BMJ

BMJ 2012;345:e4692 doi: 10.1136/bmj.e4692 (Published 24 July 2012)

RESEARCH

Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24 827 people: systematic review and meta-analysis

Seena Fazel *Wellcome Trust senior research fellow in clinical science*[1], Jay P Singh *postdoctoral research fellow*[2], Helen Doll *statistician*[3], Martin Grann *professor*[4]

[1]Department of Psychiatry, University of Oxford, Warneford Hospital, Oxford OX3 7JX, UK; [2]Department of Mental Health Law and Policy, University of South Florida, Tampa, FL, USA; [3]Department of Population Health and Primary Care, University of East Anglia, Norwich, UK; [4]Centre for Violence Prevention, Karolinska Institutet, Stockholm, Sweden

Abstract

Objective To investigate the predictive validity of tools commonly used to assess the risk of violence, sexual, and criminal behaviour.

Design Systematic review and tabular meta-analysis of replication studies following PRISMA guidelines.

Data sources PsycINFO, Embase, Medline, and United States Criminal Justice Reference Service Abstracts.

Review methods We included replication studies from 1 January 1995 to 1 January 2011 if they provided contingency data for the offending outcome that the tools were designed to predict. We calculated the diagnostic odds ratio, sensitivity, specificity, area under the curve, positive predictive value, negative predictive value, the number needed to detain to prevent one offence, as well as a novel performance indicator—the number safely discharged. We investigated potential sources of heterogeneity using metaregression and subgroup analyses.

Results Risk assessments were conducted on 73 samples comprising 24 847 participants from 13 countries, of whom 5879 (23.7%) offended over an average of 49.6 months. When used to predict violent offending, risk assessment tools produced low to moderate positive predictive values (median 41%, interquartile range 27-60%) and higher negative predictive values (91%, 81-95%), and a corresponding median number needed to detain of 2 (2-4) and number safely discharged of 10 (4-18). Instruments designed to predict violent offending performed better than those aimed at predicting sexual or general crime.

Conclusions Although risk assessment tools are widely used in clinical and criminal justice settings, their predictive accuracy varies depending on how they are used. They seem to identify low risk individuals with high levels of accuracy, but their use as sole determinants of detention, sentencing, and release is not supported by the current evidence. Further research is needed to examine their contribution to treatment and management.

Introduction

With the increasing recognition of the public health importance of violence,[1 2] the prediction of violence, or violence risk assessment, has been the subject of considerable clinical and research interest. Since the late 1980s, such assessment has mostly been conducted by structured instruments after several studies found unstructured clinical opinion to have little evidence in support.[3] Recent surveys have estimated that over 60% of general psychiatric patients are routinely assessed for violence risk,[4] rising to above 80% in forensic psychiatric hospitals.[5]

The widespread use of these tools has been partly driven by public concern about the safety of mentally ill patients,[6] research evidence that severe mental illness is associated with violence,[7-9] and clinical practice guidelines in some countries, including the United Kingdom and United States,[10 11] recommending violence risk assessment for all patients with schizophrenia. Furthermore,

Correspondence to: S Fazel seena.fazel@psych.ox.ac.uk

Extra material supplied by the author (see http://www.bmj.com/content/345/bmj.e4692?tab=related#webextra)

Web figure 1: Results of a systematic search conducted to identify replication studies of nine commonly used risk assessment tools
Web figure 2: Distribution of effect sizes (reported as Cohen's d) in studies included and not included in the current systematic review
Web appendix: Studies included in quantitative synthesis
Web table 1: Comparison of summary accuracy estimates produced by three types of risk assessment tools when moderate risk individuals were classified as low risk
Web table 2: Metaregression analyses within three classes of risk assessment tools
Web figure 3: Summary receiver operator characteristic curve from bivariate analysis of individual violence risk assessment tools
Web figure 4: Summary receiver operator characteristic curve from bivariate analysis of individual sexual risk assessment tools
Web figure 5: Summary receiver operator characteristic curve from bivariate analysis of individual criminal risk assessment tools

criminal justice systems in many countries have welcomed the use of risk assessment to assist sentencing and release decisions. Risk assessment has been used to inform indeterminate sentencing in the UK,[12] and has become a largely uncontested part of an expanded criminal justice process in the US.[13] Furthermore, a 2004 survey reported that of the 32 US states where parole is an option, 23 had used such instruments as part of these decisions.[14]

The current group of risk assessment tools either provide a probabilistic estimate of violence risk in a specified time period (actuarial instruments), or allow for a professional judgment to be made on risk level (for example, low, moderate, or high) after taking into account the presence or absence of a predetermined set of factors (structured clinical judgment instruments). Over 150 of these structured measures currently exist,[15] and are starting to be implemented in low and middle income countries.[16 17]

However, these tools are time consuming and resource intensive, typically taking many hours to complete by a multidisciplinary group of professionals.[18] They can also be expensive; training is required for most tools, and payment is often needed for individual use. Further, and more importantly, the instruments' predictive accuracy remains a source of considerable uncertainty, with some reviews recommending their use in clinical and correctional settings and others finding that they lead to an unacceptably high number of false positive decisions.[18-22] Expert opinion is equally divided.[23-25]

We have therefore conducted a systematic review and meta-analysis of the predictive accuracy of the most commonly used risk assessment instruments. To consistently report outcomes for individual studies, we requested tabular data from primary authors. We have synthesised these data across a range of accuracy estimates, one of which was developed for the purposes of this review.

