A systematic review of risk assessment strategies for populations at high risk of engaging in violent behaviour: update 2002–8

R Whittington, JC Hockenhull, J McGuire, M Leitner, W Barr, MG Cherry, R Flentje, B Quinn, Y Dundar and R Dickson
A systematic review of risk assessment strategies for populations at high risk of engaging in violent behaviour: update 2002–8

R Whittington,1* JC Hockenhull,2 J McGuire,3 M Leitner,4 W Barr,1 MG Cherry,2 R Flentje,4 B Quinn,1 Y Dundar2 and R Dickson2

1Health and Community Care Research Unit, University of Liverpool, Liverpool, UK
2Liverpool Reviews and Implementation Group, University of Liverpool, Liverpool, UK
3Clinical Psychology, University of Liverpool, Liverpool, UK
4Infotech UK Research (Medical Division of ER&IC Ltd), Cheshire, UK

*Corresponding author

Declared competing interests of authors: none

Published October 2013
DOI: 10.3310/hta17500

This report should be referenced as follows:


Health Technology Assessment is indexed and abstracted in Index Medicus/MEDLINE, Excerpta Medica/EMBASE, Science Citation Index Expanded (SciSearch®) and Current Contents®/Clinical Medicine.
Criteria for inclusion in the Health Technology Assessment journal
Reports are published in Health Technology Assessment (HTA) if (1) they have resulted from work for the HTA programme, and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in Health Technology Assessment are termed ‘systematic’ when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

HTA programme
The HTA programme, part of the National Institute for Health Research (NIHR), was set up in 1993. It produces high-quality research information on the effectiveness, costs and broader impact of health technologies for those who use, manage and provide care in the NHS. ‘Health technologies’ are broadly defined as all interventions used to promote health, prevent and treat disease, and improve rehabilitation and long-term care.

The journal is indexed in NHS Evidence via its abstracts included in MEDLINE and its Technology Assessment Reports inform National Institute for Health and Care Excellence (NICE) guidance. HTA research is also an important source of evidence for National Screening Committee (NSC) policy decisions.

For more information about the HTA programme please visit the website: www.hta.ac.uk/

This report
The research reported in this issue of the journal was commissioned and funded by the HTA programme on behalf of NICE as project number 08/101/99. The protocol was agreed in May 2009. The assessment report began editorial review in October 2012 and was accepted for publication in March 2013. The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The HTA editors and publisher have tried to ensure the accuracy of the authors’ report and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this report.

This report presents independent research funded by the National Institute for Health Research (NIHR). The views and opinions expressed by authors in this publication are those of the authors and do not necessarily reflect those of the NHS, the NIHR, NETSCC, the HTA programme or the Department of Health. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, NETSCC, the HTA programme or the Department of Health.

© Queen’s Printer and Controller of HMSO 2013. This work was produced by Whittington et al. under the terms of a commissioning contract issued by the Secretary of State for Health. This issue may be freely reproduced for the purposes of private research and study and extracts (or indeed, the full report) may be included in professional journals provided that suitable acknowledgement is made and the reproduction is not associated with any form of advertising. Applications for commercial reproduction should be addressed to: NIHR Journals Library, National Institute for Health Research, Evaluation, Trials and Studies Coordinating Centre, Alpha House, University of Southampton Science Park, Southampton SO16 7NS, UK.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Prepress Projects Ltd, Perth, Scotland (www.prepress-projects.co.uk).
Editor-in-Chief of Health Technology Assessment and NIHR Journals Library

Professor Tom Walley  Director, NIHR Evaluation, Trials and Studies and Director of the HTA Programme, UK

NIHR Journals Library Editors

Professor Ken Stein  Chair of HTA Editorial Board and Professor of Public Health, University of Exeter Medical School, UK

Professor Andree Le May  Chair of NIHR Journals Library Editorial Group (EME, HS&DR, PGfAR, PHR journals)

Dr Martin Ashton-Key  Consultant in Public Health Medicine/Consultant Advisor, NETSCC, UK

Professor Matthias Beck  Chair in Public Sector Management and Subject Leader (Management Group), Queen's University Management School, Queen's University Belfast, UK

Professor Aileen Clarke  Professor of Health Sciences, Warwick Medical School, University of Warwick, UK

Dr Tessa Crilly  Director, Crystal Blue Consulting Ltd, UK

Dr Peter Davidson  Director of NETSCC, HTA, UK

Ms Tara Lamont  Scientific Advisor, NETSCC, UK

Dr Tom Marshall  Reader in Primary Care, School of Health and Population Sciences, University of Birmingham, UK

Professor Elaine McColl  Director, Newcastle Clinical Trials Unit, Institute of Health and Society, Newcastle University, UK

Professor William McGuire  Professor of Child Health, Hull York Medical School, University of York, UK

Professor Geoffrey Meads  Honorary Professor, Business School, Winchester University and Medical School, University of Warwick, UK

Professor Jane Norman  Professor of Maternal and Fetal Health, University of Edinburgh, UK

Professor John Powell  Consultant Clinical Adviser, National Institute for Health and Care Excellence (NICE), UK

Professor James Raftery  Professor of Health Technology Assessment, Wessex Institute, Faculty of Medicine, University of Southampton, UK

Dr Rob Riemsma  Reviews Manager, Kleijnen Systematic Reviews Ltd, UK

Professor Helen Roberts  Professorial Research Associate, University College London, UK

Professor Helen Snooks  Professor of Health Services Research, Institute of Life Science, College of Medicine, Swansea University, UK

Please visit the website for a list of members of the NIHR Journals Library Board: www.journalslibrary.nihr.ac.uk/about/editors

Editorial contact: nihredit@southampton.ac.uk

NIHR Journals Library www.journalslibrary.nihr.ac.uk
Abstract

A systematic review of risk assessment strategies for populations at high risk of engaging in violent behaviour: update 2002–8

R Whittington,1* JC Hockenhull,2 J McGuire,3 M Leitner,4 W Barr,1 MG Cherry,2 R Flentje,4 B Quinn,1 Y Dundar2 and R Dickson2

1Health and Community Care Research Unit, University of Liverpool, Liverpool, UK
2Liverpool Reviews and Implementation Group, University of Liverpool, Liverpool, UK
3Clinical Psychology, University of Liverpool, Liverpool, UK
4Infotech UK Research (Medical Division of ER&IC Ltd), Cheshire, UK

*Corresponding author

Background: This review systematically examines the research literature published in the period 2002–8 on structured violence risk assessment instruments designed for use in mental health services or the criminal justice system. It adopted much broader inclusion criteria than previous reviews in the same area in order to capture and summarise data on the widest possible range of available instruments.

Objectives: To address two questions: (1) what study characteristics are associated with a risk assessment instrument score being significantly associated with a violent outcome? and (2) which risk assessment instruments have the highest level of predictive validity for a violent outcome?

Data sources: Nineteen bibliographic databases were searched from January 2002 to April 2008, including PsycINFO, MEDLINE, Cumulative Index to Nursing and Allied Health Literature, Allied and Complementary Medicine Database, British Nursing Index, International Bibliography of the Social Sciences, Education Resources Information Centre, The Cochrane Library and Web of Knowledge.

Review methods: Inclusion criteria for studies were (1) evaluation of a structured risk tool; (2) outcome measure of interpersonal violence; (3) participants aged 17 years or over; and (4) participants with a mental disorder and/or at least one offence and/or at least one indictable offence. A series of bivariate analyses using either a chi-squared test or Spearman’s rank-order correlation were conducted to explore associations between study characteristics and outcomes. Data from a subset of studies reporting area under the curve (AUC) analysis were combined to provide estimates of mean validity.

Results: For the overall set of included studies (n = 959), over three-quarters (77%) were conducted in the USA, Canada or the UK. Two-thirds of all studies were conducted with offenders who had either no formal mental health diagnosis (43%) or forensic samples with a formal diagnosis (25%). The Psychopathy Checklist-Revised was tested in the largest number of studies (n = 192). Most studies (78%) reported a statistically significant (p < 0.05) relationship between the instrument score and a violent outcome. Prospective data collection ($\chi^2 = 4.4, p = 0.035$), number of people recruited ($U = 27.8, p = 0.012$) and number of participants at end point ($U = 26.9, p = 0.04$) were significantly associated with predictive validity. For those instruments tested in five or more studies reporting AUC values, the General Statistical Information on Recidivism instrument had the highest mean AUC (0.73).
Limitations: Agreement between pairs of reviewers in the initial pilot exercises was good but less than perfect, so discrepancies may be present given the complexity and subjectivity of some aspects of violence research. Only five of the seven calendar years (2003–7) are completely covered, with partial coverage of 2002 and 2008. There is no weighting for sample or effect sizes when results from studies are aggregated.

Conclusions: A very large number of studies examining the relationship between a structured instrument and a violent outcome were published in this relatively short 7-year period. The general quality of the literature is weak in places (e.g. over-reliance on cross-sectional designs) and a vast range of distinct instruments have been tested to varying degrees. However, there is evidence of some convergence around a small number of high-performing instruments and identification of the components of a high-quality evaluation approach, including AUC analysis. The upper limits (AUC ≥ 0.85) of instrument-based prediction have probably been achieved and are unlikely to be exceeded using instruments alone.

Funding: The National Institute for Health Research Health Technology Assessment and Research for Patient Benefit programmes.
# Contents

List of abbreviations ...................................................... ix

Scientific summary ....................................................... xi

Chapter 1 Background ................................................... 1
Definition of violence ................................................... 1
What is risk assessment in relation to the criminal justice system and mental health services? ....... 1
Why is risk assessment important to the criminal justice system and mental health services? ....... 2
Approaches to predicting violence: the recent history ................................................... 3
Rationale for the review ................................................ 6

Chapter 2 Methods ....................................................... 9
Methods for reviewing clinical effectiveness ................................................... 9
Inclusion and exclusion criteria ................................................... 9
Quality assessment ....................................................... 12
Data extraction ......................................................... 12
Descriptive data ........................................................... 13
Statistical methods ....................................................... 13
Advisory panel .............................................................. 14

Chapter 3 Overview of the literature .................................... 15
Selection of included studies ................................................... 15
Quality assessment ....................................................... 15
Study characteristics ...................................................... 18
Participant characteristics .................................................... 18
Risk assessment characteristics ................................................... 23

Chapter 4 Results of analyses .............................................. 29
Bivariate analyses ....................................................... 29
Multivariate analysis ...................................................... 30
Area under the curve analysis ................................................... 31

Chapter 5 Discussion and conclusions .................................. 39
Strengths and limitations of the review ................................................... 39
Summary of key findings ...................................................... 40
Conclusions and implications for research ................................................... 43

Acknowledgements ......................................................... 45

References ................................................................. 47

Appendix 1 Search strategies .............................................. 111

Appendix 2 Area under the curve data .................................... 113

Appendix 3 Protocol ......................................................... 117
## List of abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>AUC</td>
<td>area under the curve</td>
</tr>
<tr>
<td>BDI</td>
<td>Beck Depression Inventory</td>
</tr>
<tr>
<td>BIS</td>
<td>Barratt Impulsiveness Scale</td>
</tr>
<tr>
<td>BPRS</td>
<td>Brief Psychiatric Rating Scale</td>
</tr>
<tr>
<td>BVC</td>
<td>Broset Violence Checklist</td>
</tr>
<tr>
<td>CI</td>
<td>confidence interval</td>
</tr>
<tr>
<td>CIDRRI</td>
<td>Clinical Inventory of Dynamic Reoffending Risk Indicators</td>
</tr>
<tr>
<td>CSS</td>
<td>Criminal Sentiments Scale</td>
</tr>
<tr>
<td>CTS</td>
<td>Conflict Tactics Scale</td>
</tr>
<tr>
<td>DVSI</td>
<td>Domestic Violence Screening Instrument</td>
</tr>
<tr>
<td>GSIR</td>
<td>General Statistical Information on Recidivism</td>
</tr>
<tr>
<td>HCR-20</td>
<td>Historical Clinical Risk Management-20</td>
</tr>
<tr>
<td>HTA</td>
<td>Health Technology Assessment</td>
</tr>
<tr>
<td>ICT</td>
<td>Iterative Classification Tree</td>
</tr>
<tr>
<td>ITT</td>
<td>intention to treat</td>
</tr>
<tr>
<td>LSI</td>
<td>Level of Service Inventory</td>
</tr>
<tr>
<td>MCAA</td>
<td>Measures of Criminal Attitudes and Associates</td>
</tr>
<tr>
<td>MCMI</td>
<td>Millon Clinical Multiaxial Inventory</td>
</tr>
<tr>
<td>MMPI</td>
<td>Minnesota Multiphasic Personality Inventory</td>
</tr>
<tr>
<td>NAS</td>
<td>Novaco Anger Scale</td>
</tr>
<tr>
<td>NIHR</td>
<td>National Institute for Health Research</td>
</tr>
<tr>
<td>OASys</td>
<td>Offender Assessment System</td>
</tr>
<tr>
<td>OGRS</td>
<td>Offender Group Reconviction Scale</td>
</tr>
<tr>
<td>PCL-R</td>
<td>Psychopathy Checklist-Revised</td>
</tr>
<tr>
<td>PS</td>
<td>Preliminary Scheme</td>
</tr>
<tr>
<td>RASP</td>
<td>Risk Assessment Scale for Prison</td>
</tr>
<tr>
<td>RfPB</td>
<td>Research for Patient Benefit</td>
</tr>
<tr>
<td>RM-2000</td>
<td>Risk Matrix 2000</td>
</tr>
<tr>
<td>ROC</td>
<td>receiver operating characteristic</td>
</tr>
<tr>
<td>RRASOR</td>
<td>Rapid Risk Assessment for Sexual Offender Recidivism</td>
</tr>
<tr>
<td>SAQ</td>
<td>Self-Appraisal Questionnaire</td>
</tr>
<tr>
<td>SD</td>
<td>standard deviation</td>
</tr>
<tr>
<td>SDRS</td>
<td>Short Dynamic Risk Scale</td>
</tr>
<tr>
<td>SE</td>
<td>standard error</td>
</tr>
<tr>
<td>SIGLE</td>
<td>System for Information on Grey Literature in Europe</td>
</tr>
<tr>
<td>SIR-R1</td>
<td>Statistical Information on Recidivism – Revised 1</td>
</tr>
<tr>
<td>SORAG</td>
<td>Sex Offender Risk Appraisal Guide</td>
</tr>
<tr>
<td>SPJ</td>
<td>Structured Professional Judgement</td>
</tr>
<tr>
<td>SPSS</td>
<td>Statistical Product and Service Solutions</td>
</tr>
<tr>
<td>START</td>
<td>Short-Term Assessment of Risk and Treatability</td>
</tr>
<tr>
<td>STAXI</td>
<td>State-Trait Anger Expression Inventory</td>
</tr>
<tr>
<td>SVR</td>
<td>Sexual Violence Risk</td>
</tr>
<tr>
<td>VRAG</td>
<td>Violence Risk Appraisal Guide</td>
</tr>
<tr>
<td>VRS</td>
<td>Violence Risk Scale</td>
</tr>
<tr>
<td>VSC</td>
<td>Violence Screening Checklist</td>
</tr>
</tbody>
</table>
Scientific summary

Background

This review systematically examines the research literature published in the period 2002–8 on structured violence risk assessment instruments designed for use in mental health services or the criminal justice system. Violence is a major social problem and improved assessment of those who present an above-average risk is an important goal in the overall strategy for addressing the issue. Techniques for formally assessing individual and social risk factors have developed rapidly over the past two decades from a process of unstructured clinical judgement to one of structured assessment based on empirically tested instruments. A vast number of structured risk assessment instruments relating to violence in different populations have been developed over this period and attempts have been made elsewhere to summarise aspects of the literature relating to various instruments. This review adopted much broader inclusion criteria than previously used in order to capture and summarise data on the widest possible range of available instruments.

Objectives

The objectives of the review were to address two questions: (1) what features (i.e. population, instrument, outcome measure and design aspects) are associated with a risk assessment instrument score being significantly associated with a violent outcome? and (2) which risk assessment instruments have the highest level of predictive validity for a violent outcome?

Methods

Data sources

Evidence on the relationship between scores on a structured instrument and the occurrence of violence was identified using both a comprehensive search strategy to interrogate 19 bibliographic databases and the checking of reference lists of identified reviews. The database searches covered the period from 2002 to 2008.

Inclusion criteria

The inclusion criteria for papers were purposefully broad to capture as wide-ranging a selection of relevant studies as possible. Studies had to evaluate a structured risk tool and report an outcome measure of interpersonal violence either directly (e.g. reconviction for a violent offence) or indirectly through a proxy measure (e.g. a validated anger instrument). Participants had to be aged 17 years or over and either have a mental disorder, be an offender, or have committed an indictable offence (without necessarily having been prosecuted, e.g. pre-court diversion schemes, ‘dating’ violence self-reported purely in the context of the research study).

Data extraction

Data extraction was carried out independently by nine reviewers, with regular meetings to co-ordinate activity and to explicitly cross-check extracted data. Data from each included study (n = 959) relating to study design, sample, setting, type of intervention, type of outcome and whether or not a statistically significant outcome was reported were extracted into a predefined Statistical Product and Service Solutions database. Data from a subset of studies (n = 65) which reported area under the curve (AUC) statistics were independently extracted into a Microsoft Excel (Microsoft Corporation, Redmond, WA, USA) spreadsheet by one reviewer and cross-checked.
Data synthesis
A series of bivariate analyses using either a chi-squared test or Spearman’s rank-order correlation were conducted to explore possible sources of variance in whether or not an instrument was found to have good predictive validity (i.e. scores on the instrument were significantly related to a violent outcome). Data from the AUC subset studies were combined to provide estimates of mean validity.

Results
For the overall set of included studies (n = 959), 59% adopted a cross-sectional design and 32% were longitudinal with follow-up of a single group (median follow-up period of 20.2 months).

One-third of longitudinal studies lost 20% or more of their participants between the two time points and the median sample size at end point was 146.

Beyond issues of validity, less than half of the studies (44%) reported data on reproducibility (e.g. test–retest reliability) and less than 2% examined issues of clinical utility (e.g. acceptability to users).

Over half (52%) of the studies were conducted in the USA and over three-quarters (77%) were conducted in the USA, Canada or the UK.

The mean proportion of males in study samples was 79% and the median proportion of participants classified as ‘white’ and/or ‘Caucasian’ was 61%. The median age was 34.1 years. There were nearly 100 studies of female-only samples and over 500 studies of male-only samples.

Two-thirds of all studies were conducted with offenders who had either no formal mental health diagnosis (43%) or forensic samples with a formal diagnosis (25%). Fewer studies were conducted with non-indicted perpetrators (17%) and people with a formal mental health diagnosis but no offence history (10%). With mental health samples, mixed diagnoses were the most common group (50%) and with offender samples, mixed offences were most common (34%). Where studies focused on a single diagnostic group, schizophrenia (9%) and personality disorder (9%) were most common. Where studies focused on a single offence category, sexual violence (21%) and domestic violence (21%) were more common than general violence (12%).

The following structured tools were tested in the most studies: Psychopathy Checklist-Revised (n = 192 studies); Conflict Tactics Scale (n = 56), STATIC-99 (n = 54); Historical-Clinical-Risk-20 (n = 51); Violence Risk Appraisal Guide (n = 45); State-Trait Anger Expression Inventory (n = 42); Millon Clinical Multiaxial Inventory (n = 40); Minnesota Multiphasic Personality Inventory (n = 36); Beck Depression Inventory (n = 35) and Barratt Impulsiveness Scale (n = 34).

Most studies (82%) started and ended in the same setting. The most common setting was an offenders’ institution such as prison (25% of study start points and 20% of study end points).

Most studies (83%) assessed violence outcomes using a scale score (51%) or official data such as reconviction rates (32%).

In terms of crude predictive validity, most studies (78%) reported a statistically significant (p < 0.05) relationship between the instrument score and a violent outcome. Only three variables were associated with the tendency to report such a relationship: prospective data collection (χ² = 4.4, p = 0.035), number of people recruited (U = 27.8, p = 0.012) and number of participants at end point (U = 26.9, p = 0.04). With only three relevant variables (including two which were not independent), no multivariate analysis of validity predictors was conducted.
The mean AUC value in the subset of studies reporting such values \( (n = 65) \) was 0.69 [standard deviation (SD) = 0.08] and AUC values ranged from 0.44 to 0.88. For those instruments tested in more than one study reporting AUC values, the Broeset Violence Checklist achieved the highest mean AUC value (0.81), albeit over a very short (24-hour) period. For those instruments tested in five or more studies reporting AUC values, the relative AUC values were as follows: General Statistical Information on Recidivism (0.73); Violence Risk Appraisal Guide (0.72); Sex Offender Risk Appraisal Guide (0.71); Level of Service Inventory (0.69); Psychopathy Checklist-Revised (0.69); Historical-Clinical-Risk Management-20 (0.69); Psychopathy Checklist Screening Version (0.68); STATIC-99 (0.66); and Rapid Risk Assessment for Sexual Offender Recidivism (0.64).

**Conclusions**

A very large number of studies examining the relationship between a structured instrument and a violent outcome were published in this relatively short 7-year period. The general quality of the literature is weak in places (e.g. over-reliance on cross-sectional designs, high attrition in longitudinal studies, lack of information on clinical utility, little evidence on cross-cultural transferability, avoidance of follow-up from one setting to another between start and end points, over-reliance on self-report scales for both predictor and outcomes) and a vast range of distinct instruments have been tested to varying degrees. However, there is evidence of some convergence around a small number of high-performing instruments and identification of the components of a high-quality evaluation approach, including AUC analysis. The upper limits \( (AUC \geq 0.85) \) of instrument-based prediction have probably been achieved and are unlikely to be exceeded using instruments alone.

**Recommendations for future research**

1. The small number of tools that already have demonstrable replicated efficacy should be tested out on a wider range of populations. This expansion should include empirical testing beyond North America and the parts of Europe where they have been extensively tested.

2. There should be a strong case made for expending significant effort on developing and testing any new risk assessment tools given the proliferation of tools developed over the past 15 years. There will always be a tension between a ‘one size fits all’ philosophy in which three to four dominant instruments with extensive empirical support are seen as suitable for all populations and a ‘bottom-up’ approach which recognises that many different tools (including those with minimal evidence) are needed to reflect the complexities of variations across populations.

3. Cross-sectional studies and/or studies relying purely on scale scores should be avoided. Too much of the existing literature is based on correlating a predictor and an outcome occurring simultaneously. This prevents the testing of any causal hypotheses and thus does not help in the development of theoretical frameworks for understanding violence. The problem is compounded when both the ‘predictor’ and the ‘outcome’ are measured using self-reported experiences recorded on a scale, as opposed to observable hostility or violence. The validity of such scale measures is lower than that of behavioural outcomes.

