve (AUC), positive and negative predictive values (PPV and NPV, respectively), and both the number needed to detain (NND) and the number safely screened (NSS). In addition, the diagnostic odds ratio (DOR), an outcome statistic commonly used in the medical diagnostic literature that has not been used in previous meta-analyses of the risk assessment literature, was also included. The properties of these six effect sizes are discussed in detail elsewhere (Chapter I, Section 1.4.1.1; Chapter II, Section 2.5.3).

Though considered for inclusion, sensitivity (i.e., the proportion of offenders accurately identified by the tool) and specificity (i.e., the proportion of non-offenders accurately identified by the tool) were not used as outcome measures, because the AUC provided an index of sensitivity and specificity across score thresholds. Other single effect indicators such as correlation coefficients and Cohen's *d* were also excluded as both the AUC and DOR provided global indices of tool performance.

Tests of Pooling Assumptions for Effect Sizes

Tests of assumptions were conducted to determine whether sample AUCs, PPVs, NPVs, NNDs, NSSs, or DORs could be pooled. There was a significant correlation between the sensitivities and specificities produced by the samples using both the high risk versus low/moderate risk binning strategy (r[86] = -0.54, p = 0.01) and the moderate/high risk versus low risk binning strategy (r[76] = -0.83, p = 0.01). As independent sensitivities and specificities are assumed for pooling ROC curves (Deeks, 2001), sample AUCs were not pooled.

Positive and negative predictive values as well as the number needed to detain and the number safely screened are base rate dependent statistics (Fleminger, 1997; Large et al., 2010). Therefore, pooling samples with heterogeneous base rates may produce biased summary effect estimates. The base rate of offending was calculated for the included samples of both binning strategies. The median base rate of offending in the high risk versus low/moderate risk binning strategy was 33.4% (*IQR* = 20.0-50.4%) and in the moderate/high risk versus low risk binning strategy was 33.6% (*IQR* = 20.3-50.8%). As base rates varied widely in both binning strategies, it was decided not to pool PPVs, NPVs, NNDs, or NSSs. Future research wishing to pool these effect sizes may wish to investigate instruments' validity in predicting more specific forms of offending in more specific populations. In such instances, variability in the base rate of offending may be lower. Though the AUCs, PPVs, NPVs, NNDs, and NSSs of the samples were not pooled, their medians and interquartile ranges were calculated for both binning strategies.

To pool DORs, the diagnostic odds ratios of each sample must follow a symmetrical ROC curve (Deeks, 2001). To test this assumption, the standard Moses, Littenberg, and Shapiro (1993) regression test was used. For this test, each sample's ln(DOR) was plotted against a measure of threshold:

$$\ln\left(\frac{TP}{1-TP}\times\frac{FP}{1-FP}\right)$$

Non-significant variation in diagnostic performance with threshold was found for both the high risk versus low/moderate risk binning strategy (β = 0.01, p = 0.98) as well as the moderate/high risk versus low risk binning strategy (β = -0.01, p = 0.28). These non-significant findings implied that a symmetrical ROC curve could be generated for the data, meaning that sample DORs could be pooled (Deeks, 2001). The DORs of the included samples were pooled using the Mantel-Haenszel (1959) method of combining odds ratios for fixed effects analyses and the DerSimonian-Laird (1986) method for random effects analyses (*fixed effects and random effects models* will be discussed in the next section). These methods both rely on 2 x 2 tables and are currently the fixed and random effects standards (respectively) when conducting meta-analyses of dichotomous outcomes (Deeks et al., 2006). To conduct the meta-analyses, the "metan" command in STATA/IC 10.1 for Windows (StataCorp, 2007) was used.

A Priori Outlier Check

Before data analysis began, outlier DORs were investigated. Outliers were identified as those effect sizes that were larger or smaller than 1.5**IQR* for either binning strategy (Rossi, 2010). Study authors were contacted if their sample produced an outlying effect size and were asked to confirm their 2 x 2 contingency table. In all cases (k = 14), data was confirmed to be accurate and was, therefore, included in the meta-analysis.

3.3.5.2 Choice of Fixed Effects versus Random Effects Models for Combining Odds Ratios

Traditionally, meta-analysis is conducted using one of two statistical models: fixed effects or random effects (Borenstein, Hedges, Higgins, & Rothstein, 2009). These models are based on different assumptions and weight individual effect sizes (i.e., sample DORs for the present review) differently when calculating summary effect estimates. *Fixed effects models* assume that there is one true effect size which all study samples are attempting to estimate (Deeks, 2001). These models assume that all differences in observed effects are due to sampling error and, therefore, assign weights to minimise within-study heterogeneity. Fixed effects summary effect estimates are calculated using the following equation:

$$M_{DOR} = \frac{\sum_{i=1}^{k} W_i Y_i}{\sum_{i=1}^{k} W_i}$$

where Y_i is the individual effect size and W_i is the weight applied to that effect size (i.e., 1/within-study variance). The variance of the summary effect size is then estimated by the inverse of the sum of the individual effect size weights. As they assume low levels of between-study heterogeneity, fixed effects models were used to combine sample effect sizes in the present meta-analysis when I^2 was less than 75% (Higgins et al., 2003). The I^2 statistic describes the percentage of variation across samples due to between-study variability rather than sampling error alone (Higgins & Thompson, 2002; Higgins et al., 2003).[22] As such, a resulting I^2 of 0 implies that all variability in effect estimates is produced by sampling error. As the I^2 percentage

[22] The I^2 statistic was chosen to measure of heterogeneity rather than of Cochran's (1954) Q statistic as the latter reports the presence or absence of heterogeneity but does not indicate the extent of such variability (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006).

increases, so too does the proportion of effect size variability that is due to between-study heterogeneity. Summary effect estimates with high I^2 values may be better approximated by random effects models.

Random effects models presume that each study in a meta-analysis is estimating a different true effect (Borenstein et al., 2009). They assume that, due to differences in sample demographics and study design characteristics, no one true effect size exists and differences in observed effect sizes can be attributed to both sampling error as well as between-study variability (traditionally estimated by τ^2). Therefore, random effects summary effect estimates are calculated using the following equation:

$$M_{DOR} = \frac{\sum_{i=1}^{k}(V_i\tau^2)Y_i}{\sum_{i=1}^{k}V_i\tau^2}$$

where Y_i is the individual effect size, V_i is equivalent to W_i in the fixed effects model, and τ^2 is an estimate of between-study heterogeneity. As in fixed effects models, the variance of the summary effect size is estimated by the inverse of the sum of the individual effect size weights (i.e., $V_i\tau^2$). Random effects models take into account that some effect sizes may not have been included in the meta-analysis because they were either from studies which were not identified during the systematic search process or from studies that have yet to have been conducted (Cooper & Hedges, 1994). Therefore, the findings of random effects models are considered to be more generalisable than those calculated using the fixed effects approach (Hedges & Vevea, 1998). As they assume high levels of between-study variability, random effects models were used to combine sample effect sizes when I^2 was equal to or greater than 75% (Higgins et al., 2003). In addition, as such models take into account between-

study variability, random effects models were used to investigate sources of heterogeneity.

3.3.6 Risk Assessment Tool Ranking

To assess which risk assessment tools produced the highest and lowest levels of predictive validity across outcome statistics, a ranking system was devised. First, tools were ordered from worst to best with regard to their DORs, AUCs, PPVs, and NPVs. (As NND and NSS are derived from PPV and NPV, these effect sizes were not included in the ranking system to avoid double-counting.) The instruments were given scores of +1 (poorest performance) through +9 (strongest performance) based on their ranking within an outcome statistic. If multiple tools produced the same effect size, they were given the same score. This procedure was repeated for each outcome measure for both binning strategies. The scores for each instrument were then summed across binning strategies and used as a summary performance score to determine the strongest and weakest tools.

3.3.7 Investigation of Sources of Heterogeneity

To investigate sources of heterogeneity, random effects subgroup and metaregression analyses were conducted using the DORs of the included samples. Due to the clinical heterogeneity introduced into meta-analyses by pooling instruments designed using different approaches to risk assessment to predict different forms of offending (general or violent [including sexual]) in different populations, a thorough investigation of between-study heterogeneity was conducted.

3.3.7.1 Subgroup Analysis and Metaregression

Subgroup analysis allows researchers to calculate effect sizes for subsets of samples that share a common characteristic (e.g., effect sizes produced by prospective versus retrospective samples). The summary effect estimates of these subgroups can then be compared, routinely using 95% confidence intervals (Borenstein et al., 2009). Subgroup analysis is limited to the investigation of dichotomous variables. In addition, subgroup analysis cannot assess the independence of multiple moderating variables. Random effects subgroup analyses were carried out using the DOR data and the "metan" command in STATA/IC 10.1 for Windows (StataCorp, 2007).

Metaregression investigates the relationship between sample effect sizes and one or more sample- or study-level characteristics. Unlike subgroup analysis, metaregression allows researchers to examine the influence of continuous variables on effect size and to assess the independence of multiple moderating variables modelled simultaneously (Thompson & Higgins, 2002). Random effects metaregression was used to investigate whether different sample or study characteristics were associated with the predictive validity of the risk assessment tools. In addition, a test of independence on those variables which were found to be significant at the $p < 0.10$ level was used to assess whether these variables influenced predictive validity independently of one another. Metaregression analyses were performed using the DOR data and the "metareg" command in STATA/IC 10.1 for Windows (StataCorp, 2007). Power for the metaregression analyses was investigated using G*Power 3 for Windows (Faul, Erdfelder, Lang, & Buchner, 2007).

3.3.7.2 Tool and Demographic Variables of Interest

Tool- and sample-level variables of interest during the investigation of sources of heterogeneity included type of risk assessment tool, gender, ethnicity, and age.

Type of risk assessment tool was analysed as a dichotomous variable (i.e., actuarial versus SCJ). The influence of gender was analysed as both a categorical variable (i.e., male sample data versus female sample data) and continuously (i.e., percentage of study sample that was male). The ethnic composition of samples was also analysed categorically (i.e., <90% white versus ≥90% white) and continuously (i.e., percentage of study sample that was white). The mean age of participants in a sample was analysed both as a dichotomous variable (i.e., <25 years of age versus ≥25 years)[23] and as a continuous variable (in years).

3.3.7.3 Study Design Variables of Interest

Design-related variables of interest included study setting, temporal design, length of follow-up, sample size, country of origin, type of offending, type of outcome, source of information used to score tool, tool authorship, and publication status. Study setting was explored in two ways: institutional setting (i.e., prison or psychiatric unit) versus community setting and prison setting versus psychiatric setting. Temporal design was also included as a dichotomous variable (i.e., prospective versus retrospective). Mean length of follow-up was investigated in both categorical (i.e., ≤2 years versus >2 years) and continuous (in months) forms. Sample size was included as both a dichotomous variable (i.e., <500 participants versus ≥500 participants) and as a continuous variable. Also included in the analyses were the dichotomous variables of country of origin (i.e., study conducted in North America versus Europe), type of offending being predicted (i.e., general offending versus violent offending), type of outcome (i.e., arrest, charge, conviction, or incarceration versus institutional incident), source of information used to score tool (i.e., file review

[23] The traditional age cut-off of 18 years was not used, as only SAVRY samples had a mean participant age below 18 years.

only versus other), and tool authorship (author of the English-language version of the tool under investigation as a study author versus not). To investigate publication bias, publication status was also investigated as a dichotomous variable (i.e., study published in peer-reviewed journal versus not).[24]

3.3.7.4 Summary of Investigation of Heterogeneity

As discussed in the previous chapter (Section 2.5.2), between-study heterogeneity can be divided into three subtypes: clinical heterogeneity, methodological heterogeneity, and statistical heterogeneity (Gagnier et al., in press; West et al., 2010). Using those potential sources of heterogeneity identified by the metareview, I attempted to investigate between-study variability resulting from a set of clinical covariates at the tool level, participant level, outcome level, and setting level. In addition, I attempted to measure methodological sources of between-study variability using common markers of study quality (e.g., prospective versus retrospective orientation and sample size). Finally, statistical heterogeneity was measured using the I^2 statistic.

3.4 RESULTS

3.4.1 Descriptive and Demographic Characteristics

Information was collected on 25,980 participants in 88 samples from 68 independent studies. Information from 54 (61.4%; n = 15,775) of the samples was specifically obtained from study authors for the purposes of this synthesis. Of the 25,980 participants in the included samples, 8,155 (31.4%) went on to offend. The tools with the most samples included the PCL-R (*k samples* = 20; 22.7%), the

[24] The moderating influence of the base rate of offending on sample DOR was not investigated, as the diagnostic odds ratio is independent of base rate.

Static-99 (k = 14; 15.9%), and the VRAG (k = 12; 13.6%). The majority of the samples (k = 61; 69.3%) were assessed using an actuarial tool (Table 3.2).

Content analyses revealed that the most common items included on the nine risk assessment tools were, in order of frequency, clinical factors, criminal history variables and socio-demographic characteristics. The most common clinical factors included previous and/or current substance abuse (N *tools* = 7), pro-criminal attitudes (N = 7), and previous and/or current diagnosis of a personality disorder (N = 5). The most common criminal history factors included prior conditional release failure (N = 8), previous and/or current violence (N = 7), and having a criminal record as a juvenile (N = 4). The most common socio-demographic factors included the quality of intimate relationships (N = 5), employment (N = 4), and marital history (N = 3).

Data on gender composition was available for 82 (93.2%) samples, with a trend towards predominantly male samples. Of the included samples, 80 (90.9%) were composed of over 50% male participants. Data on participants' ethnic backgrounds was available for 34 (38.6%) samples and showed a trend towards predominantly white samples. Of the included samples, 28 (31.8%) were composed of over 50% white participants. The mean age of participants was 31.6 (SD = 7.6) years. Participant psychiatric diagnosis data was available for 23 (26.1%) samples. Personality disorders were the most commonly reported diagnoses (n = 1,333; 5.1% of all participants) followed by psychotic disorders (n = 697; 2.7%), drug or alcohol abuse or dependency (n = 629; 2.4%), mood disorders (n = 127; 0.5%), conduct disorder (n = 26; 0.1%), and anxiety disorders (n = 5; 0.1%).

3.4.2 Study Design Characteristics

Regarding study setting, 37 (42.0%) samples consisted of prisoners, 33 (37.5%) of psychiatric patients, 4 (4.5%) of community persons, and 14 (15.9%) of participants from a mixture of different settings (Table 3.2). Of those 33 samples composed of psychiatric patients, 3 (3.4%) consisted of patients in general settings and 30 (34.1%) of patients in forensic settings. Prospective research methodology was used with 40 (45.5%) samples, whereas 43 (48.9%) samples were investigated retrospectively. The mean length participants were followed up was 56.3 (*SD* = 41.3) months. The mean sample size was 296 (*SD* = 422) participants. Included studies were conducted in 13 countries: Argentina (*n* = 199; 0.8% of all participants), Austria (*n* = 799; 3.1%), Belgium (*n* = 732; 2.8%), Canada (*n* = 9,112; 35.1%), Denmark (*n* = 304; 1.2%), Finland (*n* = 208; 0.8%), Germany (*n* = 1,337; 5.1%), The Netherlands (*n* = 622; 2.4%), New Zealand (*n* = 220; 0.8%), Spain (*n* = 276; 1.1%), Sweden (*n* = 2,204; 8.5%), the UK (*n* = 3,929; 15.1%), and the US (*n* = 6,038; 23.2%). Regarding the type of offending being predicted, general offending was the outcome in 46 (52.3%) samples as opposed to violent offending in 40 (45.5%) samples, or non-violent offending in 1 (1.1%) sample. The majority of samples (*k* = 56; 63.6%) used arrest, charge, conviction, or incarceration as their outcome. Of the included samples, 52 (59.1%) used file review alone to score a risk assessment tool, whereas 2 (2.3%) used information from interviews only and 16 (18.2%) used a combination of file review, self-report, and interview information. Study participants were selected using a specified sampling method in 52 (59.1%) samples. Of the 88 included samples, 64 (72.7%) were in studies published in peer-reviewed journals.

An author or translator of a risk assessment tool was also an author on a study investigating the predictive validity of that instrument in 22 studies (*k* = 26). Authors

of the English-language version of a given tool's manual were also authors of a study investigating that tool's predictive validity on 10 studies constituting 12 (13.6%) samples: 3 (3.4%) samples for the SARA, 2 (2.3%) samples for the HCR-20, 2 (2.3%) samples for the SORAG, 2 (2.3%) samples for the Static-99, 2 (2.3%) samples for the VRAG, and 1 (1.1%) sample for the SAVRY. A tool's translator was also an author of a study investigating that tool's predictive validity in 12 studies constituting 14 (15.9%) samples: 3 (3.4%) for the SVR-20, 2 (2.3%) samples for the HCR-20, 2 (2.3%) samples for the PCL-R, 2 (2.3%) samples for the SAVRY, 2 (2.3%) samples for the SORAG, 2 (2.3%) samples for the Static-99, and 1 (1.1%) sample for the VRAG. Six of the 14 journals in which the studies appeared requested that authors report any financial or non-financial conflicts of interest. None of the 22 studies contained such a disclosure.

The average study included in the meta-analysis fulfilled 17 (*SD* = 3) of the 25 STARD criteria. In the 68 studies, the criteria that were most commonly fulfilled included descriptions of the sample and setting in which a study took place (*N studies* = 68; 100.0%), descriptions of the tool whose predictive validity was being investigated (*N* = 68; 100.0%), and discussions of the clinical applicability of study findings (*N* = 68; 100.0%). The least commonly fulfilled STARD criteria included specifying sample selection criteria (*N* = 28; 41.2%), reporting the number of and training of coders (*N* = 28; 41.2%), detailing why some participants were excluded or did not complete follow-up (*N* = 28; 41.2%), reporting how indeterminate results, missing responses, and outliers were handled (*N* = 25; 36.8%), and whether there were adverse effects from the testing procedures (*N* = 0; 0.0%). There was no clear evidence of an association between STARD score and sample DOR in either binning strategy (β = -0.03, p = 0.36; β = -0.01, p = 0.87, respectively).

3.4.3 Risk Assessment Tool Performance

3.4.3.1 Median Area Under the Curve

The median AUC of all risk assessment tools combined was 0.70 (*IQR* = 0.66-0.75) (Figure 3.2). The risk assessment tools with the highest median AUCs were the SVR-20 (0.78; *IQR* = 0.71-0.83), the SORAG (0.75, *IQR* = 0.69-0.79), and the VRAG (0.74. *IQR* = 0.74-0.81) (Table 3.3).[25]

3.4.3.2 Median Positive Predictive and Negative Predictive Values

When the data was analysed using the high risk versus low/moderate risk binning strategy, the median PPV of all risk assessment tools combined was found to be 0.55 (*IQR* = 0.35-0.74) (Figure 3.3). The risk assessment tools with the highest median PPVs were the SAVRY (0.76; *IQR* = 0.42-0.85), the HCR-20 (0.71; *IQR* = 0.55-0.85), and the VRAG (0.66; *IQR* = 0.37-0.79) (Table 3.4, upper panel). The overall median NPV for this binning strategy was 0.70 (*IQR* = 0.56-0.86) (Figure 3.4). The three tools with the highest median NPVs were the Static-99 (0.82; *IQR* = 0.71-0.94), the SARA (0.79; *IQR* = 0.67-0.92), and the SAVRY (0.76; *IQR* = 0.49-0.91).

The moderate/high risk versus low risk binning strategy found the median PPV of all tools combined to be 0.47 (*IQR* = 0.24-0.67) (Figure 3.5). The risk assessment tools with the highest median PPVs were the HCR-20 (0.64; *IQR* = 0.45-0.70), the SAVRY (0.60; *IQR* = 0.27-0.73), and the PCL-R (0.52; *IQR* = 0.45-0.75) (Table 3.4, lower panel). The overall median NPV for the second binning strategy was 0.81 (*IQR* = 0.57-0.92) (Figure 3.6). The risk assessment tools with the highest

[25] As the AUC takes into account all possible cut-off scores, the two binning strategies were not necessary for the calculation of this effect size.

median NPVs were the SARA (0.96; *IQR* = 0.82-0.98), the Static-99 (0.95; *IQR* = 0.76-0.99), and the SAVRY (0.90; *IQR* = 0.74-0.95).

3.4.3.3 Median Number Needed to Detain and Number Safely Screened

The median NND of all risk assessment tools combined was 2 (*IQR* = 1-3) for the high risk versus low/moderate risk binning strategy (Figure 3.7). The risk assessment tools with the lowest median NNDs were the SAVRY (1; *IQR* = 1-3) and the HCR-20 (1; *IQR* = 1-2) (Table 3.5, upper panel). The median NSS for all instruments combined was 3 (*IQR* = 2-7) (Figure 3.8). The instruments with the highest median NSSs were the SARA (6; *IQR* = 3-16), the Static-99 (4; *IQR* = 6-24), the SAVRY (4; *IQR* = 2-12), and the VRAG (4; *IQR* = 3-6).

For the moderate/high risk versus low risk binning strategy, the median NND was 2 (*IQR* = 2-4) (Figure 3.9). The measures with the lowest median NNDs were the SAVRY (2; *IQR* = 1-4), the HCR-20 (2; *IQR* = 1-2), the PCL-R (2; *IQR* = 1-2), and the LSI-R (2; *IQR* = 1-4) (Table 3.5, lower panel). The median NSS for this binning strategy was 5 (*IQR* = 2-12) (Figure 3.10) and the tools with the highest NSSs were the SARA (27; *IQR* = 17-37), the Static-99 (24; *IQR* = 4-86), and the SORAG (12; *IQR* = 4-22).

3.4.3.4 Pooled Diagnostic Odds Ratios

When the data from the high risk versus low/moderate risk binning strategy was analysed, the random effects pooled DOR was 3.20 (95% CI = 2.58-3.96) with significant heterogeneity (I^2 = 83.4; 95% CI = 80.1-86.2) (Figure 3.11). The three tools with the highest pooled DORs were the SAVRY (6.93; 95% CI = 4.93-9.73), the VRAG (3.84; 95% CI = 2.85-5.16), and the HCR-20 (3.48; 95% CI = 2.62-4.62) (Table 3.6, upper panel). When the moderate/high risk versus low risk binning

strategy was used, the random effects pooled DOR was 3.33 (95% CI = 2.60-4.27) with significant heterogeneity (I^2 = 83.1; 95% CI = 79.4-86.1) (Figure 3.12). The three tools with the highest pooled DORs using this binning strategy were the SARA (7.87; 95% CI = 3.12-19.87), the SAVRY (6.40; 95% CI = 4.40-9.32), and the SORAG (5.54; 95% CI = 4.09-7.50) (Table 3.6, lower panel). All instruments in both binning strategies produced DORs significantly better than chance at the $p < 0.05$ level.

3.4.3.5 Risk Assessment Tool Comparison

Using the ranking system that collated results from median AUC, PPV, NPV and pooled DOR across binning strategies, the SAVRY was found to produce the highest rates of overall predictive validity (Table 3.7). The SARA and the VRAG also produced high rates. Ranked lowest were the LSI-R and the PCL-R. In both binning strategies, the pooled DOR, alone, accurately identified the three most and least accurate risk assessment tools.

3.4.4 Investigation of Sources of Heterogeneity

Random effects subgroup and metaregression analyses were conducted for both binning strategies using sample DORs (Tables 3.8 and 3.9, respectively). For a summary of these results, see Table 3.10.

3.4.4.1 High Risk versus Low/Moderate Risk Binning Strategy

Non-significant Findings

Subgroup and metaregression analyses of the sample DOR data using the high risk versus low/moderate risk binning strategy data resulted in non-significant

findings for the following variables: type of risk assessment tool, gender composition,[20] ethnic composition,[26] mean participant age, study setting, temporal design, mean length of follow-up,[27] sample size, country of origin,[28] type of outcome, source of information, and tool authorship. In addition, no evidence of publication bias was found.

Gender

A sensitivity analysis was conducted to further explore gender effects. As women-only data was available for the HCR-20, LSI-R, SAVRY, and VRAG, data on all other instruments were excluded for the men. Hence, outcomes were investigated in the male and female data using the same set of instruments. The pooled DOR of these risk assessment tools for men was 3.83 (95% CI = 2.13-6.87), and for women was 4.93 (95% CI = 2.70-9.01).

Age

Neither subgroup analysis nor metaregression found that predictive validity estimates varied with mean participant age. As all SAVRY samples (k = 9) had mean participant ages below 25 and as this instrument produced the highest predictive validity estimates across outcome measures, a sensitivity analysis was conducted in which SAVRY samples were excluded. When SAVRY samples were excluded, metaregression analysis of mean age as a continuous variable found a significant

[26] Power analyses suggested that there might not have been enough sample data to detect even large effect sizes for this variable ($B < 0.80$).
[27] Ancillary subgroup and metaregression analyses were conducted on the dichotomous form of this variable in both binning strategies using cut-off periods of 1, 3, 4, 5, and 10 years. None of these comparisons yielded a significant difference.
[28] In response to literature suggesting that using a PCL-R cut-off score of +30 produces different effect sizes in North America versus Europe (Patrick, 2006), a sensitivity analysis was conducted. Scoring the PCL-R using a cut-off score of +30 was found to produce commensurate effect sizes in North American and European studies ($\beta = 0.63$, $p = 0.44$).

trend: the higher the mean age of a sample, the higher the DOR ($\beta = 0.09$, $p = 0.02$). To further investigate this finding, subgroup analyses were conducted dividing mean participant age into three groups (<25 years, 25-40 years, and >40 years). The summary random effects DORs for these age bands were 0.79 (95% CI = 0.16-3.95), 2.86 (95% CI = 2.12-3.86), and 4.01 (95% CI = 3.00-5.35), respectively.

Type of Offending

Subgroup analysis revealed that samples reporting on violent offending produced significantly higher DORs than samples investigating general offending (Figure 3.13). Metaregression analysis confirmed this finding ($\beta = 0.81$, $p = 0.01$).

Type of Outcome

Subgroup analysis found a trend such that samples which used institutional incidents as their outcome produced higher DORs than samples which used arrest, charge, conviction, or incarceration as their outcome. Metaregression analysis confirmed this trend at the $p < 0.10$ level ($\beta = 0.98$, $p = 0.07$).

3.4.4.2 Moderate/High Risk versus Low Risk Binning Strategy

Non-significant Findings

Subgroup and metaregression analyses of the moderate/high risk versus low risk binning strategy data found that the following variables did not significantly influence effect size: type of risk assessment tool, gender composition,[20] mean participant age, study setting, temporal design, mean length of follow-up,[29] sample

[29] Ancillary subgroup and metaregression analyses were conducted on the dichotomous form of this variable in both binning strategies using cut-off periods of 1, 3, 4, 5, and 10 years. None of these comparisons yielded a significant difference.

size, country of origin, type of outcome, and source of information. Further, no evidence of publication bias was found.

Gender

As in the first binning strategy, a sensitivity analysis was conducted using only the HCR-20, LSI-R, SAVRY, and VRAG sample data. Non-significant increases in DOR were found for women as opposed to men using subgroup analysis (DOR for men = 3.21 [95% CI = 1.61-6.39] vs. DOR for women = 5.12 [95% CI = 2.97-8.83]).

Ethnicity

Subgroup analysis revealed that samples with 90% or more white individuals did not produce different DORs than samples with less than 90% white individuals. Using metaregression, the higher the proportion of white individuals in a sample, the higher the resulting DOR ($\beta = 0.02$, $p = 0.04$).

Age

As in the high risk versus low/moderate risk binning strategy, neither subgroup analysis nor metaregression revealed that DORs varied with mean participant age. When SAVRY samples were excluded, however, a significant trend was found such that the higher the mean age of participants in a sample, the higher the DOR ($\beta = 0.08$, $p = 0.04$). In addition, *post hoc* subgroup analyses were conducted on three mean participant age bands (<25 years, 25-40 years, and >40 years). The summary random effects DORs were 0.85 (95% CI = 0.16-4.54), 2.73 (95% CI = 1.99-3.74), and 4.85 (95% CI = 2.86-8.21), respectively.

Type of Offending

Subgroup analysis found that samples that used violent offending as their outcome produced higher DORs than samples that used general offending (Figure 3.14). Metaregression analysis confirmed this finding (β = 0.57, p = 0.01).

Tool Authorship

Subgroup analysis found evidence that samples from studies where an author of the tool under investigation was also a study author produced higher DORs than samples from studies conducted by independent researchers (Figure 3.15). Metaregression confirmed this finding (β = -0.87, p = 0.02). Sensitivity analyses were conducted to further explore the evidence of an authorship effect. When evidence of an authorship effect was investigated for actuarial and SCJ instruments, separately, no clear evidence of an authorship effect was found (β = -0.78, p = 0.14; β = -0.72, p = 0.17, respectively). When evidence of an authorship effect was investigated in studies published in peer-reviewed journals, no statistically significant evidence of an authorship effect was found (β = -0.74, p = 0.11). Similarly, no clear evidence of an authorship effect was found in studies from the grey literature (β = -0.52, p = 0.31). When the operational definition of "authorship" was broadened to include authors of non-English translations, there was no trend towards higher DORs in samples where a tool author was a study author (β = -0.36, p = 0.21).

3.4.4.3 Multivariate Metaregression

Metaregression was carried out on all variables that were significant at the $p < 0.10$ level to determine which, if any, produced effects independently of one another. For the high risk versus low/moderate risk binning strategy, variables at the

$p < 0.10$ level included type of outcome and type of offending. When these variables were modelled together using multivariate metaregression, type of offending remained a significant predictor of sample DOR ($\beta = 0.78$, $p = 0.01$).[30] When SAVRY data was excluded and mean age of participants (continuous), type of outcome, and type of offending were regressed together, none of the variables remained significant predictors.[31]

For the moderate/high risk versus low risk binning strategy, variables at the $p < 0.10$ level included: ethnic composition (continuous), type of offending, and tool authorship. When these variables were modelled together using multivariate metaregression, both ethnic composition ($\beta = 0.02$, $p = 0.03$) and type of offending ($\beta = 0.86$, $p = 0.04$) remained significant predictors of DOR.[32] When SAVRY data was excluded and mean age (continuous) was added to the metaregression model, only ethnic composition ($\beta = 0.02$, $p = 0.04$) remained significant.[33,34]

3.5 DISCUSSION

This large-scale meta-analysis investigated the predictive validity of nine commonly used forensic risk assessment tools: the HCR-20, LSI-R, PCL-R, SARA, SAVRY, SORAG, Static-99, SVR-20, and the VRAG. Data was collected from 68 independent studies constituting 88 samples. These samples included a total of

[30] Power analysis using a Bonferroni-corrected α-level of 0.025, the number of samples contributing to the analysis ($k = 67$), two independent variables, and a strong effect size ($R^2 = 0.25$) resulted in a *B*-level of 0.94.
[31] Power analysis using a Bonferroni-corrected α-level of 0.025, the number of samples contributing to the analysis ($k = 44$), three independent variables, and a strong effect size ($R^2 = 0.25$) resulted in a *B*-level of 0.69.
[32] Power analysis using a Bonferroni-corrected α-level of 0.025, the number of samples contributing to the analysis ($k = 34$), three independent variables, and a strong effect size ($R^2 = 0.25$) resulted in a *B*-level of 0.51.
[33] Power analysis using a Bonferroni-corrected α-level of 0.025, the number of samples contributing to the analysis ($k = 27$), three independent variables, and a strong effect size ($R^2 = 0.25$) resulted in a *B*-level of 0.37.
[34] When only mean age (continuous) and tool authorship were modelled together with SAVRY data excluded, neither variable remained a significant predictor of sample DOR.

25,980 participants from 13 countries. Previously unavailable information on over 60% of the participants was specifically obtained from study authors for the purposes of this synthesis. Predictive validity was measured using six outcome measures, including a single effect indicator not before used in a review of the forensic risk assessment literature, the diagnostic odds ratio, and a novel outcome statistic developed for this thesis, the number safely screened. In addition to investigating the general utility of risk assessment tools, this review explored four major uncertainties in the field of forensic risk assessment, including: (1) are there differences between the predictive validity of commonly used risk assessment tools, (2) do actuarial and clinically based risk measures produce different rates of predictive validity, (3) what demographic factors are associated with higher or lower rates of predictive validity, and (4) what aspects of study design influence predictive validity.

3.5.1 The General Utility of Risk Assessment Tools

A primary aim of the present meta-analysis was to investigate the general utility of forensic risk assessment tools. The average DOR produced by the risk instruments ranged from 3.20 to 3.33 depending on how risk bins were combined. As a DOR above 1.00 denotes that an instrument has the ability to distinguish between offenders and non-offenders at rates above chance (Glas et al., 2003), these findings suggest that structured risk instruments have some predictive utility. Further, these DORs are similar to those produced when predicting other behaviours. For example, using risk factors such as depression and protective factors such as social support to predict adherence with medication has produced diagnostic odds ratios of between 2 to 4 (DiMatteo, 2004; DiMatteo, Lepper, & Croghan, 2000). In addition, the use of truancy rates to predict alcohol abuse in juveniles has been found to produce DORs

between 3 to 5 (Hallfors et al., 2002). However, these effect estimates are considerably lower than those produced by commonly used medical diagnostic tests. For example, meta-analyses of the predictive accuracy of mammography report DORs above 100 (Kang, Pang, Li, Liu, & Liu, 2010) and for prostatic specific antigen testing above 8 (Wang, Sun, Pan, Guo, & Li, 2006). For more specific tests such as magnetic resonance angiography for peripheral atherosclerosis the DOR is approximately 8 (Nelemans, Leiner, de Vet, & van Engelshoven, 2000) and around 16 for screening for renal artery stenosis using duplex sonography (Williams et al., 2007) (Table 3.11).

The median AUC produced by the forensic risk assessment tools was 0.70, suggesting that the probability of a randomly selected offender having a higher test score than a randomly selected non-offender was approximately 70% (Mossman, 1994). Current guidelines suggest than such an AUC can be interpreted as being of poor to moderate strength (Douglas et al., 2010; Tape, 2006; Vranova, Horak, Kratka, Hendrichova, & Kovarikova, 2009). Positive and negative predictive value results suggested that the likelihood of a risk instrument making a false positive prediction is approximately 50% and the probability of making a false negative prediction between 20% and 30%.

Collectively, the overall performance of forensic risk assessment tools suggests that mental health professionals in psychiatric, correctional, and court settings may not wish to rely exclusively on these instruments to make decisions regarding individual liberty and public protection. Despite low positive predictive values, structured assessment tools do appear to offer some improvements in their ability to predict the likelihood of future offending compared to chance, especially in the accurate identification of low risk individuals.

3.5.2 The Comparative Predictive Validity of Risk Assessment Tools

A central finding of the present meta-analysis was that there are substantial differences between the predictive validity of individual risk assessment tools. These differences were found in all six outcome statistics used to measure predictive accuracy. For example, tools' pooled DORs varied from 1.2 to 7.9. This suggests that it may in fact matter what instrument is used for risk assessment.

The risk assessment tool that produced the highest rate of predictive validity varied slightly depending on which outcome statistic was used. Therefore, a ranking system that collated performance on the included effect sizes was developed to identify the most and least accurate measures. Overall, instruments designed to assess the risk of offending in specific populations produced higher rates of predictive validity than tools designed for more general populations. The SAVRY, an instrument designed to assess the risk of violence in adolescents, produced the highest rates of predictive validity across outcome statistics in both binning strategies. The LSI-R, a tool that was designed to predict the likelihood of general offending in adult offenders, and the PCL-R, a clinical rating scale that was not designed for the purposes of forensic risk assessment, produced the lowest rates of predictive validity. While these two measures may be administered to a broad range of participants, this appears to come at the cost of predictive accuracy. The present meta-analysis would, therefore, argue against the view of some experts that the PCL-R is unparalleled in its ability to predict future offending (Hart, 1998; Salekin, Rogers, & Sewell, 1996).

The finding that risk assessment tools designed for more specific purposes produce higher rates of predictive validity is supported by another main finding in the current report: Instruments were better at detecting risk of violent offending than general offending. Future research could examine whether tools produce even higher

rates of predictive validity when more specific forms of violent offending (e.g., sexual violence) are investigated. The results of the present review suggest that the future development of risk assessment tools could take the direction of designing measures for specific populations or more specific forms of offending.

The finding that the SAVRY has the highest rates of predictive validity may be partly due to replication samples for the SAVRY all having been conducted on adolescent offenders, the population for which the tool was designed, unlike other tools where replication samples were more varied. In addition, as the SAVRY is amongst the most specific (designed to detect violence risk in juveniles) and thorough (composed of both risk and protective factors) of the included tools, researchers may have been more attentive to using the instrument according to the protocol set forth by the instrument's authors. Despite these reservations, results suggest that, currently, the SAVRY should be routinely used when assessing violence risk in adolescents.

The finding that risk assessment tools differ in their predictive validity conflicts with a recent meta-analysis which concluded that risk measures are interchangeable in their predictive abilities (Yang, Wong, & Coid, 2010). However, the dissenting review only included four of the same instruments as the present meta-analysis (HCR-20, LSI-R, PCL-R, VRAG) and included only those studies that compared two or more instruments, resulting in the inclusion of fewer than half the number of investigations as the present review. Further, the authors limited their meta-analysis to studies published between 1999 and 2009, whereas the present review included studies published between 1995 and 2008. Finally, the dissenting review measured predictive validity using only one effect size (Cohen's *d*), thus limiting the opportunities to observe important differences in tool utility (e.g., comparative usefulness in making "rule in" and "rule out" decisions).

3.5.3 The Efficacy of Actuarial versus Clinically Based Risk Assessment

No evidence was found that, compared with SCJ tools, actuarial instruments produced better levels of predictive validity. This finding suggests that clinicians and researchers could focus on identifying which measure, actuarial or not, produces the highest rate of predictive validity for their population, outcome, and setting of interest. Additional considerations when choosing a risk measure may include the costs of training and materials, ease of use, and whether a tool is useful in making decisions regarding effective treatment and risk management. The latter has been considered a primary strength of the SCJ approach (Douglas, Cox, & Webster, 1999; Hart, 2008; Heilbrun, 1997). The relative utility of actuarial and clinically based tools may, however, be different for certain subgroups, such as sexual offenders. Meta-analyses on the accuracy of risk assessment tools for sexual offenders have found that actuarial instruments outperform measures that employ structured clinical judgement (Hanson & Morton-Bourgon, 2004, 2007). To be included in these reviews, studies had to include only sexual offenders. The present meta-analysis included samples of offenders regardless of their index offence, making it more representative of the criminal population. Future meta-analyses could investigate the relative predictive validity of actuarial and SCJ instruments for more specific forms of offending in more specific populations.

3.5.4 The Influence of Demographic Factors on Predictive Validity

The present review found some potentially interesting differences in the predictive validity of risk assessment tools according to age, ethnicity, and gender. The most consistent finding was that older age was associated with higher rates of predictive validity. This finding is not surprising in view of the fact that the mean age

of the samples was 32 years, and many of these instruments were developed in released prisoners who would have been in their late-twenties and thirties. A second finding was that there was some evidence that validity was better in those samples comprising mainly white participants. Again, most of these risk assessment tools were calibrated on samples of predominantly white participants, so this is not unexpected.

One of the implications of the age and ethnicity findings is that caution is warranted when using these tools to predict offending in samples dissimilar to their calibration samples. A recent study by Dernevik and colleagues (2010) is consistent with this view: The researchers found that instruments designed to detect recidivism risk in general offender populations performed poorly when used to predict misconduct in terrorists (Dernevik, Beck, Grann, Hogue, & McGuire, 2010). In addition, an investigation based on all sex offenders leaving prisons in Sweden found important differences in factors associated with reoffending by 10 year age bands, particularly in those aged over 60 (Fazel, Sjöstedt, Långström, & Grann, 2006).

Although some evidence was found that risk assessment tools produce higher rates of predictive validity for women than for men, the data on women was based on seven samples and, hence, should be interpreted with considerable caution. Future research should present predictive validity estimates for men and women separately (in addition to overall effect sizes) to assist reviewers in further investigating this uncertainty.