Methods

Review protocol

We followed the preferred reporting items for systematic reviews and meta-analyses statement.[26]

Risk assessment tools

We identified the nine most commonly used tools risk assessment using recent reviews[27-29] and questionnaire surveys.[30 31] Actuarial instruments included the Level of Service Inventory-Revised (LSI-R),[32] the Psychopathy Checklist-Revised (PCL-R),[33 34] the Sex Offender Risk Appraisal Guide (SORAG),[35 36] the Static-99,[37 38] and the Violence Risk Appraisal Guide (VRAG).[35 36] Structured clinical judgment tools included the Historical, Clinical, Risk management-20 (HCR-20);[39 40] the Sexual Violence Risk-20 (SVR-20);[41] the Spousal Assault Risk Assessment (SARA);[42-44] and the Structured Assessment of Violence Risk in Youth (SAVRY).[45 46] We divided tools into three categories: those designed to predict violent offending (HCR-20, SARA, SAVRY, and VRAG), sexual offending (SORAG, Static-99, and SVR-20), and any criminal offending (LSI-R and PCL-R). Although the PCL-R was originally developed to diagnose psychopathic personality disorder, it has become widely used for risk assessment purposes, as numerous studies have found the PCL-R score to be statistically significantly associated with criminal and antisocial outcomes.[47] Table 1⇓ reports additional details of all the instruments. Although these instruments were mostly designed to predict the likelihood of offending, we included violent, sexual, and antisocial outcomes (based on clinical records and other measures) even if they did not lead to convictions. For the sake of consistency, however, we refer to all outcomes as offences.

Systematic search

A systematic search was conducted to identify studies that measured the predictive validity of the nine tools. We searched the following databases between 1 January 1995 and 1 January 2011 using acronyms and full names of the instruments as keywords: PsycINFO, Embase, Medline, and US National Criminal Justice Reference Service Abstracts. Additional studies were identified through references, annotated bibliographies, and correspondence with researchers in the field. Studies in all languages and unpublished investigations were considered for inclusion. We excluded studies if they measured the predictive validity of select scales of a measure, if instruments were coded retrospectively without blinding to outcomes, or if they were calibration studies for the actuarial tools (which may give inflated effects).[48] When studies used overlapping samples, we used the sample with the largest number of participants to avoid double-counting. Using this search strategy, we identified 251validation studies (web figure 1).

To be included in the meta-analysis, studies had to report rates of true positives, false positives, true negatives, and false negatives at a given cut-off score for the outcome which the instrument was designed to predict. A pilot study showed that different score thresholds were used to classify people as being at low, moderate, or high risk of future offending. We contacted study authors and asked them to complete a standardised form if tabular data using the cut-off scores recommended in the most recent version of an instrument's manual were unavailable, or if the number of participants classified as low, moderate, or high risk was missing from a study of a structured clinical judgment tool. For publications in which multiple tools designed to predict the same outcome were tested on the same sample (eight studies), we requested tabular data for all outcomes but only included those for the tool with the fewest replication studies to increase the breadth of the findings. This procedure probably did not bias results, since χ^2 tests of differences in proportions found no differences in rates of true and false positives and true and false negatives in the tabular data obtained for included and excluded tools from the same study with the same outcome.

Standardised outcome data were available in the manuscripts of 30 eligible studies (32 samples). We requested additional data from the authors of 174 studies (330) and obtained data for 52 studies (62). Accuracy estimates from 235 of those 268 samples for which we were unable to obtain data were converted to Cohen's d using standard methods.[49-51] The median d value produced by those samples for which we could not obtain data (0.67, interquartile range 0.45 to 0.87) was similar to that of the 94 obtained samples (0.74, 0.54 to 0.95) (web figure 2 shows distribution of effect sizes). In addition, the Hodges-Lehmann percentile difference,[52] the median difference between all possible pairs of d values from the two groups, was small (0.01, 95% confidence interval 0.00 to 0.08). Finally, of the 82 studies for which tabular data was obtained, we were able to include information from 68 (73 samples; references available in web appendix), since the other 14 studies used instruments to predict outcomes other than those for which they were designed.

Data analysis

We followed the current guidance provided by the Cochrane collaboration for systematic reviews of diagnostic test accuracy.[53] The statistical methods for such reviews focus on

two statistical measures of diagnostic accuracy of the test: sensitivity (the proportion of offenders who a risk assessment tool predicted to offend) and specificity (the proportion of non-offenders who a risk assessment tool predicted would not offend). The aim of the analysis was to quantify and compare these statistics as well as the error rates (false positive and false negative diagnoses) for each type of test. The required analysis is a bivariate analysis of sensitivity and specificity for each study accounting for correlation between sensitivities and specificities.[54] The resulting model without covariates is a different parameterisation of the hierarchical summary receiver operating characteristic model.[55] We used summary receiver operator characteristic plots to display the results of each study in receiver operating characteristic space, plotting each study plotted as a single sensitivity-specificity point. Parameter estimates from the bivariate model produced a summary receiver operating characteristic curve with a summary operating point (that is, summary values for sensitivity and specificity), 95% confidence region, and 95% prediction region. We used the summary point from each curve to calculate the summary diagnostic odds ratio and both the sensitivity and specificity, each with 95% confidence intervals.

Since binary test outcomes are defined on the basis of a cut-off value for test positivity, we chose these values a priori. Risk assessment tools are predominantly used in clinical situations as instruments for identifying higher risk individuals,[19] thus, we combined participants who were classified as being at moderate or high risk for future offending and compared them with those classified as low risk. We did secondary analyses by comparing participants classified as high risk with those classified as low or moderate risk, an approach consistent with screening, and also by completely excluding those classified as moderate risk.

Accuracy estimates

We used a range of accuracy estimates to report on the predictive validity of the risk assessment tools. Firstly, the summary operating point was used to estimate the summary diagnostic odds ratio and both sensitivity and specificity. We obtained estimates for the area under the curve, positive predictive value, negative predictive value, number needed to detain, and number safely discharged from the individual sample estimates.

The diagnostic odds ratio is the ratio of the odds of a positive test result in an offender relative to the odds of a positive result in a non-offender, and is recommended for use with diagnostic instruments.[56] The area under the curve is an index of sensitivity and specificity across score thresholds, and is currently considered the accuracy estimate of choice in violence risk assessment when measuring predictive accuracy.[57] Neither the diagnostic odds ratio nor the area under the curve are affected by base rates of offending. The positive predictive value is the proportion of participants classified as at risk who go on to offend, whereas the negative predictive value refers to the proportion of those classified as not at risk who do not go on to offend. The number needed to detain is the number of people judged to be at risk who would need to be detained to prevent one incident of subsequent violence.[19 58] This outcome allows some quantification of the implications of using risk assessment tools to make detention decisions. Finally, the number safely discharged is a new performance statistic that we developed for the purposes of this review. This accuracy estimate calculates the number of participants judged to be at low risk who could be discharged into the community before a single act of violence occurs (1÷[1−negative predictive value]−1). A complement to the number needed to detain, the number safely discharged, allows researchers to quantify the implications of relying on a risk instrument to make discharge or release decisions.