4. More studies should be conducted prospectively from hospital/prison to the community to examine the potential support of risk assessment tools for discharge/release decisions. While the prevention of intra-institutional violence is important in terms of protecting staff and other patients, it is the transition from hospital or prison to the community which is of most significance for the patient/prisoner and society at large. It is also more challenging to achieve effective prediction when moving from one environment to another and methodologically more difficult to keep track of participants. But clinical decisions on release or discharge are core issues faced by professionals and better research over this transition period is essential.

5. Clinical utility of those instruments with a strong evidence base in terms of predictive validity should be assessed to contextualise this information. While good predictive validity is a core component of an effective instrument, there are a number of other aspects which must be present for the instrument to...
be considered entirely effective. These include the availability of a user manual, reasonable cost, available training, specified user competencies (including training and specified qualifications and skill levels), ease of use, appropriate administration time and recognition of protective factors. Some of these aspects can be studied as part of the overall research evaluation of specific tools.

6. The findings from the female-only studies should be examined and summarised separately. The pathways to violence and consequences following from it are likely to be different for females compared with males. Given the identification here of a large literature of female-only samples, there is scope for a powered analysis of this topic on its own to examine differences from the male-only samples.

7. A statistical procedure [similar to Cohen’s $d$ (e.g. Rice ME, Harris GT. The size and sign of treatment effects in sex offender therapy. Ann N Y Acad Sci 2003; 989:428–40; discussion paper 41–5) for intervention effect size] should be developed for aggregating across AUCs. Mean AUCs were calculated for this study in order to aggregate across replications but this is a rather simplistic approach, especially when there are few studies. Effect size for intervention studies, drawing on means and SDs, is a more robust statistic and the research effort in the area of risk assessment would benefit from a similar approach.

**Source of funding**

Funding for this study was provided by the National Institute for Health Research Health Technology Assessment and Research for Patient Benefit programmes.
Chapter 1  Background

This review systematically examines the research literature on structured violence risk assessment instruments designed for use in mental health services or the criminal justice system published between 2002 and 2008. In this introductory section, we provide context for the technical analysis by examining the concepts of violence and risk assessment in both mental health care and criminal justice settings. We also examine the key debates around the appropriate use of such approaches to reduce risk and improve opportunities for managing violent people more effectively.

This review is a companion to a previously published review that examined prevention and intervention strategies for populations at high risk of engaging in violent behaviour.1

Definition of violence

In a general sense, many would consider violence to consist of the use of physical force that is intended to hurt or injure another person.2 However, this arguably limited conceptualisation ignores the more insidious effects of non-physical violence, such as threats and intimidation. It has been suggested that there may be several approaches to the definition of violence,3 although at present there is no widely held agreement on which of these is most appropriate. In this document, we have adopted, with some modifications, the broad conception first coined in 2002 by the World Health Organization,4 which has defined violence as:

The intentional use of physical force or power, threatened or actual, against . . . another person, . . . that either results in or has a high likelihood of resulting in injury, death, psychological harm, maldevelopment or deprivation.4

This broader definition thus includes threats, intimidation and abuse (whether physical, sexual or psychological) and conceives of violence in terms of its concrete outcomes on health and well-being.

Excluded from the concept here are acts of collective violence committed as a concomitant of war, terrorism or gang conflict.

What is risk assessment in relation to the criminal justice system and mental health services?

At a theoretical level, risk assessment has been defined as: ‘the process of gathering information via personal interviews, psychological/medical testing, review of case records, and contact with collateral informants, for use in making decisions pertaining to an individual's risk and its most appropriate, effective, and proportionate prevention or minimization’ (p. 295).5

This definition clearly goes beyond the simple prediction of risk, to include action-planning designed to minimise or prevent the development of factors that give rise to risk. Over the past two decades, the activities of professional staff working in mental health services and to some extent those working with general offender populations have become increasingly dominated by the unreasonable public expectation that all risks can be not only predicted but also effectively managed. However, in addition to predicting and managing risk, staff working in mental health services face further challenges.
In terms of assessing risk of violence, they are also expected to consider their clients’ broader well-being. Beyond simply addressing risk with a seemingly coercive response, such as continued detention in a psychiatric hospital, clinicians are asked to develop strategies for positive risk management in collaboration with service users.

**Why is risk assessment important to the criminal justice system and mental health services?**

The case for improved risk assessment approaches with violent offender populations is self-evident. In criminal justice settings, forms of risk assessment have been in use for several decades. Traditionally in criminology, they were used primarily to aid decisions such as allocation to different types of prison regimes or concerning applications for parole. Hollin provides an overview of some of the methods developed in this field. They include, for example, the Salient Factor Score and other measures developed as long ago as the 1960s and 1970s. Subsequently, prediction methods also began to be used in the process of selection for different types of intervention programmes, combining information about risk of recidivism with assessment of the factors thought likely to be associated with it. The latter became known as ‘risk–needs assessment’. The evidence base supporting the value of risk-based prediction is now fairly extensive, and the use of the process has flourished vigorously since it was first introduced. Within this field, there were also specific attempts to develop methods that would be specifically helpful in predicting violence. Different variables were found to be helpful for the prediction of institutional violence and post-release recidivism, respectively.

As criminal justice has traditionally been more focused on the ‘processing’ of offenders than on the direct contact and individual assessment that would be expected in mental health services, there has traditionally been more emphasis on actuarial approaches. One result of this is that data sets pertaining to criminal justice prediction have been very much larger than those typically used in mental health research. To some extent, the process became almost automated, with risk measures developed for high-turnover tasks, and relative speed and ease of use involving minimal exercise of judgement and little specialised training. Leading examples of this include the General Statistical Information on Recidivism (GSIR) measure developed in Canada and the Offender Group Reconviction Scale (OGRS) developed in England and Wales. The second version of the latter (OGRS-2) was formatted as a computer program entailing entry of just seven pieces of information, yielding a 2-year likelihood of reconviction score. The scale is now in a third edition. In side-by-side comparisons in the UK, OGRS-2 and criminal history variables in general emerged as more accurate predictors of recidivism among forensic mental health populations than other measures specifically developed for use in those settings. Overall, therefore, there are strong reasons for reviewing the literature in these fields: to test the overall efficacy of risk assessment, with particular reference to violence, and to compare the predictive power of different categories of variables and their relative predictive accuracy in different service settings.

The issue is more contested with regard to working with people with mental health problems, as the relationship between violence and mental health is highly complex and overestimating risk can lead to unjustifiable stigmatisation and social exclusion. More than 30 years have elapsed since it was claimed that mentally ill people are no more likely to engage in violent behaviour than the general population. Since that time, a considerable amount of research has indicated that this is true for some groups but not for others. For example, research has recently suggested that there is an increased risk of violent offending in people with a serious mental disorder. Trends over time are no less problematic. More than 10 years ago, it was observed that the number of homicides in England and Wales committed by persons with serious mental disorders had steadily declined over a 38-year period. Despite this, acts of violence committed by people with mental illness remain a matter of continuing major concern to the public, as well as to service providers and policy-makers.
Recent large-scale reviews suggest that some diagnosed mental disorders, notably schizophrenia and other psychoses, are associated with an increased risk of violence. Fazel et al.\textsuperscript{22} reviewed 20 studies with an aggregate sample of 18,423 individuals and, after discounting the influence of concurrent substance abuse, found an odds ratio of 2 : 1 for the relationship between active schizophrenia and violence. Douglas et al.\textsuperscript{23} reviewed a total of 204 studies, subsuming 166 independent samples, and concluded that ‘psychosis was reliably and significantly associated with an approximately 49\% to 68\% increase in the odds of violence relative to the odds of violence in the absence of psychosis’ (p. 687).\textsuperscript{23}

Violence by people with mental health problems in the community usually generates extensive public anxiety and debate but aggressive behaviour by psychiatric inpatients can also have serious consequences. These range from interfering with the quality of the therapeutic milieu to endangering the physical safety of other patients and health-care staff. Rates of aggressive behaviour vary as a function of the way they are defined, the form the behaviour takes and the setting.\textsuperscript{24}

When measuring a broad range of problematic behaviours in people with mental health problems, rates from 13\%\textsuperscript{25,26} to 60\%\textsuperscript{27} have been reported. When violence occurs, this can be the result of both distal and proximal factors.\textsuperscript{28,29} Distal factors are those that are removed from the violent situation that is occurring, such as a history of childhood abuse. Proximal factors are those chronologically situated near the time of the violent event, such as being intoxicated or experiencing command hallucinations. Individual and personal factors are important variables in contributing to violence and have been the focus of much research, but it is also important to include situational variables,\textsuperscript{30} which can include those in the immediate environment, such as the temperature of the room or the culture within a treatment setting. A broad-based risk assessment would need to take account of all these factors.

Violence risk management in mentally ill individuals is a critical aspect of a clinician’s daily responsibilities in many psychiatric inpatient settings.\textsuperscript{31} Successful risk management in mental health care is based on the complex task of drawing together knowledge and expertise from many different sources, and risk assessment is therefore an essential skill for mental health professionals.\textsuperscript{32,33} As Whittington and Logan\textsuperscript{5} have observed, in mental health services it is an understanding of the risks posed that lies at the heart of effective risk management.

**Approaches to predicting violence: the recent history**

It is well known that approaches to risk assessment prior to the 1990s largely relied upon unstructured clinical judgement in which the subjective impression of the clinician was used to estimate the likelihood of future violence. While implicit expertise gained through extensive experience remains a key part of the risk management process, it became apparent from the 1970s onwards that unaided clinical judgement was highly subjective, prone to bias and could sometimes be no more accurate than chance.\textsuperscript{34} This led to the empirical testing of structured instruments from the 1990s onwards, a process that has grown enormously over the past decade and has developed through a number of stages.

**Actuarial approaches**

A number of risk assessment instruments (e.g. STATIC-99) were developed in the 1990s as part of the ‘first generation’ of such tools, with items derived using an actuarial approach. These tools focused particularly on risk prediction rather than on the actions that might be taken to prevent or ameliorate violent incidents. The adoption of statistical techniques such as the receiver operating characteristics (ROC) curve\textsuperscript{35} made it possible to test the effectiveness of these tools in terms of predictive validity and improved the potential to estimate this aspect. Predictive validity is only one aspect of an effective instrument, and several other dimensions (e.g. clinical utility, user acceptability\textsuperscript{36}) are also relevant, but the most effort has been invested
in testing predictive validity. The most common tools were those where the prediction was derived purely from a mathematical aggregation of empirically established risk factors – known as ‘the actuarial approach’. The preference for these methods in risk prediction may have lain in their apparent base in evidence, although their usefulness in directly supporting work with specific individuals is limited to locating that individual's propensity for violence in relation to that of a reference population. Actuarial tools are limited to static factors (e.g. age at first offence) which are not amenable to therapeutic intervention and thus are less useful than dynamic factors as a basis for clinical practice. Despite these criticisms, however, it has been shown that the accuracy of actuarial approaches to the prediction of violence in people with a mental disorder improved in the 30 years between 1970 and 2000.

Structured professional judgement

Actuarial approaches to risk assessment improved the accuracy of the prediction of risk over clinical judgement alone. However, the accurate prediction and management of risk for a specific individual is considerably more complex; estimating the probability of violence in a homogenous group is quite different from estimating it for one specified member of that group. This difficulty has been addressed in recent years by the development of approaches which include a focus on risk prevention (rather than prediction), risk formulation and risk management, with professional judgement taking account of environmental and individual factors as they pertain to a particular person. As stated by Logan:

An alternative way of deriving a risk estimate or judgement is to make a structured professional judgement about the individual's potential to be harmful in the future based on an appraisal of all the present factors. This judgement may be structured very simply by the professional appraisal of the risk factors that are present – the judgement of high, medium or low risk is deduced from the pattern of risk factors identified and the significance given to them by the practitioner undertaking the assessment. However, the risk judgement can be more substantially structured by involving a formal process of formulation which organises the information derived about prior harmful conduct into an explanation for why it happened as it did and when and therefore the circumstances in which it could potentially happen again. In such structured formulations, risk estimates or judgements (high, medium or low risk) are in fact obsolete because what is prepared is a plan of action for continuously monitoring risk and adjusting risk management. This latter process is structured professional judgement at its most refined.

It has been suggested that a major step forward in the approach was achieved with the publication of The Risk for Sexual Violence Protocol in 2003. Within a very short space of time, structured professional judgement (SPJ) had become so entrenched in service functioning that in 2006 the Scottish Risk Management Authority advised that actuarial assessments should not be used in isolation from SPJ because actuarial tools give little guidance regarding the risk level for a specific individual.

To determine an individual's risk using SPJ, the assessor(s) consider(s) a minimum set of variables, to provide an evaluation of an individual's level of risk. Logan identifies six steps in conducting a formal SPJ: (1) information gathering; (2) judging the presence of an identified risk factor; (3) judging the relevance of each identified risk factor for this client; (4) clinical judgement of protective factors; (5) selection of risk management strategies; and (6) statement of a summary judgement. The individual items that go to make up the total score are given considerable attention in the process of SPJ, it being accepted that total scores alone have limited value in drawing up a detailed risk management plan. In addition to this, the assessor also takes account of any case-specific factors or compelling evidence that may affect the individual's risk level. Thus, the assessor has the freedom to include additional information pertaining specifically to the circumstances of the individual person at risk.
While SPJ has its advocates (such as Douglas et al.48), the approach has not been without its critics, who argue that assessor discretion in the decision to include or exclude certain information is far too subjective. Harris et al.49 have stated their view that:

Research . . . shows that clinicians’ impressions of dangerousness, insight, treatment response, and so on, are, at best, very weakly related to violent recidivism. Combining actuarial scores with clinical judgments inevitably produces lower accuracy than actuarial scores alone. Therefore, we recommend that clinical judgment not be blended with actuarial scores, actuarial scores not be used only as components to clinical judgment, and clinical judgment not be used to decide which patients receive actuarial assessment.

Despite the ongoing debate, throughout the past decade SPJ has grown in its importance as an approach to the prediction, prevention and management of violence. Whereas previous approaches tended to focus only on violence prediction, SPJ has been credited with refocusing attention on prevention and management, and research has generally supported SPJ instruments in terms of their predictive validity for recidivism and their clinical value in practice.50 Unsurprisingly, tools developed during the last decade have been increasingly focused on supporting clinicians in making sound professional judgements about the best approaches to intervention. This shift away from tools that focus purely on risk prediction has been accompanied by a broadening in the criteria that researchers must apply in deciding which tools function most adequately. Although predictive validity remains important, other factors, such as cost, level of training required, availability of training manuals, ease of use and inclusion of protective factors, have grown in importance.51

The majority of instruments designed to assist in SPJ, such as the Historical Clinical Risk Management-20 (HCR-20),52 consist of both static and dynamic factors. These tools have been designed to identify the mechanisms at work within potentially violent people. They therefore provide a platform for the development of formulations specifying the precise circumstances potentially leading to an individual episode of violence and tailored interventions to reduce the level of risk, rather than solely predicting the risk of violence.

The value of SPJ in clinical practice is yet to be clearly established and the use of structured assessments in practice, while widespread and increasing, is still not universal. It is recognised that, even with the improvements incorporated into the SPJ approach, structured assessments cannot provide the whole clinical picture. Clinical judgements concerning a patient’s level of safety are notoriously complex and rarely clear-cut: what might decrease the risk of violence to the spouse of a potentially violent patient, for instance, such as the spouse moving out of their shared accommodation, may also isolate the patient and remove an important source of social support, and so lead to an increased risk of violence more generally.

**Protective factors**

It is important to note that even patients who score highly on risk assessment instruments of either type do not necessarily recidivate, and some have suggested that protective factors may play a role in this phenomenon.53 Protective factors are variables that are thought to reduce the effect of risk factors or influence the outcome independently.54 As with risk factors, protective factors can be static for an individual, such as intelligence or secure childhood attachment, or they can be dynamic, such as coping skills, social network availability and provision of supportive professional care. Some clinicians hold that protective factors are as important to clinical practice as risk factors.55 There remains considerable debate around the most appropriate conceptualisation of protective factors: some have argued for them to be thought of as the absence of a risk factor56 while others conceive of them more as being at the opposing end of a continuum from risk factors.57 Yet others have suggested that protective factors can exist in the absence of a corresponding risk factor.58 Whatever the true dynamic, there is growing evidence that
instruments utilising the concept of protective factors can have an important role in complementing SPJ assessments of risk for future violence.\textsuperscript{53,59} There are grounds for arguing that a balanced and comprehensive assessment requires consideration of both risk factors and protective factors\textsuperscript{60} and that treatment should not only aim to reduce the impact of risk factors but also reinforce protective factors where possible.\textsuperscript{61,62}

**Rationale for the review**

There are two principal reasons why the present review was planned and undertaken.

The first was that it was considered necessary and timely, given the increasing accumulation of studies pertaining to the risk assessment process. The structured violence risk assessment literature is extensive and rapidly expanding. In a 2009 literature search, for example, Singh and Fazel\textsuperscript{63} identified over 6000 relevant records, representing a 300\% increase in output over a 10-year period.

Systematic reviews and, where feasible, some form of meta-analysis are now the pre-eminent techniques for summarising evidence on a topic and are particularly valuable where evidence is proliferating so quickly that practitioners and policy-makers do not have the capacity to digest the literature. The risk assessment literature superficially appears well served in this respect, as more than 40 systematic reviews of violence risk assessment instruments have been published since 1995.\textsuperscript{63} However, the available reviews manifest two disadvantages. One is that their quality of execution is extremely variable, with many taking traditional narrative formats and not adhering to appropriate standards or procedures for locating, appraising and summarising the research evidence. The other is that their focus has typically been limited to a small range of pre-selected tools. For example, two of the largest recent reviews\textsuperscript{22,63} focus on six and nine instruments, respectively.

The second key reason for conducting the present review was therefore based on the adoption of a different, conceptually broader approach, which did not pre-select specified tools for inclusion. The review reported here is part of a larger project consisting of a suite of three reviews commissioned from 2001 onwards by the Department of Health [National Forensic Mental Health Research and Development Programme, National Institute for Health Research (NIHR) Research for Patient Benefit (RfPB); NIHR Health Technology Assessment (HTA) Programme].

An original review completed in 2002 employed ‘gold-standard’ methodology to systematically review the literature pertaining to interventions and risk assessment strategies for the management of violence in a widely defined population (offenders, people with mental health problems and offenders with mental health problems) and covered the period between 1955 and 2002. The methodology and results of this original review are available in a report\textsuperscript{64} and have been presented at a number of national and international conferences.\textsuperscript{65–69} The original review also formed the basis for national guidance on best practice in managing risk of self-harm and violence in mental health services\textsuperscript{70} and the Department of Health’s endorsement of various selected tools.\textsuperscript{71}

In 2008, the original team of reviewers was commissioned by the Department of Health to update the review using the same methods, apart from some minor adjustments, that is to say no hand search and no expert consultation in identification of papers. Some variables extracted from studies in this update had not been extracted in the original review. Four of the five expert reviewers worked on both the original review and the update.

For ease of reporting, the update was split into two reviews. First, a review of prevention and intervention strategies for populations at high risk of engaging in violent behaviour was conducted,\textsuperscript{1,72} and then the review of risk assessment strategies for violence in the same population was conducted and is reported here.
All of these reviews attempted to combine a wide set of inclusion criteria (regarding participants, interventions/assessments and outcomes) with ‘gold-standard’ levels of rigour in terms of study identification and data extraction. In particular, no exclusion criteria were set around the tools to be evaluated. Rather, a grounded ‘bottom-up’ approach was adopted to enable a sense of the entire field at this particular point in time. On these grounds, therefore, of the need for updating, and of the aim within any new review to maintain quality and breadth of coverage, it was considered that the present review was amply justified.

**Research questions**
Which structured assessment instruments within the broad domain of mental health and criminal justice have been examined in relation to violence (regardless of their prior explicit designation as a violence risk assessment instrument)?

What features (i.e. population, instrument, outcome measure and design) of these instruments are associated with a score being significantly related to a violent outcome?

Which of these instruments have the highest level of validity for predicting a violent outcome?
Chapter 2 Methods

Methods for reviewing clinical effectiveness

This review was conducted by a multidisciplinary team of reviewers with varying numbers engaged at any particular stage. The components of the review were conducted by teams of reviewers as follows: searches \((n=2)\); application of stage 1 inclusion criteria \((n=11)\); stage 2 inclusion criteria \((n=7)\); data extraction and cross-checking \((n=9)\); extraction of statistical outcomes \((n=5)\).

Search strategy

The search strategy (example shown in Appendix 1) was run on the 19 databases shown in Table 1. The first database to be searched was PsycINFO and the searches were run in April 2008. The last database to be searched was SIGLE (System for Information on Grey Literature in Europe) and the searches were carried out in November 2008. Where it was possible to limit searches, they were initially run without limits and then rerun limited to exclude ‘children OR animals OR editorials’. These results were then removed from the first searches. This method was used so that papers that had not been indexed on a term, for example ‘humans’, were not missed when running the searches.

As the searches were run, citations were imported into EndNote XIV (Thomson Reuters, CA, USA) sequentially. Owing to the limitations of EndNote XIV, duplicate references were deleted first electronically and then manually.

The reference lists of relevant reviews identified at inclusion were searched for additional relevant references.

Inclusion and exclusion criteria

The identified citations were assessed for inclusion at two stages. The criteria used are shown in Table 2.

Inclusion stage 1

As a test of inter-reviewer reliability, at stage 1 inclusion six reviewers independently applied the inclusion criteria to 200 citations and a Cohen’s kappa (Fleiss–Cuzick extension) was calculated \(\kappa = 0.63 \text{ [standard error (SE) = 0.019]}\); \(z\) \((\text{for } k = 0) = 34.24; p < 0.0001\). This result indicates satisfactory agreement on inclusion/exclusion decisions. Each new reviewer who joined the team was required to look at 100 citations that had previously been classified by a reviewer and agreement greater than 80% had to be achieved before they continued with applying the inclusion criteria. At this stage, an ‘inclusive’ approach was adopted, that is to say where there was doubt, a citation was included. Given the high level of agreement and the inclusive approach, further citations beyond the initial 200 were screened by only 1 of the 11 reviewers.

If a citation was excluded, it was possible to flag it as either a review article that needed the reference list checked (‘check’) or a paper of particular interest that should be obtained anyway (‘obtain’) to provide context for the review.

Acquiring papers

Electronic copies of papers identified during stage 1 for further examination were then downloaded where possible by the University of Liverpool’s interlibrary loans team. Where electronic copies were not available, paper copies were obtained either from the University of Liverpool’s library or through interlibrary loans at the British Library.
Inclusion stage 2

At stage 2, the inclusion/exclusion criteria were applied to the full papers identified at stage 1. To aid this process, a Microsoft Access (Microsoft Corporation, Redmond, WA, USA) database was developed using a front-page form with drop-down menus and tick boxes. It was at this stage that included papers were categorised into each of the two reviews noted above: risk or intervention. Furthermore, studies reporting not on structured risk assessment tools but rather on risk factors or unstructured tools were excluded from the review at this stage but retained for possible future analysis. This was a pragmatic decision based on the volume of literature identified, rather than a decision indicating any preference towards one method of risk prediction over another.

