11

The Scientific Validity of Current Approaches to Violence and Criminal Risk Assessment

SEENA FAZEL

I. Introduction

Criminal justice systems in many high-income countries use some form of structured risk assessment tool or instrument to inform decisions about sentencing, parole, release and probation (see van Ginneken, Chapter 2 in this volume). These tools typically consider two aspects: the future risk of an individual for re-offending and also the criminogenic needs to mitigate this future risk. One estimate is that there are more than 300 such risk assessment tools (Singh et al 2014), many of which are heavily marketed and sold commercially. In the US alone, one report based on a review from 1970 to 2012 documented that 39 states have their own risk assessment tools (Desmarais, Johnson and Singh 2016). In contrast, in England and Wales, there is one risk tool in place for prisons and probation, called OASys (Offender Assessment System), which has been revised, as its first edition was found to have poor predictive performance (Howard and Dixon 2012). Typically, such tools include a set of risk factors, which may or may not be weighted, to provide a classification of risk (such as high, medium or low), a probabilistic score (ie, a percentage probability of re-offending within a certain timeframe) or both. At its most basic, a small number of static (or unchangeable) risk factors, such as sex, age and previous offending, are used to determine high, medium or low risk, but without any information as to what these categories actually mean in terms of probabilities, data on accuracy, or how these risk factors translate into one of these categories. The increasing use of these tools has been driven by the need to provide more consistent and defensible estimates of future risk and, in tools that are more focused on needs, better matching of treatment and interventions in criminal justice with their limited resources. The needs-based approaches attempt to assess individual factors that are thought to be related to offending, such as certain

attitudes, stable accommodation, relationship problems and family support. The uptake of these tools can also be explained by research findings, which suggest in general terms that they are better at prediction than human beings (Ægisdóttir et al 2006) and that unstructured clinical judgement (or the subjective judgement of individuals without any explicit framework of assessment) may be biased for many different reasons, including recent experience, prejudice against minority groups and attitudes towards certain offences.

This chapter will present a brief overview of performance measures for risk assessment instruments and will then summarise a number of recent systematic reviews examining the accuracy of commonly used instruments. I will then identify some gaps in the field and discuss whether the current tools are fit for purpose.

II. Measuring the Statistical Performance of Risk Assessment Tools

There are two approaches to test to the performance of such instruments: discrimination and calibration. Discrimination measures a particular tool's ability to distinguish between those who have offended and those who have not by assigning a higher risk score or category to those who offend. Discrimination is tested by reporting sensitivity, specificity, positive predictive value and negative predictive value (see the definitions below), which can only be calculated at specific risk cut-offs. In addition, an overall measure of discrimination across all possible cut-offs is the area under the curve (AUC, reported as a *c* statistic or c-index in some studies), which tests the probability that a randomly selected offender has a higher score on a tool than a randomly selected non-offender. The curve is the Receiver Operating Characteristic Curve (or ROC curve), which plots true positives against true negatives. To take one example, an AUC of 0.70 is the equivalent of saying that a tool will correctly assign a higher score 70 per cent of the time to a randomly selected offender than a randomly selected non-offender. Many studies rely on simply presenting discrimination statistics, and even then, only the AUC, which on its own is uninformative. For example, a tool can correctly classify individuals into higher and lower risk groups at all possible cut-offs, but is only used at a specific cut-off, where its discrimination is much poorer. This can be exemplified in the case of a risk assessment tool that has 30 items and is scored from 0 to 30. If the tool is tested in a research study and it correctly assigns all the offenders with a score of 2 compared to all non-offenders who score 0 and 1, then it will have a perfect AUC of 1. However, the guidelines for the use of the tool state that a score of 5 and above should be used to determine high risk of offending, and therefore the AUC statistic masks its poor intended performance. If used as intended with a cut-off of 5, this would mean that everyone in the sample is assigned a low risk score, even though some of these individuals are offenders. Depending on the number of offenders and non-offenders, this would mean that the AUC is closer

to 0.5 or chance. AUCs below 0.5 are worse than chance – in other words, such models are systematically wrong. This is one of the reasons why presenting a range of performance measures is important, particularly true and false positives and negatives. Indicative values of good discrimination measures have been discussed, but there is no clear consensus (Singh 2013).