Tests of assumptions

Standard meta-analytic pooling assumptions were met for diagnostic odds ratios and both sensitivity and specificity.[59 60] Since there was a significant correlation between the sensitivities and specificities produced by the samples in each class of risk assessment tools, pooling assumptions for areas under the curve were not met.[60] In addition, because the median base rate of offending within each class of tools varied considerably (violence 32.0%, interquartile range 22.2-46.6%; sexual 16.9%, 7.4-28.2%; criminal 28.4%, 20.7-46.0%), base rate dependent statistics were not pooled (such as the positive and negative predictive values and both the number needed to detain and the number safely discharged), and medians with interquartile ranges were calculated.

Investigation of heterogeneity

The standard Q and I^2 statistics[61] do not account for heterogeneity explained by phenomena such as positivity threshold effects, and the numerical estimates of the random effect terms in the bivariate regression are not easily interpreted. Therefore, the magnitude of observed heterogeneity in meta-analyses of diagnostic accuracy is instead best determined by the scatter of points in the summary receiver operating characteristic plot and from the prediction ellipse.[53] In particular, the prediction region depicts a region within which, assuming the model is correct, we have 95% confidence that the true sensitivity and specificity of a future study should lie.[53]

Since the diagnostic odds ratios met pooling assumptions, we used random effects metaregression to investigate sources of heterogeneity between studies in sample diagnostic odds ratios for each class of tools. Metaregression investigates the relation between accuracy estimates and dichotomous or continuous sample or study characteristics.[62] We formally explored the moderating role of the following variables: sex (proportion of sample that was male), ethnicity (proportion of sample that was white), mean participant age, type of instrument (actuarial *v* structured clinical judgment), temporal design (prospective *v* retrospective), setting in which assessment was conducted (correctional, forensic psychiatric, general psychiatric, or mixture), location of offending outcome (community only *v* inside institution or other), mean length of follow-up (months), sample size, and publication status (published in peer reviewed journal *v* not). We also conducted subgroup analyses using the bivariate models on these variables. Detailed examination of the overall differences between individual instruments have been reported in a subset of the samples.[63] We did all analyses in Stata 10.2[64] using the metandi (for bivariate model meta-analysis), metan (random effects meta-analysis), and metareg (metaregression) commands.

Results

Descriptive characteristics

We collected information for 24 847 participants in 73 samples from 68 independent studies (table 2⇓). Standardised outcome information from 43 of the samples (14 798 (59.6%) participants) was not reported in manuscripts and obtained directly from study authors. Of 24 847 participants, 5879 (23.7%) offended over an average of 49.6 months (standard deviation 40.5). Studies were conducted in 13 countries: Austria,

Belgium, Canada, Denmark, Finland, Germany, the Netherlands, New Zealand, Serbia, Spain, Sweden, the UK, and the US.

Predictive validity

We found differences in estimates of predictive accuracy depending on the type of risk assessment instrument (violence, sexual, or any criminal). Overall, based on diagnostic odds ratios, violence risk assessment tools performed best, and had higher positive predictive values than tools aimed at predicting sexual offending. Risk assessment instruments for violence and sexual offending produced high sensitivities and negative predictive values. In addition, risk assessment instruments for general offending had lower diagnostic odds ratios, areas under the curve, sensitivities, and negative predictive values and higher specificities and positive predictive values than the other two classes of instrument (table 3⇓, figs 1-3⇓⇓⇓).

For assessment instruments predicting the risk of violent outcomes, the summary diagnostic odds ratio was 6.1 (95% confidence interval 4.6 to 8.1) with moderate levels of heterogeneity (individual points moderately scattered in receiver operating characteristic space, fig 1) and a median area under the curve of 0.72 (interquartile range 0.68-0.78; table 3). Of those individuals who went on violently offend, 92% (95% confidence interval 88% to 94%) had been classified as being at moderate or high risk of future violence (that is, sensitivity). Of those participants who did not go on to violently offend, 36% (28% to 44%) had been judged to be at low risk (that is, specificity). Of those predicted to violently offend, 41% did (interquartile range 27-60%; positive predictive value), which was equivalent to a median number needed to detain of two (two-four). Of those who were predicted not to violently offend, 91% did not (81-95%; negative predictive value), equivalent to a median number safely discharged of 10 (four to 18).

Similar findings were obtained when individuals judged to be at moderate risk were grouped with those judged to be at low risk for the secondary analyses, but with considerably higher specificities and lower sensitivities (web table 1). When moderate risk individuals were excluded from analyses, assessment tools for violence risk produced considerably larger summary diagnostic odds ratios (16.8, 10.8 to 26.3) and specificities (0.72, 0.63 to 0.80).

Investigation of heterogeneity

Since we saw moderate levels of heterogeneity for the instruments assessing violence risk and higher levels for instruments assessing sexual and general offending risk (scatter of points from the line being greater and the prediction ellipses larger), we did metaregression and subgroup analyses using the bivariate model to determine any possible explanations for this heterogeneity. These analyses found no evidence that sex, ethnicity, age, type of instrument, temporal design, assessment setting, location of offending outcome, length of follow-up, sample size, or publication status was associated with differences in predictive validity (web table 2). In addition, we have presented summary receiver operating characteristic curves for each type of instrument (web figures 3-5). Subtypes of tools performed similarly, lying within the 95% prediction region, with the possible exception of the SAVRY that produced higher levels of predictive accuracy than the other violence risk assessment instruments.

Discussion

This systematic review and meta-analysis examined the predictive validity of violence risk assessment tools from 73 samples involving 24 847 individuals in 13 countries. Our principal finding was that there was heterogeneity in the performance of these measures depending on the purpose of the risk assessment. If used to inform treatment and management decisions, then these instruments performed moderately well in identifying those individuals at higher risk of violence and other forms of offending. However, if used as sole determinants of sentencing, and release or discharge decisions, these instruments are limited by their positive predictive values: 41% of people judged to be at moderate or high risk by violence risk assessment tools went on to violently offend, 23% of those judged to be at moderate or high risk by sexual risk assessment tools went on to sexually offend, and 52% of those judged to be at moderate or high risk by generic risk assessment tools went on to commit any offence. In samples with lower base rates than those that contributed to the review, such as in general psychiatry, positive predictive values will probably be even lower.[25] However, negative predictive values were high, and suggest that these tools can effectively screen out individuals at low risk of future offending. Whether the cautious optimism[13] that experts have described in relation to the ability to predict violence seems justified will depend on the use to which these instruments are put.