Again, as a quality control measure, all seven of the reviewers applied the stage 2 inclusion criteria to 50 papers and a kappa score was calculated [Cohen’s kappa (Fleiss–Cuzick extension): $\kappa = 0.62$ ($SE = 0.032$); $z$ (for $k = 0$) = 19.46; $p < 0.0001$]. This result again indicated satisfactory agreement between reviewers. Investigation of individual pairs of inter-rater agreement (Table 3) revealed that one reviewer (G)
**TABLE 2  Inclusion and exclusion criteria**

<table>
<thead>
<tr>
<th>Inclusion</th>
<th>Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active diagnosis of mental illness, learning disability or personality disorder or</td>
<td>Participants are members of the general public, with no identified mental illness and no recorded violent offence and no evidence of having committed an act of violence that would constitute an indictable offence</td>
</tr>
<tr>
<td></td>
<td>Substance abuse (including alcohol abuse) in isolation from any other diagnosis of mental illness is not to be defined for the purposes of the review as an active diagnosis of mental illness</td>
</tr>
<tr>
<td></td>
<td>Substance abuse (including and separately specified as alcohol abuse) is to be identified in relation to participant characteristics for the purposes of data extraction, as it is identified in primary studies</td>
</tr>
<tr>
<td>Offender (person subject to penal sanction) or</td>
<td></td>
</tr>
<tr>
<td>Person(s) known to have committed one or more acts of aggression constituting an indictable offence (whether or not an indictment has been made)</td>
<td></td>
</tr>
<tr>
<td>Aged 17 years or older</td>
<td>Aged 16 or younger</td>
</tr>
<tr>
<td>Any assessment based on a structured instrument and linked with a violent outcome (regardless of whether the instrument is explicitly identified as purposely designed for ‘violence risk assessment’)</td>
<td>Unstructured assessments and structured assessments (including ‘violence risk assessment’ instruments) where there are no data reported on a violent outcome</td>
</tr>
<tr>
<td>Studies that focus on a main target behaviour which is not other-directed aggression (the target behaviour may be self-directed aggression) but do include an evaluation of the intervention on other-directed aggression as a subsidiary focus are to be included</td>
<td>Studies focused solely on self-directed aggression, including self-harm and suicidal behaviours, are to be excluded</td>
</tr>
<tr>
<td>Any institutional setting/location</td>
<td>Setting/location of any study is not to be regarded as grounds for excluding that study</td>
</tr>
<tr>
<td>Any community setting/location</td>
<td></td>
</tr>
<tr>
<td>Community-based ‘institutional’ settings such as outpatient clinics, A&amp;E, private practice clinics, etc. are also to be included</td>
<td></td>
</tr>
<tr>
<td>Studies conducted at ‘remote’ locations, e.g. studies evaluating interventions conducted by telephone or in writing, are also to be included</td>
<td></td>
</tr>
<tr>
<td>Any design explicitly measuring outcomes linked to a risk assessment meeting the above criteria</td>
<td></td>
</tr>
<tr>
<td>Directly observed physical or verbal aggression by person(s) with an identified mental illness</td>
<td></td>
</tr>
<tr>
<td>Directly observed physical aggression (meeting criteria for indictment) by members of the general public or current/previous offenders</td>
<td></td>
</tr>
<tr>
<td>Proxy measures of the above (including but not restricted to: self or other report of the above categories of behaviour, including reports established via clinical records; official records of offence and conviction; psychometric and other scale-based outcomes of mentations or behaviours directly relevant to aggression, e.g. BPRS measures of ‘hostility’)</td>
<td></td>
</tr>
</tbody>
</table>

continued
had poorer reliability scores but that this was as a result of their being more inclusive than the other
reviewers. Therefore, it was decided that there was high enough agreement to continue with
single-reviewer application of inclusion criteria and little risk of unjustified exclusion.

Quality assessment

Owing to the diverse nature of the papers included in this review, no appropriate methodological quality
assessment tool was available. Therefore, a range of variables selected by the research team as pertinent to
quality assessment were extracted as part of the full data extraction process (see Chapter 4, Quality assessment).

Data extraction

Data extraction was carried out independently by nine reviewers overall, with regular meetings to
co-ordinate activity and to explicitly cross-check extracted data. Data from each study relating to study
design, population, aspects of the tool(s) studied and outcomes (including the key dimension of whether
or not a statistically significant outcome was reported) were extracted into a predefined Statistical Product
and Service Solution (SPSS) (version 16, SPSS Inc., Chicago, IL, USA) database.

<table>
<thead>
<tr>
<th>Inclusion</th>
<th>Exclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outcome evaluation must be based on individual-level data</td>
<td>Evaluations based on ‘non-attributable’ rates and other summary data are to be excluded</td>
</tr>
<tr>
<td>‘Collective’ acts of aggression, such as terrorism, ‘gang’ violence, organised violent crime, football violence, drug feuds, etc., are excluded from consideration by the review where the focus of the study is on the phenomenon as a collective behaviour</td>
<td></td>
</tr>
</tbody>
</table>

| Evaluation of both imminent and non-imminent (future) violence is included within the review | Studies published in a language other than English |

<table>
<thead>
<tr>
<th>Rater</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0.618</td>
<td>0.860</td>
<td>0.753</td>
<td>0.660</td>
<td>0.711</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0.537</td>
<td>0.702</td>
<td>0.685</td>
<td>0.570</td>
<td>0.421</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.683</td>
<td>0.684</td>
<td>0.684</td>
<td>0.680</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>0.571</td>
<td>0.628</td>
<td>0.477</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>0.523</td>
<td>0.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>0.355</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The SPSS database included free-text variables, numerical variables and drop-down menus. The reviewers were trained in its use and a pilot extraction exercise was conducted. Relevant changes were made to the pilot database and then reviewers were retrained. This iterative process was repeated until the final version of the database was agreed. Ongoing support was also given to reviewers in the form of a crib sheet covering each variable, and an online ‘wiki’ forum was set up so that reviewers could post any queries for the expert reviewers to address.

Each paper was printed out and data pertaining to the basic aspects of the study were marked up on the papers and extracted into the SPSS database. The data extracted were then cross-checked by another reviewer using the marked paper and any disagreements were noted in a Microsoft Excel (Microsoft Corporation, Redmond, WA, USA) spreadsheet. The two reviewers then discussed the disagreements and where no consensus could be found a third reviewer adjudicated.

A subset of the included studies which reported area under the curve (AUC) analyses was then given to one of the expert reviewers in order for them to extract the outcome data into a Microsoft Excel spreadsheet. Data on the following variables for each tool with an AUC analysis were extracted by a single expert reviewer: tool name, AUC value, 95% confidence limits of the AUC value, nature of outcome variable (e.g. arrests) and length of follow-up. These data were then cross-checked by a second reviewer before being merged with the main study SPSS database. Subgroup analyses were not extracted where an included full analysis was reported.

Descriptive data
As a first step in analysis, descriptive analysis of frequencies for categorical variables and means, etc., for continuous variables was conducted to gain an overview of the literature. Details of key variables pertaining to quality, design characteristics, participant characteristics and risk assessment tool characteristics are tabulated and discussed in Chapter 4.

Statistical methods
In the next step, the key variables, as discussed in the descriptive section, were explored in terms of their relationship with whether or not a study reported a statistically significant result. A significant result was defined as one in which scores on the tool were associated with the occurrence of the violent outcome beyond chance levels. The alpha level throughout is 5%. These subgroup analyses are reported in Chapter 5 (see Bivariate analyses) and should be seen as hypothesis-generating rather than confirmatory analyses.

Bivariate and multivariate analyses

Bivariate analyses
A series of bivariate analyses using either a chi-squared test (for dichotomous data) or a Mann–Whitney U-test (for continuous data) were conducted to test the relationship between each relevant variable and the occurrence of a statistically significant result.

Multivariate analyses
A binary logistic regression, with categorical variables coded as ‘dummy’ (0/1 with 0 as the baseline category) and ‘whether or not a statistically significant outcome was reported’ as the dependent variable, (also coded 0/1) was planned if deemed appropriate.
Area under the curve analyses
In order to make comparisons between the relative predictive success rates in those studies in which a structured, formally developed risk assessment instrument was used, analyses focused on the reporting of the AUC statistic. The AUC was developed by extension from the model of signal-detection used in perceptual research. In that model, an accurate outcome result is considered as a signal (here, accurate prediction) that has to be distinguished from background noise (here, sources of error). The measuring tool is conceptualised as a ‘receiver’ and its degree of success in accurate prediction is calculated through an assessment of its ROC analysis. This compares the proportion of people in a sample predicted to commit violent acts who went on to do so with those who were predicted to act violently but did not. The measure has the added advantage of being independent of sample size in any given study. The ratio of true positives (accurate predictions) to false alarms (predicted to be violent but not so) is plotted on a graph with the former as the vertical axis and the later as the horizontal axis. The area between the resultant curve and the diagonal straight line that represents chance accuracy (0.50) can then be measured, producing the AUC statistic.

In the subset of studies reporting AUC data, the mean values for AUC and the 95% confidence interval (CI) boundaries were calculated for those tools tested in two or more studies. These values were also simply listed for tools tested in only one study.

Advisory panel
As this review is part of a larger project, ‘Development of Evidence Based Guidelines for the Prevention of Violence in Mental Health Settings’ (EPOV), funded by the Department of Health RfPB Programme (RfPB grant reference number PB-PG-0407–13253), the steering group for this larger project acted as an advisory panel and provided support, answered specific questions as the review progressed and commented on a draft of this report.
Chapter 3 Overview of the literature

Selection of included studies

As shown in Figure 1, the electronic searches identified 127,550 citations. After deduplication, both within and between the databases, 102,267 citations had the inclusion criteria applied at stage 1. This resulted in 96,065 citations being excluded, 246 of which were reviews.

As a result of searching the reference lists of the 246 reviews, an additional 38 references were identified. Therefore, a total of 6240 papers had the inclusion criteria applied at stage 2.

The process of applying stage 2 inclusion criteria resulted in 3760 references being excluded from both the intervention and risk reviews, 276 studies being included in the intervention review only and 936 papers having a risk focus but not testing a structured risk assessment tool as required for this review. A further 330 studies were excluded at the data extraction stage when, on closer inspection, they did not meet the inclusion criteria (this is discussed further in Chapter 6, Strengths and limitations of the review). The remaining 938 papers met the inclusion criteria for the review and data were extracted. A further eight papers were identified as reporting data from samples in other included papers. The primary paper for each study in this situation (defined as the one with the later publication date) was retained, with any additional data reported in the linked paper combined while the linked paper itself was excluded. A list of excluded papers is available on request.

Of the 930 included papers, 28 included more than one study, resulting in 959 studies having data extracted. All of the analyses in this report are reported by study rather than by paper (i.e. n = 959).

Quality assessment

Design of studies

Of the 959 studies, 563 (58.7%) were concurrent/cross-sectional group comparisons conducted at a single time point and 305 (31.8%) were cohort studies with a single group being followed up. The remaining 91 (9.5%) studies were quasi-experimental, case–control design or other designs (Table 4).

Data collection

While a large proportion of studies were cross-sectional studies that used neither retrospective nor prospective data collection (n = 463, 48.3%), the remainder were divided as set out in Table 5 below.

Reproducibility

The authors of 489 (51%) studies made reference to or made it clear whether or not reproducibility (e.g. test–retest or inter-rater reliability) had been tested. This consisted of 427 (44.5%) studies that did attempt to estimate reproducibility and 62 (6.5%) that did not.

Attempt to assess broader utility

Attempts to test the broader utility of scales (e.g. acceptability to staff, ease of use) were not stated or unclear for the vast majority of studies (845, 88.1%). Where this was noted, 13 (1.4%) studies did attempt to test the broader utility of scales, whereas 101 (10.5%) did not.
FIGURE 1 Flow diagram of inclusion of studies.

TABLE 4 Design of studies

<table>
<thead>
<tr>
<th>Study designs</th>
<th>n  (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concurrent/cross-sectional group comparison</td>
<td>563 (58.7)</td>
</tr>
<tr>
<td>Cohort design</td>
<td>305 (31.8)</td>
</tr>
<tr>
<td>Quasi-experimental design</td>
<td>23 (2.4)</td>
</tr>
<tr>
<td>Case-control design</td>
<td>19 (2.0)</td>
</tr>
<tr>
<td>Other design</td>
<td>49 (5.1)</td>
</tr>
<tr>
<td>Total</td>
<td>959 (100)</td>
</tr>
</tbody>
</table>
Length of follow-up
The total length of follow-up was reported by 912 (95.1%) studies. There was a considerable range in length of follow-up. Most studies (\(n = 608, 63.4\%\)) either were cross-sectional with no follow-up period at all or evaluated extremely short-term risk (less than 24 hours). In marked contrast, a small proportion of studies (\(n = 48, 5.0\%\)) followed up participants for 10 years or longer, with the longest follow-up being 31 years. The mean length of time for which participants were followed up across all non-cross-sectional studies (\(n = 396\)) was 1533.89 days (approximately 50.46 months) [standard deviation (SD) = 2117.95 days/69.67 months, median = 608 days/20 months] (Figure 2).

Attrition
Attrition was calculable for 939 (97.9%) of the studies (including cross-sectional studies where attrition is not an issue): 629 (65.6%) reported no attrition and 182 (18.9%) had a drop-out rate between the minimum recorded figure (0.003%) and 20% of the starting sample. The remaining 128 studies reported attrition of > 20%.

Intention to treat
Of the 959 studies included in the review, 4.5% (\(n = 43\)) were analysed on an intention-to-treat (ITT) basis, 31.2% (\(n = 299\)) were not analysed on an ITT basis and 64.3% (\(n = 617\)) did not state whether or not they were based on an ITT analysis.

<table>
<thead>
<tr>
<th>Data collection methods</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrospective</td>
<td>211 (22.0)</td>
</tr>
<tr>
<td>Prospective</td>
<td>174 (18.1)</td>
</tr>
<tr>
<td>Not stated/unclear</td>
<td>16 (1.7)</td>
</tr>
<tr>
<td>Neither</td>
<td>463 (48.3)</td>
</tr>
<tr>
<td>Both</td>
<td>95 (9.9)</td>
</tr>
<tr>
<td>Total</td>
<td>959 (100)</td>
</tr>
</tbody>
</table>

**TABLE 5** Data collection method

![FIGURE 2 Total length of follow-up in months.](image-url)
**Study characteristics**

*Number of studies*

The number of studies published was relatively steady across the years, with an average of 126 papers being published in each full year reviewed (2003–7; Figure 3).

*Country location*

Studies were conducted in 26 different countries. The largest number of studies, more than half of the total (501 studies, 52.2% of the cohort), were carried out in the USA. Canada (125, 13%) and the UK (118, 12.3%) were other major contributors, with smaller proportions reported from Australia (25, 2.6%), Germany (24, 2.5%) and Sweden (20, 2.1%). One or more studies were conducted in each of 20 other countries. A small proportion of studies (17, 1.8%) were described as multinational (i.e. participants from more than one country), and for a small minority (13, 1.4%) the country of origin was not one of those listed. Country location was not stated in 19 (2.0%) studies. The publication levels for each country can be seen in Table 6.

*Participant characteristics*

Details of the characteristics of participants included in the studies are shown in Figures 4 to 6 and Tables 7 to 11.

*Number of participants*

The number of people approached or eligible to take part in the studies was reported in 334 (34.8%) of studies and ranged between 12 and 61,321. The mean number of people approached was 1159.2 (median = 250.5, SD = 4013.8).

The number of participants enrolled was reported in 946 (98.6%) of studies and ranged between 3 and 13,601. The mean number at recruitment point (initial sample size) was considerably lower than the number of potential participants approached or eligible at 434.4 (median = 157, SD = 1133.8). Thus, on average, only slightly more than one-third (37.5%) of potential participants were successfully recruited into study samples.

![Figure 3](image-url) Number of studies by year of publication. a, 2002 and 2008 were partial years.
<table>
<thead>
<tr>
<th>Country</th>
<th>n</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>501</td>
<td>(52.2)</td>
</tr>
<tr>
<td>Canada</td>
<td>125</td>
<td>(13.0)</td>
</tr>
<tr>
<td>UK</td>
<td>118</td>
<td>(12.3)</td>
</tr>
<tr>
<td>Australia</td>
<td>25</td>
<td>(2.6)</td>
</tr>
<tr>
<td>Germany</td>
<td>24</td>
<td>(2.5)</td>
</tr>
<tr>
<td>Sweden</td>
<td>20</td>
<td>(2.1)</td>
</tr>
<tr>
<td>Netherlands</td>
<td>16</td>
<td>(1.7)</td>
</tr>
<tr>
<td>Italy</td>
<td>11</td>
<td>(1.1)</td>
</tr>
<tr>
<td>New Zealand</td>
<td>11</td>
<td>(1.1)</td>
</tr>
<tr>
<td>Spain</td>
<td>9</td>
<td>(0.9)</td>
</tr>
<tr>
<td>Belgium</td>
<td>6</td>
<td>(0.6)</td>
</tr>
<tr>
<td>Brazil</td>
<td>5</td>
<td>(0.5)</td>
</tr>
<tr>
<td>Switzerland</td>
<td>5</td>
<td>(0.5)</td>
</tr>
<tr>
<td>China</td>
<td>5</td>
<td>(0.5)</td>
</tr>
<tr>
<td>Norway</td>
<td>4</td>
<td>(0.4)</td>
</tr>
<tr>
<td>Greece</td>
<td>4</td>
<td>(0.4)</td>
</tr>
<tr>
<td>Ireland</td>
<td>3</td>
<td>(0.3)</td>
</tr>
<tr>
<td>Israel</td>
<td>3</td>
<td>(0.3)</td>
</tr>
<tr>
<td>Austria</td>
<td>3</td>
<td>(0.3)</td>
</tr>
<tr>
<td>Finland</td>
<td>3</td>
<td>(0.3)</td>
</tr>
<tr>
<td>Republic of Korea</td>
<td>3</td>
<td>(0.3)</td>
</tr>
<tr>
<td>Argentina</td>
<td>2</td>
<td>(0.2)</td>
</tr>
<tr>
<td>Denmark</td>
<td>1</td>
<td>(0.1)</td>
</tr>
<tr>
<td>Japan</td>
<td>1</td>
<td>(0.1)</td>
</tr>
<tr>
<td>France</td>
<td>1</td>
<td>(0.1)</td>
</tr>
<tr>
<td>Iceland</td>
<td>1</td>
<td>(0.1)</td>
</tr>
<tr>
<td>Multinational</td>
<td>17</td>
<td>(1.8)</td>
</tr>
<tr>
<td>Other</td>
<td>13</td>
<td>(1.4)</td>
</tr>
<tr>
<td>Not stated/unclear</td>
<td>19</td>
<td>(2.0)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>959</td>
<td>(100)</td>
</tr>
</tbody>
</table>
The number of participants at the end point of the study was reported in 944 (98.4%) of studies. The
studies reporting the final number of participants reported on a total of 339,624 individuals: the smallest
study had three participants and the largest had 11,754 participants (Figure 4). Just over one-third (35%)
of studies included 100 or fewer people. The mean number remaining at study end points was 359.8
(SD = 869.1, median = 146).

**Demographics of participants**
The percentage of male participants was reported in 889 (92.7%) studies. The majority of study
populations consisted of males, with an overall mean of 78.7% of participants being male (SD = 32.44)
(Figure 5). Nearly 100 studies of female-only samples and over 500 studies of male-only samples
were included.

Participant age parameters were reported as follows: mean age of participants (n = 777 studies, 81.02%),
median age (n = 29 studies, 3.02%), both mean and median age (n = 24 studies, 2.5%) and SD (n = 630
studies, 65.69%). The mean age in years of all participants in the studies was 34.38 years (SD = 7.50 years,
median = 34.14 years), and individual study means ranged between 17 and 85.5 years, with SDs ranging
between 0.45 and 18.10 years and study median ages ranging between 9 and 43 years. For inclusion in
the review, studies needed to report a sample mean age of 17 years or over, so some studies included
younger participants. The minimum age of participants in each study ranged between 8 and 64 years, and
the maximum ranged between 18 and 104 years. Therefore, the youngest participant in any of the studies
was 8 years and the oldest 104 years (Table 7).

**FIGURE 4** Number of participants in complete data set.

**FIGURE 5** Percentage of participants who were male.
The percentage of participants who were described as Caucasian was reported in 536 (55.9%) studies with the mean percentage of Caucasian participants in the studies being 59.5% (median = 60.8, SD = 30.9, range = 0–100%) (Figure 6).

**Population**

Five broad population categories were established on the basis of the inclusion criteria. These were (1) people with a formal diagnosis of mental disorder, (2) offenders, (3) indictable offenders (i.e. those having committed an indictable offence but not having been charged), (4) forensic participants (i.e. those with a diagnosis of mental disorder and offender/indictable offender status) and, where multiple populations were studied, (5) any combination of these groups. The numbers of studies looking at each of these population types are shown in Table 8. Studies mainly included offenders (n = 413 + 1 + 1 + 1 + 2 = 431), followed by offenders with a mental disorder (forensic) (n = 239 + 1 + 1 + 1 + 1 = 243), those reported to have committed an indictable offence (n = 158 + 33 + 14 + 1 + 1 = 207), and people with a mental disorder but no offending history (n = 93 + 33 + 3 + 1 = 130).

**Diagnosis**

Studies reporting on any individuals with a diagnosis of a mental disorder (including forensic groups) involved a range of diagnostic groups. Participants with a ‘mixed diagnosis’ (n = 186, 19.4%) were the most frequently reported, followed by patients defined as having an ‘other’ single mental health grouping (n = 54, 5.6%). With regard to specific diagnoses, participants with personality disorders only were reported in 35 (3.6%) studies, participants with a diagnosis of schizophrenia or schizo-affective disorder only were studied in 35 (3.6%) studies, and participants with a diagnosis of dementia only were studied in 1 (0.1%) study (Table 9).

There were differences between the diagnoses of participants in the mental disorder-only group (n = 130) and the forensic group (n = 243). A higher percentage of participants in the mental disorder group had a diagnosis of schizophrenia or schizo-affective disorder (n = 22, 16.9%) compared with the forensic group.
(n = 13, 5.3%), while a higher percentage of participants in the forensic group had a diagnosis of personality disorder (n = 30, 12.3%) compared with the mental disorder group (n = 5, 3.8%). There were similar percentages of participants with mixed diagnoses (mental disorder n = 70, 53.8%, forensic n = 118, 48.6%) in the two groups.