3.5.5 The Influence of Study Design Characteristics on Predictive Validity

In addition to finding that risk assessment tools predict the likelihood of violent offending more accurately than general offending, the meta-analysis found

some evidence that the type of outcome being predicted by a tool and whether or not a tool author was a study author influenced validity findings.

3.5.5.1 Type of Outcome

The present meta-analysis found evidence that risk assessment tools are more accurate at identifying individuals who will commit an institutional infraction than individuals who will be arrested, charged, convicted, or incarcerated for a crime. This finding is difficult to interpret, as the included studies inconsistently defined what behaviour constituted institutional misconduct. There is evidence from primary studies and meta-analytic research that the term "institutional incident" is more broadly operationalised than other outcomes (Guy et al., 2005; Uppal & McMurran, 2009). It is important that researchers operationally define the outcome being predicted by a risk assessment tool, as clinicians are more likely to be interested in knowing how well a given instrument predicts the likelihood of harm to others than the likelihood of more minor, possibly non-criminal offences (e.g., a patient smoking in his or her room). Future research may wish to investigate whether commonly used risk assessment tools differ in their ability to predict more versus less severe institutional infractions.

3.5.5.2 Evidence for an Authorship Effect

Univariate analyses revealed evidence of an authorship effect such that studies in which an author of the tool being investigated was also a study author produced higher rates of predictive validity than studies conducted by independent investigators. This finding was limited to only one of the binning strategies, however, and appeared to be confounded by samples' ethnic composition, mean age, and the type of offence being predicted (i.e., general or violent). These findings provide some

support for Harris and colleagues' (2010) theory that what has been referred to as an "allegiance effect" (Blair et al., 2008, p. 346) may in fact be evidence of fidelity. That is, the authorship effect may be a proxy for having used a risk assessment tool as it was designed to be used (i.e., to predict violent or general offending) in samples similar to that tool's calibration sample (i.e., older, predominantly white individuals). Multivariate metaregression should be used in future reviews which include authorship as a moderating variable so that researchers can further investigate whether tool authorship influences effect size independently of other demographic and study design characteristics.

As there was some evidence of an authorship effect, the financial and non-financial benefits that tool authors and translators may receive warrant disclosure, particularly when a journal's *Instructions to Authors* request that any potential conflicts of interest be divulged. Such disclosure has been established as a first step towards dealing with conflicts of interest in psychiatry (Fava, 2009). The present meta-analysis found that such transparency has yet to have been achieved in the forensic risk assessment literature. None of the 22 studies where tool authors or translators were also study authors reported a conflict of interest, despite 6 of the 14 journals in which they were published having requested that potential conflicts be disclosed. Apparent lack of compliance with these guidelines may have been due to study authors having chosen not to report their financial and/or non-financial interests or it may have been that journals chose not to publish a disclosure made by study authors (Krimsky & Rothenberg, 2001). To promote transparency in future research, tool authors and translators should broaden their understanding of conflict of interest policies.

3.5.6 Re-evaluating the Single Effect Indicator of Choice

In light of the findings of the present review, the use of the AUC as the preferred single effect indicator when measuring predictive accuracy may need reconsideration. As the assumptions of pooling were not met, it was not possible to statistically combine sample AUCs. Comparing tools by median AUC was not useful in identifying those instruments that performed best across outcome measures and binning strategies. The utility of the AUC is limited as a primary effect size, as it only allows for subgroup analysis to investigate sources of heterogeneity (if pooling is possible) rather than metaregression. The latter method enables researchers to explore the moderating role of continuous variables on effect size and adjust for the effects of one variable on another. In addition, as the AUC takes into consideration rates of sensitivity and specificity across score thresholds, it largely ignores the fact that many instruments have been designed to be used with a specific cut-off score.

When a ranking system was used to compare tools across all six outcome statistics, the DOR data was best able to identify the risk assessments with the highest and lowest rates of predictive validity, suggesting that the DOR may be more effective in discriminating between the diagnostic accuracy of different instruments. Further, the DOR is cut-off dependent and allows researchers to investigate sources of heterogeneity using both subgroup and metaregression analyses. Given these advantages and the statistic's ease of use, researchers should consider including the DOR in future reviews concerning forensic risk assessment.

3.5.7 Implications

According to the results of the present meta-analysis, two individuals who risk assessment tools judge to be at high risk need to be detained to prevent one

subsequent criminal incident in the community. In other words, for every individual who would go on to offend in the future who is detained to protect the public, there is someone who would not go on to offend who would also be detained. It was also concluded that of those individuals who risk assessment tools judge to be at low risk of future offending, between three and five may be released before a criminal incident occurs in the community. These results suggest that relying solely on risk assessment tools to make accurate predictions of dangerousness is problematic: for every 10 individuals detained, 5 would not have gone on to offend, and for every 10 individuals released, at least 2 will go on to commit an offence in the future.

With researchers suggesting that "a ceiling on the accuracy that can be achieved by both clinical and actuarial approaches" (Buchanan, 2008, p. 187) may have been hit, perhaps the field could benefit from improving the efficiency of the risk assessment process. A review of the literature suggests that no instruments have been developed with the explicit purpose of screening out very low risk participants prior to time-intensive and costly assessments of risk. Such screening tools could be particularly useful in populations with a low base rate of violence, such as individuals with schizophrenia (Walsh, Buchanan, & Fahy, 2002). In the next chapter, a violence screening tool for individuals diagnosed with schizophrenia will be developed. This instrument will be designed to reduce the burden of time and money on mental health services, which are currently expected to assess violence risk in every patient diagnosed with schizophrenia (American Psychiatric Association, 2004; National Institute for Health and Clinical Excellence, 2009).

3.5.8 Limitations

One limitation of the present meta-analysis is that sample data from all eligible studies was not obtainable. This may have biased estimates of clinical, methodological, and statistical heterogeneity. However, evidence was found that the effect sizes produced by the included studies were similar to those produced by the studies which were unable to be included. As a consequence of not being able to include all eligible studies, there was insufficient statistical power to examine sources of heterogeneity by individual instrument. Future reviews could attempt to improve the inclusion of more primary data from individual tool literatures. Initiatives to promote the registering of primary observational data might assist in this (Editorial, 2010). For example, the editors of leading general medical journal, *The Lancet*, recommend that observational studies be registered on a World Health Organization-compliant registry before they are conducted, that study protocols be submitted as part of the publication process, and that weblinks to primary data and study protocols be provided in published reports (Editorial, 2010). Making individual participant or tabular data from risk assessment studies publicly accessible would provide meta-analytic investigators with the opportunity to conduct more powerful analyses that could further clarify tool utility in different populations and settings.

A related limitation of the present meta-analysis is that it investigated the association between clinical covariates such as ethnic composition, mean age, and mean length of follow-up of samples and predictive validity. As only sample-level data was analysed, tool performance could not be investigated in whites versus non-whites and individuals of different ages across different lengths of follow-up. Future reviewers may wish to collect raw data from primary studies rather than outcome data for 2 x 2 tables. This would allow researchers to investigate the utility of risk

assessment tools in predicting short-term and longer-term offending in different groups using participant-level analyses.

A third limitation of the meta-analysis is that some outcomes were grouped together to reflect clinical practice. Clinical assessment is more likely to be interested in the risk of any serious offending (rather than separating out risks for sexual and violent offending). Nevertheless, some of the instruments included were designed specifically for sexual offenders, and pooling outcomes may have introduced some unaccounted for clinical heterogeneity into the analyses. Therefore, future work should investigate whether they have different rates of predictive validity when the outcome is purely sexual offending.

Another possible limitation is that studies were included regardless of their methodological quality in order to analyse a representative sample of the literature. As evidence of significant heterogeneity was found and the reporting quality of studies varied, it may be argued that having pooled results was not appropriate and that systematic review methodology should have been used and medians reported for all effect sizes (Deeks, Higgins, & Altman, 2006). To address this limitation, a thorough investigation of sources of heterogeneity was conducted. This investigation included traditional markers of methodological quality such as prospective or retrospective design and sample size. The study characteristics chosen for investigation were identified *a priori*. Nevertheless, as a number of tests of clinical, methodological, and statistical heterogeneity were conducted, the threshold for significance needs to take this into account and caution is warranted if some of the significant findings are taken on their own without considering other potentially relevant clinical factors that risk assessment instruments do not measure. This limitation was addressed by summarising significant findings in relation to different thresholds of evidentiary strength (Table 3.10).

A fifth limitation of the meta-analysis is that while a number of sources of between-study heterogeneity were investigated, it is possible that there are additional study characteristics that contribute to effect size variation that were not explored. For example, the clinical background of the individuals who administered the risk assessment tools was not included. Previous research has suggested that this may be an important moderator of risk assessment tool validity (Arkes, 1981, 1991; Nisbett & Ross, 1980; Spengler & Strohmer, 2001).

3.5.9 Conclusion

Risk assessment tools are increasingly used to make important decisions in clinical and criminal justice settings. The present meta-analysis found evidence that structured risk instruments predict the likelihood of future offending better than chance, suggesting that they have a role to play in the risk assessment process, but one that may need changing. The review found that the predictive validity of commonly used risk assessment measures varies widely. Findings suggested that tools designed for more specific purposes were more accurate at detecting individuals' risk of future offending. It was also found that the closer the demographic characteristics of the tested sample are to the original calibration sample of a tool, the higher the rate of predictive validity. As this review identified substantial variations in the predictive accuracy of commonly used instruments and heterogeneity in their validity according to different demographic factors, risk assessment procedures and guidelines by mental health services and criminal justice systems may need review. In the following chapter, an instrument will be designed for a specific psychiatric population to increase the efficiency of the risk assessment process.

Table 3.1 Characteristics of Nine Risk Assessment Tools Investigated in the Meta-analysis

Tool	Act vs. SCJ	Items	Description	Point System	Type of Offending Predicted	Authors	Domains
LSI-R	Act	54	Designed to use psychosocial status to predict the likelihood of general recidivism in adult offenders. The tool is designed to assist professionals make decisions regarding level of supervision and treatment	0 = item present 1 = item absent	General	Andrews & Bonta (1995)	(1) Criminal history (2) Leisure/Recreation (3) Education/Employment (4) Companions (5) Financial (6) Alcohol/Drug problems (7) Family/Marital (8) Emotional/Personal (9) Accommodation (10) Attitude/Orientation

PCL-R	Act	20	Designed to diagnose psychopathy as operationally defined in Cleckley's (1941) *The Mask of Sanity*	0 = item does not apply 1 = item applies to a certain extent 2 = item applies	N/A (Tool not developed for the purpose of forensic risk assessment)	Hare (1991, 2003)	(1) Selfish, callous, and remorseless use of others (2) Chronically unstable and antisocial lifestyle
SORAG	Act	14	Designed to assess the likelihood of violent (including sexual) recidivism specifically in previously convicted sex offenders	N/A (Different point values awarded for different items)	Violent	Quinsey, Harris, Rice, & Cormier (1998, 2006)	N/A
Static-99	Act	10	Designed to predict the long-term probability of sexual recidivism amongst adult male offenders who have committed a sexual offence	N/A (Different point values awarded for different items)	Sexual	Hanson & Thornton (1999) Harris, Phenix, Hanson, & Thornton (2003)	N/A
VRAG	Act	12	Designed to be used to predict risk of violence in previously violent mentally disordered offenders	N/A (Different point values awarded for different items)	Violent	Quinsey, Harris, Rice, & Cormier (1998, 2006)	N/A

HCR-20	SCJ	20	Designed to assess violence risk in criminal justice, forensic, and general psychiatric settings	0 = item not present 1 = item possibly present 2 = item definitely present	Violent	Webster, Eaves, Douglas, & Wintrup (1995) Webster, Douglas, Eaves, & Hart (1997)	(1) Historical factors (2) Clinical factors (3) Risk management factors
SARA	SCJ	20	Designed to predict future violence in men arrested for spousal assault	0 = item not present 1 = item possibly present 2 = item definitely present	Violent	Kropp & Hart (1994, 1995, 1999)	(1) Criminal history (2) Psychosocial adjustment (3) Spousal assault history (4) Alleged/current offence
SAVRY	SCJ	24	Designed to assess the risk of violence in adolescents	0 = item presents a low risk of reoffending 1 = item presents a moderate risk of reoffending 2 = item presents a high risk of reoffending	Violent	Borum, Bartel, & Forth (2002, 2003)	(1) Historical risk factors (2) Social/Contextual risk factors (3) Individual/Clinical risk factors (4) Protective factors

SVR-20	SCJ	20	Designed to predict the risk of violence (including sexual violence) in sex offenders	0 = item does not apply 1 = item possibly applies 2 = item definitely applies	Violent	Boer, Hart, Kropp, & Webster (1997)	(1) Psychosocial adjustment (2) Sexual offences (3) Future plans

Table 3.2 Descriptive and Demographic Characteristics of 88 Samples Investigating the Predictive Validity of Nine Risk Assessment Tools

Category	Subcategory	Group	Number of *k* = 88 (%)
Tool information	Tool used	HCR-20	9 (10.2)
		LSI-R	8 (9.1)
		PCL-R	20 (22.7)
		SARA	4 (4.5)
		SAVRY	9 (10.2)
		SORAG	7 (8.0)
		Static-99	14 (15.9)
		SVR-20	5 (5.7)
		VRAG	12 (13.6)
	Type of tool	Actuarial	61 (69.3)
		SCJ	27 (30.7)
Sample demographics	Male participants (per sample)	Mean (*SD*)	295 (429)
	White participants (per sample)	Mean (*SD*)	138 (118)
	Age (in years)	Mean (*SD*)	31.6 (7.6)
Study design	Study setting	Community	4 (4.5)
		Prison	37 (42.0)
		Psychiatric	33 (37.5)
		Mixed	14 (15.9)
	Temporal design	Prospective	40 (45.5)
		Retrospective	43 (48.9)
		Unstated/Unclear	5 (5.7)
	Length of follow-up (months)	Mean (*SD*)	56.3 (41.3)
	Sample size	Mean (*SD*)	296 (422)
	Country of origin	United States of America	14 (15.9)
		Canada	31 (35.2)
		United Kingdom	9 (10.2)
		Other European Union	31 (35.2)
		Latin America	2 (2.3)
		Australia/New Zealand	1 (1.1)
	Type of offending	General (any)	46 (52.3)
		Violent only	40 (45.5)
		Non-violent only	1 (1.1)
		Unstated/Unclear	1 (1.1)
	Type of outcome	Arrest/Charge/Conviction	56 (63.6)
		Institutional incident	11 (12.5)
		Mixed	16 (18.2)
		Unstated/Unclear	5 (5.7)
	Source of information used to score tool	File review	52 (59.1)
		Interview	2 (2.3)
		Mixed	16 (18.2)
		Unstated/Unclear	18 (20.5)
	Source of study	Journal article	64 (72.7)
		Conference	7 (8.0)
		Dissertation	13 (14.8)
		Government report	4 (4.5)

Note. k = number of samples; SCJ = structured clinical judgement instrument; SD = standard deviation.

Table 3.3 Median Area Under the Curve Produced by Nine Risk Assessment Tools Ranked in Order of Strength

Tool	*n*	*k*	Median AUC	*IQR*
SVR-20	380	3	0.78	0.71-0.83
SORAG	1599	6	0.75	0.69-0.79
VRAG	2445	10	0.74	0.74-0.81
SAVRY	915	8	0.71	0.69-0.73
HCR-20	1320	8	0.70	0.64-0.76
SARA	102	1	0.70	–
Static-99	8246	12	0.70	0.62-0.72
LSI-R	856	3	0.67	0.55-0.73
PCL-R	2645	10	0.66	0.54-0.68

Note. n = sample size; k = number of samples; AUC = area under the curve; IQR = interquartile range.

Table 3.4 Median Positive Predictive and Negative Predictive Values Produced by Nine Risk Assessment Tools Ranked in Order of Strength

				PPV		NPV	
Binning	Tool	*n*	*k*	Median	*IQR*	Median	*IQR*
First	SAVRY	1026	9	0.76	0.42-0.85	0.76	0.49-0.91
	HCR-20	1374	9	0.71	0.55-0.85	0.67	0.51-0.70
	VRAG	2703	12	0.66	0.37-0.79	0.74	0.62-0.84
	LSI-R	4005	8	0.57	0.44-0.70	0.53	0.44-0.65
	SARA	2305	4	0.53	0.31-0.62	0.79	0.67-0.92
	PCL-R	3854	20	0.52	0.45-0.75	0.68	0.39-0.82
	SORAG	1637	7	0.38	0.33-0.86	0.64	0.59-0.90
	Static-99	8555	14	0.33	0.18-0.56	0.82	0.71-0.94
	SVR-20	521	5	0.33	0.16-0.60	0.65	0.56-0.87
Second	HCR-20	1320	8	0.64	0.45-0.70	0.80	0.71-0.86
	SAVRY	1026	9	0.60	0.27-0.73	0.90	0.74-0.95
	PCL-R	3854	20	0.52	0.45-0.75	0.68	0.39-0.82
	LSI-R	4005	8	0.48	0.24-0.67	0.57	0.54-0.75
	SORAG	1599	6	0.40	0.19-0.71	0.88	0.77-0.96
	VRAG	2602	11	0.39	0.23-0.54	0.89	0.83-0.92
	SARA	465	3	0.37	0.10-0.60	0.96	0.82-0.98
	SVR-20	268	3	0.23	0.18-0.72	0.71	0.53-0.78
	Static-99	8097	10	0.18	0.08-0.31	0.95	0.76-0.99

Note. First = high risk versus low/moderate risk binning strategy; Second = moderate/high risk versus low risk binning strategy; n = sample size; k = number of samples; IQR = interquartile range; PPV = positive predictive value; NPV = negative predictive value.

Table 3.5 Median Number Needed to Detain and Number Safely Screened Produced by Nine Risk Assessment Tools Ranked in Order of Strength

				NND		NSS	
Binning	Tool	*n*	*k*	Median	*IQR*	Median	*IQR*
First	SAVRY	1026	9	1	1-3	4	2-12
	HCR-20	1374	9	1	1-2	3	2-4
	VRAG	2703	12	2	1-3	4	3-6
	LSI-R	4005	8	2	1-2	2	2-3
	SARA	2305	4	2	2-5	6	3-16
	PCL-R	3854	20	2	1-2	3	2-6
	SORAG	1637	7	3	1-3	3	2-10
	SVR-20	521	5	3	2-6	4	6-24
	Static-99	8555	14	4	2-6	3	2-11
Second	SAVRY	1026	9	2	1-4	11	4-21
	HCR-20	1320	8	2	1-2	5	4-7
	PCL-R	3854	20	2	1-2	3	2-6
	LSI-R	4005	8	2	1-4	2	2-4
	SARA	465	3	3	3-7	27	17-37
	SORAG	1599	6	3	2-6	12	4-22
	VRAG	2602	11	3	2-4	9	6-13
	SVR-20	268	3	4	1-6	4	2-5
	Static-99	8097	10	6	3-14	24	4-86

Note. First = high risk versus low/moderate risk binning strategy; Second = moderate/high risk versus low risk binning strategy; n = sample size; k = number of samples; IQR = interquartile range; NND = number needed to detain; NSS = number safely screened.

Table 3.6 Pooled Diagnostic Odds Ratios for Nine Risk Assessment Tools Ranked in Order of Strength

Binning	Model	Tool	*n*	*k*	DOR	95% CI
First	FE	SAVRY	1026	9	6.93	4.93-9.73
	FE	VRAG	2703	12	3.84	2.85-5.16
	FE	HCR-20	1374	9	3.48	2.62-4.62
	FE	SARA	2305	4	3.42	2.72-4.29
	RE	Static-99	8555	14	3.12	1.94-5.02
	FE	SORAG	1637	7	2.52	2.15-2.96
	RE	PCL-R	3854	20	2.08	1.14-3.81
	RE	LSI-R	4005	8	1.75	0.96-3.22
	RE	SVR-20	521	5	1.56	0.36-6.84
Second	FE	SARA	465	3	7.87	3.12-19.87
	FE	SAVRY	1026	9	6.40	4.40-9.32
	FE	SORAG	1599	6	5.54	4.09-7.50
	FE	VRAG	2602	11	5.21	3.61-7.53
	FE	HCR-20	1320	8	4.90	3.65-6.56
	FE	Static-99	8097	10	2.95	2.38-3.66
	RE	PCL-R	3854	20	2.08	1.14-3.81
	RE	LSI-R	4005	8	1.26	0.77-2.06
	RE	SVR-20	268	3	1.21	0.18-8.32

Note. First = high risk versus low/moderate risk binning strategy; Second = moderate/high risk versus low risk binning strategy; RE = random effects model where $I^2 \geq 75\%$; FE = fixed effects model (where $I^2 < 75\%$); n = sample size; k = number of samples; DOR = diagnostic odds ratio; CI = confidence interval.

Table 3.7 Summary Performance Scores of Nine Risk Assessment Tools across Four Outcome Measures

Tool	Summary Performance Score		
	First Binning Strategy	Second Binning Strategy	Total
SAVRY	31	29	60
VRAG	28	23	51
SARA	22	25	47
HCR-20	23	23	46
SORAG	20	25	45
Static-99	18	16	34
SVR-20	15	15	30
PCL-R	13	13	26
LSI-R	11	11	22

Note. First Binning Strategy = high risk versus low/moderate risk binning strategy; Second Binning Strategy = moderate/high risk versus low risk binning strategy. The four outcome measures included the area under the curve, positive predictive and negative predictive values, and diagnostic odds ratio. As the number needed to detain and the number safely screened are derivatives of the positive and negatives predictive values (respectively), they were excluded to avoid double-counting. Summary performance scores were calculated by ordering tools from poorest to strongest performance on each effect estimate. Each tool was assigned a score of +1 (poorest performance) through +9 (strongest performance) on each outcome measure. These values were then summed for each tool, yielding a composite performance score.

Table 3.8 Subgroup Analyses Investigating Sources of Heterogeneity in Replication Samples of Nine Risk Assessment Tools

	Diagnostic Odds Ratio (95% CI)	
Sample or Study Characteristic	First Binning Strategy	Second Binning Strategy
Type of risk assessment tool		
Actuarial	2.88 (2.21-3.75)	2.77 (2.08-3.69)
Structured clinical judgement	4.01 (2.81-5.71)	4.15 (2.42-6.75)
Gender		
Male sample data	3.15 (2.48-4.00)	3.35 (2.44-4.59)
Female sample data	4.93 (2.70-9.01)	5.12 (2.97-8.83)
Ethnicity		
<90% White	3.59 (2.38-5.42)	3.86 (2.50-5.96)
≥90% White	2.02 (1.03-3.94)	3.20 (1.08-9.48)
Mean age of participants [b]		
<25 years	4.40 (1.99-9.70)	4.51 (2.13-9.54)
≥25 years	3.26 (2.50-4.24)	3.21 (2.39-4.30)
Study setting		
Institution [c]	3.36 (2.58-4.39)	3.47 (2.56-4.70)
Community	3.13 (1.39-7.06)	3.00 (1.21-7.41)
Prison	3.65 (2.56-5.22)	3.20 (2.19-4.67)
Psychiatric unit	2.96 (1.96-4.45)	4.05 (2.59-6.32)
Temporal design		
Prospective	3.14 (2.31-4.28)	3.16 (2.19-4.57)
Retrospective	3.40 (2.49-4.65)	3.58 (2.63-4.90)
Mean length of follow-up		
≤2 years	2.80 (1.68-4.65)	2.16 (1.32-3.54)
>2 years	3.56 (2.73-4.65)	4.01 (3.10-5.18)
Sample size		
<500 participants	3.34 (2.67-4.18)	3.52 (2.72-4.56)
≥500 participants	2.35 (1.30-4.25)	2.21 (1.26-3.87)
Country of origin		
North America	2.78 (2.11-3.67)	3.27 (2.38-4.50)
Europe	3.85 (2.76-5.38)	3.49 (2.35-5.18)
Type of offending		
General	2.52 (1.90-3.36) [a]	2.54 (1.81-3.56) [a]
Violent	4.52 (3.59-5.70) [a]	4.78 (3.80-6.01) [a]
Type of outcome		
Arrest/Charge/Conviction	3.07 (2.37-3.97)	3.21 (2.42-4.26)
Institutional incident	7.49 (3.21-17.47)	5.30 (2.11-13.36)
Source of information		
File review only	3.28 (2.45-4.38)	3.59 (2.67-4.83)
Other	2.62 (1.64-4.20)	3.19 (1.84-5.51)
Tool authorship		
Tool author as study author [d]	4.32 (3.16-5.92)	6.22 (4.68-8.26) [a]
Tool author not study author	3.04 (2.39-3.86)	2.96 (2.27-3.86) [a]
Publication status		
Peer-reviewed journal	3.19 (2.52-4.03)	3.34 (2.53-4.42)
Other	3.22 (2.08-4.99)	3.27 (2.07-5.16)

Note. First Binning Strategy = high risk versus low/moderate risk binning strategy; Second Binning Strategy = moderate/high risk versus low risk binning strategy; CI = confidence interval.

[a] Non-overlapping confidence intervals.

[b] SAVRY samples included.

[c] Prison or psychiatric unit.

[d] Author of English-language version of risk assessment tool under investigation.

Table 3.9 Metaregression Analyses Investigating Sources of Heterogeneity in Replication Samples of Nine Risk Assessment Tools

	First Binning Strategy			Second Binning Strategy		
Sample or Study Characteristic	Power [a]	β (*SE*)	*p*	Power [a]	β (*SE*)	*p*
Type of risk assessment tool						
Actuarial vs. SCJ	0.99	0.40 (0.30)	0.18	0.98	0.35 (0.27)	0.13
Gender						
Male sample data vs. female sample data	0.73	0.80 (0.57)	0.17	0.65	0.84 (0.63)	0.19
Continuous	0.96	-0.02 (0.01)	0.10	0.93	-0.02 (0.01)	0.19
Ethnicity						
<90% White vs. ≥90% White	0.59	-0.10 (0.63)	0.87	0.48	0.11 (0.60)	0.85
Continuous	0.59	0.01 (0.21)	0.21	0.48	0.02 (0.01)	0.04 [b]
Mean age of participants [c]						
<25 years vs. ≥25 years	0.96	-0.49 (0.38)	0.20	0.93	-0.32 (0.40)	0.43
Continuous	0.96	-0.02 (0.02)	0.26	0.93	-0.01 (0.02)	0.55
Study setting						
Institution [d] vs. community	0.94	-0.45 (0.50)	0.38	0.91	-0.35 (0.56)	0.54
Prison vs. psychiatric unit	0.91	0.02 (0.36)	0.95	0.85	0.22 (0.34)	0.52
Temporal design						
Prospective vs. retrospective	0.99	0.10 (0.29)	0.74	0.97	0.07 (0.28)	0.81
Mean length of follow-up						
≤2 years vs. >2 years	0.98	0.13 (0.31)	0.69	0.96	0.48 (0.29)	0.14
Continuous	0.98	-0.01 (0.01)	0.61	0.96	-0.01 (0.01)	0.85
Sample size						
<500 participants vs. ≥500 participants	0.98	-0.48 (0.40)	0.24	0.96	-0.40 (0.40)	0.32
Continuous	0.98	-0.01 (0.01)	0.28	0.96	-0.01 (0.01)	0.32
Country of origin						
North America vs. Europe	0.99	0.38 (0.28)	0.17	0.97	0.15 (0.28)	0.59
Type of offending						
General vs. violent	0.99	0.81 (0.21)	0.01 [b]	0.98	0.57 (0.21)	0.01 [b]
Type of outcome						
Charge/Conviction/Incarceration vs. institutional incident	0.92	0.98 (0.53)	0.07	0.87	0.32 (0.54)	0.56
Source of information						
File review only vs. other	0.98	-0.32 (0.36)	0.38	0.95	-0.27 (0.39)	0.49
Tool authorship						
Tool author as study author vs. tool author not study author [e]	0.99	-0.29 (0.42)	0.49	0.98	-0.87 (0.37)	0.02 [b]
Publication status						
Peer-reviewed journal vs. other	0.99	-0.16 (0.31)	0.61	0.98	-0.05 (0.30)	0.87

Note. First Binning Strategy = high risk versus low/moderate risk binning strategy; Second Binning Strategy = moderate/high risk versus low risk binning strategy; SE = standard error; SCJ = structured clinical judgement instrument.
[a] Calculated using Bonferroni-corrected α-level of 0.004, strong effect size (R^2 = 0.25), and the number of samples contributing to the individual analysis.
[b] $p < 0.05$.
[c] SAVRY samples included.
[d] Prison or psychiatric unit.
[e] Author of English-language version of risk assessment tool under investigation.

Table 3.10 Summary of Results from Examining Sources of Heterogeneity

Significant Difference [a]	Evidence of Trends [b]	No Significant Difference [c]
Mean age of participants [d,e]	Ethnic composition [d]	Type of risk assessment tool [f]
Type of offending	Tool authorship	Gender composition
		Study setting
		Temporal design
		Mean length of follow-up
		Sample size
		Country of origin
		Type of outcome
		Source of information
		Publication status

Note. Unless otherwise noted, the classification of a variable which was analysed in both dichotomous and continuous forms concerns both its variations. Significance level set at $\alpha = 0.05$.
[a] Variables significant in both binning strategies using subgroup or metaregression analyses.
[b] Variables significant in one binning strategy using subgroup or metaregression analyses.
[c] Variables not significant in either binning strategy using subgroup or metaregression analyses.
[d] Significant only in continuous form of variable.
[e] Significant only when SAVRY samples excluded.
[f] Actuarial versus structured clinical judgement.

Table 3.11 Comparison of the Pooled Diagnostic Odds Ratios Produced by Risk Assessment Tools with Behavioural Predictors and Medical Diagnostic Tests

Predictor	Outcome	DOR	95% CI
Risk assessment tools (First)	Antisocial behaviour	3.20	2.58-3.96
Risk assessment tools (Second)	Antisocial behaviour	3.33	2.60-4.27
Depression	Adherence with medication	3.03	1.96-4.89
Emotional support	Adherence with medication	1.83	1.27-2.66
Truancy [a]	Alcohol abuse	3.50	2.55-5.19
Mammogram	Breast cancer	117.27	65.06-211.41
Prostatic specific antigen test [b]	Prostate cancer	8.44	4.45-16.00
Magnetic resonance angiogram	Peripheral atherosclerosis	7.46	2.48-22.20
Duplex sonogram [c]	Renal artery stenosis	16.00	5.10-50.60

Note. First = high risk versus low/moderate risk binning strategy; Second = moderate/high risk versus low risk binning strategy; DOR = diagnostic odds ratio; CI = confidence interval.
[a]In 9th and 10th graders.
[b]Total prostatic specific antigen (tPSA) threshold set at 4.0 ng/ml.
[c]Renal-aortic ratio used as predictor.

Figure 3.1 Results of a Systematic Search Conducted to Identify Replication Studies of Nine Commonly Used Forensic Risk Assessment Tools

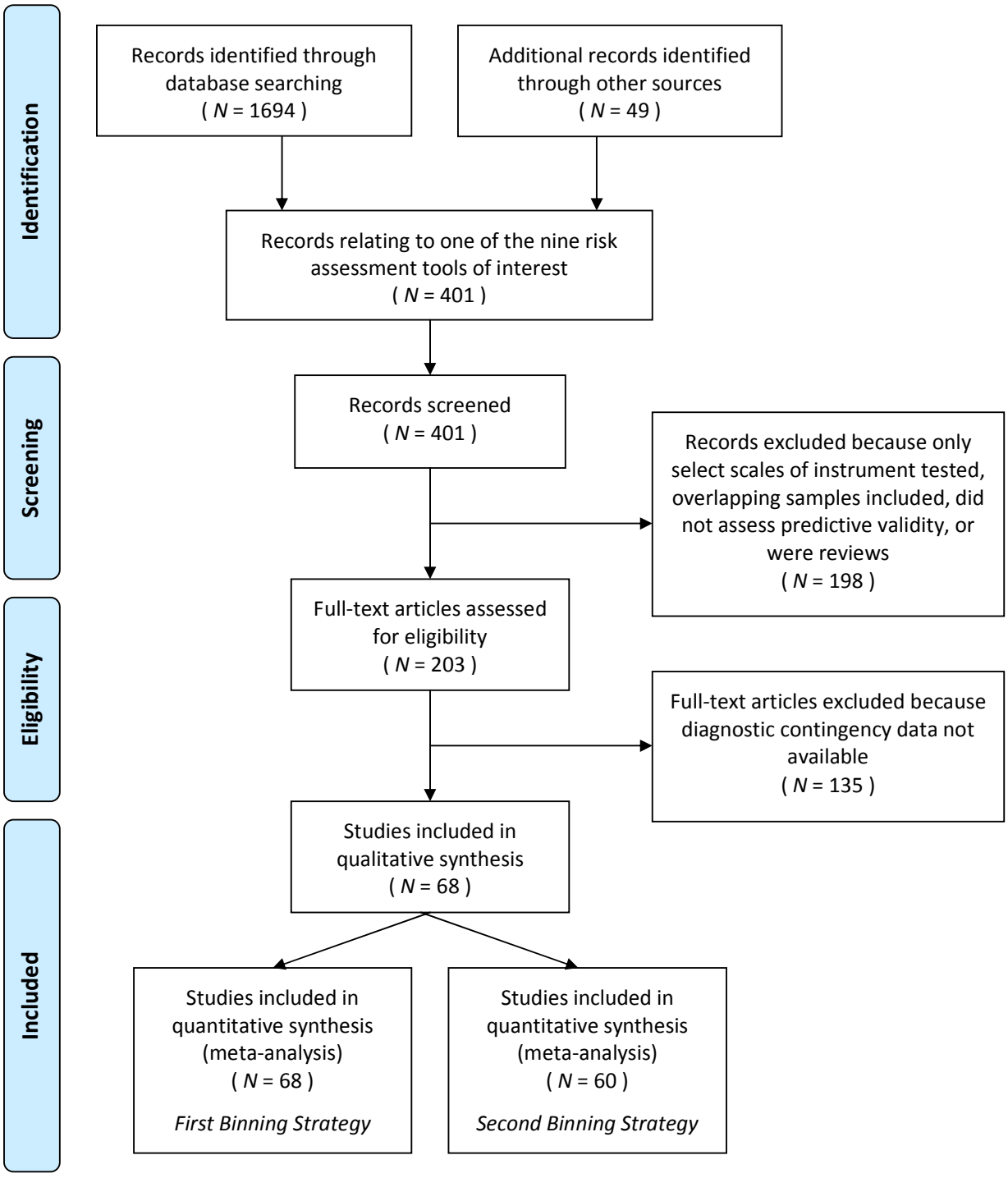


Figure 3.2 Histogram of the Areas Under the Curve Produced by Nine Commonly Used Forensic Risk Assessment Tools

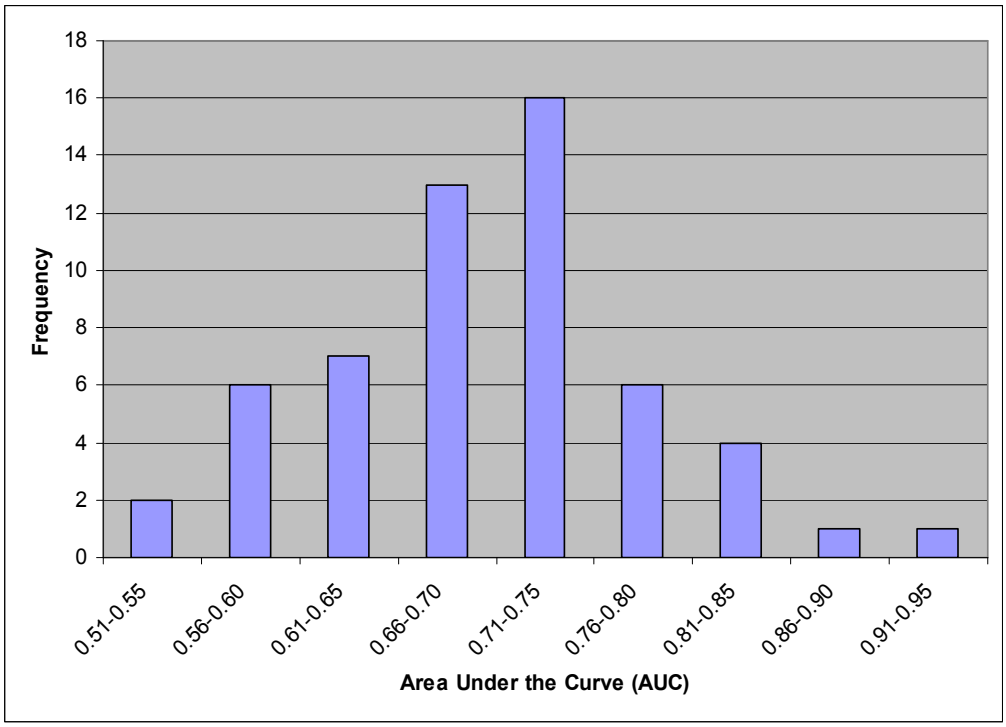


Figure 3.3 Histogram of the Positive Predictive Values Produced by Nine Commonly Used Forensic Risk Assessment Tools (High Risk versus Low/Moderate Risk)

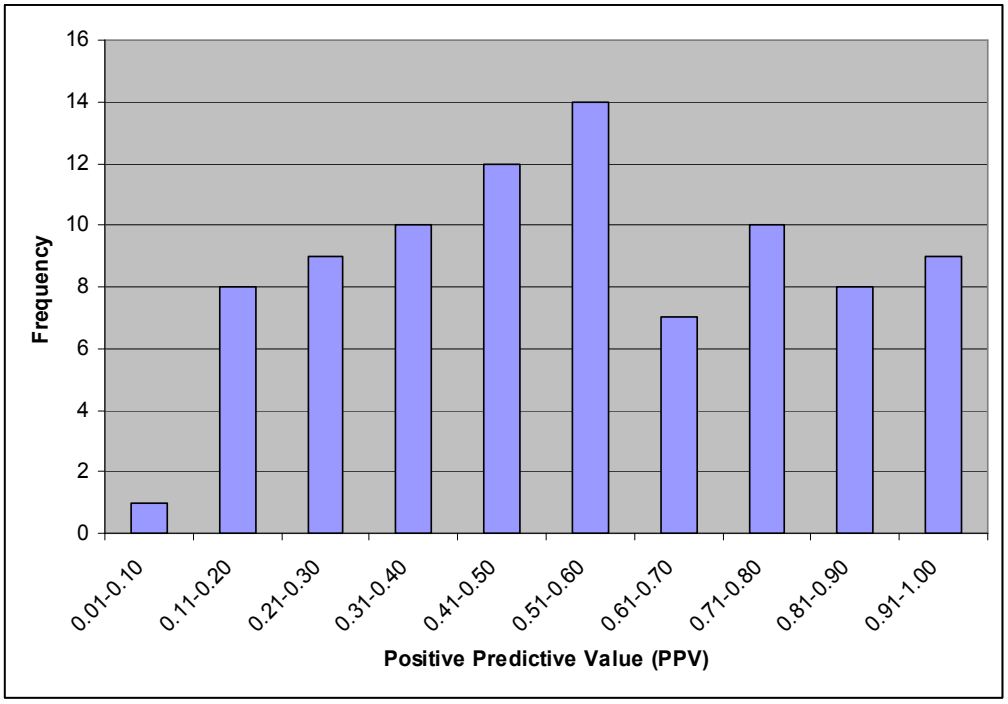


Figure 3.4 Histogram of the Negative Predictive Values Produced by Nine Commonly Used Forensic Risk Assessment Tools (High Risk versus Low/Moderate Risk)

