Violence risk assessment instruments in forensic psychiatric populations: a systematic review and meta-analysis

Maya G T Ogonah, Aida Seyedsalehi, Daniel Whiting, Seena Fazel

Summary

Background Although structured tools have been widely used to predict violence risk in specialist mental health settings, there is uncertainty about the extent and quality of evidence of their predictive performance. We aimed to systematically review the predictive performance of tools used to assess violence risk in forensic mental health, where they are routinely administered.

Methods In our systematic review and meta-analysis, we followed PRISMA guidelines and searched four databases (PsycINFO, Embase, Medline, and Global Health) from database inception to Nov 1, 2022, to identify studies examining the predictive performance of risk assessment tools in people discharged from forensic (secure) mental health hospitals. Systematic and narrative reviews were excluded from the review. Performance measures and descriptive statistics were extracted from published reports. A quality assessment was performed for each study using the Prediction Model Risk of Bias Assessment Tool. Meta-analysis was conducted on the performance of instruments that were independently externally validated with a sample size greater than 100. The study was registered with PROSPERO, CRD42022304716.

Findings We conducted a systematic review of 50 eligible publications, assessing the predictive performance of 36 tools, providing data for 10 460 participants (88% men, 12% women; median age [from 47 studies] was 35 years, IQR 33–38) from 12 different countries. Post-discharge interpersonal violence and crime was most often measured by new criminal offences or recidivism (47 [94%] of 50 studies); only three studies used informant or self-report data on physical aggression or violent behaviour. Overall, the predictive performance of risk assessment tools was mixed. Most studies reported one discrimination metric, the area under the receiver operating characteristic curve (AUC); other key performance measures such as calibration, sensitivity, and specificity were not presented. Most studies had a high risk of bias (49 [98%] of 50), partly due to poor analytical approaches. A meta-analysis was conducted for violent recidivism on 29 independent external validations from 19 studies with at least 100 patients. Pooled AUCs for predicting violent outcomes ranged from 0·72 (0·65–0·79; I^2=0%) for H10, to 0·69 for the Historical Clinical Risk Management-20 version 2 (95% CI 0·65–0·72; I^2=0%) and Violence Risk Appraisal Guide (0·63–0·75; I^2=0%), to 0·64 for the Static-99 (0·53–0·73; I^2=45%).

Interpretation Current violence risk assessment tools in forensic mental health have mixed evidence of predictive performance. Forensic mental health services should review their use of current risk assessment tools and consider implementing those with higher-quality evidence in support.

Funding Wellcome Trust.

Introduction

Violence perpetrated by individuals after hospital discharge from forensic psychiatric care is a major concern for psychiatric services, with costs to the wider economy, public health, and social care.[1] Such individuals continue to present a risk of violence following discharge, with a meta-analysis[2] finding crude violent reoffending rates of 3902 per 100 000 person-years, equivalent to approximately 4% per year. Estimates for reconviction rates for violent offences stand at approximately 12% within 2 years of discharge.[3] Those who enter forensic psychiatric services often do so as an alternative to prison, with an expectation that access to specialist mental health treatment will more effectively reduce reoffending risk.

Assessing and managing post-discharge violence risk is a core feature of forensic psychiatric care; risk assessment instruments aid in the identification of individuals at risk of violence perpetration and assist in risk management. Owing to the potential use of such tools, and with increasing recognition of the importance of predicting violence perpetration, violence risk assessment has been the subject of much clinical and research interest. More than 400 risk assessment instruments designed to assess the risk of violence and offending have been developed,[4] and many have been implemented in forensic mental health around the world. Despite their widespread use, definitive syntheses of evidence for their predictive performance are scarce. Previous reviews are outdated[5–7] or do not follow reporting

Research in context

Evidence before this study

Violence perpetrated by individuals after discharge from hospital for forensic psychiatric care is a major concern for psychiatric services. Although various risk assessment tools have been developed to assist with risk prediction and management, it is not known which tools are supported by high quality evidence and which are most accurate. We searched PubMed from database inception to Jan 10, 2022, without language restrictions using the search term (risk assess*) AND (predict* OR accura* OR "psychometric properties") AND (violen* OR crime) AND ("systematic review" OR "Meta-analy*") as a filter. We identified five relevant systematic reviews or meta-analyses that examined one or a limited number of tools, combined populations (eg, from both prison and forensic settings), combined inpatient and community outcomes, or only examined inpatient violence. We found no systematic reviews or meta-analyses specifically examining the predictive performance of risk assessment tools in forensic psychiatric patients after discharge. To develop guidance for clinicians and policy makers on which risk assessment instruments should be considered in forensic mental health, clarification of the predictive performance of individual instruments is necessary.

Added value of this study

This comprehensive synthesis of the performance of 36 risk assessment instruments used in forensic mental health settings in 10 460 participants found that most research on these tools is at high risk of bias and has only reported one performance measure—the area under the receiver operating characteristic curve (AUC)—rather than a basic suite of tests of discrimination (sensitivity and specificity) and calibration. Estimates of the pooled AUC for predicting violent recidivism ranged from 0·64 to 0·72. Furthermore, there was little research on women in forensic psychiatric populations (only two studies assessed risk assessment tools in female-only samples) and there was no research in low-income or middle-income countries.

Implications of all the available evidence

In forensic mental health services, the implementation of high quality risk assessment tools can complement clinical decision making. However, these tools should not be used to inform decisions of lengths of stay without validations showing high sensitivity, and they should only be used to inform discharge planning and resource allocation when validations show high levels of specificity. Benchmarks for acceptable sensitivity and specificity need clarification. This meta-analysis suggests current practice needs review. In the case of new services, including in low-income and middle-income countries, risk assessment tools that are introduced should have validation studies in support, with information on discrimination (including rates of false positives and negatives) and calibration provided.

guidelines, and uncertainties remain regarding the use of risk assessment instruments in new settings and populations. Furthermore, previous reviews typically combine inpatient violence with post-release community violence.[5,7,8] As violence base rates, key risk factors, and the type of intervention differ between inpatient samples and individuals in the community, a review that disaggregates inpatient and post-discharge samples is necessary. Previous research[9] has sought to review the performance of risk assessment instruments in inpatient samples; however, to our knowledge, there are no syntheses specifically assessing tool performance in predicting post-release violence.

Despite its importance, very few instruments have been designed for violence risk assessment in forensic psychiatric patients after hospital discharge. Current clinical guidelines recommend using structured assessment tools, such as the Psychopathy Checklist—Revised (PCL-R), Psychopathy Checklist—Screening Version (PCL-SV), and the (HCR-20), for use in forensic populations, which can also assist in incorporating personality difficulties into risk assessment.[10] US guidelines cite risk assessment tools as a useful memory aid, but do not recommend specific tools.[11] Similarly, European guidelines recommend the use of structured professional judgement risk assessment instruments in

forensic psychiatry, but do not suggest specific tools.[12] To develop such guidance, research synthesis that examines the performance of individual instruments is necessary.

Previous reviews assessing risk assessment tools in psychiatry have tended to include mostly men.[2,5,10] Women constitute approximately 5–18% of forensic psychiatric patients in Europe,[13,14] and have differential pathways to crime and forensic services compared with men, including risk, markers such as childhood victimisation,[15,16] emotional and cognitive dysfunction,[17] relational difficulties, and intimate partner violence.[18] Thus, it is important to know whether commonly used risk assessment instruments are applicable to women.

This study aims to systematically review and meta-analyse the performance of risk assessment instruments used to predict interpersonal violence and crime in forensic psychiatric patient samples after discharge.