Comparisons with other medical tools

Any comparison of these risk assessment scores with other common medical diagnostic and prognostic tools poses several difficulties. Firstly, comparison with diagnostic tools is mostly inappropriate because risk assessment instruments attempt to predict the likelihood of a future outcome, whereas diagnostic instrument attempt to detect the presence of a current condition. Secondly, although it may be possible to compare performance statistics of these tools with those estimating, for example, cardiovascular risk, the implications of positive predictive values need to be considered in evaluating any comparisons. Violence risk assessment potentially leads to detention of individuals for longer than necessary, with its related economic,[65] social,[66] and civil rights consequences.[67] By comparison with common medical prognostic tools, it is possible to argue that the predictive accuracy of violence risk assessment needs to be higher because of these consequences, which extend beyond the person to other people. On the other hand, it is precisely because of the risks to other people that low positive predictive values may not be as important as the ability of these instruments to predict those that are not at risk. Our introduction of a novel performance measure, the number safely discharged, could help quantify this in future research.

Despite these caveats, the areas under the curve found in this review (0.66 to 0.74) were not dissimilar to those found in studies examining scores from the most validated cardiovascular risk scheme in predicting cardiovascular disease events. Areas under the curve from the Framingham scoring system range from 0.57 to 0.86, the SCORE from 0.65 to 0.85, and QRISK from 0.76 to 0.79.[68] Many of these studies report associations between predicted and observed risks,[69] which may be helpful for future research in violence risk assessment. Finally, the standard by which these instruments are compared will differ depending on their setting. In forensic psychiatry, a more meaningful comparison will be with unstructured clinical judgment, and clinical trials are needed to test whether structured risk assessment reduces adverse outcomes.

Clinical implications

One implication of these findings is that, even after 30 years of development, the view that violence, sexual, or criminal risk can be predicted in most cases is not evidence based. This message is important for the general public, media, and some administrations who may have unrealistic expectations of risk prediction for clinicians.[70] This expectation is not as high in other medical specialties, in which the expectation that the doctor will identify the individual patient who will have an adverse event is not a primary issue whereas psychiatry, in many countries such as the UK, has developed a culture of inquiries.[71]

A second and related implication is that these tools are not sufficient on their own for the purposes of risk assessment. In some criminal justice systems, expert testimony commonly use scores from these instruments in a simplistic way to estimate an individual's risk of serious repeat offending.[67] However, our review suggests that risk assessment tools in their current form can only be used to roughly classify individuals at the group level, and not to safely determine criminal prognosis in an individual case. This approach is mostly used in forensic psychiatry in the UK and other western countries, where they form part of a wider clinical assessment process. These instruments may also assist in developing risk management plans in selected high risk groups, as suggested by recent clinical guidelines in England and Wales.[72] Furthermore, they are preferable to unstructured clinical judgment owing to their increased transparency and reliability.

Another implication is that actuarial instruments focusing on historical risk factors perform no better than tools based on clinical judgment, a finding contrary to some previous reviews.[21 73] Finally, our review suggests that these instruments should be used differently. Since they had higher negative predictive values, one potential approach would be to use them to screen out low risk individuals. Researchers and policy makers could use the number safely discharged to determine the potential screening use of any particular tool, although its use could be limited for clinicians depending on the immediate and service consequences of false positives. A further caveat is that specificities were not high—therefore, although the decision maker can be confident that a person is truly low risk if screened out, when someone fails to be screened out as low risk, doctors cannot be certain that this person is not low risk. In other words, many individuals assessed as being at moderate or high risk could be, in fact, low risk. Ultimately, however, what constitutes an appropriate balance between the ethical implications of detaining people based on the predictive ability of these tools and the need for public protection will primarily be a political consideration.

Comparison with other studies

Previous meta-analyses on risk assessment have focused on comparing instruments with one another, or measuring how individual tools perform across sexes and ethnic groups.[74] A systematic review published in 2001 examined the accuracy of violence risk assessment in high risk groups,[19] and was based on 21 studies. It estimated that six people needed to be detained to prevent one violent offence, compared with our current review's estimate of two people needing detention. This difference was despite the median base rate of violence being similar in both reviews (current review, 32% (interquartile range 22-46%) *v* 2001 review, 26%, 15-41%). Unlike the previous report, the present meta-analysis focused on structured assessment instruments and included both institutional and community samples. The current report reviewed more than three times as many studies as the 2001 review and a recent meta-analysis that only compared head to head investigations of tool use.[75]

Strengths and limitations

The strengths of the current review include the incorporation of new tabular data, the reporting of multiple accuracy estimates, and a meta-analysis using bivariate models. We received new tabular data for 14 798 people (60% of people included in the review), and hence have reported a considerable amount of new data. Finally, by using a range of accuracy estimates, we have attempted to minimise biases that may be associated with reporting only one of them.

Limitations include that we solely examined the predictive qualities of these risk assessment tools, and did not account for their potential role in informing management and reminding clinicians to enquire about potentially important prognostic and modifiable factors.[76] In addition, we found moderate to high levels of heterogeneity. Heterogeneity was to be expected, in view of the different types of samples included in the primary studies (from prison, secure hospitals, and general psychiatric hospitals) and outcomes measured.[77 78] We explored sources of heterogeneity and found no clear trends. Investigating heterogeneity in diagnostic odds ratios meant that incidence of the outcome was accounted for. One possible source of heterogeneity was the potential effects of intervention after a risk assessment, particularly in people deemed high risk. We compared diagnostic odds ratios between prospective and retrospective studies that would be expected, to some extent, to measure this, since high risk participants identified in prospective studies would probably have been enrolled in interventions designed to reduce violence risk. However, we found no differences in metaregression or subgroup analysis. Nevertheless, clinical trials are needed directly to test the possible effects of intervention. Although we tested for publication status and found no clear patterns, we cannot exclude the possibility that such bias could exist in the studies that we were unable to include. Registers of such investigations would assist future reviews.[79] In addition, few samples reported on women and, thus, this review was underpowered to examine whether predictive validity was different from men.