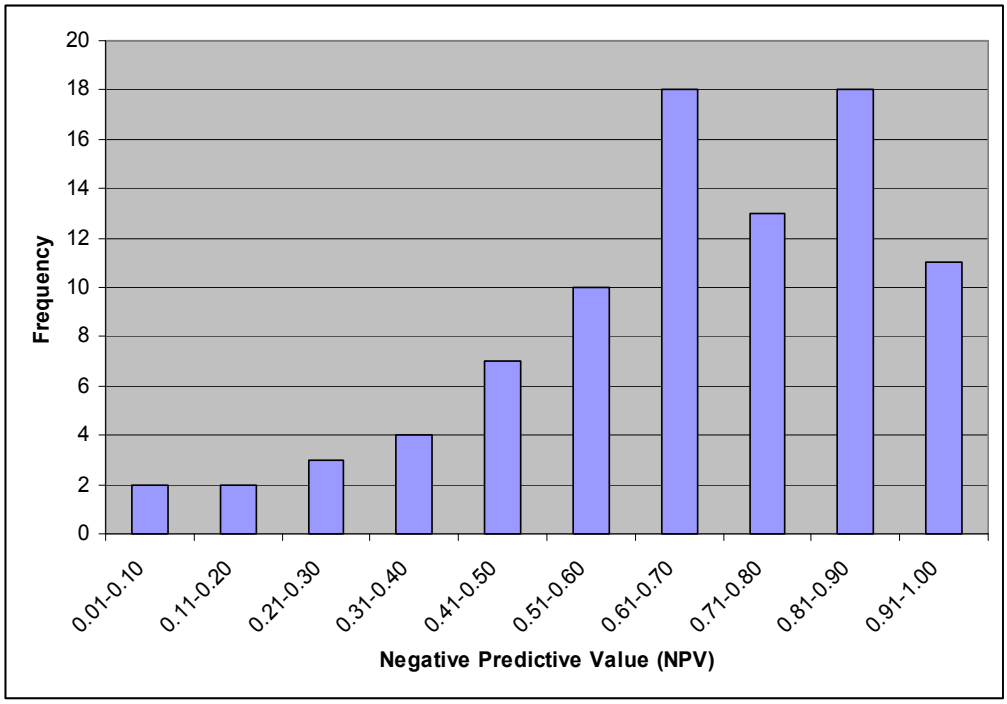


Figure 3.5 Histogram of the Positive Predictive Values Produced by Nine Commonly Used Forensic Risk Assessment Tools (Moderate/High Risk versus Low Risk)

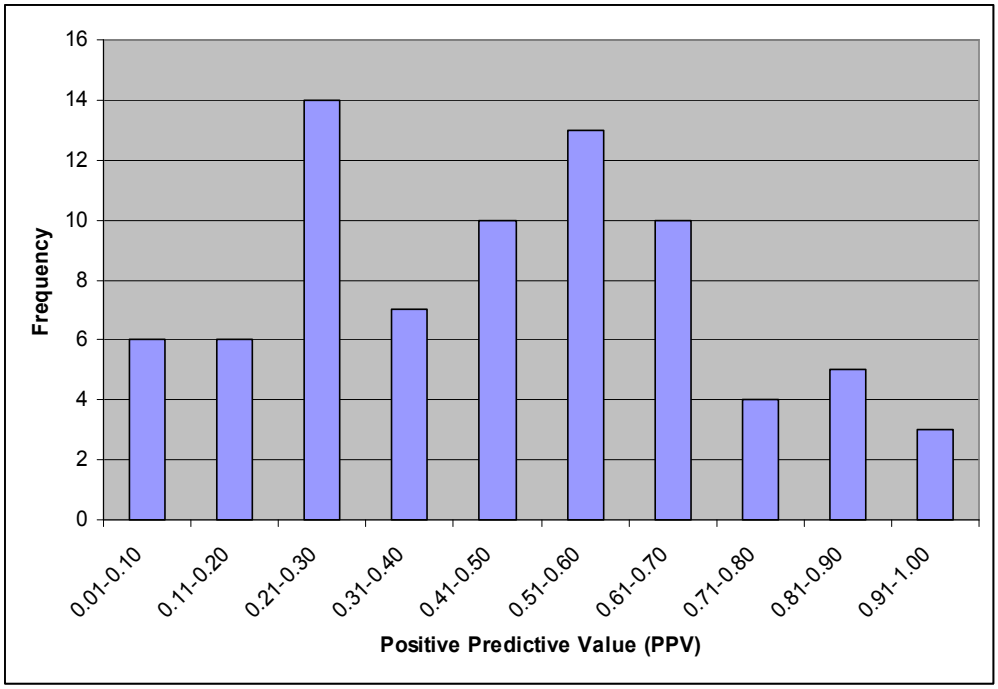


Figure 3.6 Histogram of the Negative Predictive Values Produced by Nine Commonly Used Forensic Risk Assessment Tools (Moderate/High Risk versus Low Risk)

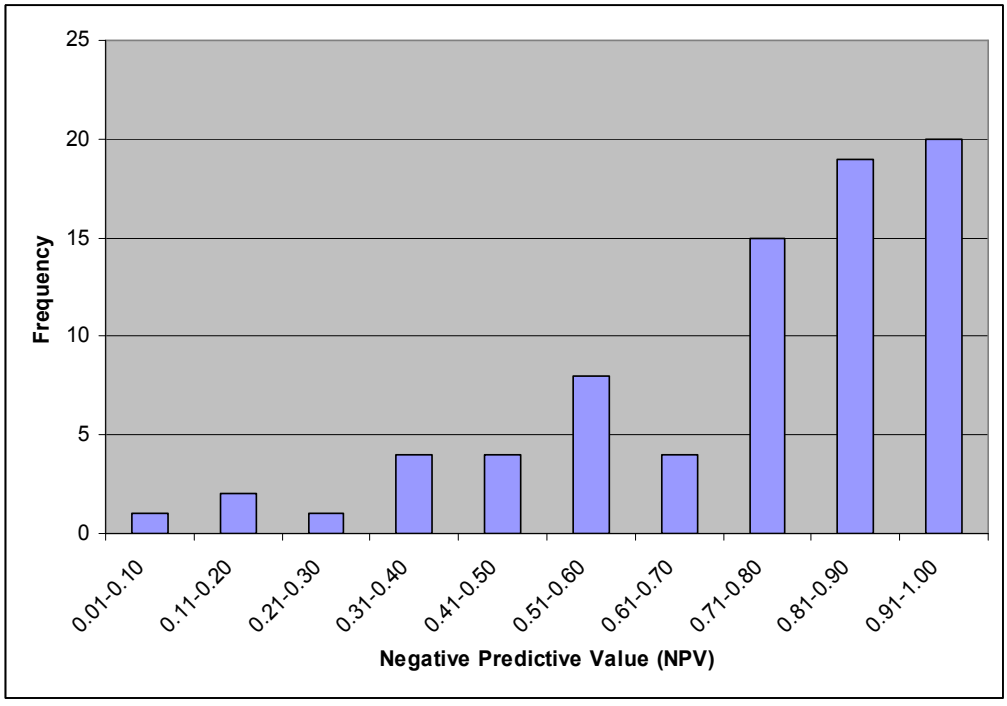


Figure 3.7 Histogram of the Numbers Needed to Detain Produced by Nine Commonly Used Forensic Risk Assessment Tools (High Risk versus Low/Moderate Risk)

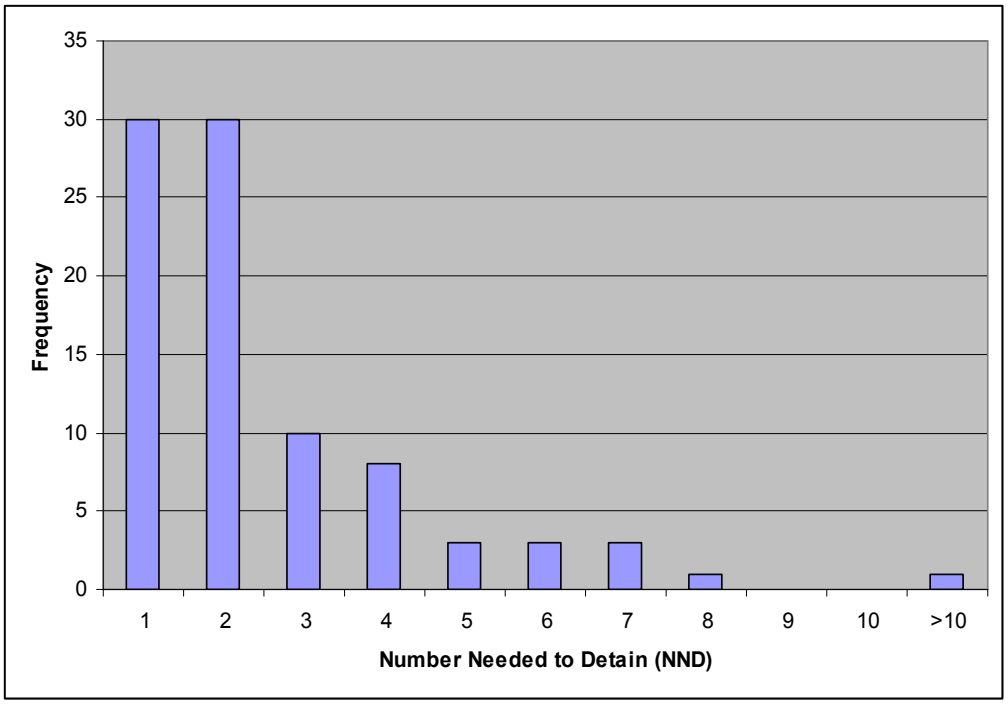


Figure 3.8 Histogram of the Numbers Safely Screened Produced by Nine Commonly Used Forensic Risk Assessment Tools (High Risk versus Low/Moderate Risk)

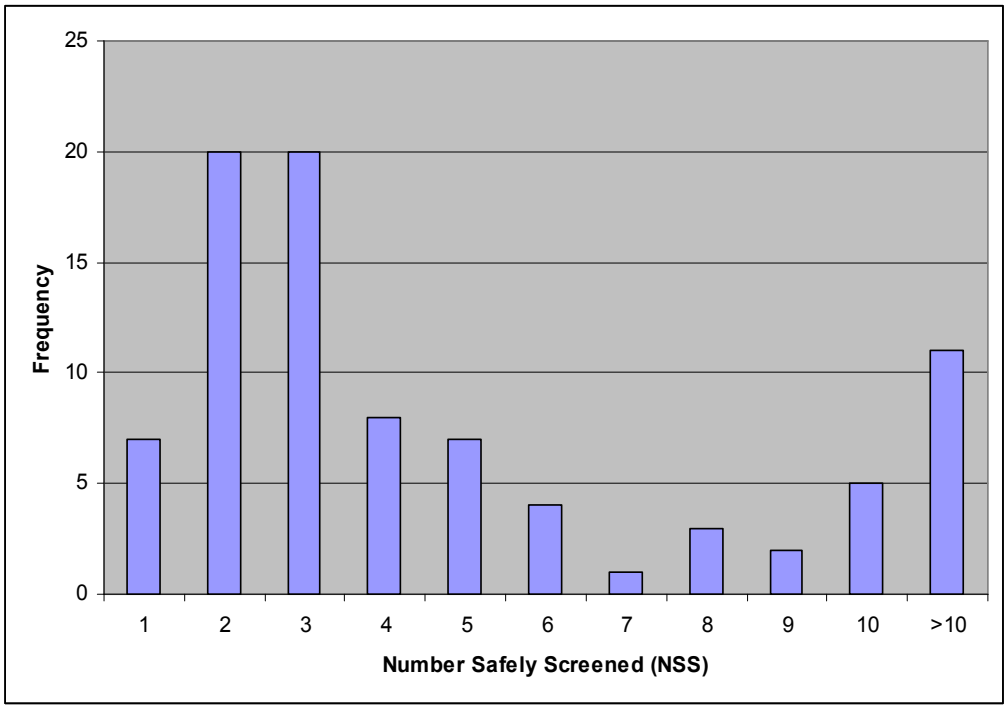


Figure 3.9 Histogram of the Numbers Needed to Detain Produced by Nine Commonly Used Forensic Risk Assessment Tools (Moderate/High Risk versus Low Risk)

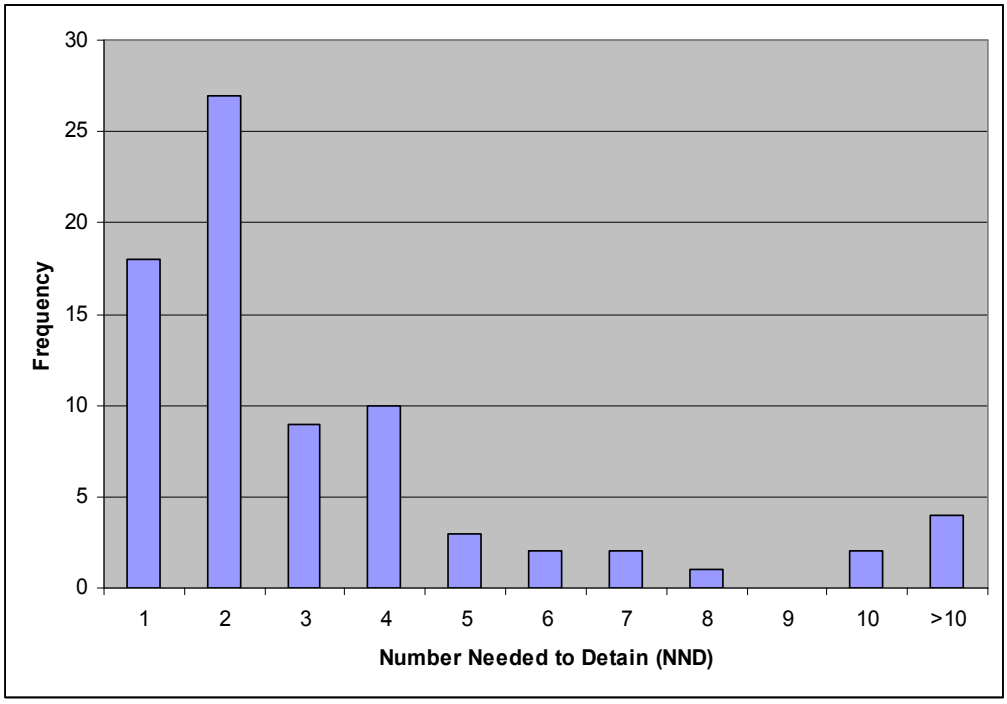


Figure 3.10 Histogram of the Numbers Safely Screened Produced by Nine Commonly Used Forensic Risk Assessment Tools (Moderate/High Risk versus Low Risk)

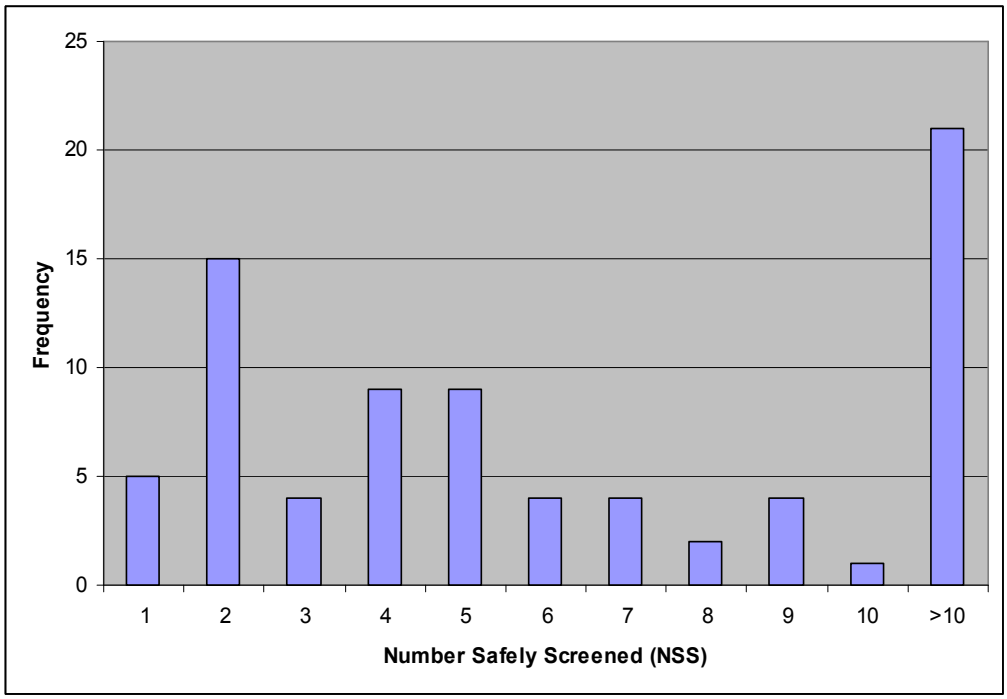


Figure 3.11 Forest Plot of the Diagnostic Odds Ratios Produced by Nine Commonly Used Forensic Risk Assessment Tools (High Risk versus Low/Moderate Risk)

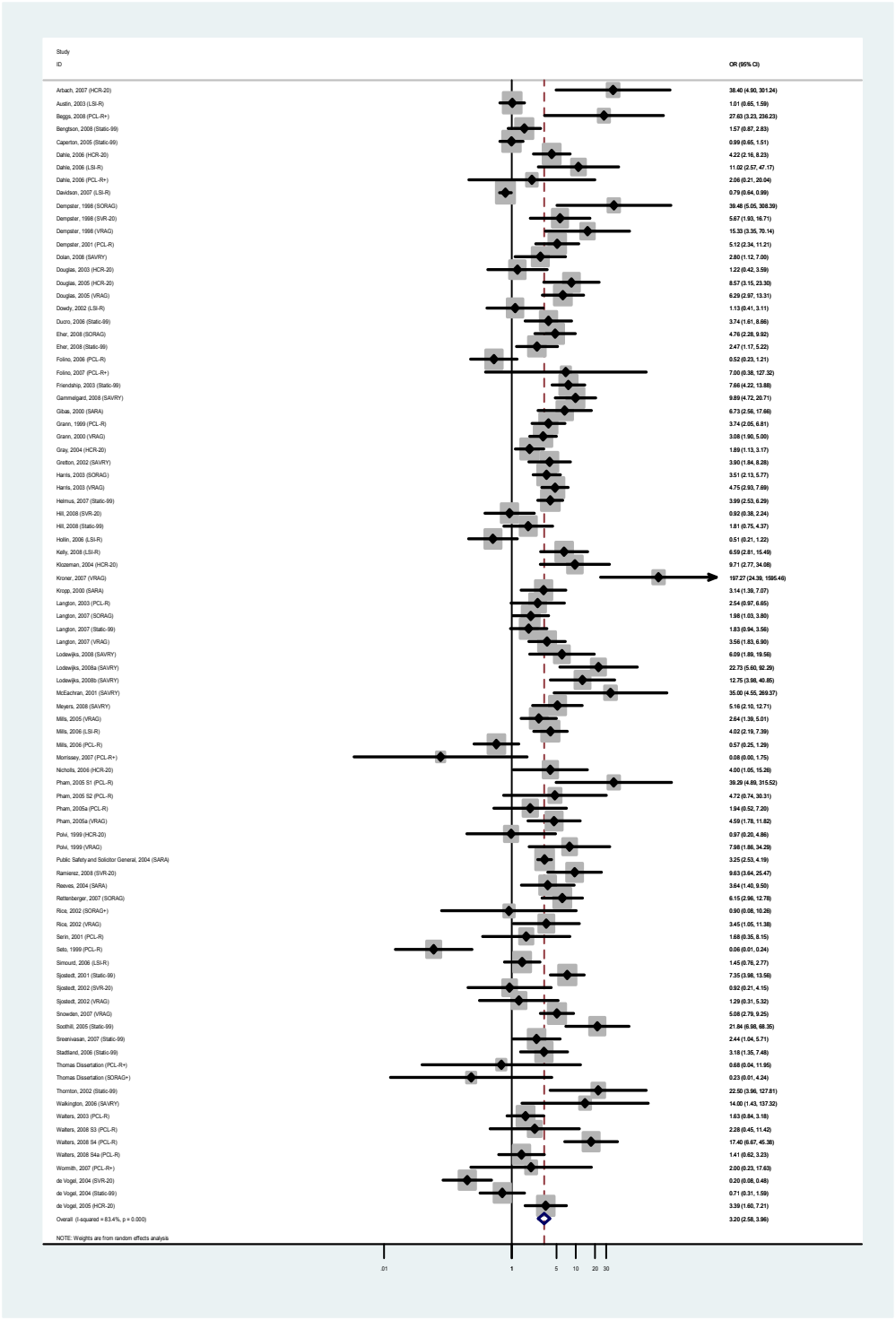


Figure 3.12 Forest Plot of the Diagnostic Odds Ratios Produced by Nine Commonly Used Forensic Risk Assessment Tools (Moderate/High Risk versus Low Risk)

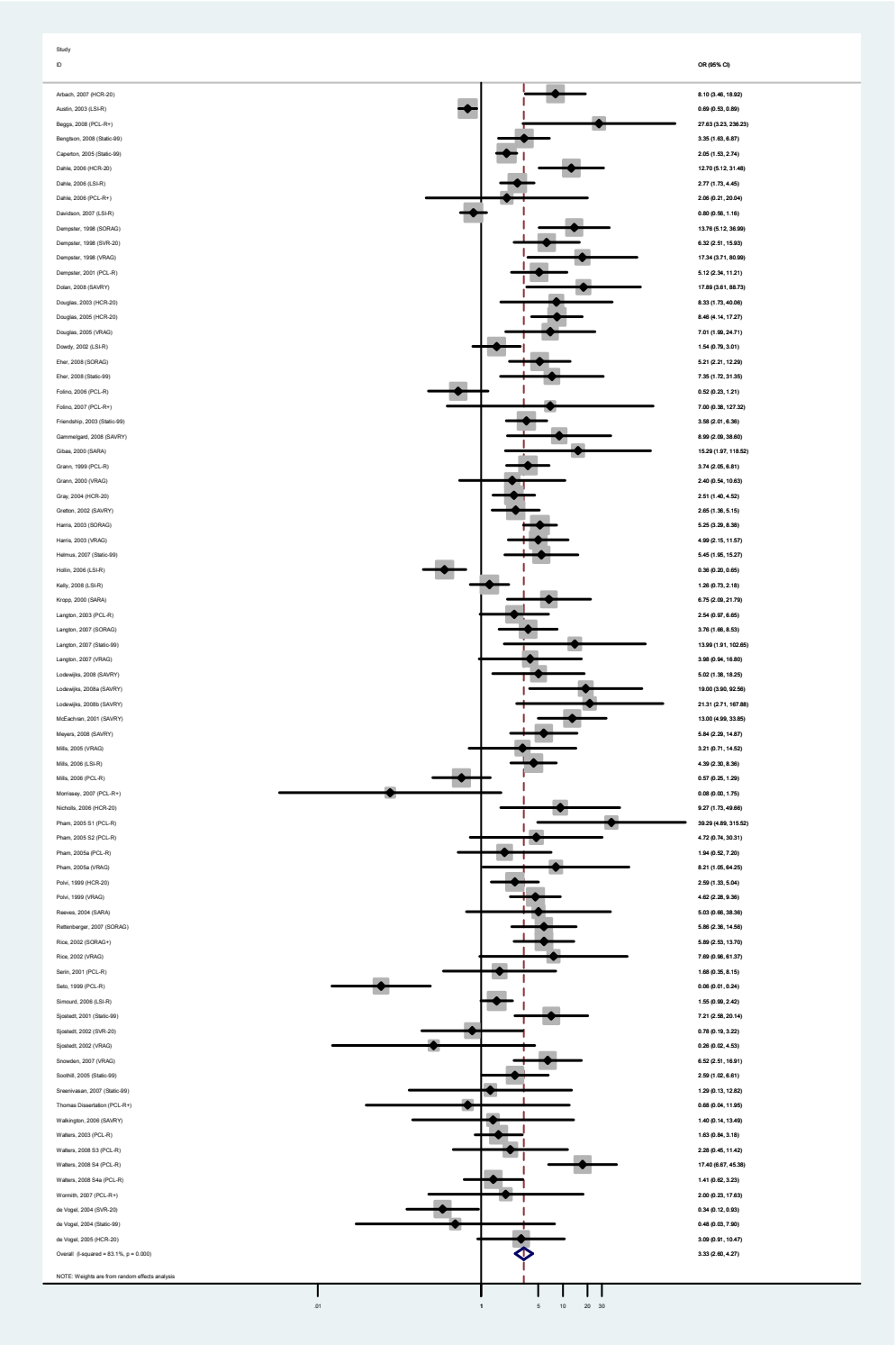


Figure 3.13 Forest Plot of the Diagnostic Odds Ratios Produced by Nine Commonly Used Forensic Risk Assessment Tools Used to Predict General versus Violent Offending (High Risk versus Low/Moderate Risk)

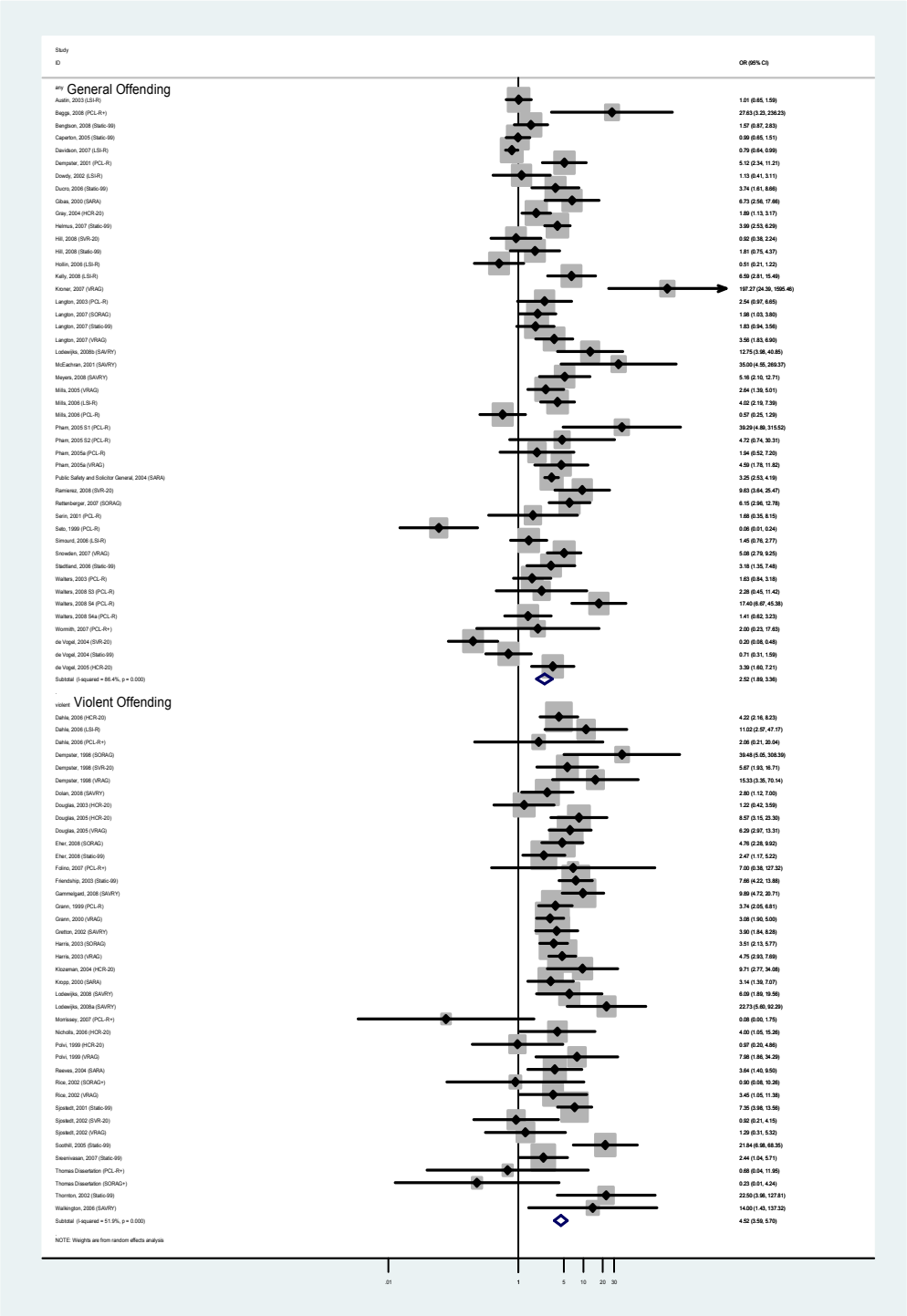


Figure 3.14 Forest Plot of the Diagnostic Odds Ratios Produced by Nine Commonly Used Forensic Risk Assessment Tools Used to Predict General versus Violent Offending (Moderate/High Risk versus Low Risk)

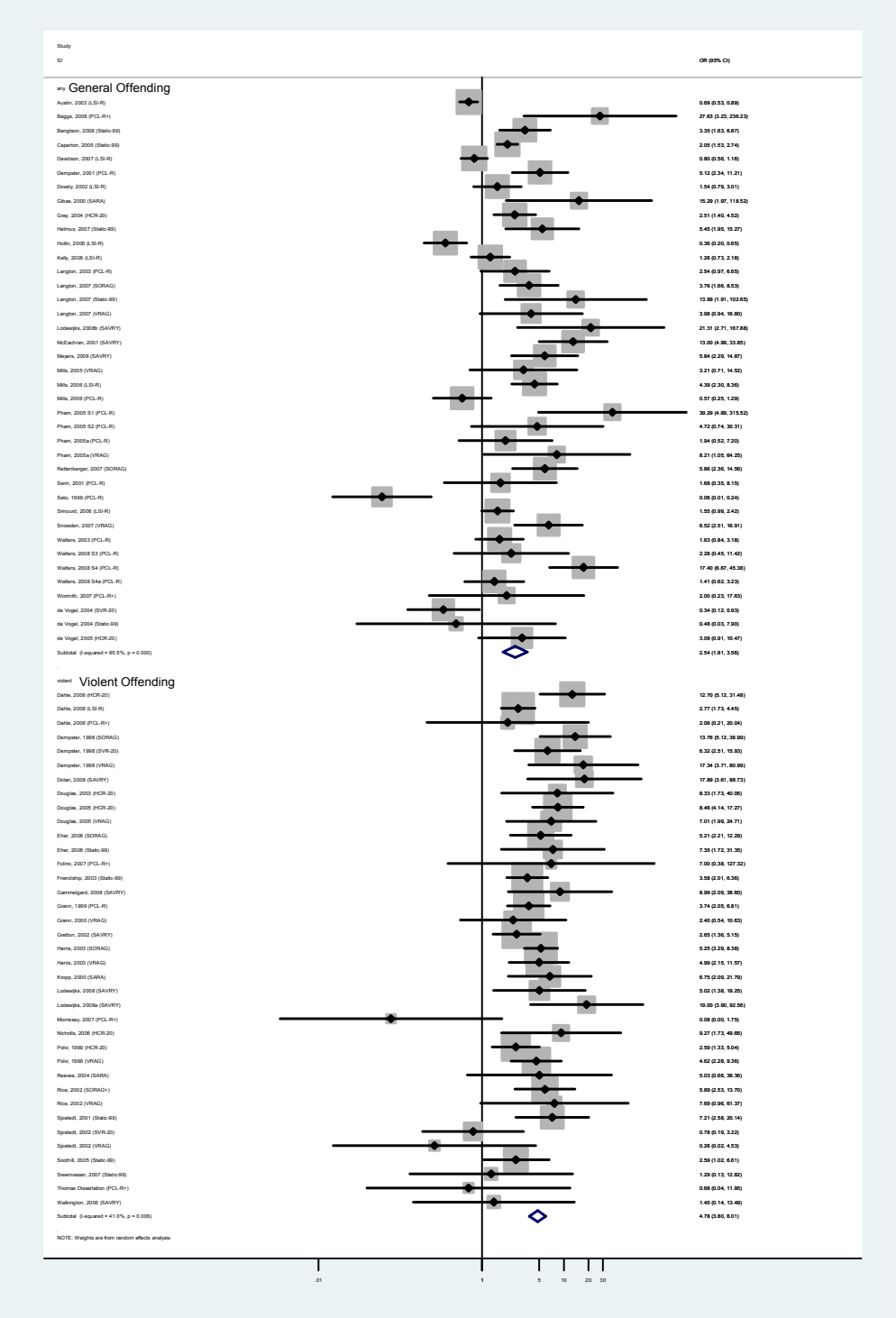
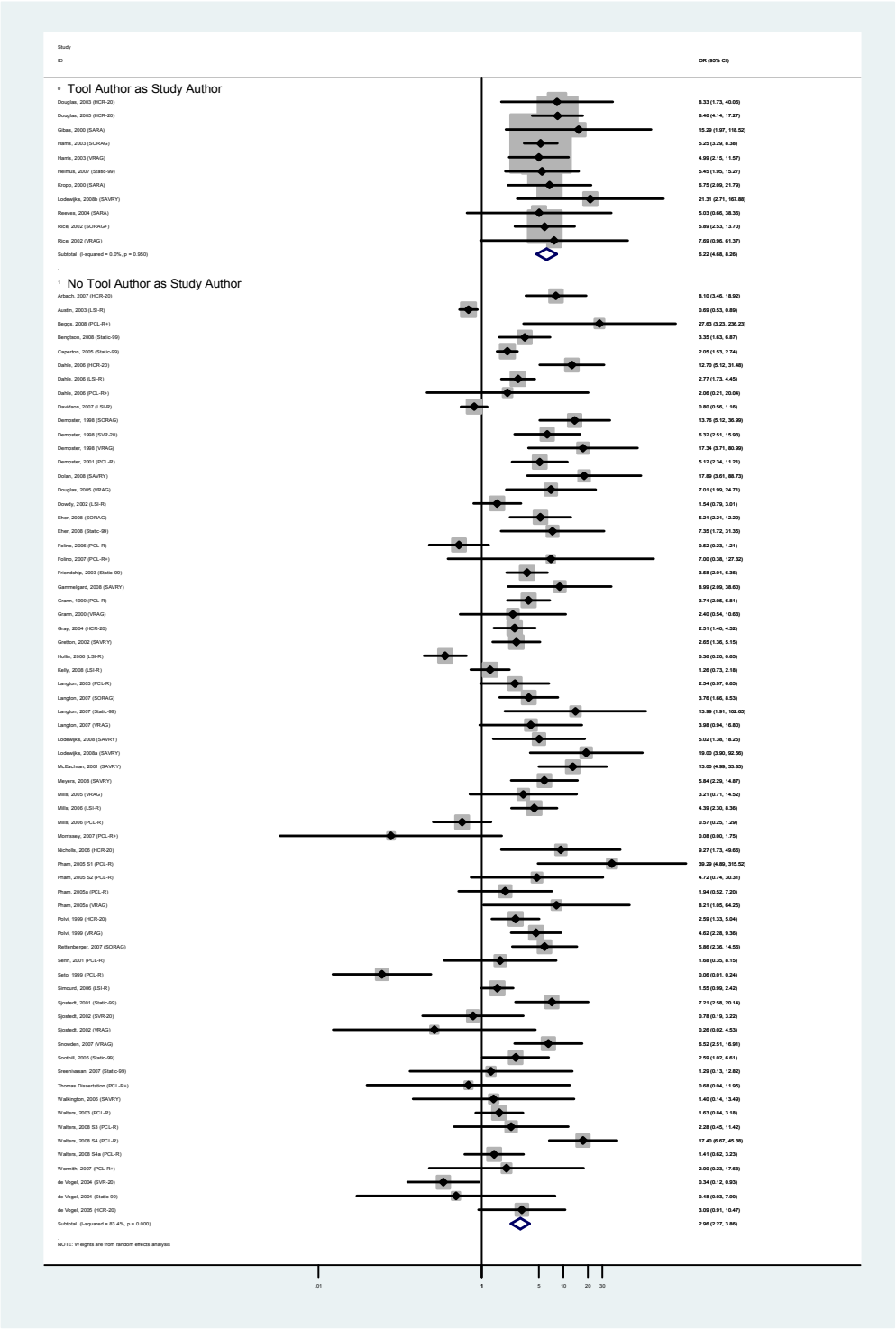


Figure 3.15 Forest Plot of the Diagnostic Odds Ratios Produced by Nine Commonly Used Forensic Risk Assessment Tools When a Tool Author was a Study Author versus Not (Moderate/High Risk versus Low Risk)



***

Chapter IV:
Developing a Violence Screening Tool for Schizophrenia

4.1 ABSTRACT

Clinical guidelines recommend that violence risk be assessed in schizophrenia. Current approaches employ detailed assessments of dangerousness for all patients, which can be costly and time-consuming. To assist in this process, a simple tool was developed to screen out individuals with schizophrenia at very low risk of violent offending. A national cohort of 13,806 individuals with hospital discharge diagnoses of schizophrenia was followed for up to 33 years for violent crime. A number of demographic, socio-economic and clinical risk factors were analysed using Cox regression to develop the screening tool, the predictive validity of which was measured using six outcome statistics. The instrument was calibrated on 6,903 participants and cross-validated using three independent replication samples of 2,301 participants each. Regression analyses resulted in a tool composed of five items: male gender, previous criminal conviction, young age at assessment, comorbid alcohol abuse, and comorbid drug abuse. At 5 years after discharge, the instrument had a negative predictive value of 0.99 (95% CI = 0.98-0.99), with only 57 of the 3,908 individuals who the tool screened out going on to be convicted of a violent offence. Weighting items by their regression coefficients or hazard ratios did not significantly improve predictive accuracy. Screening out patients who are at very low risk of violence prior to more detailed clinically based assessment may be an efficient and scalable approach for violence risk assessment in schizophrenia.

4.2 INTRODUCTION

As the number of forensic hospital beds has doubled in many Western countries over the past two decades (Priebe et al., 2008), valid risk assessment procedures are needed for psychiatric populations. The meta-analysis' findings that risk assessment tools developed for use with more specific populations produce higher rates of predictive validity and that instruments perform best when administered to samples similar to their calibration samples suggest that measures designed for specific diagnostic groups may produce higher rates of sensitivity and specificity than tools developed for heterogeneous psychiatric populations. This suggestion is supported by recent systematic reviews that have argued for the importance of violence risk assessment in psychiatric populations and have noted the need for diagnosis-specific assessments of dangerousness (Kumar & Simpson, 2005; Turgut et al., 2006; Woods & Ashley, 2007).[35] Despite these recommendations, the systematic searches conducted as part of this thesis identified no risk assessment tools designed to predict the likelihood of community violence in a specific diagnostic group. As reviews of the media literature (Klin & Lemish, 2008; Levey & Howells, 1994) and surveys conducted in Western countries (Luty, Fekadu, & Dhandayudham, 2006; National Alliance on Mental Illness, 2008; Pescosolido & Boyer, 1999; Pescosolido et al., 2010) have identified individuals with schizophrenia as the hospitalised group most commonly associated with violence, the present investigation focused on the assessment needs of this psychiatric population.

Evidence for a positive association between schizophrenia and violence has been found in large-scale epidemiological investigations in different countries with varying designs and outcomes (Fazel, Gulati, et al., 2009). Although schizophrenia is

[35] These three systematic reviews were identified and described in the metareview (Chapter II, Section 2.4.5.5).

consistently associated with an increased risk of violence compared with general population controls, most individuals with schizophrenia are not dangerous (Walsh et al., 2002). A recent population study estimated rates of violent crime to be approximately 10-15% after diagnosis (Fazel, Långström, et al., 2009) and cohort studies have reported rates between 8-14% (Brennan, Mednick, & Hodgins, 2000; Räsänen et al., 1998; Soyka, Graz, Bottlender, Dirschedl, & Schoech, 2007). Nevertheless, current treatment guidelines in the UK and US recommend that violence risk be assessed for all individuals diagnosed with schizophrenia (American Psychiatric Association, 2004; National Institute for Health and Clinical Excellence, 2009), and over 120 instruments have been developed to assist in the prediction of such behaviour in general and institutionalised populations (Chapter II, Section 2.4.2). Available instruments focus on identifying those patients who are at highest risk of violence and typically require several hours to administer, placing a considerable burden on the resources of mental health services (Large et al., 2011; Murrie, Cornell, & McCoy, 2005; Viljoen et al., 2010).

One approach to potentially improve the risk assessment process is to develop instruments that can be used prior to in-depth assessments of dangerousness to screen out patients who are at very low risk of future violence. This approach is similar to that used by many diagnostic screening tools in physical medicine and allows for clinical resources, including more detailed risk assessment, to be targeted at those more likely to be in need of intervention. In addition, using such an approach would assist in risk management by highlighting needs that could be the basis for treatment in those who are not screened out.