Methods

Search strategy and selection criteria

We report our findings according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; appendix pp 11–14).[19] Based on recent systematic reviews and international surveys,[4,20–23] the 15 most commonly used violence risk prediction instruments in forensic psychiatric samples were

See **Online** for appendix

identified to inform the search strategy, although the search was not limited to these 15 instruments.

A systematic search was conducted to identify studies that measured the performance of risk assessment instruments in predicting the outcome of interpersonal violence and crime in forensic psychiatric samples post-discharge. Four databases (PsycINFO, Embase, Medline, and Global Health) were searched separately from their start date until Nov 1, 2022. No limits, restrictions (including by language), or published search filters were used (appendix p 15). The first 100 results on Google Scholar and the reference list of previous systematic reviews were browsed to try and identify additional studies.

We included studies of forensic psychiatric patients or psychiatric patients admitted to secure units following violent or criminal incidents, and studies that assessed the predictive performance of a risk assessment instrument at predicting post-discharge interpersonal violence and crime using at least one commonly accepted performance metric (sensitivity, specificity, positive predictive value, negative predictive value, c-index or area under the receiver operating characteristic curve [AUC], and calibration). Both retrospective cohort and prospective cohort studies were included. Systematic, narrative, and book reviews were excluded from the review. Non-nested case-control studies were excluded as they cannot be used to estimate absolute risk, leading to incorrect estimates of baseline hazard. Sexual offenders referred for civil commitment were excluded from this review (appendix p 2).

MGTO screened the titles and abstracts of all identified studies, with 10% double screened by reviewer AS to ensure adequate interrater reliability. Cohen's κ was used to calculate inter-rater agreement,[24] which was 0·95, indicating almost perfect agreement between the two raters.[25] Any disagreements were resolved via consensus. Reference lists of retained studies were hand searched to identify additional studies.

In the prespecified study protocol, the main outcome was defined using the umbrella term interpersonal violence and crime. However, to be consistent with how the outcome was defined in the papers reviewed, we modified it to recidivism, subcategorised into violent, general, and sexual recidivism.

Data analysis

One reviewer (MGTO) extracted study characteristics and summary estimates and a second reviewer (AS) independently verified a random 10% subset of full-text articles; any uncertainties were referred to the senior author (SF). Disagreements were resolved via consensus. Individual studies could report on more than one risk assessment instrument, so we extracted information on each instrument. Studies often reported multiple types of interpersonal violence or crime (eg, violent and general recidivism), so all outcome measures were extracted from each paper. If multiple publications were identified corresponding to a single study, the most complete report

was chosen for data extraction and supplemented using data from associated publications. Authors were contacted when insufficient information was in the publication.

Following Cochrane guidelines,[26,27] a meta-analysis of the predictive performance of externally validated risk assessment instruments was conducted (appendix p 3). When a tool had been validated at least three times for the outcome, we applied a random effects model, using the inverse-variance method, for pooling the logit transformation of the AUC and CIs. The predictive performance of each risk assessment instrument was pooled across all external validation studies regardless of study design (eg, including both retrospective and prospective cohort studies), as recommended for meta-analyses of prediction model studies.[28] To reduce bias, only independent validation studies with a sample size that is consistent with adequate statistical power were included in the primary analysis. We set this threshold at n=100, as a balance between the current methodological recommendations for minimum event numbers for validation studies[29] and excluding too large a proportion of existing literature. As a secondary analysis, we provide a narrative summary of the predictive performance of all studies, irrespective of authorship, validation, or sample size, and a post-hoc subgroup of analysis of AUCs by prediction horizon (categorised as up to 12 months, 1–5 years, and more than 5 years). Analyses were performed with R version 4.1.0 using the metafor[30] and meta-package[31,32] (appendix p 3).

The Prediction Model Risk of Bias Assessment Tool (PROBAST),[33] designed to provide guidance on the quality assessment for systematic reviews of studies investigating diagnostic and prognostic prediction models, was adapted and provided a risk of bias rating for each study, with low, high, or unclear risk of bias categorisations (appendix p 3). The tool consists of four domains (participants, predictors, outcomes, and analysis) containing 20 signalling questions to facilitate a risk of bias assessment. A domain where all signalling questions are answered as yes or probably yes is judged as having a low risk of bias. Any answers of no or probably no for one or more questions results in a high risk of bias in that domain. The overall risk of bias is judged as high if at least one domain is rated as being at high risk of bias. MGTO assessed the risk of bias for each included study based on PROBAST guidelines, with 10% re-rated by reviewer AS. Any disagreements were resolved by consensus.

This study was registered with PROSPERO, CRD42022304716.

Role of the funding source

Results

After 4842 unique records were screened, 50 studies met inclusion criteria (figure 1; appendix pp 17–28) for the

Identification of studies via databases and registers

6605 records identified
    1755 from Embase
    1123 from MEDLINE
    3644 from PsycINFO
    83 from Global Health

1763 records removed before screening
    1444 duplicate records removed by Endnote
    319 additional duplicates identified by handsearching

4842 records screened

4619 records excluded (not relevant)

223 reports sought for retrieval

9 reports not retrieved

214 reports assessed for eligibility

166 reports excluded after full-text review
    20 repetitive publications or overlapping samples
    71 non-forensic psychiatric sample or did not analyse forensic psychiatric cohort separately
    26 failed to report outpatient outcomes (at all or independently)
    31 did not meet pre-determined statistical standards for assessing predictive validity
    3 case-control study design
    4 insufficient outcome measure
    5 not risk assessment instruments
    6 not original empirical data

Identification of studies via other methods

12 records identified
    1 Google Scholar
    11 reference lists and external sources

12 reports sought for retrieval

12 reports assessed for eligibility

10 reports excluded
    3 non-forensic psychiatric sample
    2 did not analyse forensic psychiatric cohort separately
    3 did not meet predetermined statistical standards for assessing predictive validity
    1 matched case-control study design
    1 overlapping samples