We thank the following study authors for providing tabular data for the analyses: April Beckmann, Sarah Beggs, Susanne Bengtson Pedersen, Klaus-Peter Dahle, Rebecca Dempster, Mairead Dolan, Kevin Douglas, Reinhard Eher, Jorge Folino, Monica Gammelgård, Robert Hare, Grant Harris, Leslie Helmus, Andreas Hill, Hilda Ho, Clive Hollin, Christopher Kelly, Drew Kingston, P. Randy Kropp, Michael Lacy, Calvin Langton, Henry Lodewijks, Jan Looman, Karin Arbach Lucioni, Jeremy Mills, Catrin Morrissey, Thierry Pham, Charlotte Rennie, Martin Rettenberger, Marnie Rice, Michael Seto, David Simourd, Gabrielle Sjöstedt, Jennifer Skeem, Robert Snowden, Cornelis Stadtland, David Thornton, Jodi Viljoen, Vivienne de Vogel, Zoe Walkington, and Glenn Walters.

Contributors: SF devised and coordinated the project, assisted in data acquisition and interpretation, and drafted and revised the manuscript. JPS assisted in data acquisition, performed the statistical analyses, assisted in interpreting results, and assisted in drafting and revising the report. HD assisted in statistical analysis and critically revised the manuscript for important intellectual content. MG assisted in interpreting results and critically revising the manuscript for important intellectual content. SF and JPS had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis, and will act as guarantors.

Funding: SF is funded by the Wellcome Trust.

What is already known on this topic

Instruments based on structured risk assessment predict antisocial behaviour more accurately than those based on unstructured clinical judgment

More than 100 such tools have been developed and are increasingly used in clinical and criminal justice settings

Considerable uncertainty exists about how these tools should be used and for whom

What this study adds

The current level of evidence is not sufficiently strong for definitive decisions on sentencing, parole, and release or discharge to be made solely using these tools

These tools appear to identify low risk individuals with high levels of accuracy, but have low to moderate positive predictive values

The extent to which these instruments improve clinical outcomes and reduce repeat offending needs further research

Competing interests: All authors have completed the Unified Competing Interest form at www.icmje.org/coi_disclosure.pdf (available on request from the corresponding author) and declare: SF is funded by the Wellcome Trust; no financial relationships with any organisations that might have an interest in the submitted work in the previous 3 years; no other relationships or activities that could appear to have influenced the submitted work.

Ethical approval: No ethics approval was sought because only secondary data were used.

Data sharing: Data sharing: No additional data available.