Further, an instrument may be accurate in identifying risk groups, but may do so in a way that is very different from their real offending rates. In such a case, a tool may estimate rates of offending of 10 per cent to higher-risk offenders compared to 9 per cent to lower-risk offenders, and hence perfectly discriminates between these two groups. But if the higher-risk offenders are more likely to offend at rates of around 40 per cent and the lower-risk offenders at 1 per cent, then it is very poorly calibrated and has little if any practical utility as a prediction model (Lindhiem et al 2018). Calibration refers to the agreement between observed outcomes (ie, offending) and predictions from a particular tool. For example, if there is a prediction of a 30 per cent re-offending risk following release from prison in one year, the observed frequency for re-offending should be around 30 out of 100 released prisoners with such a prediction.

Sensitivity (the proportion of people who have offended that an instrument correctly classified as high risk) needs to be high if the aim is to screen individuals for a disease (eg, for further costly or more invasive investigations) and is important from a public policy perspective, as the consequences of 'missing' an individual who offends needs to be considered. The corollary of sensitivity is the false negative rate (which is calculated as 1-sensitivity) – the proportion of individuals who commit crimes that the tool misses. A false negative rate of, say, five per cent is equivalent to the tool not correctly identifying five out of every 100 individuals who have offended. Specificity (the proportion of individuals who have not offended that are correctly identified) should be high if the implications of being labelled high risk are harmful (eg, longer sentences or preventative detention). The false positive rate is the inverse of specificity (1-specificity) – the proportion of people that the tool incorrectly estimates will commit crimes. The relative proportion of true and false positive and negative rates will be determined by a range of legal, ethical and political concerns. Low false negative and positive rates will clearly be preferred, but a high false positive rate could be acceptable if the consequences of being labelled higher risk are not harmful. To exemplify this, if a tool does not miss individuals who re-offend on release (a low false negative rate), but also identifies many people as high risk who do not re-offend (a high false positive rate), this is less concerning if the consequences for those incorrectly identified as high risk are not harmful, such as additional support on release. Where it will be problematic is if the high-risk group have their prison sentences extended.

These decisions will need some alternatives to consider, such as the relative balance without using such tools or when two approaches can be compared. Some tools have tried to maximise the combination of sensitivity and specificity by adjusting cut-off points (eg, looking at the inflection point of a ROC curve – the

point at which there is a change in concavity of the curve). Here, researchers would look at the best possible cut-off by examining the inflection point. By finding the inflection point, this will translate into a cut-off to the nearest whole number for a tool that has the best discrimination for that particular sample being studied. The problem with this approach is that it is unlikely to be applicable to other samples, and pre-specifying a cut-off is preferable methodologically. In other words, taking this approach to identifying the best cut-off statistically will likely only apply to the specific sample being studied rather than new populations.

Some commentators have suggested that positive predictive value (PPV – the proportion of people that a tool identifies as high risk that actually offend) and negative predictive value (NPV – the proportion that are identified as low risk that do not offend) are more relevant to criminal justice as it is how these tools are used in practice (Buchanan and Leese 2001; Coid, Ullrich and Kallis 2013). The main limitation with this approach is that these two measures, alongside sensitivity and specificity, are also sensitive to the base rate, so the PPV will be low if the rate of offending in the population of interest is low, and the NPV will be high. Nevertheless, the NPV is increasingly important in some countries where decarceration is a public policy priority, which provides information on the proportion of prisoners that can be safely released (ie, not re-offend within a specified time period). It is also important for some populations such as juveniles and women, where prison should be avoided if possible, due to secondary effects on education, work, family and social networks, and mental health (Abram et al 2015). Sensitivity, specificity, PPV and NPV will change if a tool's cut-off changes – if the threshold for high risk increases, then sensitivity and NPV will decrease, and correspondently specificity and PPV will increase. This is one reason why the AUC is often presented as a summary statistic as it presents measures of discrimination (sensitivity and 1-specificity) at all possible cut-offs. At the same time, using AUCs to compare risk tools is problematic as very different numbers of false negative and false positive predictions resulting from different shapes of receiver operating curve may have the same overall AUC (Mallett et al 2012).