50 studies included in review
19 independent studies included in meta-analysis
    17 violent
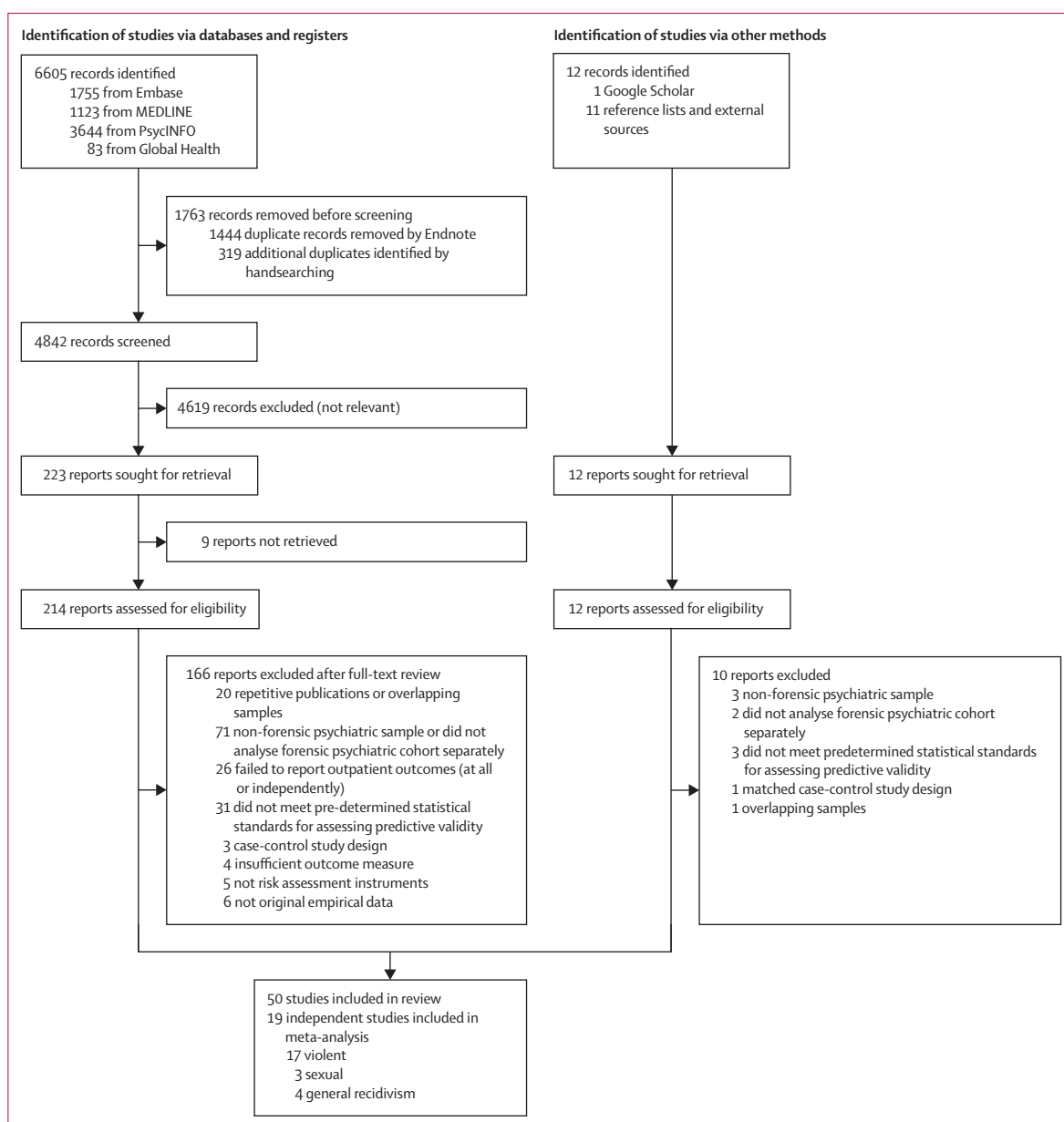   3 sexual
    4 general recidivism

Figure 1: PRISMA flow diagram

systematic review. Of these 50 studies, 33 were retrospective cohort, 16 were prospective cohort, and one was a randomised controlled trial (appendix pp 17–28). These studies included 10 460 participants (mean 209, range 45–2248). Based on 47 studies with age information, the median age was 35 years (IQR 33–38). Studies were conducted in 12 countries—Austria,[34] Australia,[35–37] Belgium,[38–42] Canada,[43–51] Denmark,[52–54] Finland,[50,55] Germany,[50,56] Japan,[57] Netherlands,[3,58–67] Sweden,[50,68–74] the UK,[75–80] and the USA[81,82]—all high-income economies (appendix p 4).[83]

The outcome of post-discharge interpersonal violence and crime was most often measured by new criminal offences or recidivism, as reported in 47 (94%) of the 50 studies and 9966 (95%) participants. Typically, recidivism was defined as new convictions (28 [56%] of 50) or criminal charges (five [10%] of 50), as recorded in criminal or police databases. One study defined recidivism on the basis of sentencing data, and six studies did not report how recidivism was defined. Three studies measured post-discharge interpersonal violence and crime using informant or self-report data on physical aggression or violent behaviour. For results synthesis, outcomes were categorised into violent recidivism, general or any recidivism, and sexual recidivism, in line with current research. Shared
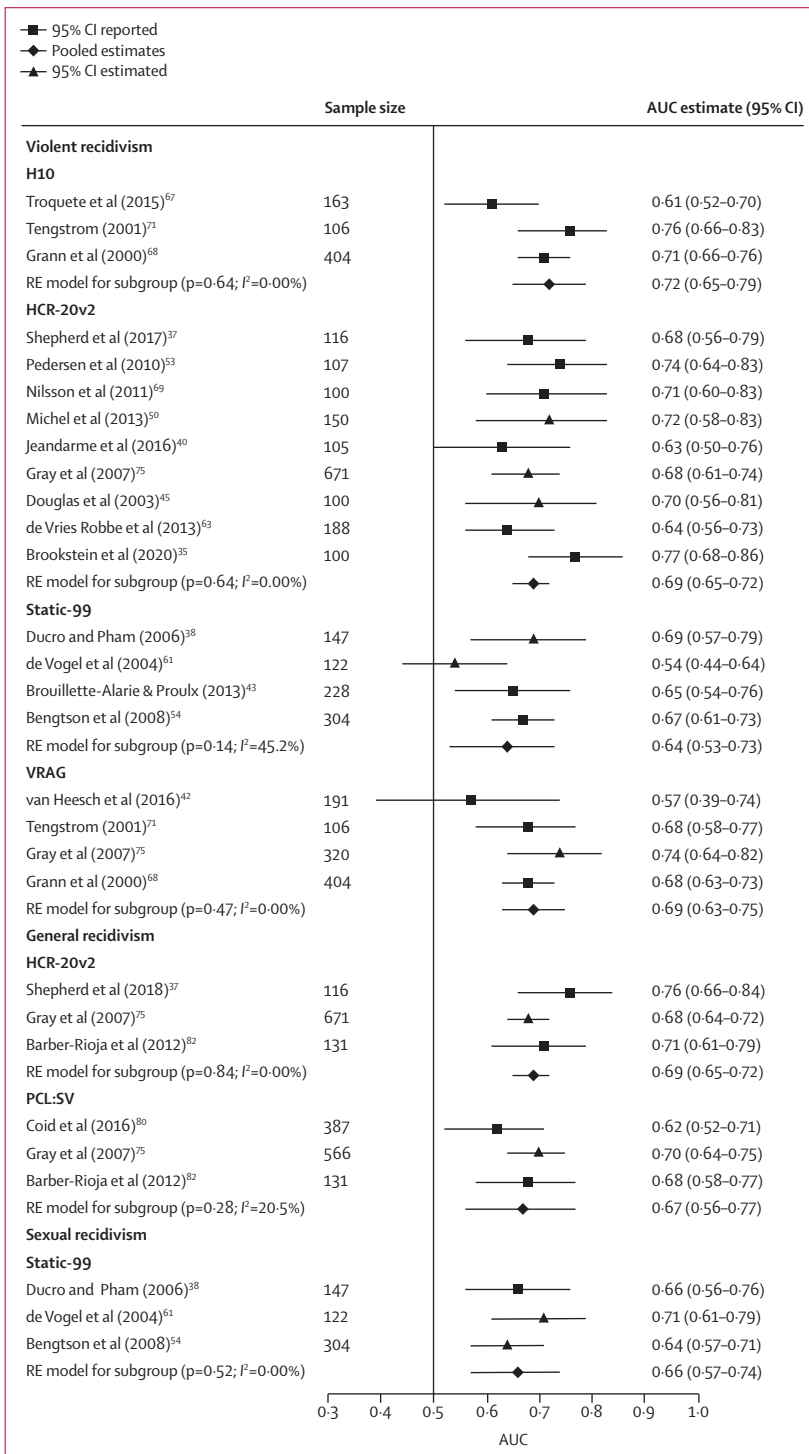
outcome definitions allowed for comparison of the predictive performance across studies. 27 [54%] of 50 studies assessed multiple types of recidivism; in these

cases, all types of recidivism were extracted and included in analysis. Sexual offences were most often included in definitions of violent recidivism, so violent recidivism (including sexual offences) was extracted and included in analysis.