1 Brundtland GH. Violence prevention: a public health approach. *JAMA* 2002;288:1580.
2 Krug EG, Mercy JA, Dahlberg LL, Zwi AB. The world report on violence and health. *Lancet* 2002;360:1083-8.
3 Ægisdóttir S, White MJ, Spengler PM, Maugherman AS, Anderson LA, Cook RS, 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.
4 Higgins N, Watts D, Bindman J, Slade M, Thornicroft G. Assessing violence risk in general adult psychiatry. *Psychiatr Bull* 2005;29:131-3.
5 Khiroya R, Weaver T, Maden T. Use and perceived utility of structured violence risk assessments in English medium secure forensic units. *Psychiatr Bull* 2009;33:129-32.
6 Pescosolido BA, Martin JK, Long JS, Medina TR, Phelan JC, Link BG. "A disease like any other"? A decade of change in public reactions to schizophrenia, depression, and alcohol dependence. *Am J Psychiatry* 2010;167:1321-30.
7 Wallace C, Mullen P, Burgess P, Palmer S, Ruschena D, Browne C. Serious criminal offending and mental disorder. *Br J Psychiatry* 1998;172:477-84.
8 Fazel S, Lichtenstein P, Grann M, Goodwin GM, Langstrom N. Bipolar disorder and violent crime: new evidence from population-based longitudinal studies and systematic review. *Arch Gen Psychiatry* 2010;67:931-38.
9 Fazel S, Långström N, Hjern A, Grann M, Lichtenstein P. Schizophrenia, substance abuse, and violent crime. *JAMA* 2009;301:2016-23.
10 National Institute for Health and Clinical Excellence. Core interventions in the treatment and management of schizophrenia in primary and secondary care. NICE, 2009.
11 American Psychiatric Association. Practice guidelines for the treatment of patients with schizophrenia. APA, 2004.
12 Harrison K. Dangerous offenders, indeterminate sentencing, and the rehabilitation revolution. *J Soc Welfare Fam Law* 2010;32:423-33.
13 Simon J. Reversal of fortune: the resurgence of individual risk assessment in criminal justice. *Ann Rev Law Soc Sci* 2005;1:397-421.
14 Harcourt B. Against prediction: profiling, policing, and punishing in an actuarial age. University of Chicago Press, 2007.
15 Singh JP, Serper M, Reinharth J, Fazel S. Structured assessment of violence risk in schizophrenia and other psychiatric disorders: a systematic review of the validity, reliability, and item content of 10 available instruments. *Schizophr Bull* 2011;37:899-912.
16 Gu Y, Hu Z. More attention should be paid to schizophrenic patients with risk of violent offences. *Psychiatry Clin Neurosci* 2009;63:592-3.
17 Jovanović AA, Toševski DL, Ivkonvić M, Damjanović A, Gašić MJ. Predicting violence in veterans with posttraumatic stress disorder. *Vojnosanit Pregl* 2009;66:13-21.
18 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.
19 Buchanan A, Leese M. Detention of people with dangerous severe personality disorders: a systematic review. *Lancet* 2001;358:1955-9.
20 Campbell MA, French S, Gendreau P. The prediction of violence in adult offenders: A meta-analytic comparison of instruments and methods of assessment. *Crim Justice Behav* 2009;36:567-90.
21 Hanson RK, Morton-Bourgon KE. The accuracy of recidivism risk assessments for sexual offenders: a meta-analysis of 118 prediction studies. *Psychol Assess* 2009;21:1-21.
22 Large MM, Ryan CJ, Singh SP, Paton MB, Nielssen OB. The predictive value of risk categorization in schizophrenia. *Harv Rev Psychiatry* 2011;19:25-33.
23 Maden A. Standardised risk assessment: why all the fuss? *Psychiatr Bull* 2003;27:201-4.
24 Mullen PE. Schizophrenia and violence: from correlations to preventative strategies. *Adv Psychiatr Treat* 2006;12:239-48.
25 Szmukler G. Violence risk prediction in practice. *Br J Psychiatry* 2001;178:84-8.
26 Moher D, Liberati A, Tetzlaff J, Altman DG, PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA Statement. *PLoS Med* 2009;6:e1000097.
27 Bonta J. Offender risk assessment: guidelines for selection and use. *Crim Just Behav* 2002;29:355-79.
28 Doren DM. Evaluating sex offenders: a manual for civil commitments and beyond. Sage, 2002.
29 Kemshall H. Risk assessment and management of known sexual and violent offenders: a review of current issues. UK Home Office, 2001.
30 Archer RP, Buffington-Vollum JK, Stredny RV, Handel RW. A survey of psychological test use patterns among forensic psychologists. *J Pers Assess* 2006;87:84-94.
31 Lally SJ. What tests are acceptable for use in forensic evaluations?: a survey of experts. *Prof Psychol Res Pract* 2003;34:491-8.
32 Andrews DA, Bonta J. LSI-R: the level of service inventory-revised. Multi-Health Systems, 1995.
33 Hare RD. The Hare psychopathy checklist-revised (PCL-R). Multi-Health Systems, 1991.
34 Hare RD. The Hare psychopathy checklist-revised. 2nd ed. Multi-Health Systems, 2003.
35 Quinsey VL, Harris GT, Rice ME, Cormier CA. Violent offenders: appraising and managing risk. American Psychological Association, 1998.
36 Quinsey VL, Harris GT, Rice ME, Cormier CA. Violent offenders: appraising and managing risk. 2nd ed. American Psychological Association, 2006.
37 Harris AJR, Phenix A, Hanson RK, Thornton D. Static-99 coding rules: revised 2003. Solicitor General Canada, 2003.
38 Hanson RK, Thornton D. Static-99: Improving actuarial risk assessments for sex offenders. Department of the Solicitor General of Canada, 1999.
39 Webster CD, Douglas KS, Eaves D, Hart SD. HCR-20: assessing risk for violence (version 2). Simon Fraser University, Mental Health, Law, and Policy Institute, 1997.
40 Webster CD, Eaves D, Douglas KS, Wintrup A. The HCR-20 scheme: the assessment of dangerousness and risk. Forensic Psychiatric Services Commission of British Columbia, 1995.
41 Boer DP, Hart SD, Kropp PR, Webster CD. Manual for the sexual violence risk-20. Professional guidelines for assessing risk of sexual violence. Simon Fraser University, Mental Health, Law, and Policy Institute, 1997.
42 Kropp PR, Hart SD, Webster CD, Eaves D. Manual for the spousal assault risk assessment guide. British Columbia Institute on Family Violence, 1994.
43 Kropp PR, Hart SD, Webster CD, Eaves D. Manual for the spousal assault risk assessment guide. 2nd ed. British Columbia Institute on Family Violence, 1995.
44 Kropp PR, Hart SD, Webster CD, Eaves D. Spousal assault risk assessment guide (SARA). Multi-Health Systems, 1999.
45 Borum R, Bartel P, Forth A. Manual for the structured assessment of violence risk in youth (SAVRY). University of South Florida, 2002.
46 Borum R, Bartel P, Forth A. Manual for the structured assessment of violence risk in youth (SAVRY): version 1.1. University of South Florida, 2003.
47 Leistico A, Salekin R, 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.
48 Blair PR, Marcus DK, Boccaccini MT. Is there an allegiance effect for assessment instruments? Actuarial risk assessment as an exemplar. *J Clin Psychol* 2008;15:346-60.
49 Cohen J. Statistical power analysis for the behavioral sciences. 2nd ed. Erlbaum, 1988.
50 Rosenthal R. Parametric measures of effect size. In: Cooper H, Hedges LV, eds. The handbook of research synthesis. Sage, 1994.
51 Ruscio J. A probability-based measure of effect size: robustness to base rates and other factors. *Psychol Meth* 2008;13:19-30.
52 Hodges JL, Lehmann EL. Estimates of located based on rank tests. *Ann Math Stat* 1963;34:598-611.
53 Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Analysing and presenting results. In: Deeks JJ, Bossuyt PM, Gatsonis C, eds. Cochrane handbook for systematic reviews of diagnostic test accuracy. Cochrane Collaboration, 2010. http://srdta.cochrane.org/.
54 Reitsma JB, Glas AS, Rutjes AW, Scholten RJ, Bossuyt PM, Zwinderman AH. Bivariate analysis of sensitivity and specificity produces informative summary measures in diagnostic reviews. *J Clin Epidemiol* 2005;58:982-90.
55 Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat Med* 2001;20:2865-84.
56 Glas AS, Lijmer JG, Prins MH, Bonsel GJ, Bossuyt PM. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epidemiol* 2003;56:1129-35.
57 Mossman D. Assessing predictions of violence: being accurate about accuracy. *J Consult Clin Psychol* 1994;62:783-92.
58 Fleminger S. Number needed to detain. *Br J Psychiatry* 1997;171:287.
59 Moses LE, Littenberg B, Shapiro D. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytical approaches and some additional considerations. *Stat Med* 1993;12:1293-316.
60 Deeks J. Systematic reviews of evaluation of diagnostic and screening tests. In: Egger M, Smith GD, Altman DG, eds. Systematic reviews in healthcare: meta-analysis in context. BMJ Publishing Groups, 2001.
61 Higgins JPT, Thompson S, Deeks J, Altman D. Measuring inconsistency in meta-analyses. *BMJ* 2003;327:557-60.
62 Thompson SG, Higgins JPT. How should meta-regression analyses be undertaken and interpreted? *Stat Med* 2002;21:1559-73.
63 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.
64 StataCorp. Stata statistical software: release 10.1. StataCorp LP, 2007.