### Offences
The index offences of participants differed greatly between the three sample groups of offenders, forensic patients and indictable offenders. Offender participants (n = 430) had been charged with predominantly a mixed group of offences (n = 167, 38.8%), followed by the specific offences of sex offending (n = 120, 27.9%) and domestic violence (n = 89, 20.7%). For studies including forensic participants (n = 241), mixed groups of offences were again most frequently reported (n = 121, 50.2%), followed by sex offending (n = 46, 19.1%). In the indictable group (n = 207), domestic violence was the most frequently reported offence type (n = 89 studies, 43.0%), followed by general violence (n = 62 studies, 30.0%) (Table 10). In total, 75 (8.7%) studies did not report the index offences of participants.

### Table 8: Number and percentage of studies reporting each population group

<table>
<thead>
<tr>
<th>Category of participants</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Offender only</td>
<td>413 (43.1)</td>
</tr>
<tr>
<td>Forensic only</td>
<td>239 (24.9)</td>
</tr>
<tr>
<td>Indictable offences only</td>
<td>158 (16.5)</td>
</tr>
<tr>
<td>Mental disorder only</td>
<td>93 (9.7)</td>
</tr>
<tr>
<td>Mental disorder and indictable offences</td>
<td>33 (3.4)</td>
</tr>
<tr>
<td>Offender and indictable offences</td>
<td>14 (1.5)</td>
</tr>
<tr>
<td>Offender and forensic</td>
<td>1 (0.1)</td>
</tr>
<tr>
<td>Forensic and indicable offences</td>
<td>1 (0.1)</td>
</tr>
<tr>
<td>Other combinations</td>
<td>3 (0.3)</td>
</tr>
<tr>
<td>All of the categories</td>
<td>1 (0.1)</td>
</tr>
<tr>
<td>Not stated/unclear</td>
<td>3 (0.3)</td>
</tr>
<tr>
<td>Total</td>
<td>959 (100)</td>
</tr>
</tbody>
</table>

### Table 9: Number and percentage of participants within each diagnostic group

<table>
<thead>
<tr>
<th>Diagnostic groups</th>
<th>Mental disorder, n (%)</th>
<th>Forensic, n (%)</th>
<th>Total, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schizophrenia or schizo-affective only</td>
<td>22 (16.9)</td>
<td>13 (5.3)</td>
<td>35 (9.4)</td>
</tr>
<tr>
<td>Dementia only</td>
<td>1 (0.8)</td>
<td>0 (0.0)</td>
<td>1 (0.3)</td>
</tr>
<tr>
<td>Personality disorder only</td>
<td>5 (3.8)</td>
<td>30 (12.3)</td>
<td>35 (9.4)</td>
</tr>
<tr>
<td>Other single mental health grouping</td>
<td>25 (19.2)</td>
<td>29 (11.9)</td>
<td>54 (14.5)</td>
</tr>
<tr>
<td>Mixed diagnostic groups</td>
<td>70 (53.8)</td>
<td>118 (48.6)</td>
<td>186 (50.4)</td>
</tr>
<tr>
<td>No specific diagnoses given</td>
<td>3 (2.3)</td>
<td>13 (5.3)</td>
<td>17 (4.3)</td>
</tr>
<tr>
<td>Not stated/unclear</td>
<td>4 (3.1)</td>
<td>40 (16.5)</td>
<td>53 (11.8)</td>
</tr>
<tr>
<td>Total</td>
<td>130 (100)</td>
<td>243 (100)</td>
<td>373 (100)</td>
</tr>
</tbody>
</table>
Substance abuse was poorly reported in most studies, with only 350 (36.5%) papers reporting whether or not current substance abuse was identified in participants. Of the papers reporting on substance abuse, 29 studies (8.3%) reported no substance abuse in their sample, 20 (5.7%) identified illicit drug abuse, 62 (17.7%) identified alcohol abuse and 184 (52.6%) identified both alcohol and drug abuse. A further 55 (15.7%) studies identified some form of substance abuse, but did not report on the nature of it (i.e. if it was drug or alcohol abuse). Further details of the number and percentage of all studies are shown in Table 11.

### Risk assessment characteristics

#### Types of risk assessment

A very wide range of tools have been studied in the literature. The presence of six tools in particular was specifically coded during data extraction, based on their relevance to policy-makers in previously commissioned research by the team71 (Table 12). By far the most widely used of these was the Hare Psychopathy Checklist-Revised (PCL-R) with 192 (20%) studies reporting its use. The Offender Assessment System (OASys) was not used in any study.

The majority of studies (n = 854, 89.1%) tested another tool either alongside one or more of these specific tools or not. The 10 most frequently reported of these ‘other tools’ are shown in Table 13. It can be seen that the three most frequently reported ‘other’ tools were the Conflict Tactics Scale (CTS), the State-Trait Anger Expression Inventory (STAXI) and the Millon Clinical Multiaxial Inventory (MCMI).

#### Setting

The start and end settings of studies are shown in Table 14. The term ‘setting’ here refers to the location where the risk assessment was conducted and, in the case of the ‘community’ category, under what conditions, that is to say under a probation order or under the compulsory supervision of a mental health practitioner or neither (e.g. a self-referring person concerned about their propensity for violence who is offered a risk assessment). The most frequently reported setting was an offenders institution such as a prison (start setting n = 244, 25.4%; end setting n = 189, 19.7%), followed by the community (start setting n = 159, 16.6%; end setting n = 160, 16.7%), and community probation (start setting n = 107, 11.2%; end setting n = 103, 10.7%). The majority of studies had the same start and end settings (n = 790, 82.4%). Of the 169 (17.6%) studies reporting different start and end settings, 56 (33.1%) began in offenders institutions and 29 (17.2%) in secure forensic mental health settings. Ninety-five (56.2%) of the 169 studies ended in mixed settings and 31 (18.3%) did not state the end setting.
### TABLE 11 Number and percentage of studies reporting on substance abuse

<table>
<thead>
<tr>
<th>Substance abuse</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Both illicit drug use and alcohol abuse identified</td>
<td>184 (19.2)</td>
</tr>
<tr>
<td>Alcohol abuse identified</td>
<td>62 (6.5)</td>
</tr>
<tr>
<td>Substance not specified</td>
<td>55 (5.7)</td>
</tr>
<tr>
<td>No substance abuse identified</td>
<td>29 (3.0)</td>
</tr>
<tr>
<td>Illicit drug use identified</td>
<td>20 (2.1)</td>
</tr>
<tr>
<td>Not stated or unclear</td>
<td>609 (63.5)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>959 (100)</td>
</tr>
</tbody>
</table>

### TABLE 12 Risk assessment tools studied

<table>
<thead>
<tr>
<th>Risk assessments used</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCL-R</td>
<td>192 (20)</td>
</tr>
<tr>
<td>STATIC-99</td>
<td>54 (5.6)</td>
</tr>
<tr>
<td>HCR-20</td>
<td>51 (5.3)</td>
</tr>
<tr>
<td>VRAG</td>
<td>45 (4.7)</td>
</tr>
<tr>
<td>SVR-20</td>
<td>12 (1.3)</td>
</tr>
<tr>
<td>OASys</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Other</td>
<td>854 (89.1)</td>
</tr>
</tbody>
</table>


### TABLE 13 Other risk assessment tools studied

<table>
<thead>
<tr>
<th>Risk assessments used</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTS</td>
<td>56 (5.8)</td>
</tr>
<tr>
<td>STAXI</td>
<td>42 (4.4)</td>
</tr>
<tr>
<td>MCMI</td>
<td>40 (4.2)</td>
</tr>
<tr>
<td>MMPI</td>
<td>36 (3.8)</td>
</tr>
<tr>
<td>BDI</td>
<td>35 (3.6)</td>
</tr>
<tr>
<td>BIS</td>
<td>34 (3.5)</td>
</tr>
<tr>
<td>RRASOR</td>
<td>25 (2.6)</td>
</tr>
<tr>
<td>MAST</td>
<td>24 (2.5)</td>
</tr>
<tr>
<td>LSI</td>
<td>24 (2.5)</td>
</tr>
<tr>
<td>BPRS</td>
<td>21 (2.2)</td>
</tr>
</tbody>
</table>

BDI, Beck Depression Inventory; BIS, Barratt Impulsiveness Scale; BPRS, Brief Psychiatric Rating Scale; LSI, Level of Service Inventory; MAST, Michigan Alcohol Screening Test; MMPI, Minnesota Multiphasic Personality Inventory; RRASOR, Rapid Risk Assessment for Sexual Offender Recidivism.
**Sources of outcome data**

Considering that official data, such as convictions, capture only a subset of all violent outcomes, studies evaluating risk assessment tools should also use a range of data sources where possible to establish outcomes most accurately. There was indeed a wide range of such sources used, with scale scores being used in almost half of the studies \( (n = 490, 51.1\%) \). This was followed by official data \( (n = 307, 32\%) \), ‘other measures’ (e.g. composite measures; \( n = 164, 17.1\% \)) and routinely collected data (e.g. inpatient violent incident reports; \( n = 95, 9.9\% \)). Self-report data and reports of significant others were used in 53 (5.5%) and 18 (1.9%) studies, respectively (Table 15).

**Above-chance prediction of violence**

A significant result, as defined in Statistical methods, was reported in 745 (77.7%) of studies and a further 12 (1.3%) found mixed results. No significant results were reported in 66 (6.9%) of studies. Other types of non-probabilistic statistics such as AUC were reported in 106 (11.1%) of studies (Table 16).
## TABLE 14 Start and end settings of studies

<table>
<thead>
<tr>
<th>Start settings</th>
<th>End settings</th>
<th>Community</th>
<th>Community probation</th>
<th>Community mental health</th>
<th>Outpatient living in community</th>
<th>Offenders’ institution, e.g. prison</th>
<th>Secure forensic mental health</th>
<th>Secure non-forensic inpatient ward</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community</td>
<td>Community</td>
<td>148</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Community probation</td>
<td>0</td>
<td>94</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Community mental health</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Outpatient living in community</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>19</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Offenders’ institution, e.g. prison</td>
<td>4</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>188</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Secure forensic mental health</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>60</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>Secure non-forensic inpatient ward</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Forensic mental health (not secure)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Open inpatient hospital ward</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>A&amp;E or psychiatric emergency service</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Nursing home</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mixed settings</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Not stated or unclear</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total n (%)</td>
<td>160 (16.7)</td>
<td>103 (10.7)</td>
<td>27 (2.8)</td>
<td>22 (2.3)</td>
<td>189 (19.7)</td>
<td>63 (6.6)</td>
<td>14 (1.5)</td>
<td></td>
</tr>
</tbody>
</table>

Note: studies with differing start and end settings are shown in shaded cells. A&E, accident and emergency.
<table>
<thead>
<tr>
<th>Start settings</th>
<th>Forensic mental health (not secure)</th>
<th>Open inpatient hospital ward</th>
<th>A&amp;E or psychiatric emergency service</th>
<th>Nursing home</th>
<th>Mixed settings</th>
<th>Other</th>
<th>Not stated or unclear</th>
<th>Total n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Community</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>5</td>
<td>159 (16.6)</td>
<td></td>
</tr>
<tr>
<td>Community probation</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>107 (11.2)</td>
<td></td>
</tr>
<tr>
<td>Community mental health</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>20 (2.1)</td>
<td></td>
</tr>
<tr>
<td>Outpatient living in community</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>32 (3.3)</td>
<td></td>
</tr>
<tr>
<td>Offenders’ institution, e.g. prison</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>38</td>
<td>0</td>
<td>8</td>
<td>244 (25.4)</td>
<td></td>
</tr>
<tr>
<td>Secure forensic mental health</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>0</td>
<td>2</td>
<td>89 (9.3)</td>
<td></td>
</tr>
<tr>
<td>Secure non-forensic inpatient ward</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>14 (1.5)</td>
<td></td>
</tr>
<tr>
<td>Forensic mental health (not secure)</td>
<td>12</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>0</td>
<td>2</td>
<td>20 (2.1)</td>
<td></td>
</tr>
<tr>
<td>Open inpatient hospital ward</td>
<td>0</td>
<td>22</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>30 (3.1)</td>
<td></td>
</tr>
<tr>
<td>A&amp;E or psychiatric emergency service</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>6 (0.6)</td>
<td></td>
</tr>
<tr>
<td>Nursing home</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1 (0.1)</td>
<td></td>
</tr>
<tr>
<td>Mixed settings</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>133</td>
<td>5</td>
<td>141 (14.7)</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>48</td>
<td>4</td>
<td>59 (6.2)</td>
<td></td>
</tr>
<tr>
<td>Not stated or unclear</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>33</td>
<td>37 (3.9)</td>
<td></td>
</tr>
<tr>
<td>Total n (%)</td>
<td>12 (1.3)</td>
<td>22 (2.3)</td>
<td>6 (0.6)</td>
<td>1 (0.1)</td>
<td>228 (23.8)</td>
<td>48 (5.0)</td>
<td>64 (6.7)</td>
<td>959 (100.0)</td>
</tr>
</tbody>
</table>

Note: studies with differing start and end settings are shown in shaded cells. A&E, accident and emergency.
### TABLE 15  Outcome measure used

<table>
<thead>
<tr>
<th>Measures of outcome</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale score</td>
<td>490 (51.1)</td>
</tr>
<tr>
<td>Official data</td>
<td>307 (32.0)</td>
</tr>
<tr>
<td>Other measure</td>
<td>164 (17.1)</td>
</tr>
<tr>
<td>Routinely collected data</td>
<td>95 (9.9)</td>
</tr>
<tr>
<td>Self-report</td>
<td>53 (5.5)</td>
</tr>
<tr>
<td>Report of significant other</td>
<td>18 (1.9)</td>
</tr>
<tr>
<td>Nurse-, doctor- or other staff-observed</td>
<td>10 (1.0)</td>
</tr>
<tr>
<td>Clinical judgement of improvement</td>
<td>2 (0.2)</td>
</tr>
</tbody>
</table>

### TABLE 16  Any significant result found?

<table>
<thead>
<tr>
<th>Any significant result?</th>
<th>n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>745 (77.7)</td>
</tr>
<tr>
<td>No</td>
<td>66 (6.9)</td>
</tr>
<tr>
<td>Yes and no</td>
<td>12 (1.3)</td>
</tr>
<tr>
<td>Other type of statistics, e.g. AUC</td>
<td>106 (11.1)</td>
</tr>
<tr>
<td>Unclear</td>
<td>11 (1.1)</td>
</tr>
<tr>
<td>No statistics</td>
<td>19 (2.0)</td>
</tr>
<tr>
<td>Total</td>
<td>959 (100.0)</td>
</tr>
</tbody>
</table>
Chapter 4 Results of analyses

Bivariate analyses

To explore the relationship between study variables and whether or not a study reported a statistically significant result, a series of bivariate analyses were conducted using the variables described in the previous section. The 823 studies reporting significant \((n = 745 + 12)\) or non-significant results \((n = 66)\) were included in these analyses. Given the heterogeneity of the studies, it is necessary to emphasise that these analyses are exploratory and designed to generate further hypotheses rather than in any way confirmatory.

Population

The population group chosen for evaluation did not have a relationship with a significant outcome effect although analyses focused on people with a mental disorder were overall more likely to report a significant outcome than those not focused on this population \((96.3\% \text{ vs. } 91.3\%)\). This association just failed to reach statistical significance \(n = 823, \chi^2 = 3.056, p = 0.051; \text{ Table 17}\).

Demographics

No association was found between a statistically significant study result and the key demographic variables of study mean age, proportion of males and proportion of participants who were Caucasian \(\text{ Table 18}\). In other words, a study was no more likely to report an effect (i.e. a statistically significant relationship between predictor and violent outcome) according to the sample profile in terms of sex, age or ethnicity.

Setting

The ‘setting’ of a study can be categorised in a number of ways (e.g. start setting, end setting, change between start and end, settings based on usual transition through the health-care system, etc.). Based on outcomes from a previous review\(^6\) we chose to simplify the comparisons drawn here, to reflect the initial choice of broad outcome setting identified in the analyses at their start point (i.e. start setting categorised into mental health, offenders’ institution, community and ‘other’). Categorised in this way, there were no statistically significant differences in respect of the outcomes based on start setting \(\text{ Table 19}\).

Outcome measure

With regard to statistically significant findings in relation to outcome measures, a significant finding occurred most frequently in studies using a scale score as an outcome measure \((n = 420)\). This was followed by those using official data \((n = 204)\), ‘other measures’ \((n = 133)\), and routinely collected data \((n = 61)\) \(\text{ Table 20}\). Due to small numbers in some cells it was not possible to calculate a chi-squared test for this analysis.

Study quality indicators

The only ‘quality’ variables that showed statistically significant associations with whether or not the study reported a significant result were whether or not data were collected prospectively \(\text{ Table 21}\), the number of participants recruited and the number of participants included in the final analysis \(\text{ Table 22}\).

Intention-to-treat analysis

Whether or not the analysis was an ITT had no association with whether or not the study reported a statistically significant result, as shown in \(\text{ Table 21}\); however, as one or more cells had an expected count of < 5, the chi-squared test could not be calculated.
RESULTS OF ANALYSES

TABLE 17 Number of analyses reporting a statistically significant outcome by population

<table>
<thead>
<tr>
<th>Populations</th>
<th>N</th>
<th>Significant outcome, n (%)</th>
<th>No significant outcome, n (%)</th>
<th>$\chi^2$</th>
<th>p-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental disorder only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>107</td>
<td>103 (96.3)</td>
<td>4 (3.7)</td>
<td>3.056</td>
<td>0.051</td>
</tr>
<tr>
<td>No</td>
<td>716</td>
<td>654 (91.3)</td>
<td>62 (8.7)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Offenders only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>372</td>
<td>339 (91.1)</td>
<td>33 (8.9)</td>
<td>0.667</td>
<td>0.245</td>
</tr>
<tr>
<td>No</td>
<td>451</td>
<td>418 (92.7)</td>
<td>33 (7.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forensic patients</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>202</td>
<td>185 (91.6)</td>
<td>17 (8.4)</td>
<td>0.570</td>
<td>0.456</td>
</tr>
<tr>
<td>No</td>
<td>621</td>
<td>572 (92.1)</td>
<td>49 (7.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indictable offenders only</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>196</td>
<td>183 (93.4)</td>
<td>13 (6.6)</td>
<td>0.671</td>
<td>0.256</td>
</tr>
<tr>
<td>No</td>
<td>627</td>
<td>574 (91.5)</td>
<td>53 (8.5)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Using SPSS.

Design
Studies following a single cohort design or a cross-sectional design were compared and no difference was found in the proportion of studies reporting a significant result (see Table 21).

Data collection
Where applicable, studies that collected data prospectively were more likely to report a statistically significant result than studies that collected the data retrospectively ($\chi^2 = 4.457$, $p = 0.035$) (see Table 21).

Reproducibility
There was no statistically significant difference in the frequency of studies reporting or not reporting a significant result according to their reported attempt to measure the reliability of tools (see Table 21).

Sample size and loss to follow-up
The mean number of participants approached to participate did not differ between studies reporting and not reporting a significant result ($U = 3913; p = 0.643$) (see Table 22). However, the mean number of participants recruited ($U = 27,809$, $p = 0.012$) and included in the final analyses ($U = 26,995; p = 0.04$) was significantly higher in studies reporting a significant result.

The median drop-out rate was 7.6% and was not associated with whether or not a study reported a significant result ($U = 25,190; p = 0.120$) (see Table 22).

Length of follow-up
The mean length of follow-up was not associated with whether or not a study reported a significant result ($U = 20,773; p = 0.055$) (see Table 22).

Multivariate analysis
As only three variables showed a statistically significant association with whether or not a study reported a significant result (i.e. prospective data collection, number of people recruited and the final number of participants) and two of these were not independent, multivariate analyses were deemed inappropriate.
Area under the curve analysis

Of the 959 studies available in the review, only 65 (6.77%) reported AUC statistics. Fewer still, only 35, reported lower and upper confident limits for the AUC obtained. The mean AUC across all 65 studies was just under 0.69 (SD = 0.077). Distributions of mean AUCs across all 65 studies are shown in Figure 7. The overall distribution of AUCs showed a slight negative skew (~0.258), as can be seen in Figure 7. The lowest single AUC reported was 0.44 and the highest was 0.88. For those studies reporting 95% CIs, the mean lower AUC CI was 0.56 and the mean upper AUC CI was 0.73.

The 65 available studies where an AUC statistic was reported were based on research with a total of 31 named risk assessment instruments. Table 23 shows findings for the 18 scales for which there was more than one study reporting AUC statistics, in descending order by number of AUC results.

The highest mean AUC reported was for the Broset Violence Checklist (BVC) at 0.815, but this was for an extremely short follow-up interval of just 1 day, as this instrument was designed for prediction of violence over relatively short periods in hospital inpatient settings. For the 10 scales regarding which four or more
RESULTS OF ANALYSES

**TABLE 19** Number of analyses reporting a statistically significant outcome by setting

<table>
<thead>
<tr>
<th>Settings</th>
<th>N</th>
<th>Significant outcome, n (%)</th>
<th>No significant outcome, n (%)</th>
<th>$\chi^2$</th>
<th>p-value$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mental health (including forensic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>113</td>
<td>106 (93.8)</td>
<td>7 (6.2)</td>
<td>0.591</td>
<td>0.442</td>
</tr>
<tr>
<td>No</td>
<td>710</td>
<td>651 (91.7)</td>
<td>59 (8.3)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Penal institution (excluding forensic)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>206</td>
<td>189 (91.7)</td>
<td>17 (8.3)</td>
<td>0.020</td>
<td>0.887</td>
</tr>
<tr>
<td>No</td>
<td>617</td>
<td>568 (92.1)</td>
<td>49 (7.9)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Community</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>290</td>
<td>262 (90.3)</td>
<td>28 (9.7)</td>
<td>1.624</td>
<td>0.202</td>
</tr>
<tr>
<td>No</td>
<td>533</td>
<td>495 (92.9)</td>
<td>38 (7.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>183</td>
<td>171 (93.4)</td>
<td>12 (6.6)</td>
<td>0.682</td>
<td>0.409</td>
</tr>
<tr>
<td>No</td>
<td>640</td>
<td>586 (91.6)</td>
<td>54 (8.4)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$^a$ Using SPSS.