36 risk assessment instruments were examined (appendix p 16). Each risk assessment instrument had between one and five studies assessing predictive validity, apart from the PCL-SV with six studies, the Static-99 and Violence Risk Appraisal Guide (VRAG) with eight studies, the PCL-R with 13 studies, and the HCR-20 version 2 with 16 studies. The most common performance statistic reported was the AUC. Nine studies assessed tool sensitivity, specificity, or positive and negative predictive values.[35,40,42,56,68,69,71,74,80] One study reported calibration.[74] In 46 studies, follow-up periods were reported, which ranged from 3 months to 16 years; nine (20%) of 46 studies reported predictive performance at different timepoints. To produce a weighted average of the predictive performance of risk assessment instruments, a meta-analysis was performed.

For the primary outcome of violent recidivism, we meta-analysed 20 independent external validations with more than 100 participants drawn from 17 separate studies (figure 2). These validations assessed four tools for violent recidivism (H10, HCR-20 version 2, Static-99, and VRAG). The HCR-20 version 3 did not meet criteria for meta-analysis. The HCR-20 version 2 was assessed in nine studies with AUC ranging from 0·63 to 0·77, and a pooled estimate of 0·69 (95% CI 0·65—0·72; I^2=0%). Four studies investigated the VRAG, with AUC ranging from 0·57 to 0·74, and a pooled estimate of 0·69 (95% CI 0·63—0·75; I^2=0%). Four studies examined the Static-99, with AUC ranging from 0·54 to 0·69, and a pooled estimate of 0·64 (95% CI 0·53–0·73; I^2=45%). Three studies assessed the H10, a subscale of the HCR-20, with AUC ranging from 0·61 to 0·76, and a pooled estimate of 0·72 (95% CI 0·65—0·79; I^2=0%). For general recidivism (any criminal behaviour), six validations (four studies) testing two tools (HCR-20 version 2 and PCL-SV) had sufficient independent external validation studies to be meta-analysed, with a pooled AUC of 0·69 (HCR-20 version 2; 95% CI 0·65–0·72; I^2=0%) and 0·67 (PCL-SV; 95% CI 0·56–0·77; I^2=21%). Only one tool (Static-99) that examined sexual recidivism was meta-analysed (three studies), with a pooled AUC of 0·66 (95% CI 0·57–0·74; I^2=0%).

In our secondary analysis, we examined all 50 eligible studies, irrespective of authorship, validation, or sample size. In 46 studies looking at violent recidivism, the AUCs ranged from 0·39 to 0·86. The risk assessment tools most commonly investigated were the PCL-R (12 studies), HCR-20 version 2 (12 studies), and VRAG (eight studies; appendix pp 5–6). 20 studies examined general recidivism and 11 reports examined sexual

	Sample size		AUC estimate (95% CI)
■ 95% CI reported			
◆ Pooled estimates			
▲ 95% CI estimated			
Violent recidivism			
H10			
Troquete et al (2015)[67]	163		0·61 (0·52–0·70)
Tengstrom (2001)[71]	106		0·76 (0·66–0·83)
Grann et al (2000)[68]	404		0·71 (0·66–0·76)
RE model for subgroup (p=0·64; I^2=0·00%)			0·72 (0·65–0·79)
HCR-20v2			
Shepherd et al (2017)[37]	116		0·68 (0·56–0·79)
Pedersen et al (2010)[53]	107		0·74 (0·64–0·83)
Nilsson et al (2011)[69]	100		0·71 (0·60–0·83)
Michel et al (2013)[50]	150		0·72 (0·58–0·83)
Jeandarme et al (2016)[40]	105		0·63 (0·50–0·76)
Gray et al (2007)[75]	671		0·68 (0·61–0·74)
Douglas et al (2003)[45]	100		0·70 (0·56–0·81)
de Vries Robbe et al (2013)[63]	188		0·64 (0·56–0·73)
Brookstein et al (2020)[35]	100		0·77 (0·68–0·86)
RE model for subgroup (p=0·64; I^2=0·00%)			0·69 (0·65–0·72)
Static-99			
Ducro and Pham (2006)[38]	147		0·69 (0·57–0·79)
de Vogel et al (2004)[61]	122		0·54 (0·44–0·64)
Brouillette-Alarie & Proulx (2013)[43]	228		0·65 (0·54–0·76)
Bengtson et al (2008)[54]	304		0·67 (0·61–0·73)
RE model for subgroup (p=0·14; I^2=45·2%)			0·64 (0·53–0·73)
VRAG			
van Heesch et al (2016)[42]	191		0·57 (0·39–0·74)
Tengstrom (2001)[71]	106		0·68 (0·58–0·77)
Gray et al (2007)[75]	320		0·74 (0·64–0·82)
Grann et al (2000)[68]	404		0·68 (0·63–0·73)
RE model for subgroup (p=0·47; I^2=0·00%)			0·69 (0·63–0·75)
General recidivism			
HCR-20v2			
Shepherd et al (2018)[37]	116		0·76 (0·66–0·84)
Gray et al (2007)[75]	671		0·68 (0·64–0·72)
Barber-Rioja et al (2012)[82]	131		0·71 (0·61–0·79)
RE model for subgroup (p=0·84; I^2=0·00%)			0·69 (0·65–0·72)
PCL:SV			
Coid et al (2016)[80]	387		0·62 (0·52–0·71)
Gray et al (2007)[75]	566		0·70 (0·64–0·75)
Barber-Rioja et al (2012)[82]	131		0·68 (0·58–0·77)
RE model for subgroup (p=0·28; I^2=20·5%)			0·67 (0·56–0·77)
Sexual recidivism			
Static-99			
Ducro and Pham (2006)[38]	147		0·66 (0·56–0·76)
de Vogel et al (2004)[61]	122		0·71 (0·61–0·79)
Bengtson et al (2008)[54]	304		0·64 (0·57–0·71)
RE model for subgroup (p=0·52; I^2=0·00%)			0·66 (0·57–0·74)

AUC axis: 0·3 0·4 0·5 0·6 0·7 0·8 0·9 1·0

***Figure 2:* Meta-analysis of independent validation studies with a sample size of more than 100 participants**
Meta-analyses by outcome type (violent, general, and sexual recidivism) and by risk assessment instrument. AUC=area under curve. HCRv2=Historial, Clinical, Risk Management-20 version 2. PCL:SV=Psychopathy Checklist—Screening Version. RE=random effects. VRAG=Violence Risk Appraisal Guide.

recidivism (appendix pp 7–8). For the nine studies that reported paired measures of classification for violent recidivism, sensitivity values ranged from 0·33 to 0·80, specificity from 0·55 to 0·85, positive predictive value (PPV) from 0·30 to 0·74, and negative predictive value (NPV) from 0·70 to 0·96. For paired measures of classification for general recidivism, sensitivity values ranged from 0·60 to 0·89, specificity from 0·52 to 0·72, PPV from 0·00 to 0·28, and NPV from 0·71 to 0·97. When investigating performance by prediction horizons in a subgroup analysis, there was no difference in AUCs (appendix p 10).

The risk of bias was high for almost all studies (49 [98%] of 50; figure 3). For individual PROBAST domain ratings, see the appendix (pp 29–30). High rates of bias risk were primarily due to the high risk of bias in the PROBAST analysis domain (49 [98%] of 50) and failure to evaluate the performance appropriately by not assessing instrument calibration (appendix pp 29–30). Only one study of FoVOx[74] reported calibration metrics. Most studies (47 [94%] of 50) externally validated at least one existing risk assessment instrument in a wholly independent sample (appendix p 9).