65 Tyrer P, Duggan C, Cooper S, Crawford M, Seivewright H, Rutter D, et al. The successes and failures of the DSPD experiment: the assessment and management of severe personality disorder. *Med Sci Law* 2010;50:95-9.
66 Szmukler G. Risk assessment: 'numbers' and 'values'. *Psychiatr Bull* 2003;27:205-207.
67 Janus E. Sexually violent predator laws: psychiatry in service to a morally dubious enterprise. *Lancet* 2004;3664:50-1.
68 Cooney MT, Dubina A, Graham I. Value and limitations of existing scores for the assessment of cardiovascular risk. *J Am Coll Cardiol* 2009;54:1209-27.
69 Eichler K, Puhan MA, Steurer J, Bachmann LM. Prediction of first coronary events with the Framingham score: a systematic review. *Am Heart J* 2007;153:722-31.
70 Geddes J. Suicide and homicide by people with mental illness. *BMJ* 1999;318:1225-6.
71 Crichton JHM. A review of published independent inquiries in England into psychiatric patient homicide, 1995-2010. *J Forensic Psychiatr Psychol* 2011;22:761-89.
72 National Institute for Health and Clinical Excellence. Antisocial personality disorder: treatment, management and prevention. NICE, 2010.
73 Hanson RK, Morton-Bourgon K. Predictors of sexual recidivism: an updated meta-analysis. Public Works and Government Services Canada, 2004.
74 Singh JP, Fazel S. Forensic risk assessment: a metareview. *Crim Justice Behav* 2010;37:965-88.
75 Yang M, Wong SCP, Coid J. The efficacy of violence prediction: a meta-analytic comparison of nine risk assessment tools. *Psychol Bull* 2010;136:740-67.
76 Gilligan DG. Violence prediction and management [electronic response to Maden A, Scott F, Burnett R, Lewis GH, Skapinakis P. Offending in psychiatric patients after discharge from medium secure units: prospective national cohort study]. *BMJ* 2004. www.bmj.com/rapid-response/2011/10/30/violence-prediction-and-management.
77 Higgins JPT. Heterogeneity in meta-analysis should be expected and appropriately identified. *Int J Epidemiol* 2008;37:1158-60.
78 Davies S, Clarke M, Duggan C. Offending in psychiatric patients after discharge from medium secure units: Conviction rate may be misleading [letter]. *BMJ* 2004;329:684.4.
79 Editorial. Should protocols for observational research be registered? *Lancet* 2010;375:348.
80 Hart SD, Kropp PR, Hare RD. Performance of male psychopaths following conditional release from prison. *J Consult Clin Psychol* 1988;56:227-32.

Accepted: 15 June 2012

Cite this as: *BMJ* 2012;345:e4692

Tables

Table 1| Characteristics of nine included risk assessment tools

Instrument type and name	No of items	Population	Outcome	Current manual
Actuarial				
LSI-R*	54	Adult offenders	Criminal offending	Andrews and Bonta (1995)[32]
PCL-R†	20	Non-specific	Not applicable‡	Hare (2003)[33,34]
SORAG	14	Sexual offenders	Sexual offending	Quinsey et al (2006)[35,36]
Static-99§	10	Sexual offenders	Sexual offending	Harris et al (2003)[37,38]
VRAG	12	Mentally disordered violent offenders	Violent offending	Quinsey et al (2006)[35,36]
Structured clinical judgment				
HCR-20	20	Psychiatric patients	Violent offending	Webster et al (1997)[39,40]
SVR-20	20	Sexual offenders	Sexual offending	Boer et al (1997)[41]
SARA	20	Spousal assaulters	Violent offending	Kropp et al (1999)[42-44]
SAVRY	24	Adolescent offenders	Violent offending	Borum, Bartel, and Forth (2003)[45,46]

*Low and low to moderate risk categories combined to make low risk bin. Moderate to high and high risk categories combined to make high risk bin.

†Psychopathic patients (score >30) considered high risk group, non-psychopathic patients (<30) considered low risk group. PCL-R scores are included in SORAG, VRAG, HCR-20, and SVR-20, and thus the predictive validity of these instruments designed for different outcomes is correlated.

‡PCL-R was designed as a personality assessment. It started to be used as a risk instrument to predict criminal offending from 1988 onwards.[80]

§Moderate-low and moderate-high risk categories combined to make moderate risk bin.

Table 2| Descriptive and demographic characteristics of samples investigating predictive validity of risk assessment tools designed to predict violent, sexual, and criminal outcomes. Data are no (%) of samples unless stated otherwise. SD=standard deviation

Category and group	Violent (n=30)	Sexual (n=20)	Criminal (n=23)
Source of study			
Journal article	21 (70)	18 (90)	18 (78)
Conference	4 (13)	0	0
Dissertation	4 (13)	2 (10)	3 (13)
Government report	1 (3)	0	2 (8)
Tool information			
Type of tool			
Actuarial	9 (30)	16 (80)	23 (100)
Structured clinical judgment	21 (70)	4 (20)	0
Tool used			
HCR-20	9 (30)	—	–
LSI-R	—	—	11 (48)
PCL-R	—	—	12 (52)
SARA	3 (10)	—	—
SAVRY	9 (30)	—	—
SORAG	—	3 (15)	—
Static-99	—	13 (65)	—
SVR-20	—	4 (20)	—
VRAG	9 (30)	—	—
Demographic (mean (SD) in sample)			
Male participants (no)	137 (98)	519 (713)	409 (590)
White participants (no)	92 (49)	201 (185)	213 (165)
Age (years)	28.3 (10.0)	39.7 (4.0)	35.2 (4.6)
Study design			
Sample size (mean (SD))	148 (94)	510 (681)	439 (720)
Assessment setting			
Correctional	9 (30)	12 (60)	21 (91)
Forensic psychiatric	11 (37)	6 (30)	0
General psychiatric	5 (17)	0	0
Mixed	3 (10)	1 (5)	2 (9)
Unstated or unclear	2 (7)	1 (5)	0 (0)
Location of outcome			
Community	21 (70)	18 (90)	22 (96)
Intra-institutional	6 (20)	0	1 (4)
Mixed	3 (10)	2 (10)	0
Temporal design			
Prospective	12 (40)	5 (25)	14 (61)
Retrospective	17 (57)	15 (75)	9 (39)
Not stated or unclear	1 (3)	0	0
Length of follow-up (months; mean (SD))	39.4 (29.6)	82.4 (50.4)	33.9 (24.8)
Source of outcome			
Criminal register	16 (53)	17 (85)	17 (74)
Institutional records	6 (20)	0	1 (4)
Collateral report	2 (7)	0	0
Mixed	6 (20)	3 (15)	5 (22)