4.2.1 Objectives

The objective of the present study was to develop a screening tool to efficiently identify individuals with schizophrenia who are at very low risk of violence after hospital discharge. The instrument was developed and cross-validated using a cohort of 13,806 discharged patients with schizophrenia, with a potential follow up of 33 years. The aim was to create a tool composed of routinely available factors that was potentially scalable as it does not require specific training or additional costs. These properties have been identified as important for any instrument aiming to be clinically relevant (Webster & Polvi, 1995).

4.3 METHOD

4.3.1 Study Protocol

The Standards for Reporting of Diagnostic Accuracy Studies (STARD) Statement (Bossuyt et al., 2003), a 25-item checklist of reporting characteristics (Appendix H), was followed.

4.3.2 Data Registries

Data was collected from the following nationwide population-based registers in Sweden: the *Hospital Discharge Registry* (HDR; held at the National Board of Health and Welfare), the *Migration Register* (Statistics Sweden), the *Cause of Death Register* (National Board of Health and Welfare), the *National Crime Register* (National Council for Crime Prevention), the *Education Register* (Statistics Sweden), the *National Censuses from 1970 and 1990* (Statistics Sweden), and the *Multi-Generation Register* (Statistics Sweden) (Table 4.1). Merging these datasets was

possible as all national registers use residents' (including immigrants) same 10-digit personal identification numbers. An independent government organisation (Statistics Sweden) merged the registers and, upon assigning each participant a unique case number, destroyed the coding sheet that linked case numbers and personal identification numbers.

It was decided to use Swedish registers due to their national coverage and historically high quality (Garpenby & Carlsson, 1994). Further, as the prevalence of schizophrenia and the rate of violent crime in Sweden do not differ substantially from other countries in Western Europe (Dolmén, 2001; Wittchen & Jacobi, 2005), findings using these registers are thought to have some generalisability (Anwar, Långström, Grann, & Fazel, 2011; Fazel, Långström, et al., 2009). Register data was made available by my supervisor, Dr. Seena Fazel, who had, in collaboration with colleagues from the Karolinska Institutet in Stockholm, Sweden, been given ethics approval to conduct epidemiological investigations into the association between schizophrenia and violence (2005/174/31/4).

4.3.3 Participants

Using the HDR, all individuals aged 15 years (the age of criminal responsibility) and older who had been admitted to hospital for assessment and/or treatment in Sweden and had been discharged between January 1, 1973 and December 31, 2004 with a diagnosis of schizophrenia were identified. The HDR is a high quality register with less than 1% of hospital discharges missing personal identification numbers (Fazel & Grann, 2006). HDR diagnoses of schizophrenia have a high concordance rate with those produced by the *Operational Checklist for Psychotic*

Disorders (OPCRIT; κ = 0.74-0.76; Dalman, Broms, Cullberg, & Allebeck, 2002)[36] and are estimated to correspond with *Diagnostic and Statistical Manual of Mental Disorders* (DSM-IV) diagnoses of schizophrenia in 86% of cases (Ekholm et al., 2005). Such findings have led to this registry having been used in a number of epidemiological studies on psychotic populations (e.g., Fazel & Grann, 2006; Zammit et al., 2003).

Diagnoses of schizophrenia were made using the *International Classification of Diseases* (ICD). Individuals discharged from hospital between 1973-1986 received diagnoses from the ICD-8 (code 295), between 1987-1996 from the ICD-9 (code 295), and between 1997-2004 from the ICD-10 (code F20). Only those individuals who had been diagnosed with schizophrenia on at least two occasions (N = 13,806) were included in order to improve specificity and reduce clinical heterogeneity (Reutfors et al., 2010).[37] That is, by using this approach, participants were less likely to have other disorders such as drug-induced psychosis or bipolar disorder.

Participants were followed until their first conviction for a violent crime or until they were censored due to emigration (data obtained from the Migration Register), death (data obtained from the Cause of Death Register), or the end of the follow-up period (December 31, 2004).[38]

---

[36] The OPCRIT is a 90-item checklist of signs and symptoms and a suite of computer programs that are used to generate diagnoses of severe psychiatric disorders according to the operational criteria defined by 12 classification systems, including the DSM-III-R and the ICD-10 (McGuffin, Farmer, & Harvey, 1991).

[37] The data was not available to determine how many patients had received only one discharge diagnosis of schizophrenia between January 1, 1973 and December 31, 2004. However, a previous study that used an overlapping dataset of Swedish patients with discharge diagnoses of schizophrenia between January 1, 1973 and December 31, 2006 (Fazel, Långström, et al., 2009) noted that 56% of individuals had received only one diagnosis. Given that the present cohort was composed of 13,806 participants, this suggests that approximately 7,731 patients with only one discharge diagnosis of schizophrenia were excluded from the calibration and cross-validation samples. For a discussion of how this may have affected findings, see Section 4.5.5.

[38] Both the Migration Register and the Cause of Death Register are high quality registers with completeness exceeding 99% (National Board of Health and Welfare, 2009; Statistics Sweden, 2010).

4.3.4 Calibration and Cross-validation Samples

The population cohort was randomly divided into two samples of 6,903 participants using the "runiform" command in STATA/IC 10.1 for Windows (StataCorp, 2007). One of the samples was used to develop the screening tool and the other to cross-validate it. As the use of regression to develop risk instruments has been criticised for exaggerating predictive validity estimates in calibration samples (Blair et al., 2010; Pedhazur, 1997), the cross-validation sample was randomly split into three independent subsamples, each consisting of 2,301 participants. Cross-validating the tool using these subsamples provided the opportunity to assess shrinkage effects.

4.3.5 Definition of Violence

Violent conviction data was extracted from the National Crime Register for the years 1973-2004. The National Crime Register is a high quality register, missing only 0.05% of personal identification numbers between 1988-2000 (Fazel & Grann, 2006). Violent crime was defined according to the Swedish Penal Code as having received a conviction for homicide, assault, robbery, arson, any sexual offence (including rape, sexual coercion, child molestation, indecent exposure, or sexual harassment), illegal threats, or intimidation. Attempted and aggravated forms of these offences were included while property, traffic, and drug offences were not. An alternative would have been to have included only contact offences in the operational definition of violence; however, the inclusion of noncontact offences is consistent with previous risk assessment research (Fazel et al., 2006; Hanson & Morton-

Bourgon, 2004; Mills & Kroner, 2006).[39] Participants were coded as either having or not having a conviction for a violent crime during follow-up.

Conviction was chosen as the outcome variable as, in accordance with the Swedish Criminal Code, individuals are convicted as guilty regardless of mental illness. Thus, conviction data includes individuals who are transferred to forensic hospital as well as those who receive custodial sentences. Conviction data also includes cases in which the prosecutor decided to caution or fine defendants. In addition, plea-bargaining is not permitted in Sweden. Therefore, conviction data accurately reflects resolved criminality.

4.3.6 Outcome Measures

Using large samples to develop actuarial instruments has been recommended because of the instability of these tools when applied to individuals (Hart, Michie, & Cooke, 2007). As the calibration and cross-validation samples were several orders of magnitude larger than previous investigations of violence risk in schizophrenia (e.g., Brennan et al., 2000; Lindqvist & Allebeck, 1990; Räsänen et al., 1998; Soyka et al., 2007), more precise effect estimates were able to be obtained. Primary outcome statistics used to measure predictive validity in the present investigation included the negative predictive value (NPV), the positive predictive value (PPV), the diagnostic odds ratio (DOR), and the area under the curve (AUC). A sensitivity analysis was also carried out using the number safely screened (NSS) and the number needed to detain (NND). These six outcome statistics were selected as they provide measures of global utility (DOR and AUC) as well as usefulness in making "rule in" (PPV and NND) and

[39] Had only contact offences been included in the operational definition of violence, the base rate of offending would have decreased, likely resulting in an instrument with a higher false positive rate and a lower false negative rate. However, the instrument's calibration would not have taken into account serious crimes such as arson and verbal threats of violence.

"rule out" (NPV and NSS) decisions. The properties of these six effect sizes are discussed in detail elsewhere (Chapter I, Section 1.4.1.1; Chapter II, Section 2.5.3).

4.3.7 Risk Factors

A recent survey of forensic mental health professionals reported that the most common difficulty encountered when using risk assessment tools is not having access to the file information needed to score items (Viljoen et al., 2010). As over 80% of clinicians reported having access to individuals' mental health records and basic criminal histories (Viljoen et al., 2010), the inclusion of risk factors that could be scored using such file information was prioritised. These routinely collected and accessible variables included *male gender* (Swanson et al., 2006), *age at assessment* (Swanson et al., 2006), *previous criminal conviction* (Swanson et al., 2006), *comorbid alcohol abuse* (Fazel, Långström, et al., 2009), and *comorbid drug abuse* (Fazel, Långström, et al., 2009). These variables were selected as the primary item content, as gender, young age, and past offending are the most frequently cited correlates of criminal conviction (Bonta et al., 1998; Buchanan & Leese, 2006), and substance abuse has been identified as a key mediator in the relationship between schizophrenia and violence (Fazel, Gulati, et al., 2009; Fazel, Långström, et al., 2009).

Information on gender was extracted from the HDR. The age at which participants in the calibration sample were discharged from hospital with a second diagnosis of schizophrenia ranged from 15 to 54 years (*Mean* = 28.9; *SD* = 7.2). The Chi-squared Automatic Interaction Detector (CHAID) method adjusted by time at risk was used to identify the age cut-off that best classified participants as being at risk of

violent conviction (Kass, 1980).[40] Previous criminal conviction referred to any crime (violent or not) before hospital discharge, and was gathered from the National Crime Register.[41] Data on participants' hospital admissions during the years 1973-2004 were extracted from the HDR and examined for evidence of principal or comorbid diagnoses of alcohol abuse (ICD-8: 303; ICD-9: 303, 305.0; ICD-10: F10, except x.5) and drug abuse (ICD-8: 304; ICD-9: 304, 305.9; ICD-10: F11-F19, except x.5).[42] This information was used as a marker of substance abuse comorbidity.

Secondary risk factors considered for inclusion on the screening tool included *low level of education* (Cannon et al., 2002; Heinrichs & Sam, 2010), *having a parent who was convicted of a violent offence* (Monahan et al., 2000), and *having a parent who was diagnosed with alcohol abuse* (Monahan et al., 2000; Pilowsky, Keyes, & Hasin, 2009). These were considered secondary as such information is only available to approximately 15% of clinicians conducting risk assessments (Viljoen et al., 2010). Additional socio-demographic variables on which information was available included personal and household income, residence in an urban area, marital status, and having children. As information needed to score these variables was only available from the 1970 and 1990 National Censuses, it could rarely be established what an offender's

---

[40] An alternative to regression methodology, the CHAID technique is a type of decision tree analysis used to determine the relative strength of multiple independent variables in predicting a dichotomous outcome (Kass, 1980). As participants ranged from 15 to 54 years old at hospital discharge, a block of 40 dummy-coded age variables (e.g., below age 22 years at assessment versus 22 years and above) was entered into a CHAID model with violent conviction as the outcome. The dummy-coded variable that constituted the first node on the decision tree (i.e., the age threshold that best classified participants as being at risk of violent conviction) was selected for use in the "young age at assessment" item.

[41] An *a priori* test of multicollinearity was conducted using participants from the calibration sample to determine whether to include any previous conviction and previous violent conviction as separate items on the screening tool. Multivariate Cox regression determined that the items were collinear at the $p < 0.05$ level such that previous general conviction accounted for the variance explained by previous violent conviction. Therefore, any previous conviction was included as an item on the screening tool whereas previous violent conviction was not.

[42] An *a priori* test of multicollinearity was carried out using participants from the calibration sample to determine whether to include comorbid drug abuse and comorbid alcohol abuse as separate items or to collate them into a single substance abuse variable. Multivariate Cox regression revealed that the items accounted for variance independently of one another at the $p < 0.05$ level and, therefore, both were included separately on the tool.

income was, whether he or she resided in an urban area, and whether he or she was married and/or had children at the time of his or her crime. Therefore, these factors were excluded. Additional familial factors on which information was available included having a mother or father who was diagnosed with schizophrenia or who either attempted or committed suicide. As a search of the literature could not locate evidence that these variables are associated with violent offending in individuals diagnosed with schizophrenia, they were not considered for inclusion.

Participants' highest level of completed education was collated from the 1970, 1990, and 2004 Education Registers, and a binary comparison was made between those who had completed compulsory school (a nine year comprehensive school for children between the ages of 7 to 16) and those who had not.[43] Consistent with previous risk assessment research, this information was used as a marker of early-onset behavioural problems (Farrington, 1989; Lösel & Bender, 2003).

Personal identification numbers were used to extract data on participants' parents from the Multi-Generation Register. This register connects each person born in Sweden (beginning in 1933) and ever registered as living in the country (after 1960) to their parents. Similar data exists for immigrants who, along with one or both parents, became Swedish citizens before 18 years of age. Fathers are defined in the Multi-Generation Register as the mother's spouse when the individual was born or as the person acknowledged by the mother to be the birth father. Parents were linked to the National Crime Register to extract data on whether individuals had a father or mother who had previously been convicted of a violent crime.[44] In addition, parents

[43] No discharged patients were convicted of a violent offence before the age of 17 years, meaning that all had the opportunity to complete compulsory school before being adjudicated.

[44] An *a priori* test of multicollinearity was conducted using participants from the calibration sample to determine whether to include maternal previous violent conviction and paternal previous violent conviction as separate items or to collate them into a single parental previous violent conviction

were linked to the HDR to extract data on diagnoses of alcohol abuse (ICD-8: 303; ICD-9: 303, 305.0; ICD-10: F10, except x.5).[45]

4.3.8 Missing Data

No information was missing on participant gender, age at assessment, previous criminal conviction, comorbid alcohol or drug abuse, having a parent who was convicted of a violent offence or having a parent who was diagnosed with alcohol abuse. Given the high quality of the HDR, the National Crime Register, and the Multi-Generation Register, this was not surprising.

Of the 13,806 participants in the cohort, 890 (6.4%) were missing information on whether they had completed compulsory school or not. Of those 12,916 participants for whom education information was available, 760 (5.9%) had not completed compulsory school. Therefore, 53 (5.9%) of the 890 participants on whom education information was not available were randomly selected using the "runiform" command in STATA/IC 10.1 for Windows (StataCorp, 2007) and were coded as not having completed compulsory school. All other participants were coded as compulsory school completers.[46]

4.3.9 Developing the Screening Tool

All variables were dummy-coded into binary variables, enabling the construction of a unit scored measure (i.e., an instrument where the presence of a risk

---

variable. Multivariate Cox regression found that the variables accounted for variance independently of one another at the $p < 0.05$ level and, therefore, were both considered as secondary items.

[45] An *a priori* test of multicollinearity was carried out using participants from the calibration sample to determine whether to include maternal alcohol abuse and paternal alcohol abuse as separate items or to collate them into a single parental alcohol abuse variable. Multivariate Cox regression found that the variables accounted for variance independently of one another at the $p < 0.05$ level and, therefore, were both considered as secondary items.

[46] Of those 53 participants randomly selected to be coded as compulsory school non-completers, 26 (49.1%) were in the calibration sample and 27 (50.9%) were in the cross-validation samples.

factor is scored as +1 and the absence as +0). Sequential Cox regression was used to develop the screening device. *Sequential regression* is a multivariate technique that prioritises a set of variables known to be associated with the outcome of interest (Tabachnick & Fidell, 2001). Secondary variables are then added in a step-by-step manner and evidence for improved predictive validity examined (Dahle, 2006; Hollin & Palmer, 2006). Sequential regression was selected so the inclusion of risk factors that could be scored using routinely available file information could be prioritised. Developing an instrument that uses readily accessible file information could limit the strain on mental health professionals who would otherwise have to spend time collecting information from potentially unreliable patients or third-party sources. Alternatives to sequential regression such as stepwise modelling, in which independent variables are automatically selected on the sole basis of pre-specified statistical criteria (e.g., adjusted R^2 or Bayesian information criteria [BIC] thresholds), would not have guaranteed the inclusion of these factors (Tabachnick & Fidell, 2001). Cox proportional hazards modeling was used as time at risk varied by participant, and alternative regression methodologies designed for dichotomous outcomes (e.g., logistic regression) would not have taken this into account (Chapter I, Section 1.3.1).[47]

Analyses took place in two stages: In the first stage, the routinely accessible variables were combined into a five-item model and regressed with violent conviction as the dependent variable. Provided that all variables accounted for variance independently of one another, rates of true positives (TP), false positives (FP), true

[47] The underlying assumption of Cox regression is that changes in independent variables produce proportionate changes in the hazard function independently of time (Altman, 1991). This assumption was tested with participants from the calibration sample and the five routinely accessible variables. Using the "stphtest" command in STATA/IC 10.1 for Windows (StataCorp, 2007), the calibration data was found to meet the proportionality assumption, $\chi^2(5, n = 6903) = 3.52$, $p = 0.62$. Therefore, proportional hazards modelling was deemed to be an appropriate method for developing the screening tool.

negatives (TN), and false negatives (FN) were calculated at each risk score at 1, 2, and 5 years follow-up. Participants who were not at risk for 1, 2, or 5 years (respectively) due to emigration, death, or end of follow-up were excluded from these calculations.

To select a cut-off score for the tool, the relative costs of false negatives and false positives were taken into consideration. The societal and political costs associated with screening out a patient who would go on to commit a violent crime (i.e., a false negative) were considered to outweigh the costs associated with conducting a more detailed risk assessment for a patient who would not go on to commit a violent offence (i.e., a false positive). Therefore, the ideal cut-off point was identified as the risk score that most closely met a sensitivity to specificity ratio of 2:1 (Smits, 2010). Using this cut-off, the instrument's NPV, PPV, and DOR were calculated at each length of follow-up, as was AUC using receiver operating characteristic (ROC) curve analysis.

In the second stage, each secondary variable was regressed with violent conviction as the dependent variable and ranked in order of their resulting hazard ratios. Variables were then added one at a time to the five-item model in order of hazard ratio and evidence for multicollinearity was assessed at each step. Provided the added factor independently accounted for variation in the dependent variable, the revised tool's predictive validity was assessed at 1, 2, and 5 years follow-up using NPV, PPV, DOR, and AUC (via ROC curve analysis). This process was repeated for each additive iteration. Effect estimates were compared with those produced by the five-item model at the same length of follow-up using 95% confidence intervals. Additional comparisons were made using the χ^2 test of differences between proportions, Breslow and Day's (1987) χ^2 test of heterogeneity between DORs, and

Hanley and McNeil's (1983) *z* test for differences in AUC. It was decided to include these standard statistical tests in addition to comparing 95% confidence intervals so that the significance of differences in effect sizes could be reported using *p*-values.

Analyses were conducted using SPSS 17.0.1 for Windows (SPSS Inc, 2009), STATA/IC 10.1 for Windows (StataCorp, 2007), and MedCalc 11.3.8.0 for Windows (MedCalc Software, 2010).

4.3.10 Item Weighting

The screening tool that produced the highest rates of predictive validity using the fewest number of items was then tested to determine whether applying weights to item responses would produce higher rates of predictive validity. Items were weighted by their unstandardised beta coefficients and hazard ratios as calculated during multivariate Cox regression.

Unstandardised beta coefficients represent independent variables' contributions to the prediction of variance in the dependent variable. Methodologists have argued that as independent variables are weighted by their beta coefficients during multivariate regression analyses, tool items should also be weighted by these parameters (Guilford, 1941). Therefore, each dummy-coded item response was multiplied by its respective unstandardised beta coefficient.[48] Rates of TP, FP, TN, and FN were calculated for the resulting tool at each risk score for 1, 2, and 5 years follow-up to identify the cut-off score that produced a sensitivity to specificity ratio nearest 2:1, and, using that cut-off, the instrument's NPV, PPV, and DOR were calculated for each length of follow-up. AUCs were also calculated using ROC curve analysis.

[48] The unstandardised beta coefficient of each independent variable was multiplied by a constant of +10.0 to increase scale (Guilford, 1941).

In the context of the present study, the *hazard ratio* is the ratio between the predicted hazard (i.e., likelihood per unit time) of violent conviction in individuals with a given characteristic compared to the hazard of violent conviction in individuals without that characteristic. If an independent variable does not affect the incidence of violent conviction, the hazard ratio will be 1.0. Hazard ratios above 1.0 can be interpreted as evidence that the presence of a characteristic is associated with a higher incidence of violent conviction, whereas hazard ratios below 1.0 can be interpreted as evidence that participants with that characteristic have a lower incidence of violent conviction. Each dummy-coded item response was multiplied by its respective hazard ratio as calculated during multivariate Cox regression. Rates of TP, FP, TN, and FN were calculated for the instrument at each risk score at 1, 2, and 5 years follow-up to identify the ideal cut-off score. Using this cut-off, the tool's NPV, PPV, and DOR were calculated, as was the AUC using ROC curve analysis.

Also considered for use in weighting items were strategies developed by Nuffield (1982) and Gagliardi (2004), both of which use base rate information to weight item responses. In the Nuffield approach, the base rate of violent conviction is calculated for the calibration sample as a whole. The base rate of violent conviction is then determined for each item response (e.g., the rate of violence in men). A weight of +1 or -1 is assigned for each difference (positive or negative, respectively) of 5% from the overall base rate.

As the Nuffield approach has been known to produce equal weights for all item responses, Gagliardi and colleagues (2004) designed a modified version of this method. First, the 99% CI of the calibration sample's base rate of violent conviction is calculated. The base rate of each item response is then determined. Item responses are assigned scores of -1, 0, or +1 depending on whether the base rate for the response

falls below the 99% CI (score -1), within the 99% CI (score 0), or above the 99% CI (score +1). As the base rate of violent conviction was low (Table 4.2), it was decided not to weight items using base rate information.

The NPVs, PPVs, DORs, and AUCs of the beta coefficient- and hazard ratio-weighted tools were compared with those produced by the unit scored version of the instrument at each length of follow-up using 95% confidence intervals as well as tests of differences between effect sizes. Analyses were conducted using STATA/IC 10.1 for Windows (StataCorp, 2007) and MedCalc 11.3.8.0 for Windows (MedCalc Software, 2010).

4.3.11 Cross-validation

The unweighted or weighted version of the screening tool that produced the highest rates of predictive validity was then cross-validated using the three replication subsamples. The effect estimates produced by the tool in each independent subsample at 1, 2, and 5 years follow-up were compared with those produced by the calibration sample at the same length of follow-up using 95% confidence intervals and tests of differences to examine evidence of shrinkage. Analyses were conducted using STATA/IC 10.1 for Windows (StataCorp, 2007) and MedCalc 11.3.8.0 for Windows (MedCalc Software, 2010).

4.3.12 Power Analyses

4.3.12.1 Cox Regression

In order to assess the level of statistical power that would be achieved when using Cox regression to develop the screening tool, *a priori* power analyses were conducted using Study Size 2.0.4 for Windows (Olofsson, 2010). The statistical

power of Cox regression analyses is dependent in part on the α-level used to reject the null hypothesis as well as sample size. An α-level of 0.05 and the sample size of the calibration sample (n = 6,903) were used in the power calculations.

The power of Cox regression analyses also depends on the base rate of the outcome of interest (i.e., violent conviction) and the proportion of cases that exhibit predictive attributes (e.g., male gender). A 12.0% base rate of violent conviction and a 14.4% predictor prevalence rate were used in the power calculations. These are relatively conservative assumptions, since the base rate of violent conviction was more than 12.0% in the calibration and two of the cross-validation samples, and a number of the predictor variables were present in more than 14.4% of the cases (Table 4.2).

Finally, the power achieved during Cox regression analyses depends on the level of interdependence between independent variables and the magnitude of the effect sizes of interest (i.e., hazard ratios). With regard to the association between variables in the multivariate model, power was computed assuming moderate (R^2 = 0.16) and strong (R^2 = 0.25) relationships between predictors (Cohen, 1988). Power was computed for three benchmark hazard ratios (1.50, 1.75, and 2.00), which represent the lower range of effect sizes calculated for the independent variables when adjusted by age and gender (Table 4.2).

Thus, power was computed for six conditions, all of which assumed an α-level of 0.05, a sample size of 6,903, a 12.0% base rate of violent conviction, and a 14.4% predictor prevalence rate:

(1) Moderate R^2, Hazard Ratio = 1.50: Power = 0.98

(2) Moderate R^2, Hazard Ratio = 1.75: Power = 0.99

(3) Moderate R^2, Hazard Ratio = 2.00: Power = 0. 99

(4) Strong R^2, Hazard Ratio = 1.50: Power = 0. 99

(5) Strong R^2, Hazard Ratio = 1.75: Power = 0. 99

(6) Strong R^2, Hazard Ratio = 2.00: Power = 0. 99

In summary, the probability of rejecting a false null hypothesis during Cox regression analyses exceeded 95%, a level of power considered stringent (Aberson, 2010). As such, the calibration sample was large enough to detect even modest effect sizes and was considered appropriate for use in developing the screening tool.

4.3.12.2 Tests of Differences

As general guidelines regarding what constitutes a small, moderate, or large difference in NPVs, PPVs, DORs, or AUCs do not exist, power analyses were not conducted for the χ^2 tests of differences between proportions, the Breslow-Day χ^2 tests of heterogeneity between DORs, or the Hanley-McNeil z tests for differences in AUCs. Given that a number of such tests were conducted to compare the effect sizes produced by the calibration and cross-validation samples as well as the unit scored and weighted instruments at different lengths of follow-up, non-significant findings should be approached with due caution.

4.4 RESULTS

4.4.1 Description of the Samples

The population cohort consisted of 13,806 individuals with two or more hospital discharge diagnoses of schizophrenia received between January 1, 1973 and December 31, 2004. The cohort was randomly divided into a calibration sample (n = 6,903) and three cross-validation subsamples (n = 2,301 for each).

The calibration sample was composed of 4,392 (63.6%) men and 2,511 (36.4%) women (Table 4.2). The mean length of follow-up was 12.5 years (*SD* = 8.0). Of the 6,903 participants, 887 (12.9%) were convicted of a violent offence by December 31, 2004. Each conviction could potentially include more than one crime. There were a total of 2,648 violent crimes in the 887 convictions. The most common violent crime was assault (*k crimes* = 1,237), followed by illegal threats (*k* = 830), sexual offences (*k* = 167), and robbery (*k* = 156). There were 117 convictions for homicide in the calibration sample (base rate = 0.1%).

The first cross-validation sample was composed of 1,480 (64.3%) men and 821 (35.7%) women. Participants were followed for a mean of 12.4 years (*SD* = 8.1). Of the 2,301 participants in the sample, 309 (13.4%) were convicted of a violent offence during follow-up. There were a total of 503 violent crimes in the 309 convictions. The most common violent crime was assault (*k* = 224), followed by illegal threats (*k* = 166), sexual offences (*k* = 37), and robbery (*k* = 34). There were 16 convictions for homicide in this subsample (base rate = 0.1%).

The second cross-validation sample was composed of 1,504 (65.4%) men and 797 (34.6%) women. The mean length of follow-up was 12.4 years (*SD* = 8.1). Of the 2,301 participants in the sample, 321 (14.0%) were convicted of a violent offence by the end of the follow-up period. The 321 convictions were for 565 violent crimes, most commonly assault (*k* = 257), illegal threats (*k* = 176), sexual offences (*k* = 40), and robbery (*k* = 36). There were 26 convictions for homicide in this cross-validation sample (base rate = 0.1%).

The third cross-validation sample was composed of 1,515 (65.8%) men and 786 (34.2%) women. Participants were followed for a mean of 12.1 years (*SD* = 8.0). Of the 2,301 participants, 275 (12.0%) were convicted of a violent offence by

December 31, 2004. There were a total of 454 violent crimes in the 275 convictions. The most common violent crime was assault (k = 215), followed by illegal threats (k = 140), sexual offences (k = 65), and robbery (k = 29). There were 18 convictions for homicide (base rate = 0.1%).

4.4.2 Developing the Screening Tool

Using the CHAID method, 32 years was identified as the age cut-off that best classified participants as being at risk of violent conviction. Therefore, the "young age at assessment" variable considered as positives all participants aged 31 years and below when they were discharged from hospital.

Multivariate Cox regression revealed that the five routinely available risk factors (i.e., male gender, previous criminal conviction, young age at assessment, comorbid alcohol abuse, and comorbid drug abuse) were significant independent predictors of the incidence of violent conviction (Table 4.3). Backwards stepwise Cox regression analyses revealed no significant evidence of item interactions accounting for variance independently of these factors.

The TP, FP, TN, and FN rates were calculated for the five-item tool at each risk score at 1, 2, and 5 years follow-up (Table 4.4). The optimal cut-off score was identified as +2. Of the 2,359 participants who scored below this threshold, 2,353 were not convicted of a violent offence within 1 year after hospital discharge (NPV = 0.99; 95% CI = 0.99-1.00). The screening tool's PPV at 1 year after discharge was 0.01 (95% CI = 0.01-0.02), the DOR was 3.79 (95% CI = 1.65-8.72), and the AUC was 0.67 (95% CI = 0.59-0.74). There was a trend towards PPVs, DORs, and AUCs increasing over 2 and 5 years follow-up, while NPVs remained at 99% (Table 4.5). To investigate the instrument's performance under varying base rate

conditions within this population, positive and negative predictive values were measured at 1, 2, and 5 years at base rates of 0%, 2%, 4%, 6%, 8%, and 10% using the sensitivities and specificities produced by the tool at each respective length of follow-up (Table 4.6). While PPVs increased markedly with increasing base rates, NPVs remained above 95%.

Univariate Cox regression analyses revealed that all five secondary variables (i.e., non-completion of compulsory school, having a father or mother who was previously convicted of a violent offence, and having a father or mother who was diagnosed with alcohol abuse) were significant predictors of the incidence of violent conviction.[49] The variable that produced the highest hazard ratio was having a mother with a previous violent conviction (HR = 3.56) followed by having a father with a previous violent conviction (HR = 2.92), having a mother who was diagnosed with alcohol abuse (HR = 1.85), non-completion of compulsory school (HR = 1.57), and having a father who was diagnosed with alcohol abuse (HR = 1.37). When the secondary variables were added to the five-item model in this order, only having a father with a previous violent conviction was found to be a significant independent predictor of the incidence of violent conviction (Table 4.3). Therefore, a six-item unit scored tool was constructed and rates of TP, FP, TN, and FN were calculated at each risk score (Table 4.7). The cut-off score that produced a sensitivity to specificity ratio closest to 2:1 was identified as +2. The NPVs, PPVs, DORs, and AUCs produced by the six-item instrument at 1, 2, and 5 years post-discharge did not differ significantly from those produced by the five-item tool at each respective length of follow-up (Table 4.8; Appendix I).

[49] At the $p < 0.05$ level.

4.4.3 Item Weighting

Weights were assigned to the five-item tool's item responses using the unstandardised beta coefficients and hazard ratios calculated during multivariate Cox regression (Table 4.9). There was no clear evidence to suggest that weighting items resulted in higher rates of predictive validity than unit scoring at 1, 2, or 5 years follow-up (Table 4.10; Appendix I).

4.4.4 Cross-validation

The unit scored five-item instrument was then cross-validated (Table 4.5). Statistical tests revealed no significant differences between the effect estimates produced by the tool in the calibration and cross-validation samples at each length of follow-up, suggesting no clear shrinkage effects (Appendix I). Of the 2,288 participants in the replication samples who were classified as low risk, 2,280 did not go on to be convicted of a violent offence within 1 year of hospital discharge (NPV = 0.99; 95% CI = 0.99-1.00). Similar to the calibration sample, there was a trend towards PPVs, DORs, and AUCs increasing over time, while NPVs remained at or above 98% at 2 and 5 years of follow-up.

4.4.5 Sensitivity Analyses

To further explore clinical utility, the screening tool's NSS and NND were calculated for the overall cohort of 13,806 participants at 1, 2, and 5 years follow-up. The instrument produced an NSS of 332 (95% CI = 200-500) at 1 year post-discharge, 202 (95% CI = 143-333) at 2 years post-discharge, and 69 (95% CI = 56-91) at 5 years post-discharge. The tool produced an NND of 91 (95% CI = 83-100) at 1 year

post-discharge, 48 (95% CI = 45-50) at 2 years post-discharge, and 19 (95% CI = 18-19) at 5 years post-discharge.

4.5 DISCUSSION

Using data collected from Swedish national registers over three decades, violence risk was investigated in 13,806 hospitalised patients with schizophrenia in order to develop a simple screening tool that could identify individuals who would not go on to be convicted of a violent offence after discharge. The instrument was composed of five routinely available risk factors: male gender, previous criminal conviction, young age at assessment, comorbid alcohol abuse, and comorbid drug abuse. Predictive validity analyses found that the tool could be accurately used to make "rule out" decisions (i.e., identifying who will not go on to violently offend), suggesting potential utility as a screening tool that could be used as part of a stepped approach to risk assessment. In this approach, very low risk patients are screened out, allowing for more detailed and resource-intensive assessment and potential intervention in screen positive patients. The reported tool requires no training to use and, hence, is potentially scalable.

4.5.1 Comparison with Available Instruments

While several violence risk assessment instruments have been developed for general psychiatric populations (Monahan et al., 2000; Roaldset, Hartvig, & Bjørkly, 2010; Watts et al., 2004), no tools have been designed to predict the likelihood of community violence in individuals with a specific psychiatric diagnosis. As epidemiological investigations have shown that the base rate of violence varies by diagnosis (Arsenault, Moffitt, Caspi, Taylor, & Silva, 2000; Brennan et al., 2000;

Eronen, Angermeyer, & Schulze, 1998; Hodgins, 1992; Stuart & Arboleda-Florez, 2001; Tiihonen, Isohanni, Räsänen, Koiranen, & Moring, 1997), general screening approaches are likely to lose sensitivity and specificity. Evidence from the meta-analysis conducted in the previous chapter suggests that developing risk assessment tools for more specific populations may improve predictive validity. The most specific instrument identified during the systematic searches in this thesis was an actuarial tool designed to predict assault in individuals diagnosed with psychotic disorders (Wootton et al., 2008) (Table 4.11). The calibration study for this instrument, a two-year clinical trial examining different models of community care in UK inner cities, also found that basic demographic information was helpful in risk prediction (Walsh et al., 2004; Wootton et al., 2008). Unlike the screening tool developed in the present study, however, this measure was designed to identify high risk individuals and, hence, produces a notably higher rate of false negatives (13% at 2 year follow-up compared with less than 1% in the present study).

4.5.2 Unit Scoring versus Weighting Tool Items

Weighting items on the screening instrument was not found to improve predictive validity compared with a simple unit scoring method (i.e., presence of a risk factor is scored as +1 and the absence as +0). Findings support the view that weighting items during prediction procedures is not appropriate when dealing with highly specific populations, as the mechanism through which risk factors lead to violence is still unclear (Cohen, 1990; Grann & Långström, 2007; Wainer, 1976). Developing item weights using an instrument's calibration sample may also result in pronounced shrinkage effects when that instrument is used in practice (Wang &

Stanley, 1970). Little evidence of shrinkage was found when the tool was scored using a simple presence versus absence coding strategy.

4.5.3 Inclusion of Clinical Override

The screening tool developed in the present study does not necessarily undermine the role of clinicians in the violence risk assessment process. Rather, by identifying patients for whom detailed assessment may not be needed in relation to violence risk, the instrument complements clinical judgement. Nevertheless, a number of risk factors that this tool does not measure may have clinical significance (e.g., poor social problem solving, treatment non-completion, and sensation seeking; Egan, Charlesworth, Richardson, Blair, & McMurran, 2001; McMurran, Blair, & Egan, 2002; McMurran & Theodosi, 2007) and warrant more comprehensive assessment to determine risk level. Therefore, a clinical override option has been included on the instrument's coding sheet (Appendix J). However, as previous research on whether the inclusion of clinical override on actuarial scales improves predictive validity has produced mixed findings (Dawes, Faust, & Meehl, 1989; Quinsey et al., 2006; Webster, Harris, Rice, Cormier, & Quinsey, 1994), future research will need to investigate the comparative utility of the screening tool both with and without this option.

4.5.4 Implications

When used as part of a stepped approach to violence risk assessment, screen out tools are likely to save psychiatric services considerable resources, though future research will have to test this in practice. Current guidelines recommend the use of risk assessment tools to supplement clinical judgement when predicting patient

dangerousness (American Psychiatric Association, 2004; National Institute for Health and Clinical Excellence, 2009). Recent UK and US surveys have found that instruments that employ structured clinical judgement, where clinicians use empirically-based risk and protective factors to guide their predictions as to whether an individual will be violent, are amongst the most commonly used tools (Archer et al., 2006; Khiroya et al., 2009). With the time involved in administering such tools (familiarising oneself with the patient's case, collecting information from multiple sources to score items, conducting interviews, scoring the tool, and making a clinical judgement regarding risk level) and the costs involved (attending training sessions, purchasing tool manuals, and paying for each coding sheet), violence risk assessment, as often currently conducted, costs mental health services significant amounts of time and money. In their survey of forensic mental health professionals, Viljoen and colleagues (2010) reported that using a structured tool to conduct a risk assessment takes an average of 15 hours and costs service providers an average of $100 USD per hour to complete. Given these costs, screening tools that require little training, are based on items that can be scored using routinely available file information, and can reduce caseload are potentially attractive.

It has been argued that the accurate prediction of high risk individuals may not be possible in populations with low base rates of violence (Maden, 2003; Szmukler, 2001). Attempting to predict low base rate behaviours commonly results in high rates of false positives and may have the unintended consequence of stigmatising patients (Large et al., 2011). Such stigma can lead to problems with social adjustment (Vauth, Kleim, Wirtz, & Corrigan, 2007), increased symptom severity (Mak & Wu, 2006), and a reduced likelihood of seeking help from psychological services (Vogel, Wade, & Haake, 2006). The use of screening tools as part of a stepped approach to violence

risk assessment may reduce the number of false positive predictions made during detailed assessments. By accurately screening out a large proportion of low risk patients who would otherwise undergo resource-intensive risk assessment, the base rate of violence in the remainder is artificially increased. To illustrate this effect, at five years after hospital discharge, using the screening tool would have artificially increased the base rate of violence in the remainder of the calibration sample by 38.7% (Table 4.4). Such increases in base rate have the potential to increase the positive predictive values produced by risk assessment procedures already in place. Thus, the use of a stepped approach maximises negative predictive power during the screening stage and then maximises positive predictive power during the detailed assessment stage.

The results of the present investigation suggest that reporting a single effect size does not provide an adequate picture of a risk assessment tool's predictive validity. Global effect estimates such as the AUC and the DOR do not measure the relative utility of an instrument in making "rule in" versus "rule out" decisions. Had only AUCs or DORs have been reported for the simple tool in this study, different conclusions could have emerged about how the tool performed. AUCs below 0.70 may have been interpreted as evidence that the screening tool lacked any utility as a violence risk assessment instrument (Sjöstedt & Grann, 2002; Tape, 2006), whereas DORs above 5 may have been interpreted as evidence that the screen performed well in identifying both high and low risk patients. It was only by examining the PPVs, NPVs, NNDs, and NSSs of the tool that a clear picture emerged of how the instrument performed. Future studies investigating the predictive validity of forensic risk assessment tools should consider reporting a single effect indicator as well as at least one "rule in" and "rule out" effect size (Chapter I, Section 1.4.1.1).

The findings of the present investigation parallel those of recent studies in both psychiatric and non-psychiatric populations that have found that a small group of historical risk factors may be able to predict the likelihood of violent behaviour just as well as (or, in some cases, better than) complex risk assessment schemes (Buchanan & Leese, 2006; Coid et al., 2007; Gray et al., 2004; Heinrichs & Sam, 2010; Walsh et al., 2004). These sets of historical factors are presented alongside those in the simple screening tool in Table 4.11. Future studies that measure the predictive validity of structured risk assessment schemes may wish to statistically compare their findings not only with chance but also with the accuracy produced using basic demographic factors.