In 47 studies, participant sex was specified or could be inferred from ward type. Of these studies, 8710 (88%) of the 9905 participants recruited were men, and 22 (47%) of the 47 studies only recruited male participants. Two studies recruited women only,[59,66] and one recruited an even ratio of men to women.[51] The two female-only studies had sample sizes of less than 100 (n=71 and n=45) so were not included in the primary analyses. The female-only samples typically reported lower AUCs than the mostly male samples when assessing the predictive performance of the HCR-20, PCL-R, and Short-Term Assessment of Risk and Treatability instruments (appendix pp 5–7). Of those studies with a mixed sample, only two studies[44,51] disaggregated predictive validity by sex, with no clear evidence that tools were more predictive when stratified by sex.

Discussion

This systematic review and meta-analysis examined the predictive performance of risk assessment instruments for violent and criminal outcomes among individuals discharged from forensic psychiatric hospitals. 50 studies were included, involving 10 460 participants from 12 countries. Overall, findings were mixed, with wide variation in study quality, outcome reporting, and the predictive performance of the tools studied. Almost all studies (49 [98%] of 50) were assessed as high risk of bias. Numerous implications follow for the clinical use of these tools, and future research examining the predictive performance of these tools should apply methods that address the methodological limitations of previous studies.

First, the number of performance measures reported by included studies was small; most studies (41 [82%] of 50)
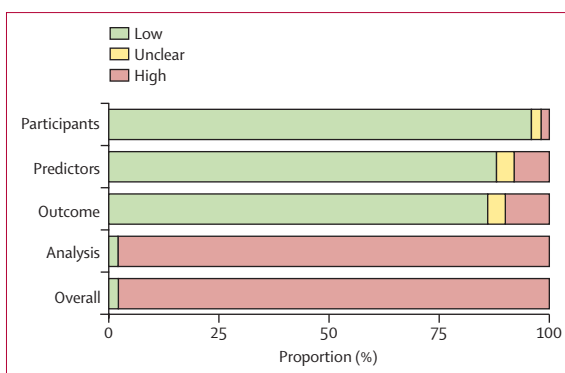
Figure 3: **Risk of bias among risk assessment instruments**
Unweighted bar plot of the distribution of risk-of-bias judgements across all studies within each bias domain.

reported only the AUC. This discrimination metric gives the probability that for a hypothetical pair of individuals who do and do not develop the outcome of interest (ie, violence or crime perpetration), the tool will assign a higher risk estimate to the individual with the outcome. Importantly, a tool can perform well in measures of discrimination even if it is systematically off-target in its predictions, for example by overestimating risk for everyone. This calibration is a crucial consideration if these absolute risk estimations are used to make clinical decisions. Studies therefore also need to report calibration (the agreement between expected and observed probabilities), which was reported in only one study.[74]

The reporting of paired measures of classification, such as sensitivity and specificity, was also low, presented in nine (18%) of 50 studies. Tool sensitivity and specificity are important when assessing the use of a tool to make specific clinical decisions. The preferred weighting of sensitivity (ie, minimising false negatives) and specificity (ie, minimising false positives) depends on the severity of the outcome being predicted and the nature of the intervention being linked to the assessment. In forensic mental health settings, tools with high sensitivity will be the most relevant from a public safety perspective and could garner more political support, whereas tools with high specificity might best protect the rights of patients.[84] More research is required to determine which risk assessment instruments exhibit optimal levels of sensitivity and specificity.

Another common methodological problem was small sample sizes, with 19 (38%) of 50 studies recruiting fewer than 100 participants (appendix pp 17–28). As recidivism outcomes are not common in these samples,[85] small samples lead to uncertain risk estimates. However, obtaining large samples of forensic psychiatric patients is not feasible in many settings. Using multisite or nationwide registry data could address this limitation. Another finding of the current review is that most tools were independently externally validated (although not six newer tools). Independent validations should be

prioritised because authorship bias can influence reporting of findings.[86]

The results from the meta-analysis provide evidence on the predictive performance of the most common risk assessment instruments. We found that the H10, HCR-20 version 2, VRAG, and Static-99 perform similarly at predicting violent recidivism, with overlapping CIs and pooled AUCs ranging from 0·64 to 0·72. A secondary analysis of all 50 eligible studies, irrespective of authorship, validation, or sample size, clarified the heterogeneity in performance between different studies and instruments. Although most studies reported AUCs of approximately 0·70, some studies (especially those with small sample sizes) reported a predictive performance not different to chance. This finding suggests that the appraisal of these risk assessment studies needs to pay particular attention to sample size.

The wide range in prediction horizons (follow-up periods ranged from 3 months to 16 years) between studies could moderate predictive performance. Although we did not find evidence of this in our post-hoc analysis (appendix p 10), future work should define outcome windows more clearly. Furthermore, most current literature does not consistently report or account for the provision and intensity of ongoing community supervision, potential readmission to hospital, or incarceration during follow-up. All these factors could potentially limit opportunities to reoffend and reduce outcome rates. Moreover, in contexts where violence risk assessment instruments are used continuously to assess and manage the likelihood of violent outcomes, it is difficult to distinguish whether false-positive errors (ie, a risk assessment predicted that someone is high risk of violence, when they did not have a violent outcome) occurred or whether the management strategies implemented following a risk assessment effectively decreased risk. Therefore, predictive performance will probably be attenuated in circumstances of active risk management. Treatment and risk management procedures should be reported to allow accurate appraisal of risk prediction potential. Current literature often does not acknowledge how the population and context in which risk assessment instruments are developed and validated contribute to predictive performance and its validity in risk management procedures.

Included studies considered a wide range of instruments to assess the risk of interpersonal violence and crime, and tools were often examined for predicting different outcomes to those they were originally developed to assess. For example, the Static-99 (and its revisions), a tool developed to predict sexual recidivism, has been used to assess violent and general recidivism. The second most frequently studied tool was the PCL-R, which was developed to assess psychopathy. As tool predictive performance is highly dependent on population and setting,[87] accuracy will be poorer if instruments are not used as intended.[88] The overlapping content between some risk assessment instruments might be relevant to comparing their performance. In addition, some items might be redundant, as they were not tested in multivariate models.

We found that the H10—a HCR-20 subscale that includes 10 historical factors—performed similarly to the HCR-20 full scale. Risk assessments are resource intensive,[22] so if short-form or tool subscales are equally predictive, using these in clinical services could constitute a more efficient use of resources.