The other key measure of a tool's performance is calibration. This asks how closely the tool's predicted risk matches the observed risk. For example, a tool that predicts a 20 per cent chance of offending in a particular sample, but only 10 per cent actually offended, is poorly calibrated. Calibration can be examined graphically by plotting predicted risk versus observed offending behaviour or through statistical tests to measure the typical level of miscalibration, such as the Brier test or HL statistic (Lindhiem et al 2018). Calibration is the key performance measure if only probability scores are used – as the discrimination measures are only possible if there are a limited number of cut-offs. One important area of contention relevant to calibration is the group to individual problem, and proponents of this view have argued that it is not possible to apply group information to individuals due to a lack of precision, also known as the G2I ('group to individual') problem. The argument is put forward that when an actuarial tool provides a probability score

of 30 per cent, applying this to an individual is subject to the potentially large variation underlying the probability score. Hence, this view proposes that 30 per cent actually means 10–50 per cent for an individual and so is not informative. However, this position is based on a misunderstanding of statistics – all individual predictions are based on group data, and their precision will be a consequence of sample size (Imrey and Dawid 2015). The probability score of 30 per cent for a risk assessment tool can be interpreted by stating that it refers to an individual with the same risk factor profile who will on average re-offend at a rate of 30 per cent.

III. The Overall Performance of Currently Used Risk Assessment Tools

So what do we know about the performance of currently used tools in criminal justice? There have been a number of systematic reviews that have outlined their performance. Interestingly, none of them has reported calibration statistics, as it seems that this is very rarely reported in the research literature. In fact, one 2013 review of how AUCs were presented in 50 studies did not report one calibration metric (Singh et al 2013). The review by Yang and colleagues in 2010 looked at head-to-head comparisons of nine violence risk assessment tools and identified 28 studies in no more than 7,221 individuals, which reported AUCs and a measure of effect size (Cohen's d). It concluded that there was little difference in the included risk assessment measures, which varied in AUCs between 0.65 and 0.71 (Yang, Wong and Coid 2010). A later and more comprehensive review of an overlapping but different set of nine instruments identified 73 studies including 24,827 people (Fazel et al 2012). This review presented a broader range of discrimination statistics, and also separately by violent offending and any criminal offending. The findings were different by type of predicted outcome – for violent crime, sensitivity was high (0.92) and specificity was low (0.36), with moderate PPV (41 per cent) and high NPV (91 per cent). For any offending, sensitivity was low (0.41) and specificity was high (0.80), with moderate PPV (52 per cent) and NPV (76 per cent). In terms of AUCs, for violent offending it was 0.72 and for criminal offending it was 0.66. Overall, these are mixed discrimination metrics – moderate AUCs and NPVs – which suggests that their use in practice needs to reflect these differing performance metrics. One possibility is to screen out low-risk offenders. Another is to only use these tools as adjuncts in the decision-making process due to positive predictive values of around 40–50 per cent. Finally, due to the low specificity of violence risk assessment, they should only be used when the consequences of high-risk categories are non-harmful interventions, such as additional management or treatment. Another way of looking at these findings is to focus on false negative and false positive rates – for tools predicting violent outcomes, it was 8 per cent and 64 per cent, respectively, while for tools predicting any criminal outcomes (such as the Level of Service Inventory (LSI-R)), it was 59 per cent false

negative and 20 per cent false positive. If the implications of false positive rates are not harmful, this would suggest that instruments predicting violent outcomes should be prioritised over those focusing on any crime. In other words, this review found that the balance between false negatives (low for tools focusing on violent crime, but more than 50 per cent for tools with any crime outcomes) and false positives (high for tools focusing on violent crime, but lower for those predicting any crime) favours the violence risk assessment tools if the consequences of false positive (ie, being labelled high risk and not re-offending) are not harmful. The 59 per cent false negative rate for tools predicting any crime is arguably too high for their widespread use in criminal justice.

A third notable review summarised research on the predictive validity of 19 instruments used in US corrections from 1970 to 2012 (Desmarais, Johnson and Singh 2016). This review underscores the problems with the reporting of this literature. It found that only summary statistics were presented and solely for general recidivism (as distinct from violent recidivism). The median AUC of these tools typically ranged from 0.64 to 0.71 for new offences, and in real-life settings, the LSI-R, which is a commonly used tool, had an AUC of 0.63 and the RMS an AUC of 0.66. As with the other reviews, no information on calibration was reported, which is problematic as all the 19 included tools provide probabilistic scores of re-offending (and, in some cases, parole violations).