**TABLE 20** Statistically significant finding in relation to outcome measure

<table>
<thead>
<tr>
<th>Outcome measures</th>
<th>Significant outcome, n (%)</th>
<th>No significant outcome, n (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scale score</td>
<td>420 (92.9)</td>
<td>32 (7.1)</td>
</tr>
<tr>
<td>Official data</td>
<td>204 (90.3)</td>
<td>22 (9.7)</td>
</tr>
<tr>
<td>Other</td>
<td>133 (89.3)</td>
<td>16 (10.7)</td>
</tr>
<tr>
<td>Routinely collected data</td>
<td>61 (88.4)</td>
<td>8 (11.6)</td>
</tr>
<tr>
<td>Self-report</td>
<td>37 (88.1)</td>
<td>5 (11.9)</td>
</tr>
<tr>
<td>Report of significant other</td>
<td>6 (75.0)</td>
<td>2 (25.0)</td>
</tr>
<tr>
<td>Nurse, doctor or other staff-observed</td>
<td>6 (85.7)</td>
<td>1 (14.3)</td>
</tr>
<tr>
<td>Clinical judgement of improvement</td>
<td>1 (100.0)</td>
<td>0 (0.0)</td>
</tr>
</tbody>
</table>

**TABLE 21** Number of analyses reporting a statistically significant outcome by study quality indicator (categorical variables)

<table>
<thead>
<tr>
<th>Variables</th>
<th>N</th>
<th>Significant outcome, n (%)</th>
<th>No significant outcome, n (%)</th>
<th>$\chi^2$</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITT analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>31</td>
<td>29 (93.5)</td>
<td>2 (6.5)</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>No</td>
<td>256</td>
<td>238 (93.0)</td>
<td>18 (7.0)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-sectional or single-cohort design</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cross-sectional</td>
<td>533</td>
<td>496 (93.1)</td>
<td>37 (6.9)</td>
<td>1.030</td>
<td>0.193</td>
</tr>
<tr>
<td>Single cohort</td>
<td>208</td>
<td>189 (90.9)</td>
<td>19 (9.1)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Prospective data collection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>205</td>
<td>189 (92.2)</td>
<td>16 (7.8)</td>
<td>4.457</td>
<td>0.035</td>
</tr>
<tr>
<td>No</td>
<td>171</td>
<td>146 (85.4)</td>
<td>25 (14.6)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reproducibility tested</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>374</td>
<td>347 (92.8)</td>
<td>27 (7.2)</td>
<td>0.595</td>
<td>0.440</td>
</tr>
<tr>
<td>No</td>
<td>449</td>
<td>410 (91.3)</td>
<td>39 (8.7)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NA, not applicable.
TABLE 22 Number of analyses reporting a statistically significant outcome by study quality indicator (continuous variables)

<table>
<thead>
<tr>
<th>Variables</th>
<th>Statistical test</th>
<th>Significant outcome</th>
<th>No significant outcome</th>
<th>U</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of eligible people approached</td>
<td>Mean</td>
<td>1154.7</td>
<td>1366.6</td>
<td>3913</td>
<td>0.643</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>4215.9</td>
<td>3685.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistic</td>
<td>158.3</td>
<td>27.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.298</td>
<td>0.821</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistic</td>
<td>11.5</td>
<td>5.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.149</td>
<td>0.421</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of people recruited</td>
<td>Mean</td>
<td>421.9</td>
<td>394.76</td>
<td>27,809</td>
<td>0.012</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1074.7</td>
<td>1487.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistic</td>
<td>63.4</td>
<td>58.314</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.178</td>
<td>0.599</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistic</td>
<td>7.2</td>
<td>7.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.089</td>
<td>0.304</td>
<td></td>
<td></td>
</tr>
<tr>
<td>No. of participants</td>
<td>Mean</td>
<td>407.5</td>
<td>351.2</td>
<td>26,995</td>
<td>0.04</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>1493.1</td>
<td>845.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistic</td>
<td>7.435</td>
<td>7.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.304</td>
<td>0.089</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistic</td>
<td>57.185</td>
<td>61.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.599</td>
<td>0.178</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentage attrition</td>
<td>Mean</td>
<td>4.4</td>
<td>7.4</td>
<td>25,190</td>
<td>0.120</td>
</tr>
<tr>
<td></td>
<td>SD</td>
<td>12.3</td>
<td>16.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Kurtosis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistic</td>
<td>10.4</td>
<td>8.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.604</td>
<td>0.178</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Skewness</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Statistic</td>
<td>3.3</td>
<td>2.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SE</td>
<td>0.306</td>
<td>0.089</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
AUC statistics were found. Figure 8 shows the distribution of AUCs obtained. The mean AUC reported for all purely ‘static’ risk assessment measures [e.g. GSIR, Offender Group Reconviction Scale (OGRS), Violence Risk Appraisal Guide (VRAG)] was slightly higher than the corresponding figure for those combining static and dynamic risk factors (Figure 9).

Table 24 shows corresponding findings for those scales (arranged in alphabetic order by title) where AUC data were available from a single study only. Of the 12 scales where this applied, only seven (62%) report 95% CIs.

The highest and second highest scores found here, for the Brief Psychiatric Rating Scale (BPRS) (0.88) and the Novaco Anger Scale (NAS) (0.82), were both based on comparatively short follow-up periods (in both cases 84 days). The lowest AUC found, for the Clinical Inventory of Dynamic Reoffending Risk Indicators (CIDRRI), is below the level that would be expected by chance (0.5). It is difficult to draw any meaningful conclusions from single studies, given that [as can be seen in Appendix 2 (see Table 26)] where there are several studies of an instrument, there is always heterogeneity among the AUCs obtained, and what appears to be a strong result from one study may be contrasted with an opposite effect in another.
Further analyses were conducted on the relationship between AUCs and other variables. There was a negative, but low and non-significant, correlation between length of follow-up and mean AUC for the 63 studies where both types of data were available ($r = –0.187$, $p = 0.142$). In other words, not surprisingly, there was a weak trend such that on average, lengthier follow-up intervals were associated with lower AUCs. There were no significant associations between year of publication or features of study design, such as rated design quality or use of ITT analysis, and AUC statistics. Nor were there any significant associations between the mean age of samples, and sample sizes either at recruitment or end point, and AUC values. There was a slight but non-significant trend towards higher AUCs reported from criminal justice compared with health settings (0.694 vs. 0.673), though the number of the former was comparatively low.

Some comparisons were made by type of instrument. One kind of comparison that is important to make is between those assessments based entirely on ‘static’ variables, which use an actuarial approach, and those which also incorporate ‘dynamic’ or clinical variables and rely to some extent on structured clinical judgments. Given their complexity, the various risk assessment instruments studied here can be grouped in a variety of ways and, in a further analysis, they were classified into three principal categories, depending on the data source and approach adopted. Group 1 ($n = 17$) consisted of combined static-dynamic

<table>
<thead>
<tr>
<th>Scale names</th>
<th>No. of studies reporting AUC</th>
<th>Mean follow-up (days)</th>
<th>Mean AUC</th>
<th>No. of studies reporting 95% CI</th>
<th>Mean lower 95% CI</th>
<th>Mean upper 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>STATIC-99</td>
<td>17</td>
<td>2297</td>
<td>0.6588</td>
<td>11</td>
<td>0.5927</td>
<td>0.7755</td>
</tr>
<tr>
<td>HCR-20</td>
<td>16</td>
<td>1445</td>
<td>0.6859</td>
<td>6</td>
<td>0.5817</td>
<td>0.7933</td>
</tr>
<tr>
<td>VRAG</td>
<td>12</td>
<td>1356</td>
<td>0.7223</td>
<td>4</td>
<td>0.6500</td>
<td>0.8375</td>
</tr>
<tr>
<td>PCL-R</td>
<td>10</td>
<td>2152</td>
<td>0.6902</td>
<td>4</td>
<td>0.5825</td>
<td>0.7775</td>
</tr>
<tr>
<td>PCL:SV</td>
<td>10</td>
<td>931</td>
<td>0.6830</td>
<td>6</td>
<td>0.5733</td>
<td>0.7833</td>
</tr>
<tr>
<td>GSIR</td>
<td>5</td>
<td>1154</td>
<td>0.7256</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>LSI</td>
<td>5</td>
<td>1684</td>
<td>0.6918</td>
<td>1</td>
<td>0.60</td>
<td>0.75</td>
</tr>
<tr>
<td>RRASOR</td>
<td>5</td>
<td>1803</td>
<td>0.6420</td>
<td>4</td>
<td>0.5275</td>
<td>0.7100</td>
</tr>
<tr>
<td>SORAG</td>
<td>5</td>
<td>1759</td>
<td>0.7120</td>
<td>3</td>
<td>0.6333</td>
<td>0.7900</td>
</tr>
<tr>
<td>BVC</td>
<td>4</td>
<td>1</td>
<td>0.8150</td>
<td>3</td>
<td>0.7367</td>
<td>0.8767</td>
</tr>
<tr>
<td>DVSİ</td>
<td>2</td>
<td>390</td>
<td>0.6600</td>
<td>1</td>
<td>0.70</td>
<td>0.72</td>
</tr>
<tr>
<td>OGRS</td>
<td>2</td>
<td>2007</td>
<td>0.7150</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>RASP</td>
<td>2</td>
<td>803</td>
<td>0.6950</td>
<td>2</td>
<td>0.6750</td>
<td>0.7100</td>
</tr>
<tr>
<td>RM-2000</td>
<td>2</td>
<td>4143</td>
<td>0.7750</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>SAQ</td>
<td>2</td>
<td>1095</td>
<td>0.7050</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>SVR-20</td>
<td>3</td>
<td>3697</td>
<td>0.6233</td>
<td>1</td>
<td>0.58</td>
<td>0.78</td>
</tr>
<tr>
<td>VRS</td>
<td>2</td>
<td>2007</td>
<td>0.6365</td>
<td>2</td>
<td>0.5500</td>
<td>0.7250</td>
</tr>
<tr>
<td>VSC</td>
<td>2</td>
<td>350</td>
<td>0.5250</td>
<td>2</td>
<td>0.4300</td>
<td>0.6500</td>
</tr>
</tbody>
</table>

assessments, and included the BPRS, CIDRRI, Domestic Violence Screening Instrument (DVSI), HCR-20, Iterative Classification Tree (ICT), Level of Service Inventory-Revised, PCL-R, Psychopathy Checklist: Screening Version, Preliminary Scheme (PS), Risk Assessment Scale for Prison (RASP), Rapid Risk Assessment for Sexual Offender Recidivism (RRASOR), Short Dynamic Risk Scale (SDRS), Short-Term Assessment of Risk and Treatability (START), Sexual Violence Risk (SVR), Violence Risk Scale (VRS) and Violence Screening Checklist (VSC). Group 2 (n = 9) consisted of actuarial instruments exclusively or primarily relying on static predictors, and included GSIR, OGRS, STATIC-99, STATIC-2000, Risk Matrix 2000 (RM-2000), Statistical Information on Recidivism (SIRRI), Sex Offender Risk Appraisal Guide (SORAG), Violent Offender Risk Assessment Scale (VORAS) and VRAG. Group 3 (n = 4) consisted of solely self-report methods, and included the Criminal Sentiments Scale (CSS), Measures of Criminal Attitudes and Associates.

FIGURE 8 Area under the curve by study for scales with ≥ 4 reported AUCs. (a) STATIC-99; (b) HCR-20; (c) VRAG; (d) PCL-R; (e) Psychopathy Checklist: Screening Version; (f) GSIR; (g) BVC; (h) Level of Service Inventory; (i) Rapid Risk Assessment for Sexual Offender Recidivism; and (j) Sex Offender Risk Appraisal Guide.
The results of this are shown in graphical form in Figure 9.

Given the very short time scale (1 day) used in the study of the BVC, it was excluded from this analysis. As earlier, we should also keep in mind that the data for two apparently top-performing instruments included in Figure 9, the BPRS and the NAS, are in both cases extracted from studies with much shorter time scales (84 days) than for other evaluations. It should also be borne in mind that the mean AUCs shown in Figure 9 are based on very different numbers of studies for different instruments (ranging from 1 to 17 as shown in Table 23 and Table 24).

There is an apparent trend such that if (given their shorter follow-up period) the BPRS and NAS are excluded, the majority of the actuarial predictors are in the upper portion of Figure 9, while the lower two-thirds are taken up with static-dynamic predictors, while self-report scales are mainly just above the mid point of the distribution. The mean AUCs for the three groups of instruments just defined were, respectively, for static-dynamic predictors, 0.650; for static/actuarial predictors, 0.709; and for self-report predictors, 0.708. Return of the BPRS to the static-dynamic list raises the mean AUC slightly to 0.664 and addition of the NAS to the self-report list raises the mean AUC rather more to 0.736, but again the shorter follow-up for both these instruments must be kept in mind. Unfortunately these differences cannot be tested statistically as for some instruments in some studies there is only a single AUC score.
### Table 24: Area under the curve for instruments where there was a single study only

<table>
<thead>
<tr>
<th>Scale names</th>
<th>Follow-up (days)</th>
<th>AUC</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPRS</td>
<td>84</td>
<td>0.88</td>
<td>0.79</td>
<td>0.96</td>
</tr>
<tr>
<td>CIDDRI</td>
<td>2272</td>
<td>0.44</td>
<td>0.31</td>
<td>0.56</td>
</tr>
<tr>
<td>CSS</td>
<td>1635</td>
<td>0.71</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>ICT</td>
<td>140</td>
<td>0.70</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>MCAA</td>
<td>610</td>
<td>0.71</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>NAS</td>
<td>84</td>
<td>0.82</td>
<td>0.72</td>
<td>0.93</td>
</tr>
<tr>
<td>PS</td>
<td>365</td>
<td>0.71</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>STATIC-2000</td>
<td>5950</td>
<td>0.70</td>
<td>0.55</td>
<td>0.86</td>
</tr>
<tr>
<td>SDRS</td>
<td>365</td>
<td>0.72</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>SIRRI</td>
<td>1241</td>
<td>0.76</td>
<td>0.67</td>
<td>0.85</td>
</tr>
<tr>
<td>START</td>
<td>319</td>
<td>0.65</td>
<td>0.57</td>
<td>0.72</td>
</tr>
<tr>
<td>VORAS</td>
<td>2811</td>
<td>0.61</td>
<td>0.53</td>
<td>0.69</td>
</tr>
</tbody>
</table>

BPRS, Brief Psychiatric Rating Scale; CIDDRI, Clinical Inventory of Dynamic Reoffending Risk Indicators; CSS, Criminal Sentiments Scale; ICT, Iterative Classification Tree; MCAA, Measures of Criminal Attitudes and Associates; NAS, Novaco Anger Scale; PS, Preliminary Scheme; SDRS, Short Dynamic Risk Scale; SIRRI, Statistical Information on Recidivism; SVR, Sexual Violence Risk; VORAS, Violent Offender Risk Assessment Scale.
Chapter 5  Discussion and conclusions

Strengths and limitations of the review

The main strength of this review is its breadth in terms of the inclusion criteria and thus the very large collection of studies that have been gathered and analysed. The total evidence base of nearly 1000 studies is larger than any previous review of this literature and, given that it covered only a 7-year period, reflects the enormous interest over the past 10 years among clinicians and researchers in developing and testing structured risk assessment instruments for this topic. This large sample of studies maximised statistical power for the analyses described in Chapter 4 and thus minimised the likelihood that the insignificant results are type 2 errors.

Each study was also analysed across a wide range of relevant variables so that the review combines great breadth and depth simultaneously. The broad definition of violence adopted for the review makes it possible to draw comparisons across evidence in different subtypes of violent populations so that work, for instance, with perpetrators of domestic violence can be compared with that of more traditional forensic populations. In this way, some attempt can be made to integrate knowledge across domains and between clinical and criminological literatures to counter the current fragmentation which some have criticised.1005 Despite the large volume of material to be digested, ‘gold-standard’ systematic review procedures were adhered to throughout and monitored by having an experienced professional reviewer as one of the team.

A number of weaknesses should be noted in order to contextualise the findings. First, despite systematic cross-checking, given the scale of the review and the number of extractors working upon it, there are inevitably some discrepancies and missing data within the data set. Agreement between pairs of reviewers in the initial exercises was good but less than perfect. Violence is a complex phenomenon and there is scope for debate and subjectivity in applying definitions (e.g. are agitation or irritability forms of aggression?) even when inclusion criteria are clearly specified. Secondly, only five of the seven calendar years (2003–7) are completely covered, with partial coverage of 2002 and 2008. Thirdly, as with most systematic reviews, the data values are extracted per study rather than being raw data for individual participants in the studies. Thus, there is no weighting for sample or effect sizes when results from studies are aggregated. Linked to this, unlike the parallel intervention review,1 which used standard meta-analysis to examine combining effect sizes, there is no equivalent procedure for combining data across instrument evaluation studies. The mean AUCs reported in Table 24, for instance, take no account of study sample size and thus AUCs obtained from small and large studies are treated as equivalent. This further emphasises that the conclusions here are tentative and suggestive rather than in any way conclusive. Finally, given the broad scope of the review and the extreme heterogeneity of the included instruments and outcome measures, it was not considered meaningful to conduct either a meta-analysis or an analysis of publication bias.

A large number of studies (330) were excluded ‘post hoc’ at the data extraction stage (i.e. having already passed through stage 1 and stage 2 inclusion). At first sight, this may seem an unusually large volume of studies to be excluded after the selection process was supposed to be completed. However, given the scale of the review, this group of 330 studies constitutes only about 5% of studies evaluated at stage 2 inclusion and in a more typically sized review this proportion would equate to one or two papers. While it is not ideal, the 5% figure is acceptable given the complexity of the area and the large team of reviewers.
Summary of key findings

Characteristics of the overall literature

A number of important findings can be highlighted from the review. It is worth noting that nearly half of the literature was based on offenders with no diagnosed mental health problems (i.e. prisoners and those on probation) and another one-quarter was on offenders with mental health problems. The literature on those with diagnosable mental health problems but no offence history (circa 10%) was relatively small in comparison with these two large groups. The capacity to generalise from offender to non-offender populations and to draw conclusions for general (e.g. acute) mental health services is thus somewhat limited. Where mental health problems were present, there was relatively little attempt to study specific diagnostic groups apart from the 70 studies that focused exclusively on people with schizophrenia or a personality disorder (with or without an offending history). The offender groups tended to be heterogeneous in terms of the type of offence that had been committed, though domestic violence and sex offending (20% each of offender studies) were clearly distinct.

As previously stated, the violence risk assessment literature is very large and, while growth is not accelerating as it did in the 1990s, it is still growing at the rate of about 150 new studies per year (see Figure 3). Over 300,000 people have been assessed in the studies included in the review. Several hundred tools have been evaluated but the study design quality of the evaluations is very variable and many tools are only evaluated in a small number of studies. There are 11 instruments (see Tables 12 and 13) which have been evaluated in relation to a violent outcome in 25 or more studies [PCL-R, CTS, STATIC-99, HCR-20, VRAG, STAXI, MCMI, Minnesota Multiphasic Personality Inventory (MMPI), Becks Depression Inventory (BDI), Barratt Impulsiveness Scale and RRASOR] and some of these (MCMI, MMPI and BDI) are not traditionally thought of as violence risk assessment instruments. These instruments have been subjected to rigorous testing but, apart from the analysis in Chapter 5 (see Area under the curve analysis), which examines AUCs for 5 of the 11 (STATIC-99, HCR-20, VRAG, PCL-R and RRASOR), this is not to say that they have demonstrable accuracy in terms of prediction. Beyond the well-tested 11, there are numerous other instruments, each with a rationale in terms of being distinctive but lacking a substantial evidence base. The usefulness of this ‘long tail’ of relatively untested instruments is debatable. It could be argued that researchers and clinicians should focus all their evaluation efforts on a small group of tried-and-tested instruments in order to improve those which are emerging as ‘the best’. On the other hand, this ‘one-size-fits-all’ approach could be seen as too restrictive, especially with such heterogeneous populations of violence perpetrators. Those working with domestic violence perpetrators, for instance, might resist pressure to incorporate, say, the HCR-20 into their practice despite it having a strong evidence base, because it does not relate to aspects of the problem that seem relevant to them in their work.

Far too much (nearly 60%) of the literature consists of cross-sectional studies and many of these involved correlating two sets of scale scores. Such studies do not constitute a rigorous test of predictive accuracy, are unclear about causal pathways, are prone to confounding and rely on self-report outcome measures which are only a proxy for real violence. They are relatively easy to conduct, hence their popularity, but the literature has essentially become cluttered with such studies which contribute little to the evidence base. Prospective studies using ‘real’ measures of violence such as incident reports or arrest records should be encouraged although the scale data could be collected at the same time to provide some opportunities for validation. The preponderance of cross-sectional studies is also apparent in Table 15 where most studies (cross-sectional or prospective) started and ended in the same setting (80%). A key potential contribution of structured risk assessment is in supporting decision-making about discharge or release from prison or hospital into the community but there were relatively few studies attempting this tracking process across settings after discharge. Only 10 studies started in an offender institution and ended in the community and the equivalent numbers for discharge from secure forensic settings and open inpatient wards into the community were two and five, respectively. While intra-institution evaluations are useful for the protection of prison and hospital staff, triggers for violence in the community are likely to be very different from those within an institution, and so this is a more stringent test of predictive accuracy and more reassuring to policy-makers and the public.
Fewer than half of the studies attempted to estimate the reliability of their adopted measures and a tiny proportion examined the clinical utility of the instrument. Joliffe et al. note that predictive validity is only one of many aspects which are relevant to designing an effective instrument and clinical utility (e.g. ease of use in clinical practice) is a key aspect. A tool that is highly effective in terms of predictive accuracy may achieve this accuracy on the basis of being unwieldy and impractical in most clinical settings. There may be a pay-off between precision and real-world utility in which some aspects of predictive accuracy have to be sacrificed in order for practitioners to adopt the instrument in everyday work.

Follow-up periods in prospective studies varied enormously from a few hours (e.g. the BVC) to beyond 10 years. Clearly, different tools have different purposes and it is not possible to say that a high AUC value obtained over 10 years is inherently better than the same AUC value obtained over a few hours. Each tool must be suitable for the population and problem for which it is designed and the variables in each case are likely to be very different.

The literature is dominated by studies conducted in the USA, as with much social science research, and over three-quarters were conducted in three countries (USA, Canada and the UK). This is not a healthy situation given the relevance of social and cultural context to the occurrence of violence and clearly there is a need for further replications in continental Europe, Australasia and beyond to the large populations in China and India. While the definition of low-level violence, and the factors relevant to a high-risk profile, may vary across cultures, various instruments have been validated in and perform adequately in different countries.

End point sample sizes were, on average, substantial (mean = 360) and much higher than in the equivalent violence intervention literature. This is presumably because conducting a risk assessment is relatively straightforward compared with delivering an intervention over a sustained period of time which may, in turn, explain why this literature is approximately four times bigger than the parallel intervention literature. This large sample size is a strength of the risk literature and should be maintained in future work.

Unsurprisingly, there was a preponderance of studies focused on male violence with over 500 studies focusing exclusively on a male sample. However, there were a number of studies focusing on mixed groups in terms of sex and nearly 100 studies with exclusively female samples. Given the differential pathways to male and female violence there is clearly a substantial literature on female samples which is worth examining in its own right. A female subgroup analysis or female–male comparative analysis could be conducted using the same approach as pursued here for the overall analysis. A small number of studies (<10) focused exclusively on non-Caucasian populations but these may have included a diversity of non-Caucasian ethnic groups and so there is less scope for separate analysis of these groups.