Included studies were typically of male-dominated samples (88% of included participants), and 47% of studies recruited exclusively male participants. The generalisability of tools to female patients could be hampered by differences in the baseline rates of violence between men and women, and in risk factor association with violence. Domains of risk relevant for female populations might include intimate partner violence,[18] sex work,[89] pregnancy at a young age,[90] and self-harm.[91] Interestingly, the FAM—designed to supplement the HCR-20 version 2 for female forensic patients—did not yield higher AUC values in a female-only sample compared with the HCR-20 version 3.[59]

Poor predictive performance of an assessment tool, or lack of validation, has direct clinical and ethical implications for patients. Discharge from secure settings could be expedited or delayed depending on the findings of a risk assessment instrument. Inaccurate risk prediction resulting in extended detention can be harmful. Furthermore, inaccurate prediction resulting in the discharge of a patient who subsequently reoffends will disrupt the clinical care of the patient and harm public health and safety. Therefore, to avoid these negative outcomes, ensuring that the risk assessment tools implemented in clinical practice have high predictive performance and perform consistently in varied contexts is important for ethical conduct. Some research studies have labelled AUC values in categories, but this is not recommended.[92] Instruments can be compared on discriminative ability using AUC values, with the best AUC implying greater discrimination performance; however, additional measures are needed to verify the potential clinical positive effect. In the case of this review, the risk assessment instrument with the best predictive performance was the H10 (pooled AUC of 0·72). One overall implication is that risk assessment instruments do discriminate better than chance, and other research suggests that tools are typically more accurate than unstructured clinical decision making, especially in predicting violence.[93] What constitutes adequate performance is dependent on the context of application;[94] therefore, the AUC values of risk assessment instruments in this field cannot, in a straightforward manner, be compared with other domains (eg, in cancer or cardiovascular medicine), since the predictors and outcomes are different.

For forensic mental health services, the findings suggest that, as a minimum, risk assessment tools should be used to complement clinical decision making; they should not be used to inform decisions about length of stay without validations showing high sensitivity. With evidence of high specificity, risk assessment tools can be included in discussions about discharge planning, particularly about how to allocate follow-up resources. New tools should not be introduced to services without information on these classification measures and calibration. In addition, any tools developed without multivariable models or inclusion of the strongest risk factors (ie, age, sex, and previous antisocial or violent behaviour) are very unlikely to be accurate in new settings. Other considerations include the tool being developed on forensic mental health populations, evidence of feasibility and acceptability, the individual variables making up the tool being weighted, and internal validation.[95] Furthermore, instruments that facilitate formulation as part of the risk assessment procedure might aid clinicians in creating effective risk management plans that are sensitive to risk erosion, actively mitigate risk, and avert violence.

To our knowledge, this is the first comprehensive systematic review and meta-analysis of the performance of risk assessment instruments for post-discharge interpersonal violence and crime in forensic mental health specifically. One of the review's limitations is that the risk of bias (PROBAST) tool was developed for prediction models in general medicine, rather than risk assessment instruments in forensic mental health, and the threshold for bias might be too low for studies reporting on crime outcomes. This limitation has resulted in high risk of bias across studies, with little granularity on how bias differs between studies; and a sensitivity analysis was not possible to examine whether risk of bias was associated with the predictive validity of the tools.

Taken together, the findings provide some directions for future work. From a methodological perspective, future research should be adequately powered and report multiple estimates of predictive performance to inform clinical decisions. These metrics should include calibration, sensitivity, specificity, and positive and negative predictive values. All available research was from high-income countries. Future research should seek to include women and samples from low-income and middle-income countries, and investigate the incremental value of novel risk factors. Another research direction is to examine to what extent performance can be improved by incorporating novel risk factors. For example, markers of previous neurological damage or head trauma could improve performance, as neurobiological stress and injury can cause emotional and behavioural changes, increasing the risk of violence.[96]

In summary, the performance of current tools at predicting the risk of interpersonal violence and crime in forensic mental health is mixed, with performance varying between instruments. Most investigations solely reported the AUC for model performance, and failed to present other key measures, such as calibration, sensitivity, and specificity. Higher quality risk assessment tools could contribute to better risk management in forensic mental health.

References
1   Senior M, Fazel S, Tsiachristas A. The economic impact of violence perpetration in severe mental illness: a retrospective, prevalence-based analysis in England and Wales. *Lancet Public Health* 2020; **5:** e99–106.
2   Fazel S, Fimińska Z, Cocks C, Coid J. Patient outcomes following discharge from secure psychiatric hospitals: systematic review and meta-analysis. *Br J Psychiatry* 2016; **208:** 17–25.
3   Bogaerts S, Spreen M, Ter Horst P, Gerlsma C. Predictive validity of the HKT-R risk assessment tool: two and 5-year violent recidivism in a nationwide sample of Dutch forensic psychiatric patients. *Int J Offender Ther Comp Criminol* 2018; **62:** 2259–70.
4   Singh JP, Desmarais SL, Hurducas C, et al. International perspectives on the practical application of violence risk assessment: a global survey of 44 countries. *Int J Forensic Ment Health* 2014; **13:** 193–206.
5   Fazel S, Singh JP, Doll H, Grann M. Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis. *BMJ* 2012; **345:** e4692.
6   Singh JP, Grann M, Fazel S. A comparative study of violence risk assessment tools: a systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clin Psychol Rev* 2011; **31:** 499–513.
7   Singh JP, Desmarais SL, Van Dorn RA. Measurement of predictive validity in violence risk assessment studies: a second-order systematic review. *Behav Sci Law* 2013; **31:** 55–73.
8   Whittington R, Hockenhull JC, McGuire J, et al. A systematic review of risk assessment strategies for populations at high risk of engaging in violent behaviour: update 2002-8. *Health Technol Assess* 2013; **17:** i–xiv, 1–128.
9   Ramesh T, Igoumenou A, Vazquez Montes M, Fazel S. Use of risk assessment instruments to predict violence in forensic psychiatric hospitals: a systematic review and meta-analysis. *Eur Psychiatry* 2018; **52:** 47–53.

10 National Institute for Health and Care Excellence. Antisocial personality disorder: prevention and management. 2009. https://www.nice.org.uk/guidance/cg77 (accessed July 18, 2023).

11 Buchanan A, Binder R, Norko M, Swartz M. Psychiatric violence risk assessment. *Am J Psychiatry* 2012; **169:** 340.

12 Völlm BA, Clarke M, Herrando VT, et al. European Psychiatric Association (EPA) guidance on forensic psychiatry: evidence based assessment and treatment of mentally disordered offenders. *Eur Psychiatry* 2018; **51:** 58–73.

13 Penney SR, Seto MC, Crocker AG, et al. Changing characteristics of forensic psychiatric patients in Ontario: a population-based study from 1987 to 2012. *Soc Psychiatry Psychiatr Epidemiol* 2019; **54:** 627–38.

14 Tomlin J, Lega I, Braun P, et al. Forensic mental health in Europe: some key figures. *Soc Psychiatry Psychiatr Epidemiol* 2021; **56:** 109–17.

15 de Vogel V, Stam J, Bouman YHA, Ter Horst P, Lancel M. Violent women: a multicentre study into gender differences in forensic psychiatric patients. *J Forensic Psychiatry Psychol* 2016; **27:** 145–68.

16 DeHart DD. Women's pathways to crime: a heuristic typology of offenders. *Crim Justice Behav* 2018; **45:** 1461–82.

17 Hodgins S. Female forensic patients may be an atypical sub-type of females presenting aggressive and antisocial behavior. *Front Psychiatry* 2022; **13:** 809901.

18 Salisbury EJ, Van Voorhis P. Gendered pathways: a quantitative investigation of women probationers' paths to incarceration. *Crim Justice Behav* 2009; **36:** 541–66.

19 Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Syst Rev* 2021; **10:** 89.

20 Hurducas CC, Singh JP, de Ruiter C, Petrila J. Violence risk assessment tools: a systematic review of surveys. *Int J Forensic Ment Health* 2014; **13:** 181–92.

21 Singh JP, Desmarais SL, Otto RK, Nicholls TL, Petersen KL, Pritchard MM. The International Risk Survey: use and perceived utility of structured violence risk assessment tools in 44 countries. In: Singh JP, Bjørkly S, Fazel S, eds. International perspectives on violence risk assessment. Oxford, Oxford University Press, 2016: 101–26.