Overall, based on these recent systematic reviews of current risk assessment tools, there are major shortcomings in how these instruments are reported, with insufficient information on their performance. In addition, there are other problems with transparency (see also van Ginneken, Chapter 2 in this volume). The statistical contribution of individual risk factors to the final model, and the process by which they were chosen and categorised should be outlined. This transparency is important as it allows for the models to be critically appraised by experts, such as the nature of the sample that it was derived in, the choice of predictors and how they were categorised, the statistical power of the study, and the precision of the performance measures. This is particularly important if harm follows from a tool's use, such as longer sentences, certain interventions, and more restrictions in the community (cf Hannah-Moffat, this volume; Hester, this volume). Another problem are the potential financial and non-financial conflicts of interests among researchers in this field, and many of the tools being studied are conducted by individuals who developed or translated them (Singh, Grann et al 2013). Such potential conflicts need to be disclosed, which currently rarely occurs.

Scalability and cost need to be considered – some of the tools have commercial licences (such as the COMPAS or Correctional Offender Management Profile for Alternative Sanctions), which takes up to 60 minutes to complete. Many of these tools also assess individual needs and treatment (and linked to responsivity, which is the extent to which an intervention is responsive to the individual needs identified), and their predictive validity is one element in their potential value. However, conflating risk and needs can lead to loss of performance on risk,

and empirically robust risk calculators are required before more careful assessment of needs and treatment. Further, there have been some recent attempts to focus on causal risk factors as these will lead to reductions in recidivism once treated (Howard and Dixon 2013). However, one problem with this approach is that the most predictive factors (eg age, previous crime) are not causal, and excluding such factors will lead to poorer performance in terms of prediction. If the next stage of any risk management process is needs assessment, then identifying causal risk factors will be informative but will require different approaches (such as quasi-experimental designs and treatment trials rather than correlational studies of risk factors). Another issue is that the performance of current tools shrinks when used in real-world settings as distinct from research studies. A recent example was reported for the commonly used Psychopathy Checklist, revised edition (PCL-R). In a field trial in Belgium, its predictive validity was poor with an overall AUC of 0.63 for general recidivism and 0.57 for violent recidivism (Jeandarme, Edens et al 2017), which compares unfavourably to mostly research studies that have reported higher AUCs of 0.66-0.67 (Singh, Grann et al 2011; Yang, Wong et al 2010). The LSI-R, when used prospectively in over 22,000 prisoners in Washington State in the US, was associated with an AUC of 0.64 for violent recidivism (Barnoski and Aos 2003), which is lower than its performance in psychiatric samples and research studies. This shrinkage is a consequence of a number of methodological weaknesses in the design of these tools (see below for more on the LSI-R).

IV. A Practical Guide to Evaluate Risk Assessment Tools

So what can we make of this in practice? How can individuals in criminal justice and public policy determine whether a tool is fit for purposes? We have proposed a 10-point guide (Fazel and Wolf 2018), which I will summarise. I will start with criteria relevant to the derivation (or discovery or development) study and will then move on to criteria relating to the validation of risk assessment tools. The relevant criteria are as follows.

A. Did the Study Deriving the Tool Follow a Protocol?

This is a key component if a study is to provide an accurate representation of a tool's performance. Without a protocol, the likelihood of creating a tool that reports strong statistical performance but performs poorly in practice is high as it possible that the original methods were changed to optimise performance. The sample, candidate variables, outcome(s), follow-up periods, statistical analyses and output should all be pre-specified before any data analysis is performed. This protocol should be published, and any deviations from it in any particular study should be

clearly explained and justified (such as a predictor being dropped because of large proportions of missing data).

B. How were Candidate Variables Selected for the Tool?

The more variables that have been tested in a derivation study, particularly if the sample was not sufficiently large, the more likely the chance that associations are found, and the reported model performance will not perform well in external validation. One rule of thumb is that for each variable tested, the derivation sample should have at least 10 outcomes (Royston and Sauerbrei 2008). Further, the choice of which variables to test and how they are categorised should have followed a protocol, and multivariable regression should have been conducted to determine their independent association with the outcome (typically criminal reoffending) before inclusion in a model. Otherwise, tools will include variables that do not add incremental predictive accuracy and will lead to overcomplicated and time-consuming instruments.

C. How were Variables Weighted?

Many tools in criminal justice give equal weighting to all included items. This makes two assumptions: first, that all included predictors have the same association with the outcome; and, second, that the variables are all independently related to the outcome. In terms of weighting, previous violent crime and living in a poor neighbourhood are both associated with higher risk of crime, but they are not equally important. Tools that have not weighted individual items will perform worse (Hamilton et al 2015).