A large proportion of studies reported a significant result, that is to say that scores on one or more of the instruments being studied had an above-chance association with a violent outcome. This high success rate is suggestive of some publication bias where non-significant studies are being disproportionately rejected from publication. Such a tendency is apparent in many research literatures and can be reduced by requiring registration of studies prior to commencement. However, it should be noted that 22% (n = 205) of included studies were theses and were therefore unpublished.

Bivariate analyses

The first research question underpinning this review asked ‘what features (i.e. population, instrument, outcome measure and design) are associated with a risk assessment instrument score being significantly associated with a violent outcome?’ The bivariate analyses indicated that not many features of a study were associated with the outcome in this way. If a strict alpha level of 0.05 is imposed, only prospective study design and sample size (number recruited and number participating) emerged as significant. Two others (focusing on a mental disorder population only and length of follow-up) had borderline significant associations. Prospective designs (92%) reported significant results more frequently than non-prospective
designs (85%) and significant studies had larger sample sizes (n = 407 participants) than non-significant studies (n = 351 participants). While the latter finding may be a result of greater statistical power in the larger studies, the contribution of design to a significant outcome is harder to explain or draw conclusions from. It may be that prospective studies require greater thought in preparation (e.g. selection of relevant measures) and are better designed overall compared with the cross-sectional studies, which are sometimes little more than opportunistic data ‘fishing expeditions’. This would be supported by the borderline significance of follow-up period as sustained data collection could be a feature of high-quality studies which are better designed overall in the way suggested above (though, as previously remarked, instruments designed to assess imminent violence will inevitably have short follow-ups and should not be seen as inherently inferior). It is certainly somewhat counterintuitive to find that significant studies had longer follow-ups than non-significant studies as presumably the task of prediction becomes harder as time progresses. In fact the reverse may be true as there is more time and opportunity for the person to act violently the longer they are kept within the study.

**Area under the curve subset analysis**

The second research question asked ‘which risk assessment instruments have the highest level of predictive validity for a violent outcome?’ The BVC had the highest mean AUC of all tools tested in more than one study but its focus on imminent short-term violence prediction makes it difficult to compare with longer-term predictive tools. The three tools with at least one replication and longer follow-up (> 1 year) which had the highest mean AUCs were the RM-2000, GSIR and VRAG. However, as noted, there was great variability in AUC values across multiple studies and instruments with acceptable mean AUCs had close-to-chance AUCs in some studies (e.g. 0.58 in one study for HCR-20). On this measure, the VRAG emerged strongly in that an AUC below 0.65 has not been reported in any of the 12 studies testing this instrument.

Part of the problem in answering this question is a lack of an accepted statistical technique for aggregating findings across the various studies due to the lack of a quantitative technique equivalent to meta-analysis as used for intervention studies.

The currently most widely favoured approach to risk assessment, also believed to be the best established empirically, entails some means of combining well-validated actuarial risk instruments with structured clinical judgments, whether in mental health or criminal justice services. Professional judgement, although discredited in what was called the ‘first generation’ of risk assessment evaluations, returned to favour in the ‘third generation’ given a developing and broad recognition that it needs to be structured in some methodical and reproducible way. Monahan et al., who undertook the highly regarded MacArthur Risk Assessment Study, urged that the task of risk assessment be carried out within an explicit, systematised framework. The same general principle has also more recently been endorsed by, for example, Farrington et al. Such an approach should be informed by the use of well-tested actuarial instruments, allocating factors that appear to influence judgments of an individual’s risk level in a stepwise, sequential method. Monahan et al. called their approach to this an iterative classification tree. They further advocate that when this is done it should be repeated several times using a slightly different variable set each time. They argued that it is only when there is sufficient agreement between separate assessments, carried out by applying these principles with different combinations of predictors, that we can feel confident in the predictions that are made.

The findings of this part of the present review may cast some of the currently accepted conventions into question. Actuarial methods employing static predictor variables are usually characterised as belonging to the second generation of risk assessment tools. On balance, however, there is a small, but in practical terms potentially meaningful, advantage of these methods and of self-report instruments over the more recently developed third generation, integrative static-dynamic approaches. This may cast into doubt the value of the additional effort involved in completing instruments that require extensive application of structured clinical judgement (though arguably that retains added value for risk management purposes). It may also be that dynamic factors, while useful and potentially decisive for short-term risk assessment, have little predictive value over longer periods. That may be underlined by the poor showing of Factor 1.
(personality variables and interpersonal functioning) of the PCL-R in the recent review by Yang et al.\textsuperscript{1011} From the point of view of sheer predictive power, the familiar adage that future behaviour is best predicted by past behaviour emerges as the best supported conclusion from the present set of findings, judged at least on the basis of the AUC results. What is potentially more surprising is that self-report scales perform commensurately with static/actuarial assessments for the purposes of violence prediction.

Placed in a wider context, however, all of the assessment methods reviewed here exhibit some marked limitations. The general predictive power of all the assessments considered raises broader questions regarding whether or not the field of risk assessment research may have entered a phase of ‘diminishing returns’. The systematic application of the ideas of researchers such as the MacArthur group, which in its most advanced form was entitled the multiple-models approach, is the most elaborate framework for risk assessment developed to date. It is illustrated in a study by Banks et al.\textsuperscript{1012} Here, a number of different assessment models, each capturing a different combination of variables, are themselves integrated. An amalgamation of five models yielded an AUC of 0.88. This outcome proved as effective as another version incorporating 10 models. However, even with this apparently very impressive level of predictive accuracy, only 76.2\% of those placed in the highest of five risk categories were subsequently violent, leaving one-quarter of that subsample wrongly classified. Moving down the four categories at progressively lower levels of estimated risk, the proportion wrongly classified was progressively higher at each level.

The present findings broadly parallel a number of those obtained by previous reviewers of the field, but who for the most part have focused on different portions of it. This includes the work of Campbell et al.\textsuperscript{11} and Farrington et al.\textsuperscript{1016} who focused primarily on the criminal justice system, Leistico et al.\textsuperscript{1013} who focused specifically on the PCL-R, and Yang et al.,\textsuperscript{1011} Singh and Fazel,\textsuperscript{63} and Singh et al.\textsuperscript{1014} who focused on a mixture of samples including younger offenders.

In closing, it should be noted that there is a difficulty with some analyses which interpret the use of prediction tools in this field as being essentially similar to a diagnostic procedure as employed in medicine. Singh et al.\textsuperscript{1014} and Fazel et al.\textsuperscript{1015} adopt such an approach and refer explicitly to comparisons with other medical tools. They apply a number of procedures commensurate with those utilised in diagnosis where there is a dichotomous outcome (thus disease present/absent is equivalent to reoffending yes/no). This entails usage of a number of statistics which yield such outcomes, such as positive and negative predictive power, and the diagnostic odds ratio, a relatively novel indicator in this context which though producing an odds ratio statistic is dependent on diagnostic yes/no categorisation within the groups being compared. However, attempts to predict the likelihood of a behaviour occurring (violent or sexual assault) are of limited usefulness in binary form and prognostic estimates will be of practical use in this field only if they generate a continuous variable (risk level).

Conclusions and implications for research

The violence risk assessment literature is very large and has a number of strengths in comparison with the equivalent literature on violence interventions (e.g. larger sample sizes, longer follow-ups). It is also stronger because risk assessment instruments are clearly defined and delimited in comparison with the complexity of psychosocial interventions, and so the research effort is more focused. The more developed instruments are manualised with clear guidance on each step in the process, which again aids replication across studies and populations. That said, the greater clinical sophistication of structured professional judgement approaches entails greater scope for subjectivity in assessments (i.e. in comparison with actuarial instruments) so this aspect of scientific rigour may be reduced as these SPJ approaches develop.

1. The small number of tools which have demonstrable replicated efficacy already should be tested out on a wider range of populations. This expansion should include empirical testing beyond North America and the parts of Europe where they have been extensively tested.
2. There should be a strong case made for expending significant effort on developing and testing any new risk assessment tools given the proliferation of tools developed over the past 15 years. There will always be a tension between a ‘one-size-fits-all’ philosophy in which three to four dominant instruments with extensive empirical support are seen as suitable for all populations and a ‘bottom-up’ approach which recognises that many different tools (including those with minimal evidence) are needed to reflect the complexities of variations across populations.

3. Cross-sectional studies and/or studies relying purely on scale scores should be avoided. Too much of the existing literature is based on correlating a predictor and an outcome occurring simultaneously. This prevents the testing of any causal hypotheses and thus does not help in the development of theoretical frameworks for understanding violence. The problem is compounded when both the ‘predictor’ and the ‘outcome’ are measured using self-reported experiences recorded on a scale, as opposed to observable hostility or violence. The validity of such scale measures is lower than that for behavioural outcomes.

4. More studies should be conducted prospectively from hospital/prison to the community to examine the potential support of risk assessment tools to discharge/release decisions. While intra-institutional violence is important in terms of protecting staff and other patients, it is the transition from hospital or prison to the community which is of most significance for the patient/prisoner and society at large. It is also more challenging to achieve effective prediction when moving from one environment to another and methodologically more difficult to keep track of participants. But clinical decisions on release or discharge are the core issues faced by professionals and better research over this transition period is essential.

5. Clinical utility of those instruments with a strong evidence base in terms of predictive validity should be assessed to contextualise this information. While good predictive validity is a core component of an effective instrument, there are a number of other aspects which must be present for the instrument to be considered entirely effective. These include the availability of a user manual, reasonable cost, available training, specified user competencies (including training and specified qualifications and skill levels), ease of use, appropriate administration time and recognition of protective factors. Some of these aspects can be studied as part of the overall research evaluation of specific tools.

6. The findings from the female-only studies should be examined and summarised separately. As argued above, the pathways to violence and consequences following from it are likely to be different for females compared with males. Given the identification here of a large literature of female-only samples, there is scope for a powered analysis of this topic on its own to examine differences from the male-only samples.

7. A statistical procedure [similar to Cohen’s d (e.g. Rice and Harris1016) for intervention effect size] should be developed for aggregating across AUCs. Mean AUCs were calculated for this study in order to aggregate across replications but this is a rather simplistic approach, especially when there are few studies. Effect sizes for intervention studies, drawing on means and standard deviations, are a more robust statistic and the research effort in the area of risk assessment would benefit from a similar approach.
Acknowledgements

Ms Janet Atkinson for providing administrative support, University of Liverpool’s Inter-Library Loan team and all members of the review team.

Advisory panel members

Provided feedback to the team during the review process

- Professor Joy Duxbury, Professor of Mental Health Nursing, Faculty of Health and Social Care, University of Central Lancashire.
- Ms Kathryn Harney, Associate Director of Research, Greater Manchester West Mental Health NHS Foundation Trust.
- Ms Sue Imlack, patient representative.
- Dr Caroline Logan, Consultant Forensic and Clinical Psychologist, Greater Manchester West Mental Health NHS Foundation Trust & University of Manchester.
- Ms Ruth Sayers, patient representative.
- Professor Jenny Shaw, Professor of Forensic Psychiatry, School of Medicine, University of Manchester.

Members of the review team

Dr Wally Barr, Senior Research Fellow, Health and Community Care Research Unit.

Dr Tamara Brown, Research Fellow, Liverpool Reviews and Implementation Group.

Dr Gemma Cherry, Research Assistant, Liverpool Reviews and Implementation Group.

Ms Regina Cooney, Clerical Officer (Research), Liverpool Reviews and Implementation Group.

Ms Maxine Cromar, Research Nurse, Merseycare NHS Trust.

Professor Rumona Dickson, Director, Liverpool Reviews and Implementation Group.

Dr Yenal Dundar, Research Fellow and Specialist Registrar in Psychiatry, Liverpool Reviews and Implementation Group.

Mr Nigel Fleeman, Research Fellow, Liverpool Reviews and Implementation Group.

Ms Rachel Flentje, Research Assistant, Infotech UK Research.

Ms Juliet Hockenhull, Research Fellow, Liverpool Reviews and Implementation Group.

Dr Maria Leitner, Director, Infotech UK Research.

Ms Gerlinde Massey, Clerical Officer (Research), Liverpool Reviews and Implementation Group.

Professor James McGuire, Academic Director/Head of Course, Division of Clinical Psychology.
Ms Bev Quinn, Research Nurse, CAMEO – Early Intervention in Psychosis Team, Cambridgeshire & Peterborough NHS Trust.

Professor Richard Whittington, Health and Community Care Research Unit.

**Contributions of authors (alphabetically)**

**Dr Wally Barr**  Advice and assistance in all aspects of the review.

**Dr Gemma Cherry**  Application of inclusion criteria, data extraction, input into all aspects of review.

**Professor Rumona Dickson**  Advice and assistance in all aspects of the review.

**Dr Yenal Dundar**  Application of search strategy.

**Ms R Flentje**  Application of inclusion criteria and data extraction.

**Ms Juliet Hockenhull**  Review co-ordination, literature searches, data management, methods and input into all aspects of review.

**Dr Maria Leitner**  Advice and assistance in all aspects of review.

**Professor James McGuire**  Data analysis, advice and assistance in all aspects of the review.

**Ms B Quinn**  Application of inclusion criteria and data extraction.

**Professor Richard Whittington**  Project management, input into all aspects of the review.

All authors read and commented on draft versions of the report.

**About the Assessment Group**

The Assessment Group was based on a partnership between two research groups within the University of Liverpool.

The Liverpool Reviews and Implementation Group (LRiG) was established at The University of Liverpool in April 2001. It is a multidisciplinary research group whose purpose, in the first instance, is to conduct health technology assessments commissioned by the HTA Programme.

The Liverpool Violence (LiVio) Research Group is a multidisciplinary partnership of academics and clinicians with a commitment to both qualitative and quantitative research approaches and their application to problems of violence and self-harm linked to mental health issues in real-world settings. It has received funding from the Department of Health since 2002 both to run primary research studies in secondary mental health service settings and to conduct the preceding stage of this systematic review.

**Rider on responsibility for report**

The views expressed in this report are those of the authors and not necessarily those of the NHS, Department of Health or NIHR HTA programme. Any errors are the responsibility of the authors.
References


REFERENCES


263. Dolan M, Khawaja A. The HCR-20 and post-discharge outcome in male patients discharged from medium security in the UK. *Aggress Behav* 2004;30:469–83. [http://dx.doi.org/10.1002/ab.20044](http://dx.doi.org/10.1002/ab.20044)


271. Doyle M, Dolan M. Predicting community violence from patients discharged from mental health services. *Br J Psychiatry* 2006;189:520–6. [http://dx.doi.org/10.1192/bjp.bp.105.021204](http://dx.doi.org/10.1192/bjp.bp.105.021204)


---

© Queen’s Printer and Controller of H.M.S.O. 2013. This work was produced by Whittington et al. under the terms of a commissioning contract issued by the Secretary of State for Health. This issue may be freely reproduced for the purposes of private research and study and extracts (or indeed, the full report) may be included in professional journals provided that suitable acknowledgement is made and the reproduction is not associated with any form of advertising. Applications for commercial reproduction should be addressed to: NIHR Journals Library, National Institute for Health Research, Evaluation, Trials and Studies Coordinating Centre, Alpha House, University of Southampton Science Park, Southampton SO16 7NS, UK.


350. Friedman AS, Terras A, Glassman K. The differential disinhibition effect of marijuana use on violent behavior: a comparison of this effect on a conventional, non-delinquent group versus a delinquent or deviant group. *J Addict Dis* 2003; **22**:63–78. http://dx.doi.org/10.1030/joad.001069v22n03_06


398. Green JB. A comparison of personality characteristics between incarcerated male pedophiles and incarcerated male nonpedophiles as measured by the Minnesota Multiphasic Personality Inventory-2 (MMPI-2). PhD thesis. College Station, TX: Texas A&M University; 2003.


REFERENCES


REFERENCES


REFERENCES


http://dx.doi. org/10.1007/s10896-007-9100-4

http://dx.doi.org/10.5210/bsi.v14i2.359

http://dx.doi.org/10.1017/S0033291705006562

http://dx.doi.org/10.1016/j.paid.2006.05.009

http://dx.doi.org/10.1016/j.addbeh.2004.01.003

http://dx.doi.org/10.1016/S0278-5846(03)00107-6


http://dx.doi.org/10.1080/13552600208413329

http://dx.doi.org/10.1002/cbm.534

http://dx.doi.org/10.1177/0886260506290663

http://dx.doi.org/10.1177/0306624X03253842

http://dx.doi.org/10.1037/0002-9432.73.3.288

http://dx.doi.org/10.1080/13552600512331329540

http://dx.doi.org/10.1007/s10896-006-9037-z

http://dx.doi.org/10.1007/s10896-007-9090-2


REFERENCES


864. Straus MA, Ramirez IL. Gender symmetry in prevalence, severity, and chronicity of physical aggression against dating partners by university students in Mexico and USA. Aggress Behav 2007;33:281–90. http://dx.doi.org/10.1002/ab.20199


REFERENCES


1013. Leistico AM, Salekin RT, DeCoster J, Rogers R. A large-scale meta-analysis relating the Hare measures of psychopathy to antisocial conduct. Law Hum Behav 2008;32:28–45. http://dx.doi.org/10.1007/s10979-007-9096-6


## Appendix 1  Search strategies

### Table 25  PsycINFO search strategy (April 2008)

<table>
<thead>
<tr>
<th>#</th>
<th>Searches</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>(((Homicid* or murder* or manslaughter* or infanticid* or parricid* or assault* or ((bodily and (harm or assault)) or assail* or bugger* or sodom* or molest* or pedophil* or paedophil* or sadis* or sadomasochis* or sado-masochis* or anger* or cruel* or rapist* or (rape* and offend*)) or physical abus* or spouse abus* or partner abus* or sexual abus*) or (((dangerous* and (behavior* or behaviour* or histor* or conduct*)) or violen*)) and (risk* or predict* or anteced* or assess* or cause* or reason* or interven* or prevention* or preventing* or controlling* or manage* or treatment* or treating* or reduction* or reducing* or stop* or mental* or forensic* or psychiatric* or offend* or Axis 1 or Axis 2 or criminal* or detain* or insan* or NGRI or retard* or (learning disab* or learning-disab*)) or acquit* or (child abus* or elder abus* or hostil* or killing* or attack* or aggress*)) and (mental* or forensic* or psychiatric* or offend* or axis 1 or axis 2 or criminal* or detain* or insan* or NGRI or retard* or (learning disab* or learning-disab*)) or acquit* or disorder*)) not (cancer* or cancer or tumo* or tumour or heart* or heart)).mp.</td>
<td>22,934</td>
</tr>
<tr>
<td>2</td>
<td>limit 1 to (((100 childhood &lt;birth to age 12 yrs&gt; or 120 neonatal &lt;birth to age 1 mo&gt; or 140 infancy &lt;age 2 to 23 mo&gt; or 160 preschool age &lt;age 2 to 5 yrs&gt; or 180 school age &lt;age 6 to 12 yrs&gt; or 200 adolescence &lt;age 13 to 17 yrs&gt;) and yr=&quot;2002 - 2008&quot;)</td>
<td>5631</td>
</tr>
<tr>
<td>3</td>
<td>Limit 1 to (animals and yr=&quot;2002 - 2008&quot;)</td>
<td>235</td>
</tr>
<tr>
<td>4</td>
<td>Limit 1 to (editorial and yr=&quot;2002 - 2008&quot;)</td>
<td>332</td>
</tr>
<tr>
<td>5</td>
<td>2 or 3 or 4</td>
<td>6198</td>
</tr>
<tr>
<td>6</td>
<td>1 not 5</td>
<td>16,736</td>
</tr>
</tbody>
</table>
### Appendix 2  Area under the curve data