22 Viljoen JL, McLachlan K, Vincent GM. Assessing violence risk and psychopathy in juvenile and adult offenders: a survey of clinical practices. *Assessment* 2010; **17:** 377–95.

23 Wilson S, Singh JP, Leech N, Nicholls T. Structured instruments commonly used in violence risk assessments. In: International perspectives on violence risk assessment. New York, NY: Oxford University Press, 2016: 353–72.

24 Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas* 1960; **20:** 37–46.

25 McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med* 2012; **22:** 276–82.

26 Debray TP, Damen JA, Snell KI, et al. A guide to systematic review and meta-analysis of prediction model performance. *BMJ* 2017; **356:** i6460.

27 Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: which scale helps ensure between-study normality for the C-statistic and calibration measures? *Stat Methods Med Res* 2018; **27:** 3505–22.

28 Kreuzberger N, Damen JA, Trivella M, et al. Prognostic models for newly-diagnosed chronic lymphocytic leukaemia in adults: a systematic review and meta-analysis. *Cochrane Database Syst Rev* 2020; **7:** CD012022.

29 Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol* 2005; **58:** 475–83.

30 Viechtbauer W. Conducting meta-analyses in R with the metafor package. *J Stat Softw* 2010; **36:** 1–48.

31 Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. *Evid Based Ment Health* 2019; **22:** 153–60.

32 Schwarzer G. Meta: an R package for meta-analysis. *R News* 2007; **7:** 40–45.

33 Moons KGM, Wolff RF, Riley RD, et al. PROBAST: a tool to assess risk of bias and applicability of prediction model studies: explanation and elaboration. *Ann Intern Med* 2019; **170:** W1–33.

34 Eher R, Hofer S, Buchgeher A, Domany S, Turner D, Olver ME. The predictive properties of psychiatric diagnoses, dynamic risk and dynamic risk change assessed by the VRS-SO in forensically admitted and released sexual offenders. *Front Psychiatry* 2020; **10:** 922.

35 Brookstein DM, Daffern M, Ogloff JRP, Campbell RE, Chu CM. Predictive validity of the HCR-20^v3 in a sample of Australian forensic psychiatric patients. *Psychiatry Psychol Law* 2020; **28:** 325–42.

36 Ferguson AM, Ogloff JRP, Thomson L. Predicting recidivism by mentally disordered offenders using the LSI-R: SV. *Crim Justice Behav* 2009; **36:** 5–20.

37 Shepherd SM, Campbell RE, Ogloff JRP. The utility of the HCR-20 in an Australian sample of forensic psychiatric patients. *Psychiatry Psychol Law* 2017; **25:** 273–82.

38 Ducro C, Pham T. Evaluation of the SORAG and the Static-99 on Belgian sex offenders committed to a forensic facility. *Sex Abuse* 2006; **18:** 15–26.

39 Jeandarme I, Edens JF, Habets P, Bruckers L, Oei K, Bogaerts S. PCL-R field validity in prison and hospital settings. *Law Hum Behav* 2017; **41:** 29–43.

40 Jeandarme I, Pouls C, De Laender J, Oei T, Bogaerts S. Field validity of the HCR-20 in forensic medium security units in Flanders. *Psychol Crime Law* 2016; **23:** 305–22.

41 Pham T, Ducro C. Risk assessment in social defence: preliminary factorial analysis of the 'Sex Offender Recidivism Appraisal Guide' (SORAG) and the Static-99. *Ann Med Psychol (Paris)* 2008; **166:** 575–79.

42 van Heesch B, Jeandarme I, Pouls C, Vervaeke G. Validity and reliability of the VRAG in a forensic psychiatric medium security population in Flanders. *Psychol Crime Law* 2016; **22:** 530–37.

43 Brouillette-Alarie S, Proulx J. Predictive validity of the Static-99R and its dimensions. *J Sex Aggress* 2013; **19:** 311–28.

44 Canales DD, Campbell MA, Wei R, Totten AE. Prediction of general and violent recidivism among mentally disordered adult offenders: test of the level of service/risk–need–responsivity (LS/RNR) instrument. *Crim Justice Behav* 2014; **41:** 971–91.

45 Douglas KS, Ogloff JR, Hart SD. Evaluation of a model of violence risk assessment among forensic psychiatric patients. *Psychiatr Serv* 2003; **54:** 1372–79.

46 Harris GT, Rice ME, Cormier CA. Prospective replication of the Violence Risk Appraisal Guide in predicting violent recidivism among forensic patients. *Law Hum Behav* 2002; **26:** 377–94.

47 Hanson RK, Thornton D. Improving risk assessments for sex offenders: a comparison of three actuarial scales. *Law Hum Behav* 2000; **24:** 119–36.

48 Hogan NR, Olver ME. Static and dynamic assessment of violence risk among discharged forensic patients. *Crim Justice Behav* 2019; **46:** 923–38.

49 Jung S, Wielinga F, Ennis L. Can we keep it simple? Using the BARR-2002R with a community-based sex offender sample. *J Sex Aggress* 2018; **24:** 25–36.

50 Michel SF, Riaz M, Webster C, et al. Using the HCR-20 to predict aggressive behavior among men with schizophrenia living in the community: accuracy of prediction, general and forensic settings, and dynamic risk factors. *Int J Forensic Ment Health* 2013; **12:** 1–13.

51 Nicholls TL. Violence risk assessments with female NCRMD acquittees: validity of the HCR-20 and PCL-SV. *Diss Abstr B Sci Eng* 2004; **64:** 4055.

52 Pedersen L, Rasmussen K, Elsass P. HCR-20 violence risk assessments as a guide for treating and managing violence risk in a forensic psychiatric setting. *Psychol Crime Law* 2012; **18:** 733–43.

53 Pedersen L, Kunz C, Rasmussen K, Elsass P. Psychopathy as a risk factor for violent recidivism: investigating the psychopathy checklist screening version (PCL:SV) and the comprehensive assessment of psychopathic personality (CAPP) in a forensic psychiatric setting. *Int J Forensic Ment Health* 2010; **9:** 308–15.

54 Bengtson S. Is newer better? A cross-validation of the Static-2002 and the Risk Matrix 2000 in a Danish sample of sexual offenders. *Psychol Crime Law* 2008; **14:** 85–106.

55 Gammelgård M, Koivisto A-M, Eronen M, Kaltiala-Heino R. Predictive validity of the structured assessment of violence risk in youth: a 4-year follow-up. *Crim Behav Ment Health* 2015; **25:** 192–206.

56 Stadtland C, Kleindienst N, Kroner C, Eidt M, Nedopil N. Psychopathic traits and risk of criminal recidivism in offenders with and without mental disorders. *Int J Forensic Ment Health* 2005; **4:** 89–97.

57 Kikuchi A, Soshi T, Kono T, Koyama M, Fujii C. Validity of short-term assessment of risk and treatability in the Japanese forensic probation service. *Front Psychiatry* 2021; **12:** 645927.

58 Delforterie MJ, Hesper BL, Nijman HLI, Korzilius HPLM, Turhan A, Didden R. The predictive value of the Dynamic Risk Outcome Scales (DROS) for predicting recidivism in (forensic) patients with mild intellectual disabilities or borderline intellectual functioning. *Tijdschr Psychiatr* 2020; **62:** 1040–48.

59 de Vogel V, Bruggeman M, Lancel M. Gender-sensitive violence risk assessment: predictive validity of six tools in female forensic psychiatric patients. *Crim Justice Behav* 2019; **46:** 528–49.