D. How were Other Parameters Selected?

Other key aspects of any research study should be determined beforehand and outlined in a protocol, such as the time for follow-up for the tool. If this has not been done, to take an example, a particular tool may perform better at three years rather than one or two years, and the researchers might decide that three years is the primary outcome. The problem with this approach is that it is a form of multiple testing and the consequence will be that the tool performs considerably worse in real-world settings.

E. Has Internal Validation Been Done?

This is typically done using a method called bootstrapping, which takes a number of random samples from the dataset to provide an estimate of accuracy of performance measures.

F. Has the Tool Been Externally Validated?

This question examines whether the tool's performance has been investigated in a new sample. In many ways, this is the most important question as tools tend to perform considerably better in the sample in which they were derived (Khoury, Gwinn and Ioannidis 2010, Monahan and Skeem 2016) and an external validation is necessary to test how accurate it is. Splitting the original derivation sample into two random groups is a form of internal validation, but is not external validation due to the equal distribution of predictor variables. Such a split will lead to comparable performance because the predictors will have a very similar distribution in the derivation and the randomly split samples. To achieve this, the sample should be split on other variables, which are not related to the outcome (Fazel et al 2016).

G. Has This Validation Been Done in the Population of Interest?

Here the key issue is whether the new population for which the tool will be used has similar characteristics, risk factors, baseline risk and outcome(s) to the sample where the tool was created. This may explain why some tools, such as the PCL-R, which was not developed to predict violence risk, but to identify a form of personality disturbance, performs among the worst of commonly used tools (Singh et al 2011). In addition, this is problematic for some tools developed in selected samples of high-risk offenders (which appears to have been the case for the LSI-R) that are then used in general criminal justice samples, such as all individuals in prison or on probation. Particularly important is using the same or very similar outcome as intended because differences in outcome prevalence will inevitably lead to reductions in performance.

H. Has the Validation Been Conducted Using Robust Methods?

Validation studies should stay true to the original model and be based on a protocol, and anticipated changes should be discussed beforehand in a protocol (eg, recalibration will be considered if the underlying base rate of offending is different, and how this recalibration of the model will be tested). Otherwise, what appears to be a validation is no longer an external validation, but the derivation of a new model. The sample size is also important and should aim for at least 100 events (or outcomes) for statistical power (Collins, Ogundimu and Altman 2016). Results should be published in peer-reviewed journals, but, on its own, this is not a marker of methodological quality. Studies should provide sufficient methodological detail in order to be replicable.

I. Has the Validation Study Reported Essential Information?

As described above, tools should report both measures of discrimination (especially rates of true and false positives and negatives) and calibration (ideally with a graphical plot that compares observed with predicted risks).

J. Is the Risk Assessment Tool Useful, Feasible and Acceptable?

The tool should provide useful information, including a relevant outcome (eg, prediction of re-offending), and clearly defined risk categories. The tools and their constituent predictors should also be easy to complete, reliable and clearly defined. For example, rating scales (eg, 1–5 Likert scales) may vary between raters. The tool should have face validity by including essential items (for example, age and sex) and should justify the inclusion of other items. There are advantages in having interview-independent tools to reduce the possibility of observer bias.

If a particular tool has not been externally validated, we argue that it should not be used in practice apart from rare circumstances when alternatives are not appropriate or available, and external validation is ongoing (Fazel and Wolf 2018). And even if it has been externally validated, instruments should undergo prospective validation after implementation to monitor their ongoing accuracy.

V. Applying Quality Criteria to Individual Risk Assessment Tools

The extent to which risk assessment tools currently used in criminal justice meet these 10 criteria needs to be systematically evaluated, but few of them meet more than one or two. To take some examples of commonly used tools, on these five criteria for derivation discussed above, two such instruments, the Historical Clinical Risk Management-20 (HCR-20) and the Violence Risk Appraisal Guide (VRAG), meet few criteria. The HCR-20 chose its 20 predictors based on expert opinion in 1997 rather than a systematic review of the evidence or testing them in multivariable models (an approach the authors reported in the following way: 'What variables might clinicians and administrators consider as they attempt evaluations of risk of violence in cases where psychiatric disorders are thought to be involved?'; Webster et al 1997: 251). The derivation did not include any statistical performance measures. Each item is scored as '0' (item not present), '1' (item possibly present) or '2' (item definitely present) rather than assigning any weighting to them (Douglas and Reeves 2010). Age and sex, two of the