4.5.5 Limitations

There are several limitations of the present investigation. First, hospital data was relied on to identify schizophrenia cases. Therefore, the screen is applicable only to patients discharged from hospital, although other research has estimated that a minority of individuals in Sweden who have schizophrenia would not have been hospitalised over a 30-year period (Hansson et al., 2001). By excluding individuals with only a single discharge diagnosis, all hospitalised individuals with schizophrenia will not have been included. However, by including only those individuals with two diagnoses of schizophrenia, the calibration and cross-validation samples had the advantage of diagnostic specificity and reduce clinical heterogeneity (Reutfors et al., 2010). That is, participants were less likely to have other psychiatric diagnoses (e.g., drug-induced psychosis or bipolar disorder). Related work has shown similar base rates of violence in those with one or more diagnoses of schizophrenia (Fazel,

Långström, et al., 2009), so it is possible that the screen is generalisable to all hospitalised patients with the disorder.

A second limitation is that information from Sweden's Hospital Discharge Register was used to determine whether patients had a comorbid diagnosis of alcohol or drug abuse. The reliability of diagnoses of substance abuse in Swedish hospital registers has been found to be fair to moderate (Bergman, Belfrage, & Grann, 1999; Fazel, Långström, et al., 2009). In addition, the prevalence of substance abuse in the cohort will have been underestimated because hospitalisation for alcohol or drug abuse was used as the criteria for substance abuse-related risk factors. Future work will need to examine whether a lower threshold for substance abuse changes the predictive validity of these factors.

A third limitation of the present research is that data was not available regarding whether participants received treatment or the nature of that treatment. There is evidence to suggest that antipsychotic medication can reduce the incidence of violence in individuals diagnosed with schizophrenia (Swanson, Swartz, & Elbogen, 2004). The effect of treatment was modeled (to some extent) by calculating negative and positive predictive values at different base rates of violence. Negative predictive values remained above 95% when the base rate was increased or decreased, suggesting that the screening tool is able to accurately identify individuals at low risk of violent conviction even if treatment modifies base rates of violence (Table 4.6).

A fourth limitation of the present study is that the positive predictive values produced by the screening instrument were low. That is, patients who were not screened out by the simple tool were rarely violent. This is probably a consequence of the low base rate of violence in individuals with schizophrenia and suggests that individuals who score highly on the tool should not be considered as moderate or high

risk, but rather as warranting a more comprehensive risk assessment. These findings highlight the challenge of identifying individuals at high risk of violence in populations with low base rates of serious offending (Szmukler, 2001).

A fifth limitation is that evidence of minimal shrinkage effects may have been due to the calibration and cross-validation samples having been selected from the same population, which had a low overall base rate of violence. This low base rate may be attributed to conviction data having been used to determine whether participants offended during follow-up, as reliance on such information may have led to an underreporting of violent incidents. Some forensic researchers have advocated for the use of several sources of outcome data (e.g., criminal register, self-report, family interviews) when assessing the predictive validity of risk measures (Wootton et al., 2008). However, other work in patients with schizophrenia has suggested that conviction and self-report underestimate the base rate of violence to a similar degree (Arsenault et al., 2000). Future work will have to test the predictive validity of the screening tool in different settings (e.g., non-hospitalised patients) with different outcomes (e.g., violence not reported to the police).

A sixth limitation of the present investigation is that the screening tool does not discriminate in young men, who would all be screen positive. Therefore, it may be that the screen provides the most clinical utility in women and older men. Future research could consider more discriminating tools in young men.

4.5.6 Conclusion

The aim of this study was to develop a simple instrument to identify those individuals diagnosed with schizophrenia who are at very low risk of future violence after hospital discharge. The objective was to design a tool that used routinely

available risk factors and was straightforward to administer. With reasonably high rates of accuracy using only five items, no complicated weighting algorithm, and an easy to interpret classification system (screen out vs. continue risk assessment), findings suggest that using a stepped strategy in which very low risk patients are screened out prior to in-depth risk assessment may assist in improving the quality and efficiency of violence risk assessment in schizophrenia.

Table 4.1 Brief Descriptions of Seven National Registers Used to Develop a Violence Screening Tool for Individuals with Hospital Discharge Diagnoses of Schizophrenia

Register	Organisation Held By	Years of Coverage	Description
Hospital Discharge Registry	National Board of Health and Welfare	1973 – Present	This register contains information on hospital admissions (including admissions to secure and private hospitals) in Sweden. Covering both psychiatric and physical medicine settings, this register includes information on discharge diagnoses according to the *International Classification of Diseases* (ICD).
Migration Register	Statistics Sweden	1969 – Present	This register contains information concerning dates of immigration into and emigration out of Sweden.
Cause of Death Register	National Board of Health and Welfare	1958 – Present	Information in this register is based on the mandatory reporting of death certificates for all deceased persons in Sweden. This register contains information on the dates and causes of deaths.
National Crime Register	National Council for Crime Prevention	1973 – Present	This register contains information about every criminal conviction in Sweden, including the date and nature of offences, the number of crimes, and the type and length of sentences.
Education Register	Statistics Sweden	1970, 1990, 2004	This register contains information concerning Swedish residents' highest level of completed education. Reporting is mandatory.
National Census	Statistics Sweden	1970, 1990	National census data includes information on the employment status, income, and occupation of Swedish residents. The completion of national censuses is enforced by law.
Multi-Generation Register	Statistics Sweden	1960 – Present	This register contains information on biological and adoptive relationships for Swedish nationals and residents. The register includes the personal identification numbers of biological and (where appropriate) adoptive parents.

Table 4.2 Descriptive Characteristics of the Calibration and Cross-validation Samples of a Violence Screening Tool for Individuals with Hospital Discharge Diagnoses of Schizophrenia

Domain	Variable	Sample				Adjusted HR (95% CI) [a]
		Calibration Sample (*n* = 6903)	Cross-validation Sample 1 (*n* = 2301)	Cross-validation Sample 2 (*n* = 2301)	Cross-validation Sample 3 (*n* = 2301)	
Socio-demographic factors	Male gender, *n* (%)	4392 (63.6)	1480 (64.3)	1504 (65.4)	1515 (65.8)	3.3 (2.7-4.0)
	Age at assessment (in years), mean (*SD*)	28.9 (7.2)	28.9 (7.2)	29.2 (7.2)	29.2 (7.3)	1.9 (1.6-2.3)
	Non-completion of compulsory school, *n* (%)	2031 (29.4)	667 (29.0)	689 (29.9)	702 (30.5)	1.6 (1.4-1.8)
Individual factors	Previous criminal conviction, *n* (%)	2806 (40.7)	899 (39.1)	974 (42.3)	978 (42.5)	3.3 (2.8-3.8)
	Alcohol abuse comorbidity, *n* (%)	1078 (15.6)	353 (15.3)	353 (15.3)	331 (14.4)	2.9 (2.5-3.4)
	Drug (non-alcohol) abuse comorbidity, *n* (%)	1151 (16.7)	360 (15.6)	389 (16.9)	357 (15.5)	3.5 (3.1-4.0)
Familial factors	Father convicted of a violent offence, *n* (%)	127 (1.8)	41 (1.8)	39 (1.7)	35 (1.5)	2.3 (1.7-3.2)
	Mother convicted of a violent offence, *n* (%)	17 (0.2)	7 (0.3)	6 (0.3)	7 (0.3)	2.8 (1.4-5.3)
	Father diagnosed with alcohol abuse, *n* (%)	539 (7.8)	188 (8.2)	163 (7.1)	147 (6.4)	1.4 (1.1-1.8)
	Mother diagnosed with alcohol abuse, *n* (%)	179 (2.6)	56 (2.4)	60 (2.6)	68 (3.0)	1.8 (1.3-2.5)
Base rate of violent conviction [b]	*n* (%)	887 (12.9)	309 (13.4)	321 (14.0)	275 (12.0)	

Note. n = number of participants; HR = hazard ratio; CI = confidence interval; SD = standard deviation.
[a] Adjusted by age at assessment and gender using participants from the calibration sample. All significant at the $p < 0.05$ level.
[b] Between January 1, 1973 and December 31, 2004.

Table 4.3 Sequential Cox Regression Analyses Developing a Violence Screening Tool for Individuals with Hospital Discharge Diagnoses of Schizophrenia

Model [a]	Variable	β (*SE*)	*p*	HR
Five-item	Male gender	0.80 (0.10)	0.01	2.22
	Previous criminal conviction	0.76 (0.08)	0.01	2.14
	Young age at assessment	0.66 (0.08)	0.01	1.92
	Comorbid alcohol abuse	0.49 (0.08)	0.01	1.63
	Comorbid drug abuse	0.78 (0.08)	0.01	2.18
Five-item +	Male gender	0.80 (0.10)	0.01	2.22
Mother convicted of a violent offence	Previous criminal conviction	0.76 (0.08)	0.01	2.14
	Young age at assessment	0.65 (0.08)	0.01	1.92
	Comorbid alcohol abuse	0.49 (0.08)	0.01	1.65
	Comorbid drug abuse	0.78 (0.08)	0.01	2.18
	Mother convicted of a violent offence	0.32 (0.34)	0.34	1.38
Five-item +	Male gender	0.80 (0.10)	0.01	2.22
Father convicted of a violent offence	Previous criminal conviction	0.75 (0.09)	0.01	2.11
	Young age at assessment	0.65 (0.08)	0.01	1.91
	Comorbid alcohol abuse	0.49 (0.08)	0.01	1.64
	Comorbid drug abuse	0.77 (0.08)	0.01	2.17
	Father convicted of a violent offence	0.41 (0.16)	0.01	1.50
Five-item +	Male gender	0.80 (0.10)	0.01	2.22
Father convicted of a violent offence +	Previous criminal conviction	0.74 (0.09)	0.01	2.11
Mother diagnosed with alcohol abuse	Young age at assessment	0.65 (0.08)	0.01	1.91
	Comorbid alcohol abuse	0.49 (0.08)	0.01	1.63
	Comorbid drug abuse	0.77 (0.08)	0.01	2.16
	Father convicted of a violent offence	0.40 (0.16)	0.01	1.49
	Mother diagnosed with alcohol abuse	0.09 (0.16)	0.61	1.09
Five-item +	Male gender	0.80 (0.10)	0.01	2.22
Father convicted of a violent offence +	Previous criminal conviction	0.74 (0.09)	0.01	2.09
Non-completion of compulsory school	Young age at assessment	0.63 (0.08)	0.01	1.88
	Comorbid alcohol abuse	0.49 (0.08)	0.01	1.63
	Comorbid drug abuse	0.76 (0.08)	0.01	2.13
	Father convicted of a violent offence	0.39 (0.16)	0.01	1.48
	Non-completion of compulsory school	0.12 (0.07)	0.09	1.13
Five-item +	Male gender	0.80 (0.10)	0.01	2.22
Father convicted of a violent offence +	Previous criminal conviction	0.75 (0.09)	0.01	2.11
Father diagnosed with alcohol abuse	Young age at assessment	0.65 (0.08)	0.01	1.91
	Comorbid alcohol abuse	0.49 (0.08)	0.01	1.64
	Comorbid drug abuse	0.77 (0.08)	0.01	2.17
	Father convicted of a violent offence	0.41 (0.16)	0.01	1.51
	Father diagnosed with alcohol abuse	-0.03 (0.12)	0.83	0.98

Note. SE = standard error; HR = hazard ratio. Values based on participants from the calibration sample.
[a] All models produced Hosmer-Lemeshow goodness-of-fit statistics non-significant at the $p < 0.05$ level.

Table 4.4 Rates of True Positives, False Positives, True Negatives, and False Negatives by Risk Score for the Calibration of a Five-item Violence Screening Tool for Individuals with Hospital Discharge Diagnoses of Schizophrenia

Length of Follow-up	Risk Score	True Positives	False Positives	True Negatives	False Negatives	Sensitivity	Specificity
1 year	0	47	6598	0	0	1.00	0.00
(*n* = 6645)	1	46	5994	604	1	0.98	0.09
	2*	41	4245	2353	6	0.87	0.36
	3	26	2105	4493	21	0.55	0.68
	4	14	848	5750	33	0.30	0.87
	5	3	218	6380	44	0.06	0.97
2 years	0	93	6314	0	0	1.00	0.00
(*n* = 6407)	1	90	5741	573	3	0.97	0.09
	2*	83	4069	2245	10	0.89	0.36
	3	58	2014	4300	35	0.62	0.68
	4	29	809	5505	64	0.31	0.87
	5	9	210	6104	84	0.10	0.97
5 years	0	224	5442	0	0	1.00	0.00
(*n* = 5666)	1	216	4967	475	8	0.96	0.09
	2*	202	3481	1961	22	0.90	0.36
	3	138	1680	3762	86	0.62	0.69
	4	64	666	4776	160	0.29	0.88
	5	15	181	5261	209	0.07	0.97

Note. n = number of participants. Values based on participants from the calibration sample. Participants were excluded if they were not at risk for 1, 2, or 5 years (respectively).
* Cut-off score.

Table 4.5 Comparison of Outcome Measures Calculated during the Calibration and Cross-validation of a Five-item Violence Screening Tool for Individuals with Hospital Discharge Diagnoses of Schizophrenia

		Outcome Measure			
Length of Follow-up	Sample	NPV (95% CI)	PPV (95% CI)	DOR (95% CI) [a]	AUC (95% CI) [a]
1 year	Calibration sample (n = 6645)	0.99 (0.99-1.00)	0.01 (0.01-0.02)	3.79 (1.65-8.72)	0.67 (0.59-0.74)
	Cross-validation sample 1 (n = 2205)	0.99 (0.99-1.00)	0.01 (0.01-0.02)	5.16 (1.33-20.04)	0.74 (0.62-0.85)
	Cross-validation sample 2 (n = 2205)	0.99 (0.99-1.00)	0.01 (0.01-0.02)	2.07 (0.73-5.93)	0.66 (0.54-0.78)
	Cross-validation sample 3 (n = 2190)	0.99 (0.99-1.00)	0.01 (0.01-0.02)	5.14 (1.34-19.93)	0.62 (0.53-0.71)
2 years	Calibration sample (n = 6407)	0.99 (0.99-1.00)	0.02 (0.02-0.03)	4.58 (2.40-8.74)	0.69 (0.64-0.74)
	Cross-validation sample 1 (n = 2136)	0.99 (0.99-1.00)	0.02 (0.02-0.03)	17.89 (3.08-103.80)	0.75 (0.67-0.82)
	Cross-validation sample 2 (n = 2111)	0.99 (0.98-1.00)	0.02 (0.02-0.03)	2.56 (1.09-6.03)	0.67 (0.58-0.75)
	Cross-validation sample 3 (n = 2111)	0.99 (0.99-1.00)	0.02 (0.02-0.03)	3.12 (1.24-7.82)	0.65 (0.57-0.73)
5 years	Calibration sample (n = 5666)	0.99 (0.98-0.99)	0.06 (0.05-0.06)	5.17 (3.33-8.03)	0.69 (0.66-0.73)
	Cross-validation sample 1 (n = 1875)	0.99 (0.98-0.99)	0.06 (0.05-0.06)	4.10 (2.12-7.91)	0.68 (0.63-0.74)
	Cross-validation sample 2 (n = 1879)	0.98 (0.97-0.99)	0.05 (0.05-0.06)	3.08 (1.63-5.81)	0.68 (0.62-0.74)
	Cross-validation sample 3 (n = 1868)	0.98 (0.97-0.99)	0.05 (0.04-0.05)	2.21 (1.23-3.95)	0.66 (0.60-0.72)

Note. NPV = negative predictive value; PPV = positive predictive value; DOR = diagnostic odds ratio; AUC = area under the curve; CI = confidence interval; n = number of participants. Participants from the calibration and cross-validation samples were excluded if they were not at risk for 1, 2, or 5 years (respectively). Test statistics reported no significant differences between the effect estimates produced by the calibration and cross-validation samples.

[a] All effect sizes were significantly larger than chance at the $p < 0.05$ level.

Table 4.6 A Comparison of the Positive and Negative Predictive Values for a Five-item Violence Screening Tool for Individuals with Hospital Discharge Diagnoses of Schizophrenia across Different Base Rates of Violent Conviction

		Base Rate of Violent Conviction					
Length of Follow-up	Outcome Measure	0%	2%	4%	6%	8%	10%
1 year (n = 6645)	NPV (95% CI)	0.99 (0.99-1.00)	0.99 (0.99-1.00)	0.99 (0.98-0.99)	0.98 (0.97-0.98)	0.97 (0.97-0.98)	0.96 (0.96-0.97)
	PPV (95% CI)	0.01 (0.00-0.01)	0.03 (0.01-0.07)	0.05 (0.03-0.08)	0.08 (0.06-0.11)	0.11 (0.08-0.13)	0.13 (0.11-0.15)
2 years (n = 6407)	NPV (95% CI)	0.99 (0.99-1.00)	0.99 (0.99-1.00)	0.99 (0.98-0.99)	0.98 (0.97-0.98)	0.97 (0.97-0.98)	0.97 (0.96-0.97)
	PPV (95% CI)	0.01 (0.00-0.01)	0.03 (0.01-0.07)	0.05 (0.03-0.09)	0.08 (0.06-0.11)	0.11 (0.09-0.13)	0.13 (0.11-0.16)
5 years (n = 5666)	NPV (95% CI)	0.99 (0.99-1.00)	0.99 (0.99-1.00)	0.99 (0.98-0.99)	0.98 (0.97-0.98)	0.98 (0.97-0.98)	0.97 (0.96-0.97)
	PPV (95% CI)	0.01 (0.00-0.01)	0.03 (0.01-0.07)	0.06 (0.04-0.09)	0.08 (0.06-0.11)	0.11 (0.09-0.13)	0.14 (0.11-0.16)

Note. NPV = negative predictive value; PPV = positive predictive value; CI = confidence interval; n = number of participants. Values based on participants from the calibration sample. Participants were excluded if they were not at risk for 1, 2, or 5 years (respectively).

Table 4.7 Rates of True Positives, False Positives, True Negatives, and False Negatives by Risk Score for the Calibration of the Six-item Version of a Violence Screening Tool for Individuals with Hospital Discharge Diagnoses of Schizophrenia

Length of Follow-up	Risk Score	True Positives	False Positives	True Negatives	False Negatives	Sensitivity	Specificity
1 year	0	47	6598	0	0	1.00	0.00
(*n* = 6645)	1	47	6106	492	0	1.00	0.07
	2*	42	4691	1907	5	0.89	0.29
	3	29	2696	3902	18	0.62	0.59
	4	17	1268	5330	30	0.36	0.81
	5	7	484	6114	40	0.15	0.93
	6	2	107	6491	45	0.04	0.98
2 years	0	93	6314	0	0	1.00	0.00
(*n* = 6407)	1	91	5851	463	2	0.98	0.07
	2*	84	4498	1816	9	0.90	0.29
	3	64	2584	3730	29	0.69	0.59
	4	39	1215	5099	54	0.42	0.81
	5	19	460	5854	74	0.20	0.93
	6	4	105	6209	89	0.04	0.98
5 years	0	224	5442	0	0	1.00	0.00
(*n* = 5666)	1	218	5054	338	6	0.97	0.06
	2*	206	3873	1569	18	0.92	0.29
	3	158	2180	3262	66	0.71	0.60
	4	90	1012	4430	134	0.40	0.81
	5	40	390	5052	184	0.18	0.93
	6	8	90	5352	216	0.04	0.98

Note. n = number of participants. Values based on participants from the calibration sample. Participants were excluded if they were not at risk for 1, 2, or 5 years (respectively).
* Cut-off score.

Table 4.8 Comparison of Outcome Measures Produced by the Five- and Six-item Versions of a Violence Screening Tool for Individuals with Hospital Discharge Diagnoses of Schizophrenia

		Outcome Measure			
Length of Follow-up	Model	NPV (95% CI)	PPV (95% CI)	DOR (95% CI) [a]	AUC (95% CI) [a]
1 year (n = 6645)	Five-item	0.99 (0.99-1.00)	0.01 (0.01-0.02)	3.79 (1.65-8.72)	0.67 (0.59-0.74)
	Six-item	0.99 (0.99-1.00)	0.01 (0.01-0.02)	3.42 (1.39-8.39)	0.65 (0.58-0.72)
2 years (n = 6407)	Five-item	0.99 (0.99-1.00)	0.02 (0.02-0.03)	4.58 (2.40-8.74)	0.69 (0.64-0.74)
	Six-item	0.99 (0.99-1.00)	0.02 (0.02-0.03)	3.77 (1.92-7.41)	0.68 (0.63-0.73)
5 years (n = 5666)	Five-item	0.99 (0.98-0.99)	0.06 (0.05-0.06)	5.17 (3.33-8.03)	0.69 (0.66-0.73)
	Six-item	0.99 (0.98-0.99)	0.05 (0.05-0.06)	4.64 (2.87-7.50)	0.69 (0.65-0.72)

Note. NPV = negative predictive value; PPV = positive predictive value; DOR = diagnostic odds ratio; AUC = area under the curve; CI = confidence interval; n = number of participants. Values based on participants from the calibration sample. Participants were excluded if they were not at risk for 1, 2, or 5 years (respectively). Test statistics reported no significant differences between the effect estimates produced by the five-item and six-item versions of the screening tool.

[a] All effect sizes were significantly larger than chance at the $p < 0.05$ level.

Table 4.9 Item Response Weights Determined using Different Weighting Strategies

	Male Gender		Previous Criminal Conviction		Young Age at Assessment		Comorbid Alcohol Abuse		Comorbid Drug Abuse	
Weighting	Yes (+)	No (–)	Yes (+)	No (–)	<32 (+)	≥32 (–)	Yes (+)	No (–)	Yes (+)	No (–)
Unit scoring	+1	+0	+1	+0	+1	+0	+1	+0	+1	+0
Unstd β coef [a]	+8.0	+0	+7.6	+0	+6.6	+0	+4.9	+0	+7.8	+0
Hazard ratio	+2.2	+0	+2.1	+0	+1.9	+0	+1.6	+0	+2.2	+0

Note. Unstd β coef = unstandardised beta coefficient. Values based on multivariate Cox regression analyses using participants from the calibration sample.
[a] Beta coefficients multiplied by a constant of 10 to increase scale.

Table 4.10 Comparison of Outcome Measures Produced by a Five-item Violence Screening Tool for Individuals with Hospital Discharge Diagnoses of Schizophrenia using Different Item Weighting Strategies

			Outcome Measure			
Length of Follow-up	Cut-off Score	Weighting	NPV (95% CI)	PPV (95% CI)	DOR (95% CI) [a]	AUC (95% CI) [a]
1 year (n = 6645)	+2	Unit scoring	0.99 (0.99-1.00)	0.01 (0.01-0.02)	3.79 (1.65-8.72)	0.67 (0.59-0.74)
	+15.6	Unstd β coef	0.99 (0.99-1.00)	0.01 (0.01-0.02)	4.19 (2.23-7.88)	0.70 (0.63-0.76)
	+4.3	Hazard ratio	0.99 (0.99-1.00)	0.01 (0.01-0.02)	4.12 (2.19-7.75)	0.70 (0.63-0.76)
2 years (n = 6407)	+2	Unit scoring	0.99 (0.99-1.00)	0.02 (0.02-0.03)	4.58 (2.40-8.74)	0.70 (0.64-0.74)
	+19.1	Unstd β coef	0.99 (0.99-1.00)	0.03 (0.02-0.03)	3.54 (2.33-6.38)	0.72 (0.67-0.77)
	+4.4	Hazard ratio	0.99 (0.99-1.00)	0.03 (0.02-0.03)	3.51 (2.31-6.34)	0.72 (0.69-0.75)
5 years (n = 5666)	+2	Unit scoring	0.99 (0.98-0.99)	0.06 (0.05-0.06)	5.17 (3.33-8.03)	0.69 (0.66-0.73)
	+19.1	Unstd β coef	0.98 (0.97-0.98)	0.06 (0.06-0.07)	4.59 (3.73-7.73)	0.71 (0.68-0.74)
	+4.4	Hazard ratio	0.98 (0.97-0.98)	0.06 (0.05-0.06)	4.56 (3.71-7.68)	0.71 (0.70-0.72)

Note. NPV = negative predictive value; PPV = positive predictive value; DOR = diagnostic odds ratio; AUC = area under the curve; CI = confidence interval; n = number of participants; Unstd β coef = unstandardised beta coefficient. Values based on participants from the calibration sample. Participants were excluded if they were not at risk for 1, 2, or 5 years (respectively). Test statistics reported no evidence of significant improvements over the unit scored tool.

[a] All effect sizes were significantly larger than chance at the $p < 0.05$ level.

Table 4.11 Item Content of Structured Instruments and Multivariate Models Designed to Predict the Likelihood of Future Offending using Historical Factors

Structured Instrument			Multivariate Model		
Simple Tool	OGRS	Wootton et al. (2008)	Walsh et al. (2004)	Heinrichs & Sam (2010)	Buchanan & Leese (2006)
Male gender	Age/Gender	Male gender	Assault in past two years [b]	Male gender	Male gender
Previous criminal conviction	Sanctioning history	Assault in past two years [b]	Previous violent criminal conviction	Low level of education	Number of prior offences
Young age at assessment	Reoffending within 1 or 2 years	Young age at assessment	Receipt of special education	Age	Age at discharge
Comorbid alcohol abuse	Principal current offence	Illicit drug use in past year [b]	Comorbid alcohol abuse [b]	Non-Canadian nationality	Legal class [c]
Comorbid drug abuse	Copas rate [a]			History of street drug use	
				Non-white	
				Disadvantaged [b]	
				Unemployed	
				MSIF score	

Note. OGRS = Offender Group Reconviction Scale (Copas & Marshall, 1998); MSIF = Multidimensional Scale of Independent Functioning (Jaeger, Berns, & Czobor, 2003).
[a] ln(number of sanction occasions/ [10+ years between first and current sanction]).
[b] Item relies on patient self-report.
[c] One (or more) of the following four classes according to the Mental Health Act of 1983: mental illness, psychopathic disorder, mental impairment, or severe mental impairment.

***

Chapter V:
General Discussion

The aim of the present thesis was to investigate the utility of forensic risk assessment tools, those instruments designed to predict the likelihood of antisocial behaviour. Using systematic review approaches, major uncertainties were identified and explored, and a potentially new approach to risk assessment was introduced in the form of a violence screening tool. In this final chapter, a summary of the main findings and implications of these investigations will be provided and directions for future research will be discussed.

5.1 SUMMARY OF MAIN FINDINGS

5.1.1 Chapter II: Forensic Risk Assessment: A Metareview

The first study presented in the thesis was a metareview, a systematic summary overview of systematic reviews and meta-analyses of the field of forensic risk assessment. The aim of the metareview was to explore the methodological quality of previous reviews and to descriptively analyse their findings in order to identify key uncertainties. Epidemiological, descriptive, and reporting characteristics were extracted from 9 systematic reviews and 31 meta-analyses that were identified through a systematic search. The quality of the included reviews was investigated using the Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) Statement, a 27-item checklist of review characteristics designed to enable a transparent and consistent reporting of results.

The methodological quality of the identified systematic reviews and meta-analyses was generally poor. The average review met only two-thirds of PRISMA

criteria with few reviews reporting a replicable search strategy, approximately half of the reviews not excluding overlapping samples or investigating sources of clinical or methodological heterogeneity, and a third of the reviews not assessing publication bias. Further, the reviews reported a narrow range of effect sizes and often included a mixture of both commonly and uncommonly used risk assessment tools, making it difficult to draw conclusions about the general utility of those measures with the greatest clinical impact. The metareview also found that previous reviews of the forensic risk assessment literature have come to conflicting conclusions on a number of issues, including the comparative predictive validity of individual risk assessment tools, the efficacy of actuarial instruments versus structured clinical judgement, the influence of demographic factors and study design characteristics on predictive validity, and the relative strength of association of individual risk factors for recidivism.

5.1.2 Chapter III: A Comparative Meta-analysis of Commonly Used Risk Assessment Tools

The second study presented in the thesis was a comprehensive meta-analysis designed to further explore the uncertainties identified in the metareview. To improve quality, PRISMA guidelines were followed. Six outcome measures were calculated, each of which measured a different aspect of tool utility, and recent surveys and reviews were used to identify those risk assessment tools most commonly used in forensic practice.

A systematic search was carried out to identify replication samples for nine risk instruments, including the Level of Service Inventory – Revised (LSI-R), the Psychopathy Checklist – Revised (PCL-R), the Violence Risk Appraisal Guide (VRAG), the Sex Offender Risk Appraisal Guide (SORAG), the Static-99, the

Historical, Clinical, Risk Management – 20 (HCR-20), the Sexual Violence Risk – 20 (SVR-20), the Spousal Assault Risk Assessment (SARA), and the Structured Assessment of Violence Risk in Youth (SAVRY). As these risk assessment tools may be used for either specific case identification, whereby high specificity (or low rates of false positives) is strived for, or for screening purposes, whereby high sensitivity (or low rates of false negatives) is required, two sets of predictive validity analyses were conducted. The first set of analyses combined participants classified as low and moderate risk and compared them with those classified as high risk. The second set of analyses grouped those participants classified as moderate or high risk and compared them with participants classified as low risk. For those samples where the data needed to construct contingency tables was not available, study authors were contacted and outcome data using manual-suggested cut-off scores was requested. Using this search strategy, 68 independent studies composed of 88 samples were identified. These samples included a total of 25,980 participants from 13 countries. Previously unavailable outcome data was obtained for 54 (n = 15,775) of these samples, thus contributing a significant amount of new data to the field.

The predictive validity of the included risk assessment tools was measured using, for the first time in a meta-analysis of this literature, pooled diagnostic odds ratios (DOR). Additional outcome measures included median areas under the curve (AUC), positive predictive values (PPV), negative predictive values (NPV), numbers needed to detain (NND), and a new effect size developed as part of this thesis, the number safely screened (NSS). These outcome measures were calculated for all tools combined as well as for each instrument, individually. A ranking system was devised such that effect size results could be collated in order to determine which tools produced the highest and lowest rates of predictive validity. Finally, those potential

sources of clinical and methodological heterogeneity identified in the metareview were investigated using both subgroup analysis and metaregression. Statistical heterogeneity was measured using the I^2 statistic.

The meta-analysis found evidence that structured risk assessment tools are able to predict the likelihood of future offending with moderate accuracy, suggesting benefits to their continued use as part of the risk assessment process. However, currently used instruments were found to produce near chance positive predictive values, suggesting that their influence should be limited in making decisions related to public protection and individual liberty. No evidence was found that, compared with tools that employ structured clinical judgement, actuarial instruments produced better levels of predictive validity. However, the predictive validity of individual risk assessment measures was found to vary widely, with instruments designed for more specific populations performing better than those designed for more general purposes. Further, samples demographically similar to the calibration sample for a given instrument produced higher rates of predictive validity. Study design characteristics, including the type of offending being predicted (i.e., general versus violent) and the type of outcome being predicted (i.e., arrest, charge, conviction or incarceration versus institutional incident), were found to moderate effect size such that risk assessment tools were better at predicting violent offending and institutional incidents. Some evidence was also found that having a tool author as a study author resulted in larger effect sizes. However, this finding was partially explained by other demographic and design characteristics and might have been a marker of greater study fidelity (i.e., having used an instrument as it was designed to be used).

5.1.3 Chapter IV: Developing a Violence Screening Tool for Schizophrenia

To increase the efficiency of the risk assessment process, the final study presented in this thesis explored the feasibility of a stepped approach to risk assessment in which individuals at very low risk of future violence are screened out prior to more detailed assessment. The implementation of this stepped approach necessitates the development of highly sensitive tools that can be used to screen out low risk individuals. Such an instrument was constructed in this study.

As the metareview had identified a number of recent reviews that discussed the importance of violence risk assessment in mentally disordered populations, it was decided to develop a screening tool for use in psychiatric settings. Given the meta-analysis' finding that instruments designed for more specific populations outperform those developed for more general purposes, it was decided to construct a tool for a specific diagnostic group. Individuals with schizophrenia were chosen as the population of interest, because they are the hospitalised group that has received the most research and media attention regarding an association with violence (Fazel, Gulati, et al., 2009; Klin & Lemish, 2008). Furthermore, the low base rate of violence in this population lends itself to the principles of a stepped approach, and high-quality data on psychotic patients was available from routinely collected Swedish national registers. To improve quality, the study was designed following Standards for Reporting of Diagnostic Accuracy Studies (STARD) Statement guidelines.

The screening tool was constructed using a cohort of 13,806 patients aged 15 years and older who had been admitted to hospital for assessment and/or treatment in Sweden and had been discharged with a diagnosis of schizophrenia. These patients were followed for up to 33 years after discharge and convictions for violent offences were identified using a high-quality crime register. A number of demographic,

socio-economic and clinical risk factors were analysed using Cox regression to develop the screening tool.

Regression analyses resulted in a tool composed of five routinely available risk factors including male gender, previous criminal conviction, young age at assessment, comorbid alcohol abuse, and comorbid drug abuse. The screening instrument consistently produced sensitivities of approximately 90%, NPVs of 99%, and NSSs above 60 at 1, 2, and 5 years follow-up. No evidence of increased predictive validity was found when low level of education, parental conviction for a violent offence, or parental alcohol abuse were added as items or when items were weighted by their regression coefficients or hazard ratios. When the unweighted five-item screening tool was cross-validated, no clear shrinkage effects were found.

The findings of this investigation suggested that the simple tool could be accurately used to make "rule out" decisions (i.e., identifying who will not go on to violently offend) as part of a stepped strategy to risk assessment. As current clinical guidelines recommend the use of risk assessment tools for all patients diagnosed with schizophrenia (American Psychiatric Association, 2004; National Institute for Health and Clinical Excellence, 2009), the use of this screening instrument prior to detailed risk assessments may save mental health services considerable resources. Further, as it requires little training and is based on items that can be scored using routinely available file information, the tool is potentially scalable. Due to the limitations of the study, however, future research will need to investigate whether the instrument remains useful under different clinical (e.g., men or women participants only, White or non-White participants only, different settings) and methodological (e.g., prospective study design or smaller sample sizes) circumstances.

5.2 SUMMARY OF MAIN IMPLICATIONS

The results of the studies presented in this thesis have a number of potentially important implications for researchers, clinicians, policymakers, and patients.

5.2.1 Research Implications

There are several key research implications of the present thesis, including: (1) encouraging researchers to report a variety of effect sizes when investigating a tool's utility, (2) supporting the comparison of predictive validity estimates with standards other than chance, and (3) promoting the use of standardised quality checklists.

The findings in this thesis suggest that researchers conducting meta-analytic or primary investigations of the utility of forensic risk assessment tools should report a variety of effect sizes in order to capture different aspects of predictive validity. Outcome statistics in the prognostic prediction literature can generally be divided into three categories: (1) single effect indicators, (2) "rule in" effect sizes, and (3) "rule out" effect sizes. Single effect indicators provide global estimates of tool utility, whereas "rule in" and "rule out" effect sizes measure whether an instrument can be accurately used to identify high or low risk individuals (respectively). The findings presented in this thesis suggest that the DOR should be considered as the single effect indicator of choice rather than the AUC, particularly when conducting meta-analyses where metaregression methodology may be useful. Findings also suggest that base rate dependent "rule in" effect sizes such as the PPV and NND and "rule out" effect sizes such as the NPV and NSS may be particularly helpful in clarifying tool utility in specific populations such as psychiatric patients with different diagnoses. As all effect sizes have their limitations, using several outcome statistics to measure predictive

validity should be encouraged regardless of the study design or population under investigation.

A second main research implication of the studies presented in this thesis concerns the reference standard against which risk assessment tools are compared. Traditionally, the predictive validity estimates calculated for a risk assessment tool have been compared to those produced by chance to examine evidence of incremental validity. As recent studies have suggested that a small group of historical risk factors may be able to predict the likelihood of violent behaviour just as well as (or, in some cases, better than) complex risk assessment schemes (Chapter IV, Section 4.5.4), an alternative would be to compare the predictive validity estimates produced by risk assessment tools to those produced by readily accessible demographic risk factors for offending (e.g., male gender and young age). A second alternative would be to compare the predictive validity estimates produced by structured instruments to those of unstructured clinical judgement. As the alternative to the use of structured risk assessment instruments in clinical practice is rarely (if ever) prediction based on chance, the use of reference standards other than chance may provide more practical comparisons of predictive validity. This suggestion is consistent with intervention research which has suggested that treatment-as-usual should be used as the comparison group in randomised controlled trials instead of placebo (Burns, 2009).

Based on the findings of the studies presented in this thesis, researchers are also encouraged to make use of standardised quality checklists such as the PRISMA Statement for reviews and the Standards for Reporting of Diagnostic Accuracy Studies (STARD) Statement for primary studies. Over 200 peer-reviewed journals currently encourage the use of either checklist in their *Instructions to Authors* (PRISMA Group, 2010; STARD Group, 2008). The use of standardised checklists at

the review level would improve the generally low quality of systematic reviews and meta-analyses in the forensic risk assessment literature and ensure that replicable search strategies are included, duplicate studies or overlapping samples are addressed, clinical, methodological, and statistical heterogeneity are investigated, and publication bias is assessed. In addition to improving quality, the use of standardised checklists at the study level would also provide the authors of future meta-analyses with the information necessary to explore the potential mediating role of clinical variables such as the training and expertise of persons administering risk assessment tools on predictive validity.

In addition to these main implications, the findings presented in this thesis may have several additional suggestions for researchers. First, given the conflicting findings of previous reviews of risk factors for antisocial behaviour, updated reviews of the relative strength of risk and protective factors for offending are needed, especially for specific types of criminal (e.g., sexual offenders) and psychiatric groups (e.g., substance abusers). Such reviews may assist researchers in developing instruments that maximise sensitivity and specificity for different populations. Second, very few tools have been validated for short follow-up times such as hours, days, or weeks, which typically is the most relevant timeframe in clinical real-world decision-making situations (SBU, 2005). Future predictive validity studies may wish to investigate the utility of risk assessment tools in detecting the risk of antisocial behaviour over shorter periods of time. Third, given the finding of the meta-analysis that commonly used risk instruments are most accurate for samples of predominantly white individuals, future research may wish to investigate risk and protective factors for ethnic minorities, possibly resulting in the construction of novel instruments. Relatedly, given the meta-analysis' findings concerning the importance of participant

age in tool accuracy, future research could investigate the relative utility of risk assessment instruments in different age bands. Fourth, while evidence was found that clinicians are better at predicting institutional misconduct than community offending, there was considerable methodological heterogeneity in the definition of an "institutional incident". Therefore, future research could investigate the usefulness of risk assessment tools in predicting different forms of institutional misconduct (e.g., contact versus non-contact incidents). Fifth, in terms of reporting standards in diagnostic accuracy studies, future investigations into the predictive validity of risk assessment tools should present rates of true positives, true negatives, false positives, and false negatives at each risk score so that interested parties can calculate a variety of effect sizes and determine which cut-off score they wish to employ for their specific needs (e.g., specific case identification versus screening). Finally, to increase the transparency of study protocols and the accessibility of primary and tabular data, researchers may wish to develop a website on which observational investigations of risk assessment tool predictive validity may be registered.

5.2.2 Clinical Implications

Risk assessment procedures and guidelines by mental health services and criminal justice systems may need review in light of the findings presented in this thesis. Key clinical implications of the present thesis include: (1) cautioning against reliance on low quality and outdated literature reviews for psychoeducation and training, (2) encouraging mental health and criminal justice services to select the tool that best matches their population and outcome of interest regardless of its approach to risk assessment, and (3) promoting a stepped approach to risk assessment.

Given the methodological limitations of previous systematic reviews and meta-analyses of the forensic risk assessment literature, clinicians should be cautioned against relying on such reviews for psychoeducation and training concerning the utility of risk assessment tools. Further, the most cited systematic reviews and meta-analyses of the forensic risk assessment literature are currently between 8 to 14 years old, suggesting that clinicians and policymakers' views of the field may be based on outdated literature. Therefore, interested clinicians may wish to read specialty journals such as *Law & Human Behavior*, *Criminal Justice & Behavior*, and *The International Journal of Forensic Mental Health* which regularly publish new reviews as well as primary articles on different risk assessment instruments in a variety of populations and settings.