Table 3| Summary accuracy estimates produced by three types of tools for risk assessment

	Violent offending (n=30)*	Sexual offending (n=20)†	Criminal offending (n=23)‡
Summary estimates (95% CI) from summary receiver operating characteristic curve			
Diagnostic odds ratio	6.07 (4.58 to 8.05)	3.88 (2.36 to 6.40)	2.84 (2.09 to 3.88)
Sensitivity	0.92 (0.88 to 0.94)	0.88 (0.83 to 0.92)	0.41 (0.28 to 0.56)
Specificity	0.36 (0.28 to 0.44)	0.34 (0.20 to 0.51)	0.80 (0.67 to 0.89)
Individual study estimates (median (IQR))			
Area under the curve	0.72 (0.68-0.78)	0.74 (0.66-0.77)	0.66 (0.58-0.67)
Positive predictive value	0.41 (0.27-0.60)	0.23 (0.09-0.41)	0.52 (0.32-0.59)
Negative predictive value	0.91 (0.81-0.95)	0.93 (0.82-0.98)	0.76 (0.61-0.84)
Number needed to detain	2 (2-4)	5 (2-11)	2 (2-3)
Number safely discharged	10 (4-18)	14 (5-48)	3 (2-6)

CI=confidence interval; IQR=interquartile range; n=number of samples.

*HCR-20, SARA, SAVRY, and VRAG.

†SORAG, Static-99, and SVR-20.

‡LSI-R and PCL-R.

Figures

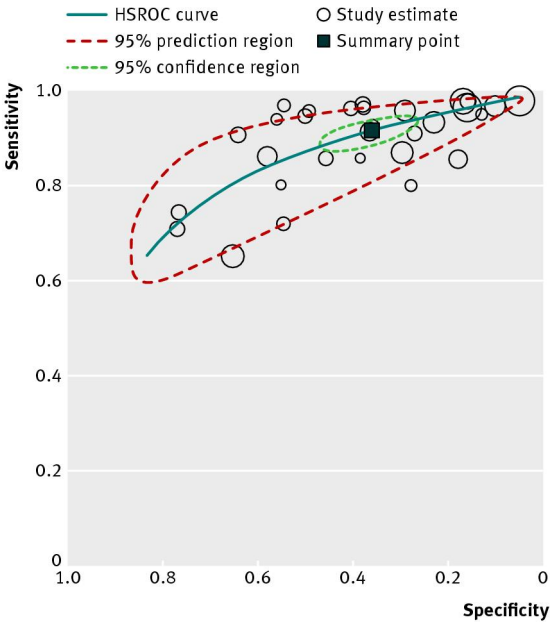

Fig 1 Summary receiver operating characteristics curve from bivariate analysis of risk assessment tools for violence offending. HSROC=hierarchical summary receiver operating curve; Summary point=best fit for sensitivity and specificity

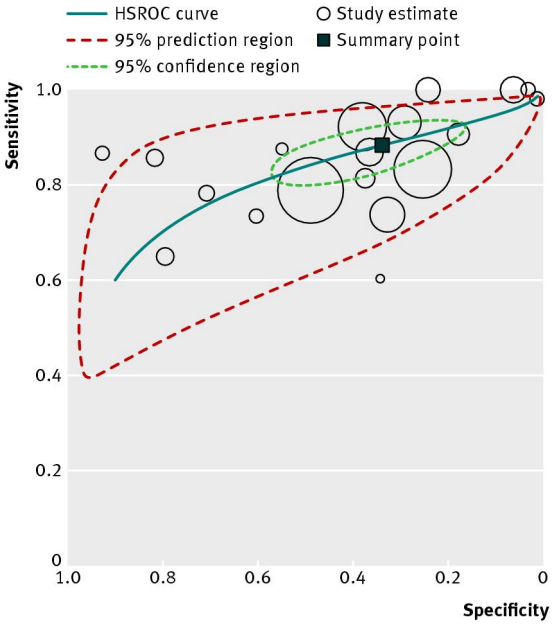

Fig 2 Summary receiver operating characteristics curve from bivariate analysis of risk assessment tools for sexual offending. HSROC=hierarchical summary receiver operating curve; Summary point=best fit for sensitivity and specificity

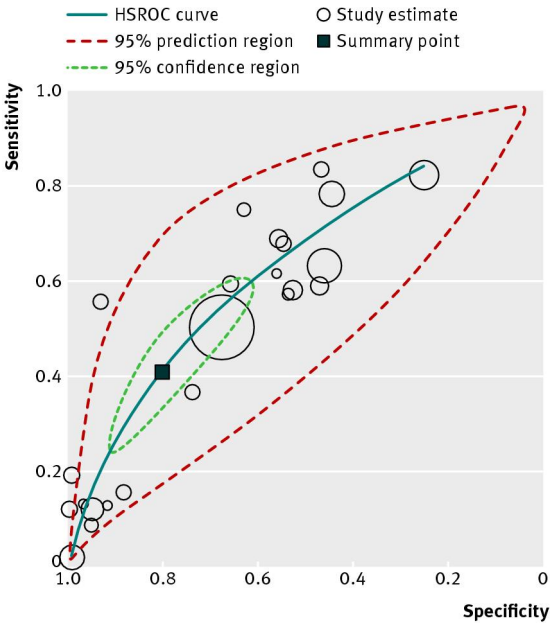

Fig 3 Summary receiver operating characteristics curve from bivariate analysis of risk assessment tools for criminal offending. HSROC=hierarchical summary receiver operating curve; Summary point=best fit for sensitivity and specificity