60 de Vogel V, van den Broek E, de Vries Robbe M. The use of the HCR-20V3 in Dutch forensic psychiatric practice. *Int J Forensic Ment Health* 2014; **13:** 109–21.

61 de Vogel V, de Ruiter C, van Beek D, Mead G. Predictive validity of the SVR-20 and Static-99 in a Dutch sample of treated sex offenders. *Law Hum Behav* 2004; **28:** 235–51.

62 de Vries Robbé M, de Vogel V, Koster K, Bogaerts S. Assessing protective factors for sexually violent offending with the SAPROF. *Sex Abuse* 2015; **27:** 51–70.

63 de Vries Robbe M, de Vogel V, Douglas KS. Risk factors and protective factors: a two-sided dynamic approach to violence risk assessment. *J Forensic Psychiatry Psychol* 2013; **24:** 440–57.

64 Hildebrand M, de Ruiter C, de Vogel V. Psychopathy and sexual deviance in treated rapists: association with sexual and nonsexual recidivism. *Sex Abuse* 2004; **16:** 1–24.

65 Nowak MJ, Nugter MA. 'I'll serve my time…' II. A replication study of recidivism in patients with a one-year hospital order. *Tijdschr Psychiatr* 2014; **56:** 237–46.

66 Schaap G, Lammers S, de Vogel V. Risk assessment in female forensic psychiatric patients: a quasi-prospective study into the validity of the HCR-20 and PCL-R. *J Forensic Psychiatry Psychol* 2009; **20:** 354–65.

67 Troquete NAC, van den Brink RHS, Beintema H, et al. Predictive validity of the Short-Term Assessment of Risk and Treatability for violent behavior in outpatient forensic psychiatric patients. *Psychol Assess* 2015; **27:** 377–91.

68 Grann M, Belfrage H, Tengstrom A. Actuarial assessment of risk for violence: predictive validity of the VRAG and the historical part of the HCR-20. *Crim Justice Behav* 2000; **27:** 97–114.

69 Nilsson T, Wallinius M, Gustavson C, Anckarsäter H, Kerekes N. Violent recidivism: a long-time follow-up study of mentally disordered offenders. *PLoS One* 2011; **6:** e25768.

70 Sjostedt G, Langstrom N. Assessment of risk for criminal recidivism among rapists: a comparison of four different measures. *Psychol Crime Law* 2002; **8:** 25–40.

71 Tengström A. Long-term predictive validity of historical factors in two risk assessment instruments in a group of violent offenders with schizophrenia. *Nord J Psychiatry* 2001; **55:** 243–49.

72 Tengström A, Grann M, Långström N, Kullgren G. Psychopathy (PCL-R) as a predictor of violent recidivism among criminal offenders with schizophrenia. *Law Hum Behav* 2000; **24:** 45–58.

73 Wallinius M, Nilsson T, Hofvander B, Anckarsäter H, Stålenheim G. Facets of psychopathy among mentally disordered offenders: clinical comorbidity patterns and prediction of violent and criminal behavior. *Psychiatry Res* 2012; **198:** 279–84.

74 Wolf A, Fanshawe TR, Sariaslan A, Cornish R, Larsson H, Fazel S. Prediction of violent crime on discharge from secure psychiatric hospitals: a clinical prediction rule (FoVOx). *Eur Psychiatry* 2018; **47:** 88–93.

75 Gray NS, Fitzgerald S, Taylor J, Macculloch MJ, Snowden RJ. Predicting future reconviction in offenders with intellectual disabilities: the predictive efficacy of VRAG, PCL-SV, and the HCR-20. *Psychol Assess* 2007; **19:** 474–79.

76 Ho H, Thomson L, Darjee R. Violence risk assessment: the use of the PCL-SV, HCR-20, and VRAG to predict violence in mentally disordered offenders discharged from a medium secure unit in Scotland. *J Forensic Psychiatry Psychol* 2009; **20:** 523–41.

77 Thomson L, Davidson M, Brett C, Steele J, Darjee R. Risk assessment in forensic patients with schizophrenia: the predictive validity of actuarial scales and symptom severity for offending and violence over 8 – 10 years. *Int J Forensic Ment Health* 2008; **7:** 173–89.

78 Vojt G, Thomson LD, Marshall LA. The predictive validity of the HCR-20 following clinical implementation: does it work in practice? *J Forensic Psychiatry Psychol* 2013; **24:** 371–85.

79 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.

80 Coid JW, Ullrich S, Kallis C, et al. Improving risk management for violence in mental health services: a multimethods approach. *Programme Grants Appl Res* 2016; **16:** 11.

81 Gonsalves VM, McLawsen JE, Huss MT, Scalora MJ. Factor structure and construct validity of the psychopathic personality inventory in a forensic sample. *Int J Law Psychiatry* 2013; **36:** 176–84.

82 Barber-Rioja V, Dewey L, Kopelovich S, Kucharski L. The utility of the HCR-20 and PCL:SV in the prediction of diversion noncompliance and reincarceration in diversion programs. *Crim Justice Behav* 2012; **39:** 475–92.

83 The World Bank. World Bank Country and Lending Groups. 2022. https://datahelpdesk.worldbank.org/knowledgebase/ articles/906519-world-bank-country-and-lending-groups (accessed July 18, 2023).

84 Douglas T, Pugh J, Singh I, Savulescu J, Fazel S. Risk assessment tools in criminal justice and forensic psychiatry: the need for better data. *Eur Psychiatry* 2017; **42:** 134–37.

85 Whiting D, Lichtenstein P, Fazel S. Violence and mental disorders: a structured review of associations by individual diagnoses, risk factors, and risk assessment. *Lancet Psychiatry* 2021; **8:** 150–61.

86 Singh JP, Grann M, Fazel S. Authorship bias in violence risk assessment? A systematic review and meta-analysis. *PLoS One* 2013; **8:** e72484.

87 Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. *Diagn Progn Res* 2022; **6:** 24.

88 Guy LS, Douglas KS, Hendry MC. The role of psychopathic personality disorder in violence risk assessments using the HCR-20. *J Pers Disord* 2010; **24:** 551–80.

89 Morgan M, Patton P. Gender-responsive programming in the justice system–Oregon's guidelines for effective programming for girls. *Fed Probat* 2002; **66:** 57–65.

90 Messer J, Maughan B, Quinton D, Taylor A. Precursors and correlates of criminal behaviour in women. *Crim Behav Ment Health* 2004; **14:** 82–107.

91 Völlm BA, Dolan MC. Self-harm among UK female prisoners: a cross-sectional study. *J Forensic Psychiatry Psychol* 2009; **20:** 741–51.

92 de Hond AAH, Steyerberg EW, van Calster B. Interpreting area under the receiver operating characteristic curve. *Lancet Digit Health* 2022; **4:** e853–55.

93 Ægisdóttir S, White MJ, Spengler PM, et al. The meta-analysis of clinical judgment project: fifty-six years of accumulated research on clinical versus statistical prediction. *Couns Psychol* 2006; **34:** 341–82.

94 Binuya MAE, Engelhardt EG, Schats W, Schmidt MK, Steyerberg EW. Methodological guidance for the evaluation and updating of clinical prediction models: a systematic review. *BMC Med Res Methodol* 2022; **22:** 316.

95 Fazel S, Wolf A. Selecting a risk assessment tool to use in practice: a 10-point guide. *Evid Based Ment Health* 2018; **21:** 41–43.

96 Silveira K, Garcia-Barrera MA, Smart CM. Neuropsychological impact of trauma-related mental illnesses: a systematic review of clinically meaningful results. *Neuropsychol Rev* 2020; **30:** 310–44.