A second clinical implication of the studies included in the present thesis is that, when deciding which risk assessment tool to administer, clinicians should focus on which instrument produces the highest rate of predictive validity for their population, outcome, and setting of interest. Additional considerations when choosing a risk measure may include the costs of training and materials, ease of use, and whether a tool is useful in making decisions regarding effective treatment and risk management. Given their focus on treatment planning, tools that adopt the structured clinical judgement approach may help to bridge the gap between accurate risk assessment and effective risk management. The Psychopathy Checklist measures, once considered to be unparalleled in their predictive validity, should now be primarily used as personality assessments as opposed to risk assessment tools.

The findings of the thesis also suggest the potential utility of a stepped approach to risk assessment in which individuals at very low risk of violence are screened out prior to detailed clinical assessment. With the time involved in

administering risk assessment tools (familiarising oneself with the patient's case, collecting information from multiple sources to score items, conducting interviews, scoring the tool, and, in some cases, making a clinical judgement regarding risk level) and the costs involved (attending training sessions, purchasing tool manuals, and paying for each coding sheet), violence risk assessment, as often conducted, can cost mental health services significant amounts of time and money. Therefore, screening tools that can reduce caseload without sacrificing accuracy are potentially attractive. In addition to saving resources, the use of a stepped strategy may reduce the number of false positive predictions made during detailed clinical assessments by artificially increasing the base rate of violence amongst those individuals who are not screened out. Finally, screening tools may also assist in developing effective risk management strategies by highlighting criminogenic needs that could be the basis for treatment (e.g., alcohol or drug abuse).

In addition to these main implications, the findings presented in this thesis may have several additional suggestions for clinicians. First, the findings of the meta-analysis that it *does* matter which instrument is used for the purposes of risk assessment and that tools perform best when used as designed suggest that clinicians who conduct risk assessments may need additional training to learn about structured measures' intended populations and outcomes. As part of this training, clinicians should be encouraged to pay stricter attention to manual-based protocols on how to administer and score items on risk assessment tools, both actuarial and clinically based. Second, practitioners could be educated about how to interpret commonly reported effect sizes such as the AUC and Cohen's *d* which may be misinterpreted as providing the clinically useful information that only base rate dependent statistics can (e.g., the positive and negative predictive values). Relatedly, clinicians should be

urged to use caution when using instruments whose predictive validity has not been empirically established. Third, given evidence from recent questionnaire surveys that the PCL-R is the most commonly used risk assessment tool in forensic evaluations in the US and UK (Archer et al., 2006; Khiroya et al., 2009; Viljoen et al., 2010) and the LSI-R is the most commonly used risk assessment tool by US parole boards (Harcourt, 2007), the finding of the meta-analysis that these instruments produced the lowest rates of predictive validity across effect sizes suggest that current risk assessment protocols in Western psychiatric and criminal justice settings may need review. The PCL-R could be used as designed, as a personality assessment, and the LSI-R as a needs assessment. For the purposes of risk assessment, perhaps tools that employ structured clinical judgement warrant consideration due to their patient-centred approach to assessment and comparable accuracy.

5.2.3 Legal Implications

The evidence presented in this thesis suggests that the field of forensic risk assessment remains, in the words of a notable paper published in 1974, at the level of "flipping coins in the courtroom" (Ennis & Litwack, 1974). The cautious optimism that experts have described in relation to the ability to predict criminal behaviour appears to be based on relying on the AUC, an index of sensitivity and specificity across all possible cut-offs, as the effect size of choice (Bonta, Harman, Hann, & Cormier, 1996; Simon, 2005). However, in the courtroom, where dichotomous decisions need to be made and positive and negative predictive values are arguably more relevant indices of tool performance, the accuracy of structured risk instruments appears to be far from perfect. Given that medico-legal decisions regarding individual liberty are frequently influenced by the judgements of structured risk measures, the

findings of the thesis may have important implications for policymakers and the courts.

In the past 15 years, legislation has been passed in countries such as the UK, Canada, the US, and Australia concerning the indeterminate imprisonment or institutionalisation of high risk prisoners and psychiatric patients for the purposes of mandated treatment (Mercado, 2006). Each of these statutes requires that it be established that the prisoner or patient in question poses a danger to others, which is frequently accomplished through the use of a structured risk assessment instrument. In 2003 new legislation was passed in the UK allowing for the indeterminate sentencing of offenders who were judged to be at high risk of future offending and whose hypothetical future offences would likely result in severe psychological and/or physical injury (Criminal Justice Act, 2003). In Canada, offenders who evidence a pattern of repetitive or persistently aggressive behaviour and are judged to be at high risk of future violence may receive the designation of *dangerous offender*, permitting indeterminate sentences or long-term community supervision orders (Heilbrun, Ogloff, & Picarello, 1999). In the United States, 20 of the 50 states currently have some form of *Sexual Violent Predator* (SVP) legislation which allows for sex offenders to be sectioned for an indeterminate period of time if judged to be at high risk of recidivism upon release. Finally, in Australia, the *Dangerous Prisoners (Sexual Offenders) Act* (2003) allows for incarcerated sexual offenders to be further detained at the end of their sentences for the purposes of public protection and continued rehabilitation. Critics of these statutes have argued that they violate civil rights, as individuals are being judged for crimes that have yet to have been committed (Birgden & Cucolo, 2010; Janus, 2004; La Fond, 2008). Nevertheless,

high court cases (e.g., *Smith v. Doe* in the US and *Attorney General v. Fardon* in Australia) have upheld these laws.

Given that establishing dangerousness is a prerequisite, court cases concerning mandated treatment, treatment guidelines, and public policy may be informed by the findings of the present thesis. While it is commonly accepted that structured approaches to risk assessment are more predictively valid than unstructured methods (Borum, 1996), this does not mean that actuarial and SCJ instruments produce highly accurate estimates of offending risk. Indeed, the probability of a false positive prediction of high risk is approximately 50% using structured instruments (Chapter III, Section 3.4.3.2), potentially resulting in the detention of individuals for longer than necessary, with its attendant economic (Tyrer et al., 2010), social (Szmukler, 2003), and civil rights consequences (Janus, 2004). If the findings of structured assessment tools are presented in court as evidence of an individual's risk of future offending, it should be made very clear that the predictive validity of these instruments is far from perfect. Novel effect sizes such as the number needed to detain and the number safely screened may allow for a more informed discussion of issues surrounding the reliance on structured prediction methods.

5.2.4 Applying Group-Level Findings to the Individual

The forensic psychiatric community continues to search for better ways of conducting research that will inform clinical practice with individual clients (Buchanan, 2008). However, research is not conducted on individuals but rather on large samples, methodology employed to ensure that robust conclusions are produced. Due to this group-based approach, predictive validity studies on risk assessment tools can only provide practitioners with information concerning the general behaviour of

actuarial and SCJ schemes rather than on their performance in predicting criminality in any specific case. Indeed, all individuals who have their risk of future offending assessed will have unique criminal histories, socio-demographic circumstances, psychiatric comorbidities, and treatment needs. While some experts have argued that this instability makes evidence-based research on the predictive validity of risk assessment tools irrelevant (Cooke & Michie, 2010; Hart, 2008; Hart et al., 2007), others have argued that vulnerability to the ecological fallacy does not mean that group-level analyses are not useful in assisting clinicians to make medico-legal decisions. To illustrate the importance of aggregated data when applied to the individual case, Grove and Meehl (1996) use the following analogy:

> [S]uppose you are a political opponent held in custody by a mad dictator. Two revolvers are put on the table and you are informed that one of them has five live rounds with one empty chamber, the other has five empty chambers and one live cartridge, and you are required to play Russian roulette. If you live, you will go free. Which revolver would you choose? Unless you have a death wish, you would choose the one with the five empty chambers. Why? Because you would know that the odds are five to one that you will survive if you pick that revolver, whereas the odds are five to one you will be dead if you choose the other one. Would you seriously think, "Well, it doesn't make any difference what the odds are. Inasmuch as I'm only going to do this once, there is no aggregate involved, so I might as well pick either one of these two revolvers; it doesn't matter which"? (pp. 305-306)

In summary, though the metareview, the meta-analysis, and the development of the simple screening instrument relied on group-based analyses to derive their conclusions, their findings still have important implications for clinical practice at the level of the individual patient or offender. Current review guidelines by systematic reviewing groups such as the Cochrane Collaboration, the Drug Effectiveness Review Project, and the Agency for Healthcare Research and Quality suggest that individual participant meta-analytic methodology would allow less biased assessment of clinical heterogeneity (West et al., 2010). This methodology is discussed further in Section 5.3.

5.2.4.1 Patient Implications

Although the ecological fallacy limits the conclusions that can be made about the predictive validity of forensic risk assessment tools for specific individuals, the findings of the thesis may still have important implications for patients. For example, the finding of the meta-analysis in Chapter III that the SAVRY, the only one of the nine researched instruments to include protective factors, produced the highest rates of predictive validity across effect sizes may result in increased interest in patients' strengths and potential for rehabilitation. In addition, the finding that instruments perform best when used on samples similar to their calibration sample may result in pre-assessment protocols designed to match individuals with risk assessment tools. Thirdly, the use of the stepped strategy proposed in Chapter IV, with its focus on predicting non-offending in low base rate populations, may decrease the stigma of dangerousness currently surrounding individuals with schizophrenia (Klin & Lemish, 2008; National Alliance on Mental Illness, 2008; Pescosolido et al., 2010). Reducing such stigma may result in increased social adjustment, decreased symptom severity, and an increased likelihood of seeking help from psychological services (Vauth et al., 2007; Mark & Wu, 2006; Vogel et al., 2006).

5.3 FUTURE DIRECTIONS FOR RESEARCH

The findings presented in this thesis suggest a number of directions for future research. Three avenues for future research include: (1) conducting individual participant meta-analyses to further explore uncertainties, (2) investigating the characteristics of clinicians who make accurate structured clinical judgements, and (3) piloting the stepped strategy to risk assessment.

5.3.1 Individual Participant Meta-analyses

While a number of meta-analyses have investigated the utility of risk assessment tools across demographics and study designs, the conclusions that can be drawn from such analyses are limited. As meta-analyses use the sample as the unit of analysis, previous reviews have been vulnerable to the "ecological fallacy" (Morgenstern, 1982, p. 1339), whereby group-level findings are falsely applied to individuals. For example, in the meta-analysis presented in Chapter III, tabular data (i.e., rates of true positives, false positives, true negatives, and false negatives) was rarely available in study manuscripts for white and non-white participants, separately. Therefore, to explore the influence of ethnicity on predictive validity, the association between the percentage of white individuals in a sample and effect size was explored as a proxy. While such analyses may have estimated the moderating influence of this demographic factor to an extent, the conclusions that can be drawn from significant or non-significant findings are limited. A potential solution would be to conduct individual participant meta-analyses in which raw data is obtained from study authors rather than tabular data or effect sizes. This would allow researchers to clarify the utility of risk assessment tools for men versus women, individuals of different ethnic backgrounds, individuals of different ages, and individuals with different psychiatric diagnoses. Individual participant meta-analytic methodology could also be used to explore the relative strength of individual risk factors for offending as well as the accuracy of risk assessment tools in making long-term versus short-term predictions of antisocial behaviour.

5.3.2 Investigating the Characteristics of Accurate Clinicians

The present thesis found evidence that risk assessment tools that employ structured clinical judgement produce commensurate rates of predictive validity to actuarial instruments. This finding could further increase the popularity of clinically based measures, which have been gaining acceptance in mental health and criminal justice settings due to their patient-centred approach to assessment and risk management (Archer et al., 2006; Khiroya et al., 2009; Viljoen et al., 2010). Future research may wish to explore which characteristics discriminate clinicians whose structured judgements produce high versus low rates of predictive validity. For example, previous studies have suggested that adequate training in probability theory and common inferential errors may increase the accuracy of clinical judgements (Arkes, 1981, 1991; Nisbett & Ross, 1980; Spengler & Strohmer, 2001). Additional clinical variables that may warrant investigation include the setting in which the professional works (e.g., emergency care, inpatient services, community contact), the professional's theoretical orientation (e.g,. cognitive, behavioural, psychodynamic), the professional's educational background (e.g., forensic coursework and specialised training workshops), the amount of supplementary information on which professionals base their structured judgements, and the professional's amount of clinical experience. Findings would have important implications for educational institutions that train mental health professionals (e.g., psychiatrists and psychiatric nurses, clinical and counselling psychologists, and social workers).

5.3.3 Piloting the Stepped Approach to Risk Assessment

Before the stepped approach to risk assessment is considered for implementation in clinical practice, it should be piloted. First, retrospective

methodology should be used to confirm that having screened out very low risk participants prior to detailed clinical assessment would have yielded similar (or possibly even higher) rates of predictive validity than when resource-intensive assessments were used to predict violence alone. Non-significant findings should be interpreted optimistically, as they would imply that using the stepped approach could have saved time and money without sacrificing accuracy. As there is evidence to suggest that the VRAG and the HCR-20 are accurate predictors of future violence in individuals with schizophrenia (Grann, Belfrage, & Tengström, 2000; Tengström, 2001) and as there is precedent for using both instruments retrospectively (Dahle, 2006; Douglas, Yeomans, & Boer, 2005; Kroner et al., 2007), the stepped strategy could be piloted using the simple tool developed in Chapter IV in the screening stage and these measures in the detailed assessment stage. If evidence is found that the stepped approach does not lead to a loss in predictive validity, the use of the simple screening tool could then be tested in practice. Such prospective investigations should include both an economic analysis (i.e., comparison of resources consumed prior to implementation and then after implementation) and qualitative analyses to explore clinicians' perceptions of the novel approach. Favourable predictive validity, economic, and qualitative findings in prospective pilot studies would increase the credibility of the stepped approach and may lead to its adoption by clinicians and mental health services.

5.4 CONCLUSION

As the prison population and number of forensic beds in Western nations continue to grow, so too will the need for valid and reliable methods of identifying individuals who will commit criminal acts. In this thesis, uncertainties concerning the

general utility of forensic risk assessment tools were identified and the predictive validity of those measures that are most commonly used in clinical practice was clarified. Findings suggest that while structured instruments predict the likelihood of offending with modest accuracy, mental health professionals in psychiatric, correctional, and court settings should not rely exclusively on these instruments to make decisions regarding individual liberty and public protection. The use of existing measures as part of a stepped strategy that employs simple screening tools may offer a way forward, with improvements in both the efficiency and accuracy of the risk assessment process.

***

REFERENCES

References marked with a ‡ are systematic reviews included in the metareview
References marked with a † are meta-analyses included in the metareview
References marked with a * are primary studies included in the meta-analysis

Aberson, C. L. (2010). *Applied power analysis for the behavioral sciences*. New York: Routledge.

Adams, C. E., Fenton, M. K. P., Quraishi, S., & David, A. S. (2001). Systematic meta-review of depot antipsychotic drugs for people with schizophrenia. *British Journal of Psychiatry, 179*, 290-299.

†Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., et al. (2006). The meta-analysis of clinical judgement project: Fifty-six years of accumulated research on clinical versus statistical prediction. *Counseling Psychologist, 34,* 341-382.

Altman, D. G. (1991). *Practical statistics for medical research.* London: Chapman and Hall.

American Psychiatric Association. (1974). *Task force report 8: Clinical aspects of the violent individual*. Washington, DC: American Psychiatric Association.

American Psychiatric Association. (2004). *Practice guideline for the treatment of patients with schizophrenia*. Arlington, VA: American Psychiatric Association.

Andrews, D. A. (1982). *The Level of Service Inventory (LSI): The first follow-up*. Toronto, ON: Ministry of Correctional Services of Ontario.

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

Andrews, D. A., & Wormith, J. S. (1984). *The Criminal Sentiments Scale*. Ottawa, ON: Correctional Services of Canada.

Anwar, S., Långström, N., Grann, M., & Fazel, S. (2011). Is arson the crime most strongly associated with psychosis? A national case-control study of arson risk in schizophrenia and other psychoses. *Schizophrenia Bulletin, 37*, 580-586.

*Arbach, K., & Pueyo, A. A. (2007). Violence risk assessment in mental disorders with the HCR-20. *Papeles del Psicologo, 28*, 174-186.

Archer, R. P., Buffington-Vollum, J. K., Stredny, R. V., & Handel, R. W. (2006). A survey of psychological test use patterns among forensic psychologists. *Journal of Personality Assessment, 87*, 84-94.

Arkes, H. R. (1981). Impediments to accurate clinical judgment and possible ways to minimize their impact. *Journal of Consulting & Clinical Psychology, 49*, 323-330.

Arkes, H. R. (1991). Costs and benefits of judgment errors: Implications for debiasing. *Psychological Bulletin, 110*, 486-498.

Arseneault, L., Moffitt, T., Caspi, A., Taylor, P., & Silva, P. (2000). Mental disorders and violence in a total birth cohort: Results from the Dunedin study. *Archives of General Psychiatry, 57*, 979-986.

*Austin, J., Coleman, D., Peyton, J., & Johnson, K. D. (2003). *Reliability and validity study of the LSI-R risk assessment instrument*. Washington, DC: Pennsylvania Board of Probation and Parole.

Baguley, T. (2009). Standardized or simple effect size: What should be reported? *British Journal of Psychology, 100*, 603-617.

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

Bauer, A., Rosca, P., Khawalled, R., Gruzniewski, A., & Grinshpoon, A. (2003). Dangerousness and risk assessment: The state of the art. *Israel Journal of Psychiatry & Related Sciences, 40,* 182-190.

Baxstrom v. Herold, 383 U.S. 107 (86 S. Ct. 760 1966).

Beck, A. T., Weissman, A., Lester, D., & Trexler, L. (1974). The measurement of pessimism: The Hopelessness Scale. *Journal of Consulting & Clinical Psychology, 42*, 861-865.

Becker, L. A., & Oxman, A. D. (2008). Overviews of reviews. In J. Higgins, & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions 5.0.1*. Chichester, UK: John Wiley & Sons.

Beech, A., Fisher, D. D., & Thornton, D. (2003). Risk assessment of sex offenders. *Professional Psychology: Research & Practice, 34,* 339-352.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., et al. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT Statement. *Journal of the American Medical Association, 276*, 637-639.

*Beggs, S. M., & Grace, R. C. (2008). Psychopathy, intelligence, and recidivism in child molesters: Evidence of an interaction effect. *Criminal Justice & Behavior, 35*, 683-695.

Bekelman, J. E., Li, Y., & Gross, C. P. (2003). Scope and impact of financial conflicts of interest in biomedical research: A systematic review. *Journal of the American Medical Association, 289*, 454-465.

Belknap, J., & Holsinger, K. (2006). The gendered nature of risk factors for delinquency. *Feminist Criminology, 1*, 48-71.

*Bengtson, S. (2008). Is newer better? A cross-validation of the Static-2002 and the Risk Matrix 2000 in a Danish sample of sexual offenders. *Psychology, Crime & Law, 14,* 85-106.

Bergman, B., Belfrage, H., & Grann, M. (1999). Mentally disordered offenders in Sweden: Forensic and general psychiatric diagnoses. *American Journal of Forensic Psychiatry, 20*, 27-37.

Bhui, H. S. (1999). Race, racism and risk assessment: Linking theory to practice with Black mentally disordered offenders. *Probation Journal, 46,* 171-181.

Birgden, A., & Cucolo, H. (2010). The treatment of sex offenders: Evidence, ethics, and human rights. *Sexual Abuse,* doi:10.1177/1079063210381412.

†Bjørkly, S. (1995). Prediction of aggression in psychiatric patients: A review of prospective prediction studies. *Clinical Psychology Review, 15*, 475-502.

†Blair, P. R., Marcus, D. K., & Boccaccini, M. T. (2008). Is there an allegiance effect for assessment instruments? Actuarial risk assessment as an exemplar. *Clinical Psychology: Science & Practice, 15*, 346-360.

‡Blank, A. (2001). Patient violence in community mental health: A review of the literature. *British Journal of Occupational Therapy, 64*, 584-589.

Bloom, H., Webster, C. D., Hucker, S., & de Freitas, K. (2005). The Canadian contribution to violence risk assessment: History and implications for current psychiatric practice. *Canadian Journal of Psychiatry, 50*, 3-11.

Boer, D. P., Hart, S. D., Kropp, P. R., & Webster, C. D. (1997). *Manual for the Sexual Violence Risk – 20. Professional guidelines for assessing risk of sexual violence*. Burnaby, BC: Simon Fraser University, Mental Health, Law, and Policy Institute.

Bonta, J. (2002). Offender risk assessment: Guidelines for selection and use. *Criminal Justice & Behavior, 29,* 355-379.

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

†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.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley.

Borum, R., Bartel, P., & Forth, A. (2002). *Manual for the structured assessment of violence risk in youth (SAVRY)*. Tampa: University of South Florida.

Borum, R., Bartel, P., & Forth, A. (2003). *Manual for the structured assessment of violence risk in youth (SAVRY). Version 1.1*. Tampa: University of South Florida.

Bossuyt, P. M., Reitsma, J. B., Bruns, D. E., Gatsonis, C. A., Glaziou, P. P., Irwig, L. M., et al. (2003). Towards complete and accurate reporting of studies of diagnostic accuracy: The STARD initiative. *British Medical Journal, 326*, 41-44.

Brennan, P. A., Mednick, S. A., & Hodgins, S. (2000). Major mental disorders and criminal violence in a Danish birth cohort. *Archives of General Psychiatry, 57*, 494-500.

Breslow, N., & Day, N. (1987). *Statistical methods in cancer research*. Lyon: International Agency for Research on Cancer.

Buchanan, A. (2008). Risk of violence by psychiatric patients: Beyond the "actuarial versus clinical" assessment debate. *Psychiatric Services, 59*, 184-190.

†Buchanan, A., & Leese, M. (2001). Detention of people with dangerous severe personality disorders: A systematic review. *Lancet, 358*, 1955-1959.

Buchanan, A., & Leese, M. (2006). Quantifying the contributions of three types of information to the prediction of criminal conviction using the receiver operating characteristic. *British Journal of Psychiatry, 188*, 472-478.

Burgess, E. W. (1928). Factors determining success or failure on parole. In A. A. Bruce, E. W. Burgess, & A. J. Harno (Eds.), *The workings of the indeterminate sentence law and the parole system in Illinois* (pp. 221-234). Springfield, IL: State Board of Parole.

Burns, T. (2009). End of the road for treatment as usual studies? *British Journal of Psychiatry, 195*, 5-6.

Buss, A. H., & Durkee, A. (1957). An inventory for assessing different kinds of hostility. *Journal of Consulting & Clinical Psychology, 21*, 343-349.

Caldwell, R. A., Bogat, G. A., & Davidson, W. S. (1988). The assessment of child abuse potential and the prevention of child abuse and neglect: A policy analysis. *American Journal of Community Psychology, 16,* 609-624.

†Campbell, M., French, S., & Gendreau, P. (2007). *Assessing the utility of risk assessment tools and personality measures in the prediction of violent recidivism for adult offenders* (Cat. No. PS3-1/2007-4E-PDF). Ottawa, ON: Department of Safety and Emergency Preparedness.

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

Campbell, M., & Gardner, M. J. (1988). Calculating confidence intervals for some non-parametric analyses. *British Medical Journal*, *296*, 1454-1456.

Cannon, M., Huttunen, M. O., Tanskanen, A. J., Arseneault, L., Jones, P. B., & Murray, R. M. (2002). Perinatal and childhood risk factors for later criminality and violence in schizophrenia. *British Journal of Psychiatry, 180*, 496-501.

*Caperton, J. D. (2005). *Predicting recidivism among sex offenders: Utility of the STATIC-99, Minnesota Sex Offender Screening Tool – Revised, and Psychopathy Checklist – Revised*. Unpublished doctoral dissertation, Sam Houston State University, Huntsville, TX.

Cipriani, A., Geddes, J. R., Furukawa, T. A., & Barbui, C. (2007). Metareview on short-term effectiveness and safety of antidepressants for depression: An evidence-based approach to inform clinical practice. *Canadian Journal of Psychiatry, 52*, 543-534.

Cleckley, H. (1941). *The mask of sanity*. St. Louis: C.V. Mosby.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10,* 101–129.

Cocozza, J. J., & Steadman, H. J. (1978). Prediction in psychiatry: An example of misplaced confidence in experts. *Social Problems, 25*, 265-270.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement, 20,* 37-46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

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

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*, 98-101.

Coid, J., Yang, M., Ullrich, S., Zhang, T., Roberts, A., Roberts, C., et al. (2007). *Predicting and understanding risk of re-offending: The Prison Cohort Study*. London: Ministry of Justice.

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

Cook, R. J., & Sackett, D. L. (1995). The number needed to treat: A clinically useful measure of treatment effect. *British Medical Journal, 310,* 452-454.

Cooke, D. J., & Michie, C. (2010). Limitations of diagnostic precision and predictive utility in the individual case: A challenge for forensic practice. *Law & Human Behavior, 34*, 259-274.

Cooper, B. S., Griesel, D., & Yuille, J. C. (2008). Clinical-forensic risk assessment: The past and current state of affairs. *Journal of Forensic Psychology Practice, 7*, 1-63.

Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis*. New York: Russell Sage Foundation.

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

Costa, P., & McCrae, R. (1992). *Revised NEO Personality Inventory (NEO–PI–R) and NEO Five-Factor Inventory (NEOFFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

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

Craig, L. A., & Beech, A. R. (2010). Towards a guide to best practice in conducting actuarial risk assessments with sex offenders. *Aggression & Violent Behavior, 15*, 278-293.

Daffern, M. (2007). The predictive validity and practical utility of structured schemes used to assess risk for aggression in psychiatric inpatient settings. *Aggression & Violent Behavior, 12*, 116-130.

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

Dalman, C., Broms, J., Cullberg, J., & Allebeck, P. (2002). Young cases of schizophrenia identified in a national inpatient register: Are the diagnoses valid? *Social Psychiatry & Psychiatric Epidemiology, 37*, 527-531.

Daniels, B. A. (2005). *Sex offender risk assessment: Evaluation and innovation*. Unpublished doctoral dissertation, Widener University, Chester, PA.

*Davidson, J. (2007). *Risky business: What standard assessments mean for female offenders*. Unpublished doctoral dissertation, University of Hawaii, Manoa, HI.

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

de Ruiter, C., & Hildebrand, M. (2007). Risk assessment and treatment in Dutch forensic psychiatry. *Netherlands Journal of Psychology, 63*, 152-160.

*de Vogel, V., & de Ruiter, C. (2005). The HCR-20 in personality disordered female offenders: A comparison with a matched sample of males. *Clinical Psychology & Psychotherapy, 12*, 226-240.

de Vogel, V., de Ruiter, C., Bouman, Y., & de Vries Robbé, M. (2007). *Handleiding bij de SAPROF: Structured Assessment of Protective Factors for Violence Risk. Versie 1.* [Manual of the SAPROF: Structured Assessment of Protective Factors for Violence Risk. Version 1]. Utrecht: Forum Educatief.

de Vogel, V., de Ruiter, C., Hildebrand, M., Bos, B., & van de Ven, P. (2004). Type of discharge and risk of recidivism measured by the HCR-20: A retrospective study in a Dutch sample of treated forensic psychiatric patients. *International Journal of Forensic Mental Health, 3,* 149-165.

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

Dean, K., Walsh, E., Morgan, C., Demjaha, A., Dazzan, P., Morgan, K., et al. (2007). Aggressive behaviour at first contact with services: Findings from the AESOP First Episode Psychosis Study. *Psychological Medicine, 37*, 547-557.

DeCoster, J. (2009). *Effect size conversion calculator*. Retrieved July 7, 2009, from http://web.me.com/rsbalkin/Site/Research_Methods_and_Statistics_files/Converting%20effect%20sizes--calculator.xls

Deeks, J. (2001). Systematic reviews of evaluation of diagnostic and screening tests. In M. Egger, G. D. Smith, & D. G. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context*. London: BMJ Publishing Groups.

Deeks, J., Higgins, J., & Altman, D. (2006). Analysing and presenting results. In J. Higgins, & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions 4.2.6.* Chichester, UK: John Wiley & Sons.

Delgado-Rodriguez, M. (2006). Systematic reviews of meta-analyses: Applications and limitations. *Journal of Epidemiology & Community Health, 60*, 90-92.

DeMatteo, D., & Edens, J. (2006). The role and relevance of the Psychopathy Checklist – Revised in court: A case law survey of US courts (1991-2004). *Psychology, Public Policy, & Law, 12,* 214-241.

DeMatteo, D., Heilbrun, K., & Marczyk, G. (2005). Psychopathy, risk of violence, and protective factors in a noninstitutionalized and noncriminal sample. *International Journal of Forensic Mental Health*, *4*, 147-157.

*Dempster, R. J. (1998). *Prediction of sexually violent recidivism: A comparison of risk assessment instruments*. Unpublished Master's thesis, Simon Fraser University, Burnaby, BC.

*Dempster, R. J. (2001). *Understanding errors in risk assessment: The application of differential prediction methodology*. Unpublished doctoral dissertation, Simon Fraser University, Burnaby, BC.

Dempster, R. J. (2003). Issues in the assessment, communication, and management of risk for violence. In W. T. O'Donohue & E. R. Levensky (Eds.), *Handbook of forensic psychology: Resource for mental health and legal professionals*. Sydney: Elsevier Academic Press.

Department of Health. (1999a). *Managing dangerous people with severe personality disorder: Proposals for policy development*. London: Department of Health.

Department of Health. (1999b). *Report of the expert committee: Review of the Mental Health Act 1983*. London: Stationery Office.

Dernevick, M., Beck, A., Grann, M., Hogue, T., & McGuire, J. (2010). The use of psychiatric and psychological evidence in the assessment of terrorist offenders. *Journal of Forensic Psychiatry & Psychology, 20,* 508-515.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*, 177-188.

Des Jarlais, D. C., Lyles, C., Crepaz, N., & the TREND Group. (2004). Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: The TREND Statement. *American Journal of Public Health, 94*, 361-366.

Deville, W. L., Buntinx, F., Bouter, L. M., Montori, V. M., de Vet, H. C. W., van der Windt, D. A. W. N., et al. (2002). Conducting systematic reviews of diagnostic studies: Didactic guidelines. *BMC Medical Research Methodology, 2,* 1-13.

DiMatteo, M. R. (2004). Social support and patient adherence to medical treatment: A meta-analysis. *Health Psychology*, *23*, 207-218.

DiMatteo, M. R., Lepper, H. S., & Croghan, T. W. (2000). Depression is a risk factor for noncompliance with medical treatment: Meta-analysis of the effects of anxiety and depression on patient adherence. *Archives of Internal Medicine*, *160*, 2101-2107.

Dixon v. Attorney General of the Commonwealth of Pennsylvania, 325 F.Supp 966 (E.D. Pa. 1971).

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

*Dolan, M., & Rennie, C. E. (2008). The Structured Assessment of Violence Risk in Youth as a predictor of recidivism in a United Kingdom cohort of adolescent offenders with conduct disorder. *Psychological Assessment, 20,* 35-46.

Dolmén, L. (2001). *Brottsligheten I olika länder.* [Criminality in different countries]. Stockholm: National Council for Crime Prevention.

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

Douglas, K. S., Blanchard, A. J. E., Guy, L. S., Reeves, K. A., & Weir, J. (2010). *HCR-20 violence risk assessment scheme: Overview and annotated bibliography*. Retrieved December 1, 2010, from www.violence-risk.com/hcr20annotated.pdf

Douglas, K. S., Cox, D. N., & Webster, C. D. (1999). Violence risk assessment: Science and practice. *Legal & Criminological Psychology, 4*, 149-184.

*Douglas, K. S., Ogloff, J. R. P., & Hart, S. D. (2003). Evaluation of a model of violence risk assessment among forensic psychiatric patients. *Psychiatric Services, 54,* 1372-1379.

Douglas, K. S., & Skeem, J. L. (2005). Violence risk assessment: Getting specific about being dynamic. *Psychology, Public Policy, & Law, 11*, 347-383.

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

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

*Dowdy, E. R., Lacy, M. G., Unnithan, N. P. (2002). Correctional prediction and the Level of Supervision Inventory. *Journal of Criminal Justice, 30,* 29-39.

*Ducro, C., & Pham, T. (2006). Evaluation of the SORAG and the Static-99 on Belgian sex offenders committed to a forensic facility. *Sexual Abuse: A Journal of Research & Treatment, 18,* 15-26.

Edens, J. (2001). Misuses of the Hare Psychopathy Checklist – Revised in court. *Journal of Interpersonal Violence, 16,* 1082-1093.

†Edens, J., & Campbell, J. (2007). Identifying youths at risk for institutional misconduct: A meta-analytic investigation of the Psychopathy Checklist measures. *Psychological Services, 4*, 13-27.

†Edens, J., Campbell, J. S., & Weir, J. M. (2007). Youth psychopathy and criminal recidivism: A meta-analysis of the Psychopathy Checklist measures. *Law & Human Behavior, 31*, 53-75.

Edens, J., Skeem, J., Cruise, K., & Cauffman, E. (2001). The assessment of juvenile psychopathy and its association with violence: A critical review. *Behavioral Sciences & the Law, 19,* 53-80.

Editorial. (2010). Should protocols for observational research be registered? *Lancet, 375*, 348.

Egan, V., Charlesworth, P., Richardson, C., Blair, M., & McMurran, M. (2001). Sensational interests and sensation seeking in mentally disordered offenders. *Personality & Individual Differences, 30*, 995-1007.

*Eher, R., Rettenberger, M., Schilling, F., & Pfafflin, F. (2008). Failure of Static-99 and SORAG to predict relevant reoffense categories in relevant sexual offender subtypes: A prospective study. *Sexual Offender Treatment, 3*, 1-14.

Ekholm, B., Ekholm, A., Adolfsson, R., Vares, M., Osby, U., Sedvall, G. C., et al. (2005). Evaluation of diagnostic procedures in Swedish patients with schizophrenia and related psychoses. *Nordic Journal of Psychiatry, 59*, 457-464.

Eronen, M., Angermeyer, M. C., & Schulze, B. (1998). The psychiatric epidemiology of violent behavior. *Social Psychiatry & Psychiatric Epidemiology, 33*, S13-S23.

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

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191.

Fava, G. A. (2009). An operational proposal for addressing conflict of interest in the psychiatric field. *Journal of Ethics in Mental Health, 4*, S1-S5.

Fazel, S., & Grann, M. (2006). The population impact of severe mental illness on violent crime. *American Journal of Psychiatry, 163,* 1397-1403.

Fazel, S., Gulati, G., Linsell, L., Geddes, J. R., & Grann, M. (2009). Schizophrenia and violence: Systematic review and meta-analysis. *PLoS Medicine*, *6,* e1000120.

Fazel, S., Långström, N., Hjern, A., Grann, M., & Lichtenstein, P. (2009). Schizophrenia, substance abuse, and violent crime. *Journal of the American Medical Association, 301,* 2016-2023.

Fazel, S., Sjöstedt, G., Långström, N., & Grann, M. (2006). Risk factors for criminal recidivism in older sexual offenders. *Sexual Abuse: A Journal of Research & Treatment, 18,* 159-167.

Federal Bureau of Investigation. (2002). *Uniform crime reports for the United States*. Washington, DC: US Government Printing Office.

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research & Practice, 40*, 532-538.

Fleminger, S. (1997). Number needed to detain. *British Journal of Psychiatry, 171*, 287.

*Folino, J., Almiron, M., & Ricci, M. A. (2007). *Factores de riesgo de recidiva violenta en mujeres filicidas*. [Violent recidivism risk factor in filicidal women]. *Vertex*, *18*, 258-267.

*Folino, J., & Castillo, J. L. (2006). *Las facetas de la psicopatia segun la Hare Psychopathy Checklist – Revised y su confiabildad*. [The facets of psychopathy described by the Hare Psychopathy Checklist – Revised and their reliability]. *Vertex, 69,* 325-330.

Forth, A., Kosson, D., & Hare, R. D. (2003). *The Hare Psychopathy Checklist: Youth Version*. New York: Multi-Health Systems.

*Friendship, C., Mann, R. E., & Beech, A. R. (2003). Evaluation of a national prison-based treatment program for sexual offenders in England and Wales. *Journal of Interpersonal Violence, 18,* 744-759.

Fujii, D., Tokioka, A., Lichton, A., & Hishinuma, E. (2005). Ethnic differences in violence risk prediction of psychiatric inpatients using the Historical Clinical Risk Management – 20. *Psychiatric Services, 56*, 711-716.

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

*Gammelgård, M., Koivisto, A. M., Eronen, M., & Kaltiala-Heino, R. (2008). The predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) among institutionalised adolescents. *Journal of Forensic Psychiatry & Psychology, 19,* 352-370.

Geddes, J., Freemantle, N., Harrison, P., & Bebbington, P. (2000). Atypical antipsychotics in the treatment of schizophrenia: Systematic overview and meta-regression analysis. *British Medical Journal, 321*, 1371-1376.

†Gendreau, P., Goggin, C., & Little, T. (1996). *Predicting adult offender recidivism: What works!* (Cat. No. JS4-1/1996-7E). Ottawa, ON: Public Works and Government Services Canada.

Gendreau, P., Goggin, C., & Smith, P. (2000). *Cumulating knowledge: How meta-analysis can serve the needs of correctional clinicians and policy-makers*. Ottawa, ON: Correctional Service of Canada.

†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 & 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.

‡Gerhold, C. K., Browne, K. D., & Beckett, R. (2007). Predicting recidivism in adolescent sexual offenders. *Aggression & Violent Behavior, 12*, 427-438.

*Gibas, A. L., Kropp, P. R., Hart, S. D., & Stewart, L. (2008, July). *Validity of the SARA in a Canadian sample of incarcerated adult males*. Paper presented at the annual conference of the International Association of Forensic Mental Health Services, Vienna, Austria.

Glas, A. S., Lijmer, J. G., Prins, M. H., Bonsel, G. J., & Bossuyt, P. M. M. (2003). The diagnostic odds ratio: A single indicator of test performance. *Journal of Clinical Epidemiology, 56,* 1129-1135.

Gottfredson, D. M., & Snyder, H. N. (2005). *The mathematics of risk classification: Changing data into valid instruments for juvenile courts* (OJJDP Publication No. 209158). Washington, DC: US Department of Justice.

Gough, H. G. (1957). *California Psychological Inventory manual*. Palo Alto, CA: Consulting Psychologists Press.

*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 & Behavior, 27,* 97-114.

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

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

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

Gray, N. S., Taylor, J., & Snowden, R. J. (2010). Predicting violence using structured professional judgment in patients with different mental and behavioral disorders. *Psychiatry Research*. doi:10.1016/j.psychres.2010.10.011

*Gretton, H., & Abramowitz, C. (2002, March). *SAVRY: Contribution of items and scales to clinical risk judgements and criminal outcomes*. Paper presented at the annual conference of the American Psychology-Law Society, Austin, TX.

Gross, A. (2008). History, race, and prediction: Comments on Harcourt's *Against Prediction*. *Law & Social Inquiry, 33*, 233-242.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficacy of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, & Law, 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*, 19-30.

Grubin, D. (1998). *Sex offending against children: Understanding the risk* (Police Research Series Paper 99). London: Home Office.

Guilford, J. P. (1941). A simple scoring weight for test items and its reliability. *Psychometrika, 6*, 367-374.

†Guy, L. (2008). *Performance indicators of the structured professional judgement approach for assessing risk for violence to others: A meta-analytic survey*. Unpublished doctoral dissertation, Simon Fraser University, Burnaby, BC.

†Guy, L., Edens, J. F., Anthony, C., & Douglas, K. S. (2005). Does psychopathy predict institutional misconduct among adults? A meta-analytic investigation. *Journal of Consulting & Clinical Psychology, 73*, 1056-1064.

Haggård-Grann, U. (2005). *Violence among mentally disordered offenders: Risk and protective factors*. Stockholm: Edita Norstedts Tryckeri.