#### Table 26  Area under the curve values and CIs by study

<table>
<thead>
<tr>
<th>Tools</th>
<th>Study ID</th>
<th>AUC (95% CI)</th>
<th>Measure</th>
<th>Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPRS</td>
<td>Doyle 2006&lt;sup&gt;372&lt;/sup&gt;</td>
<td>0.88 (0.79 to 0.96)</td>
<td>Staff observation</td>
<td>84</td>
</tr>
<tr>
<td>BVC</td>
<td>Abderhalden 2004&lt;sup&gt;37&lt;/sup&gt;</td>
<td>0.88 (0.76 to 0.99)</td>
<td>Staff observation</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Abderhalden 2006&lt;sup&gt;376&lt;/sup&gt;</td>
<td>0.86 (0.79 to 0.92)</td>
<td>Staff observation</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Almvik 2007&lt;sup&gt;293&lt;/sup&gt;</td>
<td>0.69 (0.66 to 0.72)</td>
<td>Staff observation</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Ogloff 2006&lt;sup&gt;298&lt;/sup&gt;</td>
<td>0.83&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Staff observation</td>
<td>1</td>
</tr>
<tr>
<td>CIDRRI</td>
<td>Philipse 2006&lt;sup&gt;735&lt;/sup&gt;</td>
<td>0.44 (0.31 to 0.56)</td>
<td>Reconviction</td>
<td>2272</td>
</tr>
<tr>
<td>CSS</td>
<td>Witte 2006&lt;sup&gt;283&lt;/sup&gt;</td>
<td>0.71&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction</td>
<td>1635</td>
</tr>
<tr>
<td>DVS1</td>
<td>Williams 2004&lt;sup&gt;735&lt;/sup&gt;</td>
<td>0.61&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction/other reported</td>
<td>540</td>
</tr>
<tr>
<td></td>
<td>Williams 2006&lt;sup&gt;74&lt;/sup&gt;</td>
<td>0.71 (0.7 to 0.72)</td>
<td>Reconviction</td>
<td>240</td>
</tr>
<tr>
<td>GSIR</td>
<td>Kroner 2005&lt;sup&gt;130&lt;/sup&gt;</td>
<td>0.688&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction</td>
<td>1279</td>
</tr>
<tr>
<td></td>
<td>Loza 2003&lt;sup&gt;294&lt;/sup&gt;</td>
<td>0.74&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction</td>
<td>1825</td>
</tr>
<tr>
<td></td>
<td>Mills 2004&lt;sup&gt;135&lt;/sup&gt;</td>
<td>0.79&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction</td>
<td>610</td>
</tr>
<tr>
<td></td>
<td>Mills 2006&lt;sup&gt;372&lt;/sup&gt;</td>
<td>0.67&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction</td>
<td>424</td>
</tr>
<tr>
<td></td>
<td>Witte 2006&lt;sup&gt;283&lt;/sup&gt;</td>
<td>0.74&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction</td>
<td>1635</td>
</tr>
<tr>
<td>HCR-20</td>
<td>Dahle 2006&lt;sup&gt;217&lt;/sup&gt;</td>
<td>0.65&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction</td>
<td>3650</td>
</tr>
<tr>
<td></td>
<td>De Vogel 2004&lt;sup&gt;212&lt;/sup&gt;</td>
<td>0.82&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction</td>
<td>2175</td>
</tr>
<tr>
<td></td>
<td>Dolan 2004&lt;sup&gt;164&lt;/sup&gt;</td>
<td>0.67&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction</td>
<td>1770</td>
</tr>
<tr>
<td></td>
<td>Dolan 2007&lt;sup&gt;211&lt;/sup&gt;</td>
<td>0.715 (0.6 to 0.83)</td>
<td>Casefile</td>
<td>365</td>
</tr>
<tr>
<td></td>
<td>Douglas 2003&lt;sup&gt;288&lt;/sup&gt;</td>
<td>0.7&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction/casefile</td>
<td>1290</td>
</tr>
<tr>
<td></td>
<td>Douglas 2003&lt;sup&gt;267&lt;/sup&gt;</td>
<td>0.76&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction/casefile</td>
<td>626</td>
</tr>
<tr>
<td></td>
<td>Douglas 2005&lt;sup&gt;48&lt;/sup&gt;</td>
<td>0.82 (0.75 to 0.88)</td>
<td>Reconviction</td>
<td>2810</td>
</tr>
<tr>
<td></td>
<td>Fujii 2005&lt;sup&gt;95&lt;/sup&gt;</td>
<td>0.58&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Casefile</td>
<td>192</td>
</tr>
<tr>
<td></td>
<td>Grey 2004&lt;sup&gt;176&lt;/sup&gt;</td>
<td>0.56&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Casefile</td>
<td>2190</td>
</tr>
<tr>
<td></td>
<td>Grey 2007&lt;sup&gt;394&lt;/sup&gt;</td>
<td>0.68&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction</td>
<td>1825</td>
</tr>
<tr>
<td></td>
<td>Grey 2008&lt;sup&gt;396&lt;/sup&gt;</td>
<td>0.7&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Reconviction</td>
<td>1825</td>
</tr>
<tr>
<td></td>
<td>Lindsay 2008&lt;sup&gt;972&lt;/sup&gt;</td>
<td>0.72&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Casefile</td>
<td>365</td>
</tr>
<tr>
<td></td>
<td>Macpherson 2004&lt;sup&gt;596&lt;/sup&gt;</td>
<td>0.55 (0.42 to 0.67)</td>
<td>Casefile</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>McNiel 2003&lt;sup&gt;378&lt;/sup&gt;</td>
<td>0.65 (0.54 to 0.76)</td>
<td>Staff observation</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Nicholls 2004&lt;sup&gt;393&lt;/sup&gt;</td>
<td>0.75 (0.64 to 0.87)</td>
<td>Reconviction</td>
<td>690</td>
</tr>
<tr>
<td></td>
<td>Stadtland 2005&lt;sup&gt;530&lt;/sup&gt;</td>
<td>0.65 (0.54 to 0.75)</td>
<td>Reconviction</td>
<td>3240</td>
</tr>
<tr>
<td>ICT</td>
<td>Monahan 2005&lt;sup&gt;563&lt;/sup&gt;</td>
<td>0.7&lt;sup&gt;a&lt;/sup&gt;</td>
<td>Multiple</td>
<td>140</td>
</tr>
</tbody>
</table>

Note: The table continues with additional studies and data entries.
### TABLE 26 Area under the curve values and CIs by study (continued)

<table>
<thead>
<tr>
<th>Tools</th>
<th>Study ID</th>
<th>AUC (95% CI)</th>
<th>Measure</th>
<th>Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>LSI</td>
<td>Dahle 2006·277</td>
<td>0.71*</td>
<td>Reconviction</td>
<td>3650</td>
</tr>
<tr>
<td></td>
<td>Kroner 2005·310</td>
<td>0.689*</td>
<td>Reconviction</td>
<td>1279</td>
</tr>
<tr>
<td></td>
<td>Loza 2003·394</td>
<td>0.67*</td>
<td>Reconviction</td>
<td>1825</td>
</tr>
<tr>
<td></td>
<td>Mills 2006·652</td>
<td>0.72*</td>
<td>Reconviction</td>
<td>424</td>
</tr>
<tr>
<td></td>
<td>Yessine 2006·996</td>
<td>0.67 (0.6 to 0.75)</td>
<td>Reconviction</td>
<td>1241</td>
</tr>
<tr>
<td></td>
<td>Mills 2004·575</td>
<td>0.71*</td>
<td>Reconviction</td>
<td>610</td>
</tr>
<tr>
<td></td>
<td>Doyle 2006·272</td>
<td>0.82 (0.72 to 0.93)</td>
<td>Staff observation</td>
<td>84</td>
</tr>
<tr>
<td>MCAA</td>
<td></td>
<td></td>
<td>Casefile</td>
<td>2190</td>
</tr>
<tr>
<td></td>
<td>Grey 2004·96</td>
<td>0.71*</td>
<td>Casefile</td>
<td>2190</td>
</tr>
<tr>
<td></td>
<td>Snowden 2007·920</td>
<td>0.72*</td>
<td>Reconviction</td>
<td>1825</td>
</tr>
<tr>
<td>PCL-R</td>
<td>Beggs 2008·123</td>
<td>0.71*</td>
<td>Reconviction</td>
<td>1825</td>
</tr>
<tr>
<td></td>
<td>Dahle 2006·277</td>
<td>0.69*</td>
<td>Reconviction</td>
<td>3650</td>
</tr>
<tr>
<td></td>
<td>De Vogel 2004·232</td>
<td>0.75*</td>
<td>Reconviction</td>
<td>2175</td>
</tr>
<tr>
<td></td>
<td>Douglas 2005·98</td>
<td>0.76 (0.69 to 0.83)</td>
<td>Reconviction</td>
<td>2810</td>
</tr>
<tr>
<td></td>
<td>Hildebrand 2004·986</td>
<td>0.7 (0.59 to 0.8)</td>
<td>Reconviction</td>
<td>4307</td>
</tr>
<tr>
<td></td>
<td>Kroner 2005·310</td>
<td>0.672*</td>
<td>Reconviction</td>
<td>1279</td>
</tr>
<tr>
<td></td>
<td>Loza 2003·94</td>
<td>0.67*</td>
<td>Reconviction</td>
<td>1825</td>
</tr>
<tr>
<td></td>
<td>Mills 2006·652</td>
<td>0.69*</td>
<td>Reconviction</td>
<td>424</td>
</tr>
<tr>
<td></td>
<td>Stadtland 2005·850</td>
<td>0.64 (0.54 to 0.74)</td>
<td>Reconviction</td>
<td>3240</td>
</tr>
<tr>
<td></td>
<td>Walters 2005·847</td>
<td>0.62 (0.51 to 0.74)</td>
<td>Reconviction</td>
<td>1809</td>
</tr>
<tr>
<td></td>
<td>Dolan 2006·259</td>
<td>0.65 (0.55 to 0.75)</td>
<td>Casefile</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>Dolan 2006·258</td>
<td>0.73 (0.6 to 0.86)</td>
<td>Casefile</td>
<td>365</td>
</tr>
<tr>
<td></td>
<td>Douglas 2003·267</td>
<td>0.73*</td>
<td>Reconviction/casefile</td>
<td>626</td>
</tr>
<tr>
<td></td>
<td>Douglas 2005·959</td>
<td>0.65 (0.57 to 0.73)</td>
<td>Casefile</td>
<td>365</td>
</tr>
<tr>
<td></td>
<td>Douglas 2005·94</td>
<td>0.73 (0.66 to 0.8)</td>
<td>Reconviction</td>
<td>2810</td>
</tr>
<tr>
<td></td>
<td>Edens 2006·285</td>
<td>0.76*</td>
<td>Self-report/other reported/casefile</td>
<td>350</td>
</tr>
<tr>
<td></td>
<td>Grey 2004·96</td>
<td>0.58*</td>
<td>Casefile</td>
<td>2190</td>
</tr>
<tr>
<td></td>
<td>Grey 2007·304</td>
<td>0.69*</td>
<td>Reconviction</td>
<td>1825</td>
</tr>
<tr>
<td></td>
<td>McNeil 2003·538</td>
<td>0.61 (0.5 to 0.72)</td>
<td>Staff observation</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Nicholls 2004·303</td>
<td>0.7 (0.56 to 0.84)</td>
<td>Reconviction</td>
<td>690</td>
</tr>
<tr>
<td>PS</td>
<td>Hartvig 2006·427</td>
<td>0.71*</td>
<td>Other reported</td>
<td>365</td>
</tr>
<tr>
<td>RASP</td>
<td>Cunningham 2005·222</td>
<td>0.72 (0.7 to 0.74)</td>
<td>Casefile</td>
<td>1241</td>
</tr>
<tr>
<td></td>
<td>Cunningham 2006·222</td>
<td>0.67 (0.65 to 0.68)</td>
<td>Casefile</td>
<td>365</td>
</tr>
<tr>
<td>RM-2000</td>
<td>Thornton 2003·902</td>
<td>0.81*</td>
<td>Reconviction</td>
<td>1350</td>
</tr>
<tr>
<td></td>
<td>Thornton 2003·902</td>
<td>0.74*</td>
<td>Reconviction</td>
<td>6935</td>
</tr>
<tr>
<td>RRASOR</td>
<td>Harris 2003·422</td>
<td>0.56 (0.51 to 0.62)</td>
<td>Reconviction</td>
<td>1095</td>
</tr>
<tr>
<td></td>
<td>Langstrom 2004·98</td>
<td>0.59 (0.54 to 0.65)</td>
<td>Reconviction</td>
<td>2080</td>
</tr>
<tr>
<td></td>
<td>McGrath 2007·627</td>
<td>0.58 (0.39 to 0.77)</td>
<td>Reconviction</td>
<td>2117</td>
</tr>
<tr>
<td>Tools</td>
<td>Study ID</td>
<td>AUC (95% CI)</td>
<td>Measure</td>
<td>Period</td>
</tr>
<tr>
<td>-------</td>
<td>----------</td>
<td>--------------</td>
<td>---------</td>
<td>--------</td>
</tr>
<tr>
<td>SAQ</td>
<td>Seto 2005</td>
<td>0.75*</td>
<td>Reconviction</td>
<td>1642</td>
</tr>
<tr>
<td>Sjostedt 2002</td>
<td>0.73 (0.67 to 0.8)</td>
<td>Reconviction</td>
<td>2080</td>
<td></td>
</tr>
<tr>
<td>Loza 2003</td>
<td>0.71*</td>
<td>Reconviction</td>
<td>1825</td>
<td></td>
</tr>
<tr>
<td>Loza 2005</td>
<td>0.7*</td>
<td>Reconviction</td>
<td>365</td>
<td></td>
</tr>
<tr>
<td>SDRS</td>
<td>Lindsay 2008</td>
<td>0.72*</td>
<td>Casefile</td>
<td>365</td>
</tr>
<tr>
<td>SIRRI</td>
<td>Yessine 2006</td>
<td>0.76 (0.67 to 0.85)</td>
<td>Reconviction</td>
<td>1241</td>
</tr>
<tr>
<td>SORAG</td>
<td>Duro 2006</td>
<td>0.72 (0.62 to 0.82)</td>
<td>Casefile</td>
<td>1533</td>
</tr>
<tr>
<td>Harris 2003</td>
<td>0.73 (0.68 to 0.78)</td>
<td>Reconviction</td>
<td>1095</td>
<td></td>
</tr>
<tr>
<td>Looman 2006</td>
<td>0.69*</td>
<td>Reconviction</td>
<td>1861</td>
<td></td>
</tr>
<tr>
<td>Nunes 2002</td>
<td>0.69 (0.6 to 0.77)</td>
<td>Reconviction</td>
<td>2664</td>
<td></td>
</tr>
<tr>
<td>Seto 2005</td>
<td>0.73*</td>
<td>Reconviction</td>
<td>1642</td>
<td></td>
</tr>
<tr>
<td>START</td>
<td>Nicholls 2006</td>
<td>0.65 (0.57 to 0.72)</td>
<td>Staff observation</td>
<td>318</td>
</tr>
<tr>
<td>Static-2000</td>
<td>Bengtson 2008</td>
<td>0.7 (0.55 to 0.86)</td>
<td>Reconviction</td>
<td>5949</td>
</tr>
<tr>
<td>Static-99</td>
<td>Allan 2007</td>
<td>0.62 (0.53 to 0.71)</td>
<td>Reconviction</td>
<td>2117</td>
</tr>
<tr>
<td>Beech 2002</td>
<td>0.77 (0.55 to 0.98)</td>
<td>Reconviction</td>
<td>2190</td>
<td></td>
</tr>
<tr>
<td>Beggs 2008</td>
<td>0.58*</td>
<td>Reconviction</td>
<td>1825</td>
<td></td>
</tr>
<tr>
<td>Bengtson 2008</td>
<td>0.71 (0.56 to 0.86)</td>
<td>Reconviction</td>
<td>5949</td>
<td></td>
</tr>
<tr>
<td>De Vogel 2004</td>
<td>0.54*</td>
<td>Reconviction</td>
<td>4200</td>
<td></td>
</tr>
<tr>
<td>Duro 2006</td>
<td>0.68 (0.56 to 0.8)</td>
<td>Casefile</td>
<td>1533</td>
<td></td>
</tr>
<tr>
<td>Friendship 2003</td>
<td>0.7*</td>
<td>Reconviction</td>
<td>730</td>
<td></td>
</tr>
<tr>
<td>Hanson 2006</td>
<td>0.7 (0.67 to 0.72)</td>
<td>Reconviction</td>
<td>1825</td>
<td></td>
</tr>
<tr>
<td>Harris 2003</td>
<td>0.63 (0.57 to 0.68)</td>
<td>Reconviction</td>
<td>1095</td>
<td></td>
</tr>
<tr>
<td>Langstrom 2004</td>
<td>0.72 (0.67 to 0.76)</td>
<td>Reconviction</td>
<td>2080</td>
<td></td>
</tr>
<tr>
<td>Lindsay 2008</td>
<td>0.71*</td>
<td>Casefile</td>
<td>365</td>
<td></td>
</tr>
<tr>
<td>Looman 2006</td>
<td>0.56*</td>
<td>Reconviction</td>
<td>1861</td>
<td></td>
</tr>
<tr>
<td>Nunes 2002</td>
<td>0.69 (0.6 to 0.77)</td>
<td>Reconviction</td>
<td>2664</td>
<td></td>
</tr>
<tr>
<td>Olver 2007</td>
<td>0.57 (0.51 to 0.64)</td>
<td>Reconviction</td>
<td>3650</td>
<td></td>
</tr>
<tr>
<td>Sjostedt 2002</td>
<td>0.75 (0.68 to 0.81)</td>
<td>Reconviction</td>
<td>2080</td>
<td></td>
</tr>
<tr>
<td>Stadtland 2005</td>
<td>0.71 (0.62 to 0.8)</td>
<td>Reconviction</td>
<td>3240</td>
<td></td>
</tr>
<tr>
<td>Witte 2006</td>
<td>0.56*</td>
<td>Reconviction</td>
<td>1635</td>
<td></td>
</tr>
<tr>
<td>SVR</td>
<td>Stadtland 2005</td>
<td>0.68 (0.58 to 0.78)</td>
<td>Reconviction</td>
<td>3240</td>
</tr>
<tr>
<td>SVR-20</td>
<td>Craig 2006</td>
<td>0.53*</td>
<td>Reconviction</td>
<td>3650</td>
</tr>
<tr>
<td>De Vogel 2004</td>
<td>0.66*</td>
<td>Reconviction</td>
<td>4200</td>
<td></td>
</tr>
<tr>
<td>VRAS</td>
<td>Douglas 2005</td>
<td>0.61 (0.53 to 0.69)</td>
<td>Reconviction</td>
<td>2810</td>
</tr>
<tr>
<td>VRAG</td>
<td>Douglas 2005</td>
<td>0.79 (0.72 to 0.86)</td>
<td>Reconviction</td>
<td>2810</td>
</tr>
<tr>
<td>Grey 2007</td>
<td>0.74*</td>
<td>Reconviction</td>
<td>1825</td>
<td></td>
</tr>
<tr>
<td>Harris 2003</td>
<td>0.73 (0.68 to 0.78)</td>
<td>Reconviction</td>
<td>1095</td>
<td></td>
</tr>
</tbody>
</table>

© Queen’s Printer and Controller of HMSO 2013. This work was produced by Whittington et al. under the terms of a commissioning contract issued by the Secretary of State for Health. This issue may be freely reproduced for the purposes of private research and study and extracts (or indeed, the full report) may be included in professional journals provided that suitable acknowledgement is made and the reproduction is not associated with any form of advertising. Applications for commercial reproduction should be addressed to: NIHR Journals Library, National Institute for Health Research, Evaluation, Trials and Studies Coordinating Centre, Alpha House, University of Southampton Science Park, Southampton SO16 7NS, UK.
### TABLE 26 Area under the curve values and CIs by study (continued)

<table>
<thead>
<tr>
<th>Tools</th>
<th>Study ID</th>
<th>AUC (95% CI)</th>
<th>Measure</th>
<th>Period</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Kroner 2005</td>
<td>0.746*</td>
<td>Reconviction</td>
<td>1279</td>
</tr>
<tr>
<td></td>
<td>Kroner 2007</td>
<td>0.702*</td>
<td>Reconviction</td>
<td>1741</td>
</tr>
<tr>
<td></td>
<td>Lindsay 2008</td>
<td>0.71*</td>
<td>Casefile</td>
<td>365</td>
</tr>
<tr>
<td></td>
<td>Loza 2003</td>
<td>0.63*</td>
<td>Reconviction</td>
<td>1825</td>
</tr>
<tr>
<td></td>
<td>Mills 2006</td>
<td>0.71*</td>
<td>Reconviction</td>
<td>424</td>
</tr>
<tr>
<td></td>
<td>Quinsey 2004</td>
<td>0.69*</td>
<td>Casefile</td>
<td>480</td>
</tr>
<tr>
<td></td>
<td>Snowden 2007</td>
<td>0.76*</td>
<td>Reconviction</td>
<td>1825</td>
</tr>
<tr>
<td>VRS</td>
<td>Urbanik 2006</td>
<td>0.72 (0.61 to 0.81)</td>
<td>Reconviction</td>
<td>0</td>
</tr>
<tr>
<td>VRS</td>
<td>Yessine 2006</td>
<td>0.74 (0.59 to 0.9)</td>
<td>Reconviction</td>
<td>1241</td>
</tr>
<tr>
<td>VSC</td>
<td>Dolan 2007</td>
<td>0.713 (0.6 to 0.83)</td>
<td>Casefile</td>
<td>365</td>
</tr>
<tr>
<td>VSC</td>
<td>Olver 2007</td>
<td>0.56 (0.5 to 0.62)</td>
<td>Reconviction</td>
<td>3650</td>
</tr>
<tr>
<td>VSC</td>
<td>McNiel 2003</td>
<td>0.74 (0.68 to 0.86)</td>
<td>Staff observation</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td>Nicholls 2004</td>
<td>0.31 (0.18 to 0.44)</td>
<td>Reconviction</td>
<td>690</td>
</tr>
</tbody>
</table>

* 95% CI data are missing for this value.
Appendix 3 Protocol

Risk Assessment Protocol

1. Cover Sheet

Title: Systematic Review of Risk Assessment Strategies for the Prevention, Treatment and Management of Violent Behaviour by Adults in Contact with Forensic Mental Health Services or the Criminal Justice System.

Reviewers:

Dr Wally Barr
Senior Research Fellow
Health and Community Care Research Unit
University of Liverpool

Ms. Juliet Hockenhull
Research Fellow
Liverpool Reviews and Implementation Group
University of Liverpool

Dr Maria Leitner
Director
Infotech UK Research

Professor James McGuire
Academic Director/Head of Course
Division of Clinical Psychology
University of Liverpool

Dr Richard Whittington
Reader
Health and Community Care Research Unit
University of Liverpool

Contact information for lead researcher:

Dr. Richard Whittington
Health and Community Care Research Unit (HaCCRU), School of Health Sciences, University of Liverpool

Sources of Support: Department of Health (England): National Forensic Mental Health R&D Programme (original review) and National Institute for Health Research/Research for Patient Benefit Programme (review update)

2. Background to the Review

Violent behaviour is a significant source of public and political concern, and most perpetrators will eventually come into contact with either the forensic mental health (FMH) services or the criminal justice system (CJS) (or both). This contact provides an opportunity for assessment of the individual’s risks and needs and for interventions aimed at managing violence within the institutional setting and preventing future violence within the community. Numerous risk assessment and risk management technologies have been developed over the past thirty years which are available for practitioners to deploy when working with individual perpetrators, and many of these technologies have at least a moderate evidence base. The
systematic review proposed here sets out to address the global evidence base underpinning risk assessment strategies for preventing, treating and managing violence in both FMH and CJS settings. It will be conducted in parallel with another review (submitted to the Campbell Collaboration under separate cover) addressing issues relating to interventions for violence.

Two broad categories of risk assessment research approaches can be distinguished. Unstructured risk assessment is based on establishing the relationship between a wide array of clinical, social, demographic or other factors and a violent outcome. Choice of one or more of these factors in any risk model should be driven by relevant biological, psychological and/or sociological theories but in fact may be selected for study on the basis of serendipity. At the most basic level, the size of differences in rates or scores on a single variable between violent and non-violent samples (or between violent sub-samples) are estimated and tested for statistical significance e.g. (Cantor, Kabani et al. 2008). Given the complexity of human aggression, more sophisticated multivariate models are preferable and a number of these have been developed and tested for different populations (e.g. Busby, Holman et al. 2008). Data on the variables used in this approach are usually only available retrospectively.

Structured risk assessment is a more formal approach based on psychometric theory and established methods for developing research and clinical instruments. A potentially relevant factor (e.g. anger) is operationalised, an item pool is constructed and refined through piloting, and normalised scores are obtained. Predictive validity is established through estimating the sensitivity (true positive rate) and specificity (true negative rate) of the established instrument and then, since the pay-off between the two estimates is crucial, plotting the relationship of sensitivity and specificity on a receiver-operator characteristic (ROC) curve. A very large number of instruments have been developed in this way over the past decade with varying degrees of success e.g. HCR-20 (Webster, Douglas et al. 1997). This approach has numerous advantages in terms of scientific precision over the unstructured approach, not least because the data on the variables used in this approach are usually available prospectively. However, the instruments can be laborious to complete in comparison to some of the unstructured variables and thus may be unpopular in everyday practice. Structured risk assessment tools can be further subdivided into those which are primarily actuarial and those which are based on structured clinical (or professional) judgement (SCJ). Actuarial tools used in isolation are increasingly viewed as inferior for various reasons including their tendency to emphasise unchangeable static risk factors and to ignore risk factors which do not occur commonly.