Hakeem, M. (1948). The validity of the Burgess method of parole prediction. *American Journal of Sociology, 53*, 376-386.

Hallfors, D., Vevea, J. L., Iritani, B., Cho H., Khatapoush, S., & Saxe, L. (2002). Truancy, grade point average, and sexual activity: A meta-analysis of risk indicators for youth substance use. *Journal of School health, 72*, 205-211.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148*, 839-843.

Hanson, R. K. (1997). *The development of a brief actuarial scale for sexual offense recidivism* (User report 1997-04). Ottawa, ON: Department of the Solicitor General.

Hanson, R. K. (1998). What do we know about sex offender risk assessment? *Psychology, Public Policy, & Law, 4,* 50-72.

Hanson, R. K., & Bussière, M. T. (1996). *Predictors of sexual offender recidivism: A meta-analysis* (Cat. No. JS4-1/1996-4E). Ottawa, ON: Public Works and Government Services Canada.

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

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

†Hanson, R. K., & Morton-Bourgon, K. (2004). *Predictors of sexual recidivism: An updated meta-analysis* (Cat. No. PS3-1/2004-2E-PDF). Ottawa, ON: Public Works and Government Services Canada.

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

†Hanson, R. K., & Morton-Bourgon, K. (2007). *The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis* (Cat. No. PS4-36/2007E). Ottawa, ON: Public Safety and Emergency Preparedness.

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

Hanson, R. K., & Thornton, D. (1999). *Static-99: Improving actuarial risk assessments for sex offenders* (User Report 99–02). Ottawa, ON: Department of the Solicitor General of Canada.

Hansson, L., Vinding, H. R., Mackeprang, T., Sourander, A., Werdelin, G., Bengtsson-Tops, A., et al. (2001). Comparison of key worker and patient assessment of needs in schizophrenic patients living in the community: A Nordic multicentre study. *Acta Psychiatrica Scandinavica, 103,* 45-51.

Hare, R. D. (1985). *The Psychopathy Checklist.* Vancouver, BC: University of British Columbia.

Hare, R. D. (1991). *The Hare Psychopathy Checklist – Revised*. North Tonawanda, NY: Multi-Health Systems.

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

Harrell, F. E., Lee, K. L., & Mark, D. B. (1996). Multivariate prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine, 15,* 361-387.

Harris, A. J. R., Phenix, A., Hanson, R. K., & Thornton, D. (2003). *Static-99 coding rules: Revised 2003*. Ottawa, ON: Solicitor General Canada.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (2010). Allegiance or fidelity? A clarifying reply. *Clinical Psychology: Science and Practice, 17*, 82-89.

*Harris, G. T., Rice, M. E., Quinsey, V. L., Lalumiere, M. L., Boer, D., & Lang, C. (2003). A multisite comparison of actuarial risk instruments for sex offenders. *Psychological Assessment, 15,* 413-25.

Harris, A. J. R., & Tough, S. (2004). Should actuarial risk assessments be used with sex offenders who are intellectually disabled? *Journal of Applied Research in Intellectual Disabilities, 17*, 235-241.

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

Hart, S. D. (2008). Preventing violence: The role of risk assessment and management. In A. C. Baldry, & F. W. Winkel (Eds.), *Intimate partner violence prevention and intervention* (pp. 7-18). New York: Nova Science Publishers.

Hart, S. D. (2008, July). *The Structured Professional Judgment approach to violence risk assessment: Core principles of SPJ*. Paper presented at the annual conference of the International Association of Forensic Mental Health Services, Vienna, Austria.

Hart, S. D., Cox, D., & Hare, R. D. (1995). *Psychopathy Checklist: Screening Version*. Toronto, ON: Multi-Health Systems.

Hart, S. D., Michie, C., & Cooke, D. J. (2007). Precision of actuarial risk assessment instruments: Evaluating the 'margins of error' of group v. individual predictions of violence. *British Journal of Psychiatry, 49*, S60-S65.

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

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*, 486-504.

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

Heilbrun, K. (2003). Violence risk: From prediction to management. In D. Carson & R. Bull (Eds.), *Handbook of psychology in legal contexts*. Chichester, UK: Wiley.

Heilbrun, K. (2009). *Evaluation for risk in violence in adults*. New York: Oxford University Press.

Heilbrun, K., Ogloff, J. R. P., & Picarello, K. (1999). Dangerous offender statutes in the United States and Canada: Implications for risk assessment. *International Journal of Law & Psychiatry, 22*, 393-415.

Heinrichs, R. W., & Sam, E. P. (2010). Schizophrenia and crime: How predictable are charges, convictions and violence? *International Journal of Mental Health & Addiction*, doi:10.1007/s11469-010-9308-z.

Helmus, L. (2008). *Annotated bibliography of Static-99 replications*. Retrieved March 9, 2010, from http://www.static99.org/pdfdocs/static-99annotatedbibliography.pdf

Helmus, L. (2009). *Re-norming Static-99 recidivism estimates: Exploring base rate variability across sex offender samples*. Unpublished Master's thesis, Carleton University, Ottawa, ON.

*Helmus, L., & Hanson, R. K. (2007). Predictive validity of the Static-99 and Static-2002 for sex offenders on community supervision. *Sexual Offender Treatment, 2*, 1-14.

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

Higgins, J. (2008). Heterogeneity in meta-analysis should be expected and appropriately identified. *International Journal of Epidemiology*, *37*, 1158-1160.

Higgins, J., Deeks, J., & Altman, D. G. (2008). Special topics in statistics. In J. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions 5.0.0*. London: Wiley.

Higgins, J., & Green, S. (2006). Analysing and presenting results. In J. Higgins, & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions 4.2.6.* Chichester, UK: John Wiley & Sons.

Higgins, J., & Thompson, S. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21,* 1539-1558.

Higgins, J., Thompson, S., Deeks, J., & Altman, D. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327,* 557-560.

Higgins, N., Watts, D., Bindman, J., Slade, M., & Thornicroft, G. (2005). Assessing violence risk in general adult psychiatry. *Psychiatric Bulletin, 29*, 131-133.

*Hill, A., Habermann, N., Klusmann, D., Berner, W., & Briken, P. (2008). Criminal recidivism in sexual homicide perpetrators. *International Journal of Offender Therapy & Comparative Criminology, 52,* 5-20.

Hodges, J. L., & Lehmann, E. L. (1963). Estimates of location based on rank tests. *Annals of Mathematical Statistics, 34*, 598-611.

Hodgins, S. (1992). Mental disorder, intellectual deficiency and crime: Evidence from a birth cohort. *Archives of General Psychiatry, 49*, 476-483.

Hoffman, P. B. (1983). Screening for risk: A revised salient factor score (SFS 81). *Journal of Criminal Justice, 11*, 539-547.

Hoge, R. D., & Andrews, D. A. (2002). *The Youth Level of Service/Case Management Inventory manual and scoring key*. Toronto, ON: Multi-Health Systems.

*Hollin, C. R., & Palmer, E. J. (2006). The Level of Service Inventory – Revised profile of English prisoners: Risk and reconviction analysis. *Criminal Justice & Behavior, 33,* 347-366.

‡Holtfreter, K., & Cupp, R. (2007). Gender and risk assessment: The empirical status of the LSI-R for women. *Journal of Contemporary Criminal Justice, 23*, 363-382.

Honest, H., & Khan, K. S. (2002). Reporting of measures of accuracy in systematic reviews of diagnostic literature. *BMC Health Services Research, 2*, 1-4.

Hoptman, M. J., Yates, K. F., Patalinjug, M. B., Wack, R. C., & Convit, A. (1999). Clinical prediction of assaultive behavior among male psychiatric patients at a maximum-security forensic facility. *Psychiatric Services, 50,* 1461–1466.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: *Q* statistic or I^2 index? *Psychological Methods, 11*, 193-206.

Jaeger, J., Berns, S. M., & Cozor, P. (2003). The multidimensional scale of independent functioning: A new instrument for measuring functional disability in psychiatric populations. *Schizophrenia Bulletin, 29*, 153-168.

Janus, E. (2004). Sexually violent predator laws: Psychiatry in service to a morally dubious enterprise. *Lancet, 3664*, 50-51.

Kang, M., Pang, Y., Li, J. Y., Liu, L. H., & Liu, X. T. (2010). Accuracy evaluation of mammogarphy in the breast cancer screening in Asian women: A community-based follow-up study and meta analysis. *Chinese Journal of Oncology, 32*, 212-216.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*, 119-127.

*Kelly, C. E., & Welsh, W. N. (2008). The predictive validity of the Level of Service Inventory – Revised for drug-involved offenders. *Criminal Justice & Behavior, 35,* 819-831.

Kemshall, H. (1996). *A review of research on the assessment and management of risk and dangerousness: Implications for policy and practice in the probation service*. Birmingham: Home Office Research and Statistics Directorate.

Kemshall, H. (2001). *Risk assessment and management of known sexual and violent offenders: A review of current issues*. London: Home Office.

Khiroya, R., Weaver, T., & Maden, T. (2009). Use and perceived utility of structured violence risk assessments in English medium secure forensic units. *Psychiatrist, 33,* 129-132.

Klin, A., & Lemish, D. (2008). Mental disorders stigma in the media: Review of studies on production, content, and influences. *Journal of Health Communication, 13*, 434-449.

*Kloezeman, K. C. (2004). *Violent behaviour on inpatient psychiatric units: The HCR-20 violence risk assessment scheme*. Unpublished Master's thesis, University of Hawaii, Manoa, HI.

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

Krimsky, S., & Rothenberg, L. S. (2001). Conflict of interest policies in science and medical journals: Editorial practices and author disclosures. *Science & Engineering Ethics, 7*, 205-218.

*Kroner, C., Stadtland, C., Eidt, M., & Nedopil, N. (2007). The validity of the Violence Risk Appraisal Guide (VRAG) in predicting criminal recidivism. *Criminal Behaviour & Mental Health, 17,* 89-100.

*Kropp, P. R., & Hart, S. D. (2000). The Spousal Assault Risk Assessment (SARA) guide: Reliability and validity in adult male offenders. *Law & Human Behavior, 24,* 101-118.

Kropp, P. R., Hart, S. D., Webster, C. W., & Eaves, D. (1994). *Manual for the Spousal Assault Risk Assessment guide.* Vancouver, BC: British Columbia Institute on Family Violence.

Kropp, P. R., Hart, S. D., Webster, C. D., & Eaves, D. (1995) *Manual for the Spousal Assault Risk Assessment guide* (2nd ed.). Vancouver, BC: British Columbia Institute on Family Violence.

Kropp, P. R., Hart, S. D., Webster, C. D., & Eaves, D. (1999). *Spousal Assault Risk Assessment guide (SARA)*. Toronto, ON: Multi-Health Systems.

‡Kumar, S., & Simpson, A. (2005). Application of risk assessment for violence methods to general adult psychiatry: A selective literature review. *Australian & New Zealand Journal of Psychiatry, 39*, 328-335.

La Fond, J. Q. (2008). Sexually violent predator laws and the liberal state: An ominous threat to individual liberty. *International Journal of Law & Psychiatry, 31*, 158-171.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159-174.

Långström, N. (2004). Accuracy of actuarial procedures for assessment of sexual offender recidivism risk may vary across ethnicity. *Sexual Abuse: A Journal of Research & Treatment, 16,* 107-120.

*Langton, C. M. (2003). *Contrasting approaches to risk assessment with adult male sexual offenders: An evaluation of recidivism prediction schemes and the utility of supplementary clinical information for enhancing predictive accuracy*. Unpublished doctoral dissertation, University of Toronto, Toronto, ON.

*Langton, C. M., Barbaree, H. E., Seto, M. C., Peacock, E. J., Harkins, L., & Hansen, K. T. (2007). Actuarial assessment of risk for reoffense among adult sex offenders: Evaluating the predictive accuracy of the Static-2002 and five other instruments. *Criminal Justice & Behavior, 34,* 37-59.

Large, M. M., Ryan, C. J., & Nielssen, O. B. (2010). Helpful and unhelpful risk assessment practices. *Psychiatric Services, 61*, 530.

Large, M. M., Ryan, C. J., Singh, S. P., Paton, M. B., & Nielssen, O. B. (2011). The predictive value of risk categorization in schizophrenia. *Harvard Review of Psychiatry, 19*, 25-33.

Latessa, E. J., & Lovins, B. (2010). The role of offender risk assessment: A policy maker guide. *Victims & Offenders, 5*, 203-219.

Lawson, W. B., Yesavage, J. A., Werner, P. A. (1984). Race, violence, and psychopathology. *Journal of Clinical Psychiatry, 45,* 294-297.

†Leistico, A., Salekin, R., DeCoster, J., & Rogers, R. (2008). A large-scale meta-analysis relating the Hare measures of psychopathy to antisocial conduct. *Law & Human Behavior, 32*, 28-45.

Leucht, S., Barnes, T. R. E., Kissling, W., Engel, R. R., Correll, C., & Kane, J. M. (2003). Relapse prevention in schizophrenia with new-generation antipsychotics: A systematic review and exploratory meta-analysis of randomized, controlled trials. *American Journal of Psychiatry, 160*, 1209-1222.

Levey, S., & Howells, K. (1994). Accounting for the fear of schizophrenia. *Journal of Community & Applied Social Psychology, 4*, 313-328.

Levinson, R., & Ramsay, G. (1979). Dangerousness, stress and mental health evaluations. *Journal of Health & Social Behavior, 20*, 178-187.

Lexchin, J., Bero, L. A., Djulbegovic, B., & Clark, O. (2003). Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *British Medical Journal, 326*, 1167-1170.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., et al. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: Explanation and elaboration. *British Medical Journal, 339*, B2700.

Lidz, C. W., Mulvey, E. P., & Gardner, W. (1993). The accuracy of predictions of violence to others. *Journal of the American Medical Association, 269,* 1007–1011.

Lindqvist, P., & Allebeck, P. (1990). Schizophrenia and crime: A longitudinal follow-up of 644 schizophrenics in Stockholm. *British Journal of Psychiatry, 157*, 345-350.

*Lodewijks, H. P. B., de Ruiter, C., & Doreleijers, T. A. H. (2008). Gender differences in violent outcome and risk assessment in adolescent offenders after residential treatment. *International Journal of Forensic Mental Health, 7,* 105-141.

Lodewijks, H. P. B., de Ruiter, C., & Doreleijers, T. A. H. (2010). The impact of protective factors in desistance from violence reoffending: A study in three samples of adolescent offenders. *Journal of Interpersonal Violence, 25*, 568-587.

*Lodewijks, H. P. B., Doreleijers, T. A. H., & de Ruiter, C. (2008). SAVRY risk assessment in violent Dutch adolescents: Relation to sentencing and recidivism. *Criminal Justice & Behavior, 35,* 696-709.

*Lodewijks, H. P. B., Doreleijers, T. A. H., de Ruiter, C., & Borum, R. (2008). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) during residential treatment. *International Journal of Law & Psychiatry, 31,* 263-271.

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

Loza, W., Dhaliwal, G. K., Kroner, D. G., & Loza-Fanous, A. (2000). Reliability and concurrent validity of the Self-Appraisal Questionnaire (SAQ): A tool for assessing violent and non-violent recidivism. *Criminal Justice & Behavior, 27*, 356-374.

‡Lund, C. (2000). Predictors of sexual recidivism: Did meta-analysis clarify the role of relevance of denial? *Sexual Abuse: A Journal of Research & Treatment, 12*, 275-287.

Luty, J., Fekadu, D., & Dhandayudham, A. (2006). Understanding the term "schizophrenia" by the British public. *World Psychiatry, 5*, 177-178.

Lyon, D. R., Hart, S. D., & Webster, C. D. (2001). Violence and risk assessment. In R. A. Schuller, & J. R. P. Ogloff (Eds.), *Introduction to psychology and law*: *Canadian perspectives*. Toronto, ON: University of Toronto Press.

Maden, A. (2001). Practical application of structured risk assessment. *British Journal of Psychiatry, 178*, 479.

Maden, A. (2003). Standardised risk assessment: Why all the fuss? *Psychiatrist, 25,* 129-131.

Maj, M. (2008). Non-financial conflicts of interests in psychiatric research and practice. *British Journal of Psychiatry, 193*, 91-92.

Mak, W. W., & Wu, C. F. (2006). Cognitive insight and causal attribution in the development of self-stigma among individuals with schizophrenia. *Psychiatric Services, 57,* 1800-1802.

Mantel, N., & Haenzel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*, 719-748.

†McCann, K. (2006). *A meta-analysis of the predictors of sexual recidivism in juvenile sexual offenders*. Unpublished Master's thesis, Simon Fraser University, Burnaby, BC.

McCann, K., & Lussier, P. (2008). Antisociality, sexual deviance, and sexual reoffending in juvenile sex offenders: A meta-analytical investigation. *Youth Violence & Juvenile Justice, 6*, 363-385.

*McEachran, A. (2001). *The predictive validity of the PCL:YV and the SAVRY in a population of adolescent offenders*. Unpublished doctoral dissertation, Simon Fraser University, Burnaby, BC.

McGuffin, P., Farmer, A., & Harvey, I. (1991). A polydiagnostic application of operational criteria in psychotic illness: Development and reliability of the OPCRIT system. *Archives of General Psychiatry, 48*, 764–770.

McMurran, M., Blair, M., & Egan, V. (2002). An investigation of the correlations between aggression, impulsiveness, social problem-solving, and alcohol use. *Aggressive Behavior, 28*, 439-445.

McMurran, M., & Theodosi, E. (2007). Is treatment non-completion associated with increased reconviction over no treatment? *Psychology, Crime & Law, 13*, 333-343.

McNiel, D. E., & Binder, R. L. (1995). Correlates of accuracy in the assessment of psychiatric inpatients' risk of violence. *American Journal of Psychiatry, 152,* 901–906.

MedCalc Software. (2010). *MedCalc: Version 11.3.8.0*. Mariakerke, Belgium: MedCalc Software.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and review of the evidence*. Minneapolis: University of Minnesota Press.

Menard, S. (1995). *Applied logistic regression analysis*. Thousand Oaks, CA: Sage.

Mental Health Centre Penetanguishene. (2009). *Replications of the Violence Risk Appraisal Guide or Sex Offender Risk Appraisal Guide in assessing violence risk*. Retrieved March 9, 2010, from http://www.mhcp.on.ca/Site_Published/internet/-SiteContent.aspx

Menzies, R. J., Webster, C. D., & Sepejak, D. S. (1985). The dimensions of dangerousness: Evaluating the accuracy of psychometric predictions of violence among forensic patients. *Law & Human Behavior, 9*, 35-56.

Mercado, C. C., & Ogloff, J. R. P. (2007). Risk and the preventive detention of sex offenders in Australia and the United States. *International Journal of Law & Psychiatry, 30*, 49-59.

*Meyers, J. R., & Schmidt, F. (2008). Predictive validity of the Structured Assessment of Violence Risk in Youth (SAVRY) with juvenile offenders. *Criminal Justice & Behavior, 35,* 344-355.

*Mills, J. F., Jones, M. N., & Kroner, D. G. (2005). An examination of the generalizability of the LSI-R and VRAG probability bins. *Criminal Justice & Behavior, 32,* 565-585.

*Mills, J. F., & Kroner, D. G. (2006). The effect of discordance among violence and general recidivism risk estimates on predictive accuracy. *Criminal Behaviour & Mental Health, 16,* 155-166.

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

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie, D., & Stroup, D. F. (1999). Improving the quality of reports of meta-analyses of randomised controlled trials: The QUOROM statement. *Lancet, 354*, 1896-1900.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Medicine, 6,* e1000097.

Moher, D., Tetzlaff, J., Triccol, A. C., Sampson, M., & Altman, D. G. (2007). Epidemiology and reporting characteristics of systematic reviews. *PLoS Medicine, 4,* 447-455.

Monahan, J. (1981). *The clinical prediction of violent behavior*. Rockville, MD: US Department of Health and Human Services.

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

Monahan, J., Steadman, H., Appelbaum, P., Robbins, P., Mulvey, E., Silver, E., et al. (2000). Developing a clinically useful actuarial tool for assessing violence risk. *British Journal of Psychiatry, 176*, 312-319.

Monahan, J., Steadman, H., Silver, E., Appelbaum, P., Robbins, P., Mulvey, E., et al. (2001). *Rethinking risk assessment: The MacArthur Study of Mental Disorder and Violence.* New York: Oxford University Press.

Morey, L. C. (1991). *Personality Assessment Inventory: A professional manual*. Odessa, FL: Psychological Assessment Resources.

Morgenstern, H. (1982). Uses of ecologic analysis in epidemiologic research. *American Journal of Public Health, 72*, 1336-1344.

*Morrissey, C., Hogue, T., Mooney, P., Allen, C., Johnston, S., Hollin, C., et al. (2007). Predictive validity of the PCL-R in offenders with intellectual disability

in a high secure hospital setting: Institutional aggression. *Journal of Forensic Psychiatry & Psychology, 18,* 1-15.

Moses, L. E., Littenberg, B., & Shapiro, D. (1993). Combining independent studies of a diagnostic test into a summary ROC curve: Data-analytical approaches and some additional considerations. *Statistics in Medicine, 12,* 1293-1316.

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

Murray, J., & Thomson, M. (2010). Clinical judgement in violence risk assessment. *Europe's Journal of Psychology, 1*, 128-149.

Murrie, D. C., Cornell, D., & McCoy, W. K. (2005). Psychopathy, conduct disorder, and stigma: Does diagnostic labelling influence juvenile probation officer recommendations? *Law & Human Behavior, 29*, 323-342.

National Alliance on Mental Illness. (2008). *Schizophrenia: Public attitudes, personal needs*. Arlington, VA: National Alliance on Mental Illness.

National Board of Health and Welfare. (2009). *National Cause of Death Register*. Retrieved February 11, 2009, http://www.socialstyrelsen.se/en/Statistics/statsbysubject/The_Cause_of_Death_Register.htm

National Institute for Health and Clinical Excellence. (2009). *Schizophrenia: Core interventions in the treatment and management of schizophrenia in primary and secondary care*. London: National Institute for Health and Clinical Excellence.

Nelemans, P. J., Leiner, T., de Vet, H. C. W., & van Engelshoven, J. M. A. (2000). Peripheral arterial disease: Meta-analysis of the diagnostic performance of MR angiography. *Radiology, 217*, 105-114.

*Nicholls, T. L., Ogloff, J. R. P., & Ledwidge, B. (2007). *Is the profound distrust of unbridled clinical opinion in the violence risk assessment field unfounded?* Poster presented at the 4th Annual Forensic Psychiatry Conference, Victoria, BC.

†Nikolova, N. L., Collins, M. J., Guy, L. S., Lavoie, J. A. A., Reeves, K. A., Wilson, et al. (2006, March). *HCR-20 violence risk assessment scheme: Quantitative synthesis of its application, reliability, and validity*. Poster session presented at the annual conference of the American Psychology-Law Society, St. Petersburg, FL.

Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

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

Nuffield, J. (1982). *Parole decision making in Canada: Research towards decision guidelines*. Ottawa, ON: Ministry of Supply and Services Canada.

Olofsson, B. (2010). *Study Size 2.0.4*. Frolunda, Sweden: CreoStat HB.

†Olver, M., Stockdale, K., & Wormith, J. (2009). Risk assessment with young offenders: A meta-analysis of three assessment measures. *Criminal Justice & Behavior, 36*, 329-353.

Orange County Probation Department. (1988). *Orange County California Probation Department juvenile risk/needs assessment instruments*. Anaheim, CA: Orange County Probation Department.

Otto, R. K., & Heilbrun, K. (2002). The practice of forensic psychology. *American Psychologist, 57*, 5-18.

Oxman, A. D., & Guyatt, G. H. (1988). Guidelines for reading literature reviews. *Canadian Medical Association Journal, 138*, 697-703.

Palma, S., & Delgado-Rodriguez, M. (2005). Assessment of publication bias in meta-analyses of cardiovascular diseases. *Journal of Epidemiological & Community Health, 59*, 864-869.

Parton, N. (1996). *Social theory, social change and social work*. London: Routledge.

Patrick, C. J. (2006). *Handbook of psychopathy*. New York: Guilford Press.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research*. Fort Worth, TX: Harcourt Brace.

Perlis, R. H., Perlis, C. S., Wu, Y., Hwang, C., Joseph, M., & Nierenberg, A. A. (2005). Industry sponsorship and financial conflict of interest in the reporting of clinical trials in psychiatry. *American Journal of Psychiatry, 162*, 1957-1960.

Pescosolido, B. A., & Boyer, C. A. (1999). How do people come to use mental health services? Current knowledge and changing perspectives. In A. Horwitz & T. Scheid (Eds.), *The sociology of mental illness*. New York: Cambridge University Press.

Pescosolido, B. A., Martin, J. K., Long, J. S., Medina, T. R., Phelan, J. C., & Link, B. G. (2010). "A disease like any other"? A decade of chance in public reactions to schizophrenia, depression, and alcohol dependence. *American Journal of Psychiatry, 167*, 1321-1330.

Peterson, D. R., Quay, H. C., & Cameron, G. R. (1959). Personality and background factors in juvenile delinquency as inferred from questionnaire responses. *Journal of Consulting Psychology, 23*, 395-399.

Petitti, D. B. (2001). Approaches to heterogeneity in meta-analysis. *Statistical Medicine, 20*, 3625-3633.

*Pham, T. H., Chevrier, I., Nioche, A., Ducro, C., & Reveillere, C. (2005). *Psychopathie, evaluation du risque, prise en charge*. [Psychopathy, risk assessment, and support]. *Annales Medico Psychologiques, 163,* 878-881.

*Pham, T. H., Ducro, C., Marghem, B., & Reveillere, C. (2005). *Evaluation du risque de recidive au sein d'une population de delinquants incarceres ou internes en Belgique francophone*. [Prediction of recidivism among prison inmates and forensic patients in Belgium]. *Annales Medico Psychologiques, 163,* 842-845.

Pilowsky, D. J., Keyes, K. M., & Hasin, D. S. (2009). Adverse childhood events and lifetime alcohol dependence. *American Journal of Public Health, 99*, 258-263.

*Polvi, N. H. (1999). *The prediction of violence in pre-trial forensic patients: The relative efficacy of statistical versus clinical predictions of dangerousness*. Unpublished doctoral dissertation, Simon Fraser University, Burnaby, BC.

Priebe, S., Frottier, P., Gaddini, A., Kilian, R., Lauber, C., Martinez-Leal, R., et al. (2008). Mental health care institutions in nine European countries, 2002-2006. *Psychiatric Services, 59*, 570-573.

PRISMA Group. (2010). PRISMA endorsers. In *PRIMSA: Transparent reporting of systematic reviews and meta-analyses*. Retrieved March 21, 2011, from http://www.prisma-statement.org/endorsers.htm

*Public Safety and Solicitor General (2004). *Spousal Assault Risk Assessment (SARA) and Community Risk/Needs Assessment (CRNA): Predictive efficacy and interrelationships*. Victoria, BC: Public Safety and Solicitor General.

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

Quinsey, V. L., Harris, G. T., Rice, M. E., & Cormier, C. A. (2006). *Violent offenders: Appraising and managing risk* (2nd ed.). Washington, DC: American Psychological Association.

*Ramirez, M. P., Illescas, S. R., Garcia, M. M., Forero, C. G., & Pueyo, A. A. (2008). *Predicción de riesgo de reincidencia en agresores sexuales*. [Predicting risk of recidivism in sexual offenders]. *Psicothema, 20,* 205-210.

Ransohoff, D. F., & Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine, 299*, 926-930.

Räsänen, P., Tiihonen, J., Isohanni, M., Rantakallio, P., Lehtonen, J., & Moring, J. (1998). Schizophrenia, alcohol abuse, and violent behavior: A 26-year followup study of an unselected birth cohort. *Schizophrenia Bulletin, 24,* 437-441.

†Redlak, A. (2003). *An exploratory meta-analysis of the predictor variables of juvenile sex offenders who sexually recidivate*. Unpublished doctoral dissertation, California School of Professional Psychology, Fresno, CA.

*Reeves, K. A., Kropp, P. R., & Cairns, K. (2008). *An independent validation study of the SARA*. Paper presented at the Annual Conference of the International Association of Forensic Mental Health Services, Vienna, Austria.

*Rettenberger, M., & Eher, R. (2007). Predicting reoffense in sexual offender subtypes: A prospective validation study of the German version of the Sexual Offender Risk Appraisal Guide (SORAG). *Sexual Offender Treatment, 2*, 1-12.

Rettenberger, M., Hucker, S. J., Boer, D. P., & Eher, R. (2009). The reliability and validity of the Sexual Violence Risk – 20 (SVR-20): An international review. *Sexual Offender Treatment, 4*, 1-14.

Reutfors, J., Bahmanyar, S., Jönsson, E. G., Ekbom, A., Nordström, P., Brandt, L., et al. (2010). Diagnostic profile and suicide risk in schizophrenia spectrum disorder. *Schizophrenia Research, 123*, 251-256.

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

*Rice, M. E., & Harris, G. T. (2002). Men who molest their sexually immature daughters: Is a special examination required? *Journal of Abnormal Psychology, 111,* 329-339.

Roaldset, J. O., Hartvig, P., & Bjørkly, S. (2010). V-RISK-10: Validation of a screen for risk of violence after discharge from acute psychiatry. *European Psychiatry,* doi:10.1016/j.eurpsy.2010.04.002.

Rogers, R. (2000). The uncritical acceptance of risk assessment in forensic practice. *Law & Human Behavior, 24*, 595-605.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638-641.

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis*. New York: Sage.

Rossi, R. J. (2010). *Applied biostatistics for the health sciences*. Hoboken, NJ: Wiley.

Ruddy, R., & House, A. (2005). Meta-review of high-quality systematic reviews of interventions in key areas of liaison psychiatry. *British Journal of Psychiatry, 187*, 109-120.

Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods, 13*, 19-30.

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

Schauer, F. (2003). *Profiles, probabilities, and stereotypes.* Cambridge, MA: Harvard University Press.

†Schwalbe, C. (2007). Risk assessment for juvenile justice: A meta-analysis. *Law & Human Behavior, 31*, 449-462.

†Schwalbe, C. (2008). A meta-analysis of juvenile justice risk assessment instruments: Predictive validity by gender. *Criminal Justice & Behavior, 35*, 1367-1381.

Schwalbe, C., Fraser, M. W., Day, S. H., & Arnold, E. M. (2004). North Carolina Assessment of Risk (NCAR): Reliability and predictive validity with juvenile offenders. *Journal of Offender Rehabilitation, 40,* 1-22.

Seltzer, M. L. (1971). The Michigan Alcoholism Screening Test: The quest for a new diagnostic instrument. *American Journal of Psychiatry, 127*, 1653-1658.

*Serin, R. C., Mailloux, D. L., & Malcolm, P. B. (2001). Psychopathy, deviant sexual arousal, and recidivism among sexual offenders. *Journal of Interpersonal Violence, 16,* 234-246.

Seto, M. (2005). Is more better? Combining actuarial risk scales to predict recidivism among adult sex offenders. *Psychological Assessment, 17,* 156-167.

*Seto, M., & Barbaree, H. E. (1999). Psychopathy, treatment behavior, and sex offender recidivism. *Journal of Interpersonal Violence, 14,* 1235-1248.

Sharpe, D. (1997). Of apples and oranges, file drawers and garbage: Why validity issues in meta-analysis will not go away. *Clinical Psychology Review, 17*, 881-901.

Shea, B. J., Grimshaw, J. M., Wells, G. A., Boers, M., Andersson, N., Hamel, C., et al. (2007). Development of AMSTAR: A measurement tool to assess the methodological quality of systematic reviews. *BMC Medical Research Methodology, 7*, doi:10.1186/1471-2288-7-10.

Sheldrick, C. (1999). Practitioner review: The assessment and management of risk in adolescents. *Journal of Child Psychology & Psychiatry, 40*, 507-518.

Siegel, J. M. (1986). The Multidimensional Anger Inventory. *Journal of Personality & Social Psychology, 51*, 191-200.

Simon, J. (2005). Reversal of fortune: The resurgence of individual risk assessment in criminal justice. *Annual Review of Law & Social Sciences, 1*, 397-421.

*Simourd, D. (2006). *Validation of risk/needs assessments in the Pennsylvania Department of Corrections*. Lower Allen, PA: Pennsylvania Department of Corrections.

Sjöstedt, G., & Grann, M. (2002). Risk assessment: What is being predicted by actuarial prediction instruments? *International Journal of Forensic Mental Health, 1*, 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 & Human Behavior, 25,* 629-645.

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

†Skeem, J., Edens, J. F., Camp, J., & Colwell, L. H. (2004). Are there ethnic differences in levels of psychopathy? A meta-analysis. *Law & Human Behavior, 28*, 505-527.

†Smith, P., Cullen, F., & Latessa, E. (2009). Can 14,737 women be wrong? A meta-analysis of the LSI-R and recidivism for female offenders. *Criminology & Public Policy, 8*, 183-208.

Smith, J., & Lanyon, R. I. (1968). Prediction of juvenile probation violators. *Journal of Consulting & Clinical Psychology, 32*, 54-58.

Smits, N. (2010). A note on Youden's *J* and its cost ratio. *BMC Medical Research Methodology*, *10*, 89-93.

*Snowden, R. J., Gray, N. S., Taylor, J., & MacCulloch, M. J. (2007). Actuarial prediction of violent recidivism in mentally disordered offenders. *Psychological Medicine, 37,* 1539-1549.

*Soothill, K., Harman, J., Francis, B., & Kirby, S. (2005). Identifying future repeat danger from sexual offenders against children: A focus on those convicted and those strongly suspected of such crime. *Journal of Forensic Psychiatry & Psychology, 16,* 225-247.

Soyka, M. (2002). Aggression in schizophrenia: Assessment and prevalence. *British Journal of Psychiatry, 180*, 278-279.

Soyka, M., Graz, C., Bottlender, R., Dirschedl, P., & Schoech, H. (2007). Clinical correlates of later violence and criminal offences in schizophrenia. *Schizophrenia Research, 94*, 89-98.

Spengler, P. M., & Strohmer, D. C. (2001, August). *Empirical analyses of a scientist-practitioner model of assessment*. Paper presented at the annual conference of the American Psychological Association, San Francisco, CA.

SPSS Inc. (2009). *SPSS for Windows: Release 17.0.1*. Chicago: SPSS Inc.

*Sreenivasan, S., Garrick, T., Norris, R., Cusworth-Walker, S., Weinberger, L. E., Essres, G., et al. (2007). Predicting the likelihood of future sexual recidivism:

Pilot study findings from a California sex offender risk project and cross-validation of the Static-99. *Journal of the American Academy of Psychiatry & the Law*, *35*, 454-468.

Sreenivasan, S., Kirkish, P., Garrick, T., Weinberger, L. E., & Phenix, A. (2000). Actuarial risk assessment models: A review of critical issues related to violence and sex-offender recidivism assessments. *Journal of the American Academy of Psychiatry & the Law, 28,* 438-448.

*Stadtland, C., Hollweg, M., Kleindienst, N., Dietl, J., Reich, U., & Nedopil, N. (2006). *Rueckfallprognosen bei Sexualstraftaetern - Vergleich der praediktiven Validitaet von Prognoseinstrumenten*. [Predictions of recidivism in sexual offenders: Comparison of the predictive validity of assessment tools]. *Nervenarzt, 77,* 587-595.

STARD Group. (2008). STARD news. In *STARD Statement: Standards for the Reporting of Diagnostic Accuracy Studies*. Retrieved March 21, 2011, from http://www.stard-statement.org/

StataCorp. (2007). *Stata statistical software: Release 10.1*. College Station, TX: StataCorp LP.

Statistics Sweden. (2010). *Multi-Generation Register 2009: A description of content and quality*. Örebro, Sweden: Statistics Sweden.

Steadman, H. J. (1977). A new look at recidivism among Patuxent inmates. *Bulletin of the American Academy of Law & Psychiatry, 5*, 200-209.

Steadman, H. J., & Cocozza, J. J. (1974). Some refinements in the measurement and prediction of dangerous behavior. *American Journal of Psychiatry, 131*, 1012-1014.

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., et al. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *Journal of the American Medical Association, 283*, 2008-2012.

Stuart, H. L., & Arboleda-Florez, J. E. (2001). A public health perspective on violent offenses among persons with mental illness. *Psychiatric Services, 52,* 654-659.

Sturidsson, K., Haggård-Grann, U., Lotterberg, 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*, 23-36.

Sullivan, G., Wells, K. B., Morgenstern, H., & Leake, B. (1995). Identifying modifiable risk factors for rehospitalization: A case–control study of seriously mentally ill persons in Mississippi. *American Journal of Psychiatry, 152,* 1749-1756

Swanson, J. W., Swartz, M. S., & Elbogen, E. B. (2004). Effectiveness of atypical antipsychotic medications in reducing violent behavior among persons with schizophrenia in community-based treatment. *Schizophrenia Bulletin, 30,* 3-20.

Swanson, J. W., Swartz, M. S., Van Dorn, R. A., Elbogen, E. B., Wagner, H. R., Rosenheck, R. A., et al. (2006). A national study of violent behavior in persons with schizophrenia. *Archives of General Psychiatry, 63*, 490-499.

Swedish Council on Health Technology Assessment. (2005). *Riskbedömningar inom psykiatrin. Kan våld i samhället förutsägas?* [Risk assessments in psychiatry. Is it possible to predict community violence?]. Stockholm: Swedish Council on Health Technology Assessment.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest: A Journal of the American Psychological Society, 1*, 1-26.

Szmukler, G. (2001). Violence risk prediction in practice. *British Journal of Psychiatry, 178*, 84-85.

Szmukler, G., & Holloway, F. (2000). Reform of the Mental Health Act: Health or safety? *British Journal of Psychiatry, 177*, 196-200.

Tabachnick, B., & Fidell, L. (2001). *Using multivariate statistics*. Needham Heights, MA: Allyn and Bacon.

Tape, T. (2006). The area under the ROC curve. In *Interpreting diagnostic tests*. Retrieved March 7, 2011, from http://gim.unmc.edu/dxtests/ROC3.htm

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

*Thomas, D. J. (2001). *Identifying the sexual serial killer: A comparative study of sexual serial killers, serial rapists, and sexual offenders*. Unpublished doctoral dissertation, Union Institute Graduate College.

Thompson, S. G., & Higgins, J. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, *21*, 1559-1573.

Thornberry, T. P., & Jacoby, J. E. (1979). *The criminally insane: A community follow-up of mentally ill offenders*. Chicago: University of Chicago Press.

*Thornton, D. (2002). Constructing and testing a framework for dynamic risk assessment. *Sexual Abuse: A Journal of Research & Treatment, 14,* 139-153.

Tiihonen, J., Isohanni, M., Räsänen, P., Koiranen, M., & Moring, J. (1997). Specific major mental disorders and criminality: A 26-year prospective study of the 1966 northern Finland birth cohort. *American Journal of Psychiatry, 154*, 840-845.

Townsend, J. K. (1967). The relation between Rorschach signs of aggression and behavioral aggression in emotionally disturbed boys. *Journal of Projective Techniques & Personality Assessment, 31*, 13-21.

‡Turgut, T., Lagace, D., Izmir, M., & Dursun, S. (2006). Assessment of violence and aggression in psychiatric settings: Descriptive approaches. *Klinik Psikofarmakoloji Bülteni, 16*, 179-194.

Tyrer, P., Duggan, C., Cooper, S., Crawford, M., Seivewright, H., Rutter, D., et al. (2010). The successes and failures of the DSPD experiment: The assessment and management of severe personality disorder. *Medicine Science & the Law, 50*, 95-99.

Uppal, G., & McMurran, M. (2009). Recorded incidents in a high-secure hospital: A descriptive analysis. *Criminal Behaviour & Mental Health, 19*, 265-276.

Vance, J. E., Bowen, N. K., Fernandez, G., & Thompson, S. (2002). Risk and protective factors as predictors of outcome in adolescents with psychiatric disorder and aggression. *Journal of the American Academy of Child & Adolescent Psychiatry*, *41*, 36-43.

Vandenbroucke, J. P., von Elm, E., Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., et al. (2007). Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and elaboration. *PLoS Medicine, 4*, e297.

Vauth, R., Kleim, B., Wirtz, M., & Corrigan, P. W. (2007). Self-efficacy and empowerment as outcomes of self-stigmatizing and coping in schizophrenia. *Psychiatry Research, 150,* 71-80.

Vess, J. (2008). Sex offender risk assessment: Consideration of human rights in community protection legislation. *Legal & Criminological Psychology, 13,* 245-256.

Viljoen, J. L., McLachlan, K., & Vincent, G. M. (2010). Assessing violence risk and psychopathy in juvenile and adult offenders: A survey of clinical practices. *Assessment, 17*, 377-395.

Vogel, D. L., Wade, N. G., & Haake, S. (2006). Measuring the self-stigma associated with seeking psychological help. *Journal of Counselling Psychology, 53,* 325-337.

Volavka, J., Laska, E., Baker, S., Meisner, M., Czobor, P., & Krivelevich, I. (1997). History of violent behaviour and schizophrenia in different cultures. Analyses based on the WHO study on determinants of outcome of severe mental disorders. *British Journal of Psychiatry, 171,* 9-14.

Vranova, J., Horak, J., Kratka, K., Hendrichova, M., & Kovarikova, K. (2009). ROC analysis and the use of cost-benefit analysis for determination of the optimal cut point. *Časopis Lékařů Českých, 148,* 410-415.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83,* 213-217.

*Walkington, Z., O'Keeffe, C., & Thomas, S. (2006). *Predicting violence recidivism in violent juveniles: A UK trial*. London: HM Prison Service.

Walmsley, R. (2009). *World prison population list*. London: King's College London International Centre for Prison Studies.

Walsh, E., Buchanan, A., & Fahy, T. (2002). Violence and schizophrenia: Examining the evidence. *British Journal of Psychiatry, 180,* 490-495.

Walsh, E., Gilvarry, C., Samele, C., Harvey, K., Manley, C., Tattan, T., et al. (2004). Predicting violence in schizophrenia: A prospective study. *Schizophrenia Research, 67*, 247-252.

Walters, G. (1995). The Psychological Inventory of Criminal Thinking Styles: Part I. Reliability and preliminary validity. *Criminal Justice & Behavior, 22*, 307-325.

†Walters, G. (2003a). Predicting criminal justice outcomes with the Psychopathy Checklist and Lifestyle Criminality Screening Form: A meta-analytic comparison. *Behavioral Sciences & the Law, 21*, 89-102.

†Walters, G. (2003b). Predicting institutional adjustment and recidivism with the Psychopathy Checklist factor scores: A meta-analysis. *Law & Human Behavior, 27*, 541-558.

†Walters, G. (2006). Risk-appraisal versus self-report in the prediction of criminal justice outcomes: A meta-analysis. *Criminal Justice & Behavior, 33*, 279-304.

*Walters, G., Duncan, S., & Geyer, M. (2003). Predicting disciplinary adjustment in inmates undergoing forensic evaluation: A direct comparison of the PCL-R and the PAI. *Journal of Forensic Psychiatry & Psychology, 14,* 382-393.

*Walters, G., Knight, R. A., Grann, M., & Dahle, K. P. (2008). Incremental validity of the Psychopathy Checklist facet scores: Predicting release outcome in six samples. *Journal of Abnormal Psychology, 117,* 396-405.

Walters, G., White, T. W., & Denney, D. (1991). The Lifestyle Criminality Screening Form: Preliminary data. *Criminal Justice & Behavior, 18*, 406-418.

Wang, E. W., & Diamon, P. M. (1999). Empirically identifying factors related to violence risk in corrections. *Behavioral Sciences & the Law, 17,* 377–389.

Wang, M., & Stanley, J. (1970). Differential weighting: A review of methods and empirical studies. *Review of Educational Research, 40*, 663-705.

Wang, Y., Sun, G., Pan, J. G., Guo, Z. J., & Li, T. (2006). Performance of tPSA and f/tPSA for prostate cancer in Chinese. A systematic review and meta-analysis. *Prostate Cancer & Prostate Diseases, 9*, 374-378.

Watts, D., Bindman, J., Slade, M., Holloway, F., Rosen, A., & Thornicroft, G. (2004). Clinical Assessment of Risk Decision Support (CARDS): The development and evaluation of a feasible violence risk assessment for routine psychiatric practice. *Journal of Mental Health, 13*, 569-581.

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

Webster, C. D., Eaves, D., Douglas, K. S., & Wintrup, A. (1995). *The HCR-20 scheme: The assessment of dangerousness and risk*. Vancouver, BC: Mental Health Law and Policy Institute, and Forensic Psychiatric Services Commission of British Columbia.

Webster, C. D., Harris, G. T., Rice, M. E., Cormier, C., & Quinsey, V. L. (1994). *The Violence Prediction Scheme: Assessing dangerousness in high risk men*. Toronto: Centre of Criminology.

Webster, C., & Polvi, N. (1995). Challenging assessments of dangerousness and risk. In J. Ziskin (Ed.), *Coping with psychiatric and psychological testimony*. Marina del Rey, CA: Law and Psychology Press.

Wechsler, D. (1997). *The Wechsler Adult Intelligence Scale* (3rd ed.). San Antonio, TX: The Psychological Corporation.

West, S. L., Gartlehner, G., Mansfield, A. J., Poole, C., Tant, E., Lenfestey, N., et al. (2010). *Comparative effectiveness review methods: Clinical heterogeneity*. Research Triangle Park, NC: Agency for Healthcare Research and Quality.

Westen, D., & Weinberger, J. (2004). When clinical description becomes statistical prediction. *American Psychologist, 59*, 595-613.

Whitting, P., Rutjes, A. W., Reitsma, J. B., Bossuyt, P. M., & Kleijnen, J. (2003). The development of QUADAS: A tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology, 10*, 3-25.

Williams, G. J., Macaskill, P., Chan, S. F., Karplus, T. E., Yung, W., Hodson, E. M., et al. (2007). Comparative accuracy of renal duplex sonographic parameters in the diagnosis of renal artery stenosis: Paired and unpaired analysis. *American Journal of Roentgenology, 188*, 798-811.

Wittchen, H. U., & Jacobi, F. (2005). Size and burden of mental disorders in Europe: A critical review and appraisal of 27 studies. *European Neuropsychopharmacology, 15*, 357-376.

‡Woods, P., & Ashley, C. (2007). Violence and aggression: A literature review. *Journal of Psychiatric & Mental Health Nursing, 14*, 652-660.

Wootton, L., Buchanan, A., Leese, M., Tyrer, P., Burns, T., Creed, F., et al. (2008). Violence in psychosis: Estimating the predictive validity of readily accessible clinical information in a community sample. *Schizophrenia Research, 101*, 176-184.

World Health Organization. (2004). *The world health report 2004: Changing history*. Geneva: World Health Organization.

‡Worling, J., & Långström, N. (2003). Assessment of criminal recidivism risk and adolescents who have offended sexually: A review. *Trauma Violence Abuse, 4*, 341-362.

*Wormith, J. S., Olver, M. E., Stevenson, H. E., & Girard, L. (2007). The long-term prediction of offender recidivism using diagnostic, personality, and risk/need approaches to offender assessment. *Psychological Services, 4,* 287-305.

Yang, M., Wong, S. C. P., & Coid, J. (2010). The efficacy of violent prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin, 136*, 740-767.

Young, S. (2009). *Risk assessment tools for children in conflict with the law*. Dublin: Irish Youth Justice Service.

Zammit, S., Allebeck, P., Dalman, C., Lundberg, I., Hemmingsson, T., & Lewis, G. (2003). Investigating the association between cigarette smoking and schizophrenia in a cohort study. *American Journal of Psychiatry, 160*, 2216-2221.

***

APPENDIX A Coding Sheet for Metareview Inter-rater Reliability Check

1) Type of review (tick one)
- ☐ Systematic review
- ☐ Meta-analysis

2) Type of journal (tick one)
- ☐ General
- ☐ Specialty
- ☐ N/A

3) Publishing journal/organisation: _____

4) Journal impact factor (2 year): _____

5) Citation count (as of June 2009): _____

6) Update of a previous review (tick one)
- ☐ No
- ☐ Yes

7) Eligibility criteria based on study population (tick one)
- ☐ Specific offender type
- ☐ Gender
- ☐ Other: _____
- ☐ None

8) Eligibility criteria based on study design (tick one)
- ☐ Prospective only
- ☐ Retrospective only
- ☐ Mixed
- ☐ None/Unstated

9) Eligibility criteria based on language of report (tick one)
- ☐ English only
- ☐ All languages considered
- ☐ Language criteria not reported

10) Offender status of participants in included samples (tick one)
- ☐ Offenders only
- ☐ Non-offenders only
- ☐ Both

11) Population of interest (tick one)
- ☐ Prisoners only
- ☐ Patients only
- ☐ Community only
- ☐ Mixed

12) Grey literature (conference presentations, govt reports, theses) included (tick one)
- ☐ No
- ☐ Yes

13) Number of databases searched: _____

14) Number of other sources searched: _____

15) Years of coverage reported in systematic search (tick one)
- ☐ No
- ☐ Partially (only start or end date)
- ☐ Yes

16) Search terms reported for one or more electronic databases (tick one)
- ☐ No search terms
- ☐ Keywords
- ☐ Boolean operators
- ☐ Readers referred elsewhere

17) Overlapping samples or duplicate studies considered (tick one)
- ☐ No
- ☐ Yes

18) Review flow of systematic search reported (tick one)
- ☐ No (number screened, number eligible, and number included reported not reported)
- ☐ Partially (either number screened, number eligible, or number included reported)
- ☐ Yes (number screened, number eligible, and number included reported)

19) Number of studies included: _____

20) Risk assessment tools investigated: _____

21) Outcome measures reported (e.g., *r*, *d*, AUC): _____

22) PRISMA score: _____

((*Note*: Please complete PRISMA Checklist for each study))

APPENDIX B Preferred Reporting Items for Systematic Reviews and Meta-analyses (PRISMA) Statement

Section/topic	#	Checklist item
TITLE		
Title	1	Identify the report as a systematic review, meta-analysis, or both.
ABSTRACT		
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.
INTRODUCTION		
Rationale	3	Describe the rationale for the review in the context of what is already known.
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).
METHODS		
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.

Section/topic	#	Checklist item
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.
RESULTS		
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).
DISCUSSION		
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.
FUNDING		
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.

APPENDIX C Metareview Assessment of Reporting Quality (MARQ) Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a metareview.	
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; eligibility criteria; reporting quality appraisal methods; results; and implications.	
INTRODUCTION			
Rationale	3	Describe the rationale for the metareview (e.g., conflicting findings of previous reviews or questionable methodology).	
Objectives	4	Provide an explicit statement of the purpose of the metareview (e.g., to identify uncertainties and methodological characteristics of a specific review literature).	
METHODS			
Eligibility criteria	5	Specify review characteristics (e.g., reviews concerning diagnostic accuracy with a specific population or at a certain length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	
Information sources	6	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional reviews) in the search and the dates searched.	
Systematic search	7	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	
Study selection	8	State the process for selecting reviews (i.e., screening, eligibility, included in the metareview).	
Data collection process	9	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate).	
Quality checklist	10	Indicate whether a standardised checklist of reporting characteristics (e.g., PRISMA, MOOSE) was used to assess within-review quality and where it can be accessed.	
Inter-rater reliability	11	State whether the inter-rater reliability of the review characteristics extracted from reports was measured and present relevant effect size (e.g., kappa coefficient or intraclass correlation coefficient).	
Thematic analysis	12	Specify any descriptive thematic analyses conducted to investigate the consistency of findings across reviews addressing similar areas.	

Section/topic	#	Checklist item	Reported on page #
RESULTS			
Review selection	13	Give numbers of reviews screened, assessed for eligibility, and included in the metareview, with reasons for exclusions at each stage, ideally with a flow diagram.	
Review characteristics	14	For each review, present epidemiological characteristics (e.g., impact factors, dates published), descriptive characteristics (e.g., diagnostic instruments used, populations investigated), and reporting characteristics (e.g., duplicate studies considered, publication bias assessed) for which data were extracted.	
Quality checklist	15	Report the average number of reporting characteristics met by the included reviews (with standard deviation) along with the most and least frequently met criteria.	
Thematic analysis	16	Report themes identified across reviews and present, for each, a description of the relevant reviews and their findings.	
DISCUSSION			
Summary of evidence	17	Summarise the main themes/uncertainties identified by the metareview as well as methodological weaknesses that future systematic reviews and meta-analyses could address.	
Implications	18	Consider findings' relevance to key groups (e.g., service providers, clinicians, researchers, and policymakers).	
Limitations	19	Discuss limitations of the employed methodology (e.g., overly- or under-sensitive search strategy, exclusion of unobtainable reviews).	
Conclusion	20	Provide a general interpretation of the results in the context of other evidence.	
ROLE OF EXTERNAL SOURCES			
Conflicts of interest	21	Describe any potential conflicts of interest (e.g., sources of funding and the role of funders).	

Note: Select items adapted from: Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA Statement. *PLoS Medicine*, *6*: e1000097.

APPENDIX D Risk Assessment Tools Included in the Meta-analysis

- **D.1** Level of Service Inventory – Revised (LSI-R)
- **D.2** Psychopathy Checklist – Revised (PCL-R)
- **D.3** Violence Risk Appraisal Guide (VRAG)
- **D.4** Sex Offender Risk Appraisal Guide (SORAG)
- **D.5** Static-99
- **D.6** Historical, Clinical, Risk Management – 20 (HCR-20)
- **D.7** Sexual Violence Risk – 20 (SVR-20)
- **D.8** Spousal Assault Risk Assessment (SARA)
- **D.9** Structured Assessment of Violence Risk in Youth (SAVRY)

APPRENDIX D.1 Level of Service Inventory – Revised (LSI-R)

Domain	Item
Criminal history	
1	Any prior convictions
2	Two or more prior convictions
3	Three or more convictions
4	Three or more present offenses
5	Arrested under age 16
6	Ever incarcerated upon conviction
7	Escape history from a correctional facility
8	Ever punished for institutional misconduct
9	Violation/Charge while on supervision
10	Official record of assault/violence
Education/Employment	
11	Currently employed
12	Frequently unemployed
13	Never employed for a full year
14	Ever fired
15	Education less than grade 10
16	Education less than grade 12
17	Suspended or expelled at least once
18	Participation/performance
19	Peer interactions
20	Authority interactions
Financial	
21	Financial problems
22	Reliance upon social assistance
Family/Marital	
23	Dissatisfaction with marital or equivalent situation
24	Non-rewarding relationship with parents
25	Non-rewarding relationship with relatives
26	Criminal family member or spouse
Accommodation	
27	Unsatisfactory
28	Three or more address changes last year
29	High crime neighbourhood
Leisure/Recreation	
30	Absence of recent participant in an organized activity
31	Could make better use of time

Companions	
32	A social isolate
33	Some criminal acquaintances
34	Some criminal friends
35	Few anti-criminal acquaintances
36	Few anti-criminal friends
Alcohol/Drug problem	
37	Previous alcohol problem
38	Previous drug problem
39	Current alcohol problem
40	Current drug problem
41	Law violations problem
42	Marital/family problems
43	School/work problems
44	Medical problems
45	Other alcohol/drug indicators
Emotional/Personal	
46	Moderate interference
47	Severe interference/active psychosis
48	Past mental health treatment
49	Present mental health treatment
50	Psychological assessment indicated
Attitudes/Orientation	
51	Supportive of crime
52	Unfavourable attitude toward convention
53	Poor attitude toward sentence
54	Poor attitude toward supervision

Note. Adapted from Andrews and Bonta (1995).

APPRENDIX D.2 Psychopathy Checklist – Revised (PCL-R)

Factor	Item
Factor 1	
1	Glibness/Superficial charm
2	Grandiose sense of self-worth
3	Pathological lying
4	Conning/Manipulative
5	Lack of remorse or guilt
6	Shallow affect
7	Callous/Lack of empathy
8	Failure to accept responsibility for actions
Factor 2	
1	Need for stimulation/proneness to boredom
2	Parasitic life-style
3	Poor behavioural controls
4	Early behaviour problems
5	Lack of realistic, long-term goals
6	Impulsivity
7	Irresponsibility
8	Juvenile delinquency
9	Revocation of conditional release
Non-loading items	
1	Promiscuous sexual behaviour
2	Many short-term marital relationships
3	Criminal versatility

Note. Adapted from Hare (1991).

APPRENDIX D.3 Violence Risk Appraisal Guide (VRAG)

	Item
1	Lived with both biological parents to age 16
2	Elementary school maladjustment
3	History of alcohol problems
4	Marital status
5	Previous non-violent crime (Cormier-Lang score)
6	Failure on prior conditional release
7	Age at index offence
8	Victim injury (index offence)
9	Female victim (index offence)
10	DSM-III diagnosis of personality disorder
11	DSM-III diagnosis of schizophrenia
12	Psychopathy Checklist (PCL-R) score

Note. Adapted from Quinsey, Harris, Rice, and Cormier (1998).

APPRENDIX D.4 Sex Offender Risk Appraisal Guide (SORAG)

	Item
1	Lived with both biological parents to age 16
2	Elementary school maladjustment
3	History of alcohol problems
4	Marital status
5	Previous non-violent crime (Cormier-Lang score)
6	Previous violent crime (Cormier-Lang score)
7	Number of previous convictions for sexual offences
8	History of sex offences against girls under age 14 only
9	Failure on prior conditional release
10	Age at index offence
11	DSM-III diagnosis of personality disorder
12	DSM-III diagnosis of schizophrenia
13	Phallometric test results
14	Psychopathy Checklist (PCL-R) score

Note. Adapted from Quinsey, Harris, Rice, and Cormier (1998).

APPRENDIX D.5 Static-99

	Item
1	Young (18-24 years)
2	Ever lived with intimate partner for two or more years
3	Index non-sexual violence
4	Prior non-sexual violence
5	Prior sex offences
6	Prior sentencing dates
7	Prior non-contact sex offences
8	Any unrelated victims
9	Any stranger victims
10	Any male victims

Note. Adapted from Harris, Phenix, Hanson, and Thornton (2003)

APPRENDIX D.6 Historical, Clinical, Risk Management – 20 (HCR-20)

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

Note. Adapted from Webster, Douglas, Eaves, and Hart (1997).

APPRENDIX D.7 Sexual Violence Risk – 20 (SVR-20)

Scale	Item
Psychosocial adjustment	
1	Sexual deviation
2	Victim of child abuse
3	Psychopathy
4	Major mental illness
5	Substance use problems
6	Suicidal/homicidal ideation
7	Relationship problems
8	Employment problems
9	Past nonsexual violent offences
10	Past nonviolent offences
11	Past supervision failure
Sexual offences	
12	High density sex offences
13	Multiple sex offence types
14	Physical harm to victim(s) in sex offences
15	Uses of weapons or threats of death in sex offences
16	Escalation of frequency or severity of sex offences
17	Extreme minimization or denial of sex offences
18	Attitudes that support or condone sex offences
Future plans	
19	Lacks realistic plans
20	Negative attitude toward intervention

Note. Adapted from Boer, Hart, Kropp, and Webster (1997).

APPRENDIX D.8 Spousal Assault Risk Assessment (SARA)

Scale	Item
Criminal history	
1	Past assault of family members
2	Past assault of strangers or acquaintances
3	Past violation of conditional release or community supervision
Psychosocial adjustment	
4	Recent relationship problems
5	Recent employment problems
6	Victim of and/or witness to family violence as a child or adolescent
7	Recent substance abuse/dependence
8	Recent suicidal or homicidal ideation/intent
9	Recent psychotic and/or manic symptoms
10	Personality disorder with anger, impulsivity, or behavioural instability
Spousal assault history	
11	Past physical assault
12	Past sexual assault/sexual jealousy
13	Past use of weapons and/or credible threats of death
14	Recent escalation in frequency or severity of assault
15	Past violation of “no contact” orders
16	Extreme minimization or denial or spousal assault history
17	Attitudes that support or condone spousal assault
Alleged (current) offence	
18	Severe and/or sexual assault
19	Uses of weapons or threats of death
20	Violation of “no contact” order

Note. Adapted from Kropp, Hart, Webster, and Eaves (1995).

APPRENDIX D.9 Structured Assessment of Violence Risk in Youth (SAVRY)

Scale	Item
Historical risk factors	
1	History of violence
2	History of non-violent offending
3	Early initiation of violence
4	Past supervision/intervention failures
5	History of self-harm or suicide attempts
6	Exposure to violence in the home
7	Childhood history of maltreatment
8	Parental/Caregiver criminality
9	Early caregiver disruption
10	Poor school achievement
Social/Contextual risk factors	
11	Peer delinquency
12	Peer rejection
13	Stress and poor coping
14	Poor parental management
15	Lack of personal/social support
16	Community disorganization
Individual/Clinical risk factors	
17	Negative attitudes
18	Risk taking/impulsivity
19	Substance use difficulties
20	Anger management problems
21	Low empathy/remorse
22	Attention deficit/hyperactivity difficulties
23	Poor compliance
24	Low interest/commitment to school
Protective factors	
P1	Prosocial involvement
P2	Strong social support
P3	Strong attachments and bonds
P4	Positive attitude towards intervention and authority
P5	Strong commitment to school
P6	Resilient personality traits

Note. Adapted from Borum, Bartel, and Forth (2003).

APPENDIX E Standardised E-mail Template for Meta-analysis Data Collection

UNIVERSITY OF OXFORD
DEPARTMENT OF PSYCHIATRY

WARNEFORD HOSPITAL
OXFORD
OX3 7JX

TEL: (07810) 637139
FAX: (01865) 793101
jay.singh@psych.ox.ac.uk

<<Date>>
<<Author E-mail Address>>

Dear Dr. *<<Author Last Name>>*

Systematic review of studies of risk assessment tools

We are currently conducting a meta-analysis of studies of the predictive validity of risk assessment tools.

We should be very grateful, therefore, if you would provide results from your study examining the predictive validity of the *<<Tool>>*:

<<Reference>>

Inclusion of these data from your study would assist in the appropriate interpretation of the other available evidence. We would request that outcome information be provided for the most sensitive outcome measured (i.e., *<<Outcome>>*). We've attached a simple document into which the relevant information can be entered (in case that might be helpful) and, of course, any assistance that you provide would be acknowledged in the report. Many thanks.

Yours sincerely

Jay P. Singh Seena Fazel Martin Grann

Data from: <<*Reference*>>

Number of participants in LOW risk category (scores <<*min*>> to <<*max*>>): _____

No. male No. female

_____ _____

Number of participants in LOW risk category WHO DID offend: _____

No. male No. female

_____ _____

Number of participants in MODERATE risk category (scores <<*min*>> to <<*max*>>):
_____

No. male No. female

_____ _____

Number of participants in MODERATE risk category WHO DID offend: _____

No. male No. female

_____ _____

Number of participants in HIGH risk category (scores <<*min*>> to <<*max*>>): _____

No. male No. female

_____ _____

Number of participants in HIGH risk category WHO DID offend: _____

No. male No. female

_____ _____

Area Under the Curve (AUC) by Risk Score for <<*Tool*>>: _____

Standard Error (SE) for AUC of <<*Tool*>>: _____

95% Confidence Interval for AUC of <<*Tool*>>: _____ **to** _____

Please fax to:
S Fazel, FAX: +44-1865-793101
Or e-mail: jay.singh@psych.ox.ac.uk
Or post to: S. Fazel, Warneford Hospital, Oxford OX3 7JX, UK

APPENDIX F Characteristics of Samples Missing From Second Binning Strategy

Ten (11.4%) samples provided data for the high risk versus low/moderate risk binning strategy but not the moderate/high risk versus low risk binning strategy. Sample data was missing for the Static-99 (k = 4), the SVR-20 (k = 2), the HCR-20 (k = 1), the SARA (k = 1), the SORAG (k = 1), and the VRAG (k = 1). Of the 2,881 participants in the samples, 589 (20.4%) offended. The random effects pooled DOR of the missing samples for the high risk versus low/moderate risk binning strategy was 4.38 (95% CI = 2.35-8.16).

The mean sample size was 288 (SD = 548) participants. All of the samples with missing data were composed of over 50% male participants. On average, 284 (SD = 549) male participants were included in each sample. Ethnicity data was available for half (k = 5) of the samples with missing study data. Of the samples, 4 (4.5%) were composed of over 50.0% white participants. An average of 90 (SD = 62) white participants were included in each sample. The mean age of participants was 38.7 (SD = 6.4) years.

Regarding study setting, 3 (3.4%) samples consisted of prisoners, 6 (6.8%) of psychiatric patients, and 1 (1.1%) of a mixture of prisoners and psychiatric patients. Of those 6 samples in psychiatric settings, 1 (1.1%) consisted of patients in a general setting and 5 (5.7%) in forensic settings. Prospective research methodology was used with 3 (3.4%) samples, whereas 7 (8.0%) samples were investigated retrospectively. The mean length participants were followed up was 76.3 (SD = 36.6) months. Three (3.4%) samples were conducted in North America, whereas 7 (8.0%) were conducted in Europe. Regarding type of offending, general offending was the outcome criteria in 7 (8.0%) samples as opposed to violent offending in 3 (3.4%) samples. Seven (8.0%) of the missing samples used charge, conviction, or incarceration as their outcome criteria. Of the included samples, 5 (5.7%) used file review alone to score a risk assessment tool, whereas 1 (1.1%) used information from interviews only. The remaining 4 (4.5%) samples were from studies that did not explicitly report the source of information used to score the tool(s) of interest. Finally, 7 (8.0%) samples were in studies published in peer-reviewed journals.

APPENDIX G Coding Sheet for Meta-analysis Inter-rater Reliability Check

1) Instrument of interest (tick all relevant)
- ☐ HCR-20
- ☐ LSI-R
- ☐ PCL-R
- ☐ SARA
- ☐ SAVRY
- ☐ SORAG
- ☐ Static-99
- ☐ SVR-20
- ☐ VRAG

2) Approach to risk assessment (tick one)
- ☐ Actuarial
- ☐ Structured clinical judgement

3) Source of study (tick one)
- ☐ Journal article
- ☐ Conference presentation
- ☐ Dissertation
- ☐ Master's thesis
- ☐ Government report

4) Country in which study was conducted (tick one)
- ☐ United States of America
- ☐ Canada
- ☐ United Kingdom
- ☐ Other: _____

5) Study setting (tick one)
- ☐ Community
- ☐ Prison/Correctional
- ☐ Psychiatric/Mental health
- ☐ Mixed

6) Temporal design (tick one)
- ☐ Prospective (outcome data not available when tool scored)
- ☐ Retrospective (outcome data already available when tool scored)

7) Type of offending for which data obtained (tick one)
- ☐ General (violent and non-violent)
- ☐ Violent only
- ☐ Sexual only
- ☐ Non-violent only

8) Type of outcome (tick one)
- ☐ Arrest
- ☐ Charge
- ☐ Conviction
- ☐ Incarceration
- ☐ Institutional incident

9) Source of information used to score tool (tick one)
- ☐ File review
- ☐ Interview
- ☐ Mixed

9) Tool authorship (tick all relevant)
- ☐ Author of English-language manual of tool was also study author
- ☐ Author of translation of tool manual was also study author

10) Sample size: _____

11) Percentage of sample male: _____%

12) Percentage of sample white: _____%

13) Mean age of participants: _____ years

14) Mean length of follow-up: _____ months

15) STARD score: _____

((*Note*: Please complete STARD Checklist for each study))

Outcome Measures

Instrument of interest: _____

AUC (risk score): _____; 95% CI: _____ - _____

High Risk vs. Low/Moderate Risk Binning Strategy:

		Outcome	
		Offender	Non-offender
Test Result	Positive	TP _____	FP _____
	Negative	FN _____	TN _____

Moderate/High Risk vs. Low Risk Binning Strategy:

		Outcome	
		Offender	Non-offender
Test Result	Positive	TP _____	FP _____
	Negative	FN _____	TN _____

((*Note*: Please complete multiple outcome measure sections if more than one instrument of interest included in study))

APPENDIX H Standards for Reporting of Diagnostic Accuracy Studies (STARD) Statement

Manuscript number and/or corresponding author name:		
Section and Topic	**Item #**	
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.
METHODS		Describe
Participants	3	The study population: The inclusion and exclusion criteria, setting and locations where the data were collected.
	4	Participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?
	5	Participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in item 3 and 4? If not, specify how participants were further selected.
	6	Data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?
Test methods	7	The reference standard and its rationale.
	8	Technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.
	9	Definition of and rationale for the units, cutoffs and/or categories of the results of the index tests and the reference standard.
	10	The number, training and expertise of the persons executing and reading the index tests and the reference standard.
	11	Whether or not the readers of the index tests and reference standard were blind (masked) to the results of the other test and describe any other clinical information available to the readers.
Statistical methods	12	Methods for calculating or comparing measures of diagnostic accuracy, and the statistical methods used to quantify uncertainty (e.g. 95% confidence intervals).
	13	Methods for calculating test reproducibility, if done.
RESULTS		Report
Participants	14	When study was done, including beginning and ending dates of recruitment.
	15	Clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, comorbidity, current treatments, recruitment centers).
	16	The number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test (a flow diagram is strongly recommended).
Test results	17	Time interval from the index tests to the reference standard, and any treatment administered between.
	18	Distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.
	19	A cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.
	20	Any adverse events from performing the index tests or the reference standard.
Estimates	21	Estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% confidence intervals).
	22	How indeterminate results, missing responses and outliers of the index tests were handled.
	23	Estimates of variability of diagnostic accuracy between subgroups of participants, readers or centers, if done.
	24	Estimates of test reproducibility, typically imprecision (as CV) at 2 or 3 concentrations.
DISCUSSION	25	Discuss the clinical applicability of the study findings.

APPENDIX I Statistical Tests Comparing Effect Sizes Produced by Different Versions of a Screening Tool for Individuals with Hospital Discharge Diagnoses of Schizophrenia

Calibration Sample vs. Cross-validation Samples

NPV

Calibration sample at 1 year vs. Cross-validation sample 1 at 1 year: $\chi^2 = 0.02$, $p = 0.88$
Calibration sample at 2 years vs. Cross-validation sample 1 at 2 years: $\chi^2 = 0.02$, $p = 0.88$
Calibration sample at 5 years vs. Cross-validation sample 1 at 5 years: $\chi^2 = 0.02$, $p = 0.88$
Calibration sample at 1 year vs. Cross-validation sample 2 at 1 year: $\chi^2 = 0.02$, $p = 0.88$
Calibration sample at 2 years vs. Cross-validation sample 2 at 2 years: $\chi^2 = 0.02$, $p = 0.88$
Calibration sample at 5 years vs. Cross-validation sample 2 at 5 years: $\chi^2 = 0.03$, $p = 0.87$
Calibration sample at 1 year vs. Cross-validation sample 3 at 1 year: $\chi^2 = 0.02$, $p = 0.89$
Calibration sample at 2 years vs. Cross-validation sample 3 at 2 years: $\chi^2 = 0.03$, $p = 0.89$
Calibration sample at 5 years vs. Cross-validation sample 3 at 5 years: $\chi^2 = 0.03$, $p = 0.87$

PPV

Calibration sample at 1 year vs. Cross-validation sample 1 at 1 year: $\chi^2 = 1.63$, $p = 0.18$
Calibration sample at 2 years vs. Cross-validation sample 1 at 2 years: $\chi^2 = 1.59$, $p = 0.21$
Calibration sample at 5 years vs. Cross-validation sample 1 at 5 years: $\chi^2 = 0.70$, $p = 0.40$
Calibration sample at 1 year vs. Cross-validation sample 2 at 1 year: $\chi^2 = 1.11$, $p = 0.34$
Calibration sample at 2 years vs. Cross-validation sample 2 at 2 years: $\chi^2 = 1.52$, $p = 0.22$
Calibration sample at 5 years vs. Cross-validation sample 2 at 5 years: $\chi^2 = 0.68$, $p = 0.41$
Calibration sample at 1 year vs. Cross-validation sample 3 at 1 year: $\chi^2 = 1.66$, $p = 0.17$
Calibration sample at 2 years vs. Cross-validation sample 3 at 2 years: $\chi^2 = 1.76$, $p = 0.16$
Calibration sample at 5 years vs. Cross-validation sample 3 at 5 years: $\chi^2 = 0.69$, $p = 0.40$

DOR

Calibration sample at 1 year vs. Cross-validation sample 1 at 1 year: $\chi^2 = 1.50$, $p = 0.22$
Calibration sample at 2 years vs. Cross-validation sample 1 at 2 years: $\chi^2 = -1.13$, $p = 0.29$
Calibration sample at 5 years vs. Cross-validation sample 1 at 5 years: $\chi^2 = 0.06$, $p = 0.81$
Calibration sample at 1 year vs. Cross-validation sample 2 at 1 year: $\chi^2 = 0.08$, $p = 0.77$
Calibration sample at 2 years vs. Cross-validation sample 2 at 2 years: $\chi^2 = -0.04$, $p = 0.84$
Calibration sample at 5 years vs. Cross-validation sample 2 at 5 years: $\chi^2 = -0.01$, $p = 0.98$
Calibration sample at 1 year vs. Cross-validation sample 3 at 1 year: $\chi^2 = -0.54$, $p = 0.46$
Calibration sample at 2 years vs. Cross-validation sample 3 at 2 years: $\chi^2 = -0.61$, $p = 0.44$
Calibration sample at 5 years vs. Cross-validation sample 3 at 5 years: $\chi^2 = -0.14$, $p = 0.71$

AUC

Calibration sample at 1 year vs. Cross-validation sample 1 at 1 year: $z = -1.03$, $p = 0.30$
Calibration sample at 2 years vs. Cross-validation sample 1 at 2 years: $z = -1.09$, $p = 0.28$
Calibration sample at 5 years vs. Cross-validation sample 1 at 5 years: $z = 0.26$, $p = 0.79$
Calibration sample at 1 year vs. Cross-validation sample 2 at 1 year: $z = 0.10$, $p = 0.92$
Calibration sample at 2 years vs. Cross-validation sample 2 at 2 years: $z = 0.53$, $p = 0.60$
Calibration sample at 5 years vs. Cross-validation sample 2 at 5 years: $z = 0.42$, $p = 0.67$
Calibration sample at 1 year vs. Cross-validation sample 3 at 1 year: $z = 0.76$, $p = 0.45$
Calibration sample at 2 years vs. Cross-validation sample 3 at 2 years: $z = 0.89$, $p = 0.37$
Calibration sample at 5 years vs. Cross-validation sample 3 at 5 years: $z = 0.92$, $p = 0.36$

Five-item Tool vs. Six-item Tool

NPV

Five-item tool at 1 year vs. Six-item tool at 1 year: $\chi^2 = 0.02$, $p = 0.88$
Five-item tool at 2 years vs. Six-item tool at 2 years: $\chi^2 = 0.02$, $p = 0.88$
Five-item tool at 5 years vs. Six-item tool at 5 years: $\chi^2 = -0.03$, $p = 0.87$

PPV

Five-item tool at 1 year vs. Six-item tool at 1 year: $\chi^2 = 1.13$, $p = 0.29$
Five-item tool at 2 years vs. Six-item tool at 2 years: $\chi^2 = 0.52$, $p = 0.47$
Five-item tool at 5 years vs. Six-item tool at 5 years: $\chi^2 = 0.17$, $p = 0.68$

DOR

Five-item tool at 1 year vs. Six-item tool at 1 year: $\chi^2 = -0.03$, $p = 0.87$
Five-item tool at 2 years vs. Six-item tool at 2 years: $\chi^2 = -0.16$, $p = 0.69$
Five-item tool at 5 years vs. Six-item tool at 5 years: $\chi^2 = -0.11$, $p = 0.74$

AUC

Five-item tool at 1 year vs. Six-item tool at 1 year: $z = -0.36$, $p = 0.72$
Five-item tool at 2 years vs. Six-item tool at 2 years: $z = -0.35$, $p = 0.73$
Five-item tool at 5 years vs. Six-item tool at 5 years: $z = -0.25$, $p = 0.80$

Unit Scored Tool vs. Weighted Tools

NPV

Unit scored at 1 year vs. Unstd β coef weighted tool at 1 year: $\chi^2 = 0.01$, $p = 0.97$
Unit scored at 2 years vs. Unstd β coef weighted tool at 2 years: $\chi^2 = 0.01$, $p = 0.96$
Unit scored at 5 years vs. Unstd β coef weighted tool at 5 years: $\chi^2 = -0.02$, $p = 0.96$
Unit scored at 1 year vs. HR weighted tool at 1 year: $\chi^2 = 0.01$, $p = 0.98$
Unit scored at 2 years vs. HR weighted tool at 2 years: $\chi^2 = 0.01$, $p = 0.97$
Unit scored at 5 years vs. HR weighted tool at 5 years: $\chi^2 = -0.01$, $p = 0.97$

PPV

Unit scored at 1 year vs. Unstd β coef weighted tool at 1 year: $\chi^2 = 0.01$, $p = 0.98$
Unit scored at 2 years vs. Unstd β coef weighted tool at 2 years: $\chi^2 = 0.01$, $p = 0.98$
Unit scored at 5 years vs. Unstd β coef weighted tool at 5 years: $\chi^2 = 0.01$, $p = 0.96$
Unit scored at 1 year vs. HR weighted tool at 1 year: $\chi^2 = 0.01$, $p = 0.97$
Unit scored at 2 years vs. HR weighted tool at 2 years: $\chi^2 = 0.01$, $p = 0.98$
Unit scored at 5 years vs. HR weighted tool at 5 years: $\chi^2 = 0.01$, $p = 0.96$

DOR

Unit scored at 1 year vs. Unstd β coef weighted tool at 1 year: $\chi^2 = 0.21$, $p = 0.61$
Unit scored at 2 years vs. Unstd β coef weighted tool at 2 years: $\chi^2 = -0.64$, $p = 0.40$
Unit scored at 5 years vs. Unstd β coef weighted tool at 5 years: $\chi^2 = -0.83$, $p = 0.35$
Unit scored at 1 year vs. HR weighted tool at 1 year: $\chi^2 = 0.18$, $p = 0.67$
Unit scored at 2 years vs. HR weighted tool at 2 years: $\chi^2 = -0.63$, $p = 0.42$
Unit scored at 5 years vs. HR weighted tool at 5 years: $\chi^2 = -0.88$, $p = 0.32$

AUC

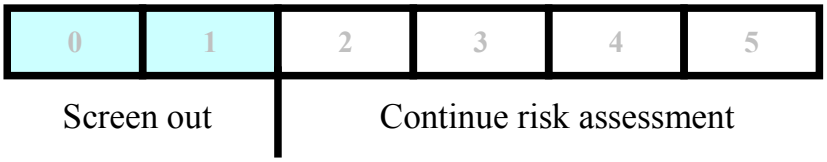
Unit scored at 1 year vs. Unstd β coef weighted tool at 1 year: $z = 0.55$, $p = 0.58$
Unit scored at 2 years vs. Unstd β coef weighted tool at 2 years: $z = 0.71$, $p = 0.48$
Unit scored at 5 years vs. Unstd β coef weighted tool at 5 years: $z = 0.60$, $p = 0.55$
Unit scored at 1 year vs. HR weighted tool at 1 year: $z = 0.54$, $p = 0.58$
Unit scored at 2 years vs. HR weighted tool at 2 years: $z = 0.72$, $p = 0.48$
Unit scored at 5 years vs. HR weighted tool at 5 years: $z = 0.64$, $p = 0.52$

APPENDIX J Coding Sheet for a Violence Screening Tool for Individuals with Hospital Discharge Diagnoses of Schizophrenia

Rule Out Screening Tool for the Evaluation of Risk in Schizophrenia (ROSTER)

	Tick one	
	YES	**NO**
1. Male gender?	+1	+0
2. Previous criminal conviction?	+1	+0
3. Below age 32 at time of assessment?	+1	+0
4. Comorbid alcohol abuse?	+1	+0
5. Comorbid drug abuse?	+1	+0

Total risk score (tick one):



Clinical Override (tick one): Yes No