After twenty years of sustained activity in this area, the primary research literature is now very large yet the evidence base for making clinical and policy decisions is often bemoaned as inadequate (Department of Health 2000) A number of systematic reviews have been conducted to summarise and integrate the findings from the literature and these provide evidence on a number of specific areas (Risk management Authority 2006). However, inevitably these reviews tend to focus on a specific instrument, and/or a specific outcome (e.g. reoffending) in various special populations (e.g. sex offenders). This review will instead adopt a more comprehensive approach by aiming to capture research on all instruments and risk factors relating to a broad range of violence-related outcomes amongst a wide FMH and CJS population. In this way it is anticipated that the fragmented clinical and criminological literatures can be reintegrated to the mutual benefit of practitioners and researchers in both settings (Hollin 2008).

This Risk Assessment review is being conducted in tandem with a review of Interventions with the same population and it is important to emphasise that the two processes should be closely linked. Estimates of predictive validity from a risk assessment tool are of little use on their own if they are not used to design and target effective interventions. This is another advantage of SCJ approaches as these are recognised as encouraging practitioners to focus on risk management and to remain flexible rather than relying on straightforward prediction based on static factors.

The protocol builds on the work of a previously completed systematic review in this area. The final report of this review has had significant influence on national policy in England and is currently flagged on the website of the Department of Health/Ministry of Justice (England) National Risk Management
Programme (CSIP/NIMHE). It also formed the basis for a set of national best practice guidelines on risk management (Department of Health 2007) and national policy guidance on selection of risk assessment tools (Leitner 2006).

3. Objectives of the review

3.1 To provide a systematic review of primary research evaluating the sensitivity, specificity and clinical utility of risk assessment strategies for the prevention of violent behaviour by people in contact with forensic mental health or criminal justice systems.

3.2 To produce a general statement about the efficacy of risk assessment for violent behaviour specifically targeted at people in this group through the synthesis of individual study results.

3.3 To examine reasons for conflicting evidence on effectiveness in this area.

4. Methods

This protocol relates to a systematic review which, in its entirety will cover the publication period from the inception of the research literature to mid-2008. The original review (covering studies published up to the end of 2002) has been completed and resulted in the inclusion of approximately 600 studies in the Liverpool Violence (LiVio) Research Archive and the construction of an associated SPSS database of extracted information on 200+ variables per study. A technical report on the original review is available (Leitner, Barr et al. 2006). The review update, covering studies published between 2002 and 2008 will conform closely to the original review methods and subsequently the findings will be combined.

Estimating the value of risk assessment tools for the assessment of violent behaviour risk is conceptually akin to assessing the value of screening and diagnostic tools designed to identify disorders of physical health. That is, the aim is to address:

- Validity, reliability and related ‘technical’ characteristics
- Sensitivity and specificity in predicting target outcomes
- Acceptability to patients and clinicians (i.e. clinical utility)

The methods identified for the review take into account these requirements and the ways in which the review process may need to depart from the more ubiquitous type of systematic review focusing on the effectiveness of interventions.

4.1 Criteria for inclusion and exclusion of studies in the review

For a study to be included in this systematic review it must have the following characteristics:

I. Participant/Population characteristics

1. The study participants must (a) have an active diagnosis of mental illness, learning disability or personality disorder, OR (b) be an offender (person subject to penal sanction), OR (c) be a person known to have committed one or more acts of aggression constituting an indictable offence (whether or not an indictment has been made). Studies will be excluded if (a) the sample participants are members of the general public, with no identified mental illness and no evidence of having committed an act of violence which would constitute and indictable offence, (b) Substance abuse (including alcohol abuse) in isolation from any other diagnosis of mental illness is not to be defined for the purposes of the review as an active diagnosis of mental illness. Substance abuse (including and separately specified as alcohol abuse) is to be identified in relation to participant characteristics for the purposes of data extraction, as it is identified in primary studies.

2. The study participants must be aged 17 years and older.
II. Risk assessment Characteristics

1. Risk assessments must be (a) any potential risk assessment instrument OR (b) any potential individual risk factor.

Studies will be excluded if:

(a) risk factors are evaluated solely in the context of predicting target behaviours other than aggression towards others

(b) focused solely on self-directed aggression, including self-harm and suicidal behaviours. (Studies which have a focus on a main target behaviour which is not other-directed aggression (the target behaviour may be self-directed aggression), but which do include an evaluation of the association of potential risk factors with other-directed aggression as a subsidiary focus are to be included)

(c) they only evaluate the impact of broad-based local or national population-level initiatives and fail to evaluate outcomes at the individual level. For example, a study evaluating poverty as a risk factor for aggression which evaluated outcomes purely by noting changes in population rates of violence across time would be excluded; a study evaluating the same potential risk factor but reporting outcomes based on the same set of individuals with behaviour evaluated before and after the initiative would be included. The key point is that the specific individuals being assessed need to be evaluated at outcome.

2. Risk assessments must be evaluated for immediate association with violent behaviour (e.g. ‘naturalistic’ evaluation in a clinical setting).

3. Risk assessments may include, but are not restricted to demographic, psychological, neurological, neuro-chemical or genetic factors, psychometric and non-psychometric scales, indices and check-lists

III. Setting/location

1. Studies must be conducted in (a) any institutional setting/location, (b) any community setting/location. Community-based ‘institutional’ settings such as out-patient clinics, A&E, private practice clinics etc. are also to be included. Studies conducted at ‘remote’ locations, for example risk evaluations conducted by telephone or in writing, are also to be included. Setting/location of any study is not to be regarded as grounds for excluding that study

IV. Study Design Characteristics

1. The study design must match one of five basic designs:

(a) Test-retest design: a structured instrument is administered on at least two occasions to the same sample and associations between scores across measurement points are reported

(b) Concurrent validity design: a structured instrument is administered on at least one occasion alongside either (i) administration of a second structured instrument or (ii) measurement of a violence-related phenomenon, and associations between the two instruments or the instrument and the phenomenon are reported

(c) Predictive validity design: a structured instrument is administered on at least one occasion and at some point subsequently either (i) a second structured instrument is administered or (ii) a violence-related phenomenon is measured, and associations between the two instruments or the instrument and the phenomenon are reported
(d) Concurrent risk factor evaluation design: a phenomenon is measured on at least one occasion alongside measurement of a violence-related phenomenon, and associations between the two phenomena are reported

(e) Predictive risk factor evaluation design: a phenomenon is measured on at least one occasion and at some point subsequently a violence-related phenomenon is measured, and associations between the two phenomena are reported

(f) Clinical utility evaluation design: a quantitative or qualitative assessment of the utility of a structured instrument is made at least once. Utility is defined as any aspect of the process of administering the instrument (e.g. time for completion)

2. Studies must use individual-level data. Studies will be excluded if (a) evaluations are based on non-attributable rates or other summary data (for example, an epidemiological study which contrasted whole-population rates before and after an educational initiative targeting violence prevention would be excluded, the same initiative with outcomes measured for an identified group of individuals with any changes in rates of aggression calculated on the basis of individual ‘scores’ or profiles would be included), OR (b) the aggressive behaviour is ‘Collective’ acts of aggression, such as terrorism, ‘gang’ violence, organised violent crime, football violence, drug feuds etc. where the focus of the study is on the phenomenon as a collective behaviour; studies focussed specifically on individual behaviour within these contexts should be included

4.2 Search strategy for identification of relevant studies

A search strategy for electronic databases (outlined in generic form below) was developed in collaboration with information technology staff from the British Library, taking into account lessons drawn from previous work in similar areas, kindly supplied to us by colleagues in the Cochrane and Campbell Collaborations. The search strategy is intentionally broad. The approach adopted for search development was the Successive Fractions approach described by (Hartley, Keen et al. 1993). Initial trials of the search strategy were carried out on the DIALOG system by British Library information staff and subsequently refined by the Review Team using MEDLINE as a search model. The search strategy is designed to be sufficiently inclusive to provide a comprehensive overview of relevant material in this area. It will be used to identify both completed and ongoing research and will encompass both primary research and review material.

4.2.1 Search term (structure modified to suit individual data sources)

(((Homicid* OR murder* OR manslaughter* OR infanticid* OR parricid* OR assault* OR (bodily AND (harm OR assault)) OR assais* OR bugger* OR sodom* OR molest* OR pedophil* OR paedophil* OR sadis* OR sadomasochis* OR sado-masochis* OR anger* OR cruel* OR rapist* OR (rape* AND offend*) OR physical abus* OR spouse abus* OR partner abus* OR sexual abus*) OR (( (dangerous* AND (behavior* OR behaviour* OR histor* OR conduct*)) OR violen*) AND (risk* OR predict* OR anteced* OR assess* OR cause* OR reason* OR interven* OR prevention* OR preventing* OR controlling* OR manage* OR treatment* OR treating* OR reduction* OR reducing* OR stop* OR mental* OR forensic* OR psychiatric* OR offend* OR Axis 1 OR Axis 2 OR criminal* OR detain* OR insane* OR NGRI OR retard* OR (learning disab* OR learning-disab*) OR acquit*)) OR ((child abus* OR elder abus* OR hostil* OR killing* OR attack* OR aggress*) AND (mental* OR forensic* OR psychiatric* OR offend* OR axis 1 OR axis 2 OR criminal* OR detain* OR insane* OR NGRI OR retard* OR (learning disab* OR learning-disab*) OR acquit*) OR disorder*))) NOT (cancer* OR cancer [mh] OR tumo* OR tumour [mh] OR heart* OR heart [mh]))

4.2.2 Electronic searches

Electronic searches are not restricted by either geographic or site location of the research or the type of publication. In the review update, studies will be restricted to those with an English language abstract and dissertations will be restricted to those available electronically. Electronic searches will be restricted to the publication period 2002–2008. The following sources will be searched:
Following a reliability exercise within the team, inclusion criteria will be applied to the search results in two stages. First, each reviewer will be allocated a subset of the citations (title, publication details and abstract) to which they will independently apply the inclusion criteria. Full-text versions of all studies deemed to meet inclusion criteria will be obtained for full review. Stage 2 will involve the application of
the inclusion criteria to full-text versions that were identified. Each paper will be looked at by one reviewer. A conservative, inclusive approach will be adopted towards doubtful studies so that reviewers will err in favour of inclusion where any uncertainty exists and decisions regarding inclusion will be made through consultation with a second reviewer.

4.2.3 Handsearching, reference lists and consultation with experts
The original review demonstrated that the benefits of handsearching 34 journals did not justify the effort involved. Therefore, in this review update, the five most relevant journals will be identified empirically and handsearched for the period 2002–2008 in order to ensure the comprehensiveness of the review and assess the reliability of the electronic search.

The Review Team will also hand search the reference lists of all systematic reviews obtained in the course of the review process.

Discussions with, and formal requests to, experts in the field – notably those who have authored reviews and/or are actively engaged in primary research – will also be used to supplement the formal searches. Finally, the Advisory Panel will be asked to review the complete list of selected material for missing studies of relevance to the review.

4.2.4 Data Management
Citations and abstracts downloaded from the electronic searches will be entered into Endnote (a data management package for bibliographic material). Material from separate databases will be combined in a composite database, prior to pre-screening for inclusion, to exclude duplicates. Citations from each data source will be catalogued separately and tagged to allow the Review Team to keep track of the relative value of each source in contributing to the final review material. As the search strategy has also been developed to inform the intervention review mentioned earlier, a tagging system will be used in the initial screening stages to track material of relevance to each review, as there will be some overlap. Separate databases will then be established for the two reviews.

4.3 Description of methods used in the component studies

Quality assessment of retrieved material
Study designs appropriate to addressing the performance of an assessment or screening tool (e.g. observational studies with comparisons drawn against a ‘gold standard’) are subject to less well-developed quality guidelines than is the case for intervention research and the area of risk assessment currently lacks a ‘gold’ standard comparator. Nevertheless, tools to support quality assessment in related areas, such as diagnostic screening, are increasingly common (e.g. (Mulrow, Linn et al. 1989; Bossuyt, Reitsma et al. 2003)) and the Review Team will explore the value of extant scales and checklists in the current context.

The preferred ‘ideal’ model would be for blind comparisons against a known reference standard whilst the use of clinical opinion without objective support and/or a case–control design would be seen as rather less optimal. Therefore, all studies will be evaluated on an 8-item, design quality scale specified in the ‘coding categories’ section below.

Experience from the earlier review process indicates that four main groups of studies are likely to be captured. Predictive validity studies will be concerned with establishing the relationship between predictors and outcome variables and thus will enable the estimation of instrument sensitivity and specificity. Other validity studies will be concerned with other aspects of validity such as construct and concurrent validity. Reliability studies will be concerned with the consistency of measurement over time and between raters. Clinical utility studies will be concerned with any aspect of instrument acceptability amongst practitioners or service users. These issues are of equal importance to the intended audiences for the review, hence the decision to keep inclusion criteria broad.
All included studies will be evaluated against the STARD criteria for reporting of studies on diagnostic criteria (Bossuyt et al., 2008).

It should be noted that the review is designed to be as comprehensive as possible and thus to capture qualitative designs. The field is likely to be dominated by lower-quality designs due to the complexity of the population and other factors so evidence must be based, with appropriate caveats, on these lower-quality designs.

4.4 Criteria for determination of independent findings
The reviewers will attempt to identify samples reported in more than one paper. Where this is detected, the most stringent test (i.e. the paper with the longest period between baseline and endpoint) will be selected for inclusion in the quantitative synthesis. Where individual studies report multiple outcomes \( k \) each of them will be coded separately for analysis. The method of computing outcomes will be coded as a method variable. Discrete analyses will be conducted across effect size measures integrating findings obtained with different measures as separate outcome variables. For all effect size measures so obtained, conversion formulae will be used to present overall findings in several ways, for example as mean effect sizes (Cohen's \( d \)), correlation coefficients (\( r \) or \( \varphi \)), and odds ratios where appropriate.

Findings utilising identical outcome variables within studies (e.g. from separate sub-samples) will be coded as independent outcome indicators and regarded as equivalent to outcome variables comparably defined from other studies. Where individual studies report a number of variables, types of outcomes will be coded and in each case mean effect sizes will only be computed for individual variables of comparable types from independent studies. Where studies report multiple outcome measures, the reviewers will identify the main effect size for one primary and one subsidiary outcome measure on the basis of the authors’ stated goals. Any additional effect sizes (either for these outcome measures or any subsidiary outcome measures) will be coded in a separate annex to the main coding form.

4.5 Data extraction
Data extraction will be performed by two coders and extracted data will be loaded onto the LiVio SPSS database holding information from the original review. For conceptual clarity the extracted variables will be grouped into the following clusters of coding categories, which will assist in defining separate analyses and inferential tests to be conducted.

- Data management cluster
- Publication cluster
- Design cluster
- Sample cluster
- Assessments cluster
- Outcomes cluster
- Results cluster

4.5.1 Data synthesis
Where available, sensitivity and specificity data will be aggregated across studies, as appropriate. A narrative synthesis of the available material will also be used to explore and outline the extent, nature and quality of the available evidence in this area. This qualitative assessment of the available data will also be used to explore any observed heterogeneity (in study or sample characteristics, study designs and outcomes) and to inform the structure for quantitative synthesis of the data, including the choice of comparisons to be made and the outcome measures amenable to quantitative treatments. It will also be used to address the issue of generalisability. The extent of observed heterogeneity will then be established quantitatively (e.g. \( Q \) or \( I^2 \)) and, where appropriate, data will be combined in quantitative synthesis as outlined below, to obtain combined estimates of predictive value for specific risk assessment tools and their associated confidence intervals. It is unlikely that individual patient data will be made available to the Review Team given the timescale of the review. Sensitivity analyses will be used to explore the robustness
of the review outcomes to changes in the underlying assumptions regarding the data and regarding the methods applied. Publication bias will be explored using funnel plots.

4.6 Statistical procedures and conventions

The statistical procedures available to the Review Team are in large part dependent on the nature of the available primary research material. For example, the primary research may present dichotomous outcomes separately in relation to the sensitivity and specificity of risk assessment tools using proportions (either absolute or as a function of predictive values) or, more reliably, using likelihood ratios. Alternatively, or in addition, the research may balance sensitivity and specificity using receiver operating characteristics (ROC) curves. In a minority of studies outcomes may be presented as continuous data, with or without thresholds, or addressed in purely qualitative terms. A further complication in this field is that there is ongoing debate over the most appropriate means of summarising results (e.g. averaging proportions or likelihood ratios; pooling odds ratios or fitting summary ROC curves). In carrying out any quantitative synthesis, the Review Team will explore the impact on outcomes of presenting and combining data in the range of ways outlined. The impact of cluster variables, as outlined in relation to data extraction, on outcome will be explored in quantitative synthesis where suitable data are available e.g. to address the appropriateness of a risk assessment tool for different sub-populations or to explore the impact of study design on the apparent value of an assessment tool. These analyses will be guided by the narrative overview of available data. Summaries of the quantitative data will be presented in both tabular and graphical form.

The most appropriate method of quantitative synthesis depends on the nature of the data identified. A final decision regarding whether quantitative synthesis is appropriate at all and, if so, which method(s) should be adopted will therefore be made once the data has been collected. Judging from the original review, binary data in quantitative synthesis can be validly presented either as odds ratios or as relative risk ratios, since the base rates for violence are generally low and both measures give comparable estimates under this condition. Absolute risk differences are less likely to be appropriate, since in the original review variation in baseline event rates was commonly found when comparing across studies, even where these used very similar measures and populations. In comparing odds ratios and relative risk, the eventual choice of effect measure for the quantitative synthesis of binary data is likely to depend on the eventual audience for the outcomes of a particular analysis. For example, physicians are more familiar with the concept of relative risk and may find results presented using this effect measure more readily interpretable. In contrast statisticians and psychologists are more familiar with odds ratios.

In quantitative synthesis of continuous data a weighted mean difference effect measure is the most likely choice, with the weight given to the mean difference in each study equal to the inverse of the variance of the effect size. However, the original review revealed that a number of otherwise comparable studies had measured outcomes using different scales. In such cases, it would be more appropriate to adopt a standardised mean difference approach (dividing the mean difference by an estimate of the within-group standard deviation to produce a unit-free standardised measure of effect). This will produce ‘equated effect sizes’. It should also be noted that a number of studies in the original review used survival curve data to summarise outcomes. In combining such studies in a quantitative synthesis, it would be most appropriate to use hazard ratios as the effect measure.

As stated above, the statistical analysis will follow C2 guidelines and report, as appropriate, quantitative synthesis of higher-quality studies, comparative groups and pre–post designs separately. Analysis of the original review studies identified an unusually high degree of heterogeneity between studies. This was sufficient in fact to rule out quantitative synthesis as an appropriate approach in all but a minority of sub-groups of the studies included. It is anticipated that the research literature since 2001 has become more coherent given the development of protocols etc. and thus that more recent studies captured in the review update will show a greater degree of homogeneity. Nevertheless it seems likely that a random effects model will be the most appropriate approach to combining data in quantitative synthesis. (Popay, Rogers et al. 1998; Thomas and Harden 2007)
The reviewers aim to present outcomes using both fixed and random effects models. This will be especially pertinent if it is found that publication bias, poor design and implementation quality remain an issue in more recent studies.

The main categories for meta-analyses that will be considered where appropriate data are available will include:

1. Test-retest reliability of the risk assessment tools (using $r$ as effect size)
2. Concurrent criterion validity studies (using either $r$ or odds ratio as effect size, depending on whether the measures are binary or continuous)
3. Predictive criterion validity studies (using either $r$ or odds-ratio as effect size, depending on whether the measures are binary or continuous)
4. Sensitivity of risk assessment tool (proportion; odds-ratio; risk-ratio)
5. Specificity of risk assessment tool (proportion; odds-ratio; risk-ratio)
6. ROC

The original review clearly identified that moderator variables in this context are confounded. Associations within and between moderators will be initially identified via tests of individual association appropriate to the variables in question (e.g. correlation coefficients for continuous variables, $\chi^2$ statistic for discrete variables). The combined impact of multiple moderator variables identified as confounded will then be modelled using suitable meta-analytic regression analyses.

The original review also identified study design (broadly described here as ‘method’) as a moderator variable. Given also a priori concerns regarding the quality of distinct designs, the reviewers intend, if sufficient resources are available, to run a set of meta-analyses weighting effect sizes by study design/‘quality’ rather than simply by sample size in order to evaluate the impact on outcomes. Following the outcome of the moderator regression analyses described above, this analysis may be redundant, in which case the plan of analysis will be adjusted accordingly.

4.7 Treatment of qualitative research

Qualitative primary research studies (or components of studies) are likely to be of comparatively little value in establishing the sensitivity and specificity of risk assessment tools. However, they will be of relevance to closely associated issues such as the acceptability (to patient and clinician) of different assessment tools and similarly will be informative in relation to the likely pragmatic constraints pertaining to assessment in different settings or to different procedures for conducting assessment. They will also be of value in providing additional contextual information to inform overall judgments of the likely ‘cost/benefit’ of using different assessment tools. The criteria for inclusion/exclusion of qualitative studies will be as outlined earlier. Quality hierarchies appropriate to this type of material will be as rigorously applied as for the quantitative literature (Popay, Rogers et al. 1998; Thomas and Harden 2007).

5. Timeframe

We intend to produce the updated review report by July 2009. The project has been funded and is currently under way, with a project timetable and milestones agreed with the funders as follows:

October 2008: identification of relevant studies completed.

December 2008: data extraction and loading completed.

March 2009: data analysis completed.

The Review Report to be provided to the funding body will serve as a focus for dissemination. Rather than breaking this large report into separate journal articles, a contract has been obtained with Cambridge University Press for production of a research monograph incorporating both this and the parallel Intervention review. Executive summaries of the report will be made available to relevant stakeholders.

6. Plans for updating the review
All search material will be maintained on Endnote. Updating and subsequent transparency will be supported by clear documentation of the search process. If the Campbell Collaboration accept the review, the expectation would be for biennial updates of the review to be carried out, providing sufficient funding or institutional support could be obtained to secure the necessary staff time.

7. Acknowledgements
We are grateful to the National Forensic Mental Health Research & Development Programme (Department of Health, England) for funding the original review. We also thank the Programme’s former Director (Kathryn Harney) for her continued interest in and support of the project and the members of the original project’s Advisory Panel (Dr Ron Blackburn; Dr David Cooke; Dr Mairead Dolan; Dr Tom Mason; Mr Cathal Meehan; Dr Malcolm Millar and Prof Jenny Shaw) for volunteering their valuable time to the project. With regard to the review update, we are also grateful to the National Institute for Health Research/Research for Patient Benefit Programme (Department of Health, England) for funding, the members of the advisory panel (Prof Jenny Shaw; Dr. Joy Duxbury; Ruth Sayers; Sue Imlack; Kathryn Harney; and Dr. Caroline Logan) and to Ms. Rumona Dickson for additional advice.

8. Statement concerning conflict of interest
None.

9. References