strongest predictors of violence that are considered important for face validity, were not included. In developing the VRAG, 42 candidate variables were collected from a single sample of 618 mentally disordered Canadian offenders (of whom 191 re-offended). Of those, 332 individuals had been admitted to a maximum-security prison, while the remaining 286 had been admitted to a secure hospital for a brief pre-trial psychiatric assessment – not a sample that will be generalisable to most prisoners. With regard to the outcome, 191 re-offenders does not provide sufficient statistical power for 42 candidate variables (Harris, Rice and Quinsey 1993), and good practice would suggest that at least double the number of re-offenders would be required for derivation. The VRAG's derivation study reports performance measures at five different cut-offs (which were not pre-specified) and does not provide an overall performance measure. As with the HCR-20, the offender's sex was not one of the variables considered and hence was not included in the final model, which consists of 12 items (that are weighted).

Two other widely used tools are difficult to evaluate due to a lack of published information about certain aspects of their derivation and original validation. The LSI-R is based on 54 dynamic items, and the OASys Violent Predictor (OVP) in England and Wales, which is given to all individuals who receive sentences of 12 months or more, is derived from the 100-item OASys (Howard and Dixon 2012). However, the LSI-R does not include some of the most powerful predictors such age or gender, and has items that appear to be unreliable psychometrically (such as 'could make better use of time', has 'very few prosocial friends' and four items on current attitudes). Importantly, the original derivation study has not been published to my knowledge. The OASys is better reported and has some selected publications explaining aspects of its derivation, but lacks detail on some key areas (Howard and Dixon 2011, 2013). At the same time, both the LSI-R and the OASys have weighting for individual predictors that were tested using logistic regression in developing the model, along with relatively simple scoring methods, and have been subject to external validation.

Putting this altogether, I would argue that the most commonly used tools in criminal justice are not fit for purpose for prediction purposes. None of them meet all the 10 tests outlined above to my knowledge, and few meet more than one or two of the criteria outlined. At the same time, some of these instruments may provide a useful framework for organising information, may act as a reminder for those working in criminal justice to assess certain risk factors and domains, and may match individuals for treatment based on needs. The first two of these justifications are arguably too high a price to pay for those instruments that are resource-intensive.

VI. The OxRec Model

After reviewing this literature, I have been part of a team that has developed the Oxford Risk of Recidivism tool (OxRec), using Swedish national data, which

provides a probabilistic score for violence and any re-offending in one and two years post-release from prison, and also low/medium and high categories based on pre-specified levels. It can be completed in 5–10 minutes using 14 routinely collected predictors and using a freely available online calculator (Fazel et al 2016). The weighting of the individual predictors and how they are combined to create a probability score has been published (with the original protocol), with a full range of discrimination and calibration statistics, making it a fully transparent risk prediction model. It has been externally validated in Sweden in more than 10,000 individuals leaving prison (Fazel et al 2016), with another recent external validation in the Netherlands (Fazel et al 2019) and some ongoing in other countries, and provides a methodological rigorous approach with which to develop risk assessment instruments. The probability score is relatively precise as it was derived based on a study of 37,100 released prisoners.

VII. Summary

In summary, I have outlined some key ways of evaluating the performance of risk assessment instruments in criminal justice and have highlighted the importance of both investigating measures of discrimination and calibration. I have outlined some systematic reviews of the field, which suggest that many current tools, such as the LSI-R and the PCL-R, have at best moderate performance in discrimination with no information on calibration. Most tools currently used in criminal justice have not been included in these reviews because research on their external validation has not been published. Further, the development of risk assessment tools in criminal justice has lagged behind methodological improvements in prognostic models in science, and particularly in medicine.

Finally, I have provided a 10-point checklist that can be used to evaluate any risk tool. On this basis, I have argued that current widely used tools should probably not be used for prediction. At the very least, their use should be reviewed in the light of the 10 tests outlined, and information that is lacking should be requested from these tool's developers and commercial entities marketing them. In terms of its implications for predictive sentencing, risk predictions from commonly-used tools – either as categories such as high, medium or low, or as probability scores – do not have a sufficient evidence-base in support that they can currently be used in court. As I have shown, the current risk assessment tools have not met some basic criteria in terms of how they were derived or in subsequent validations of their performance. Furthermore, when empirically tested on a range of performance measures and mostly in research studies, they typically lead to unacceptably high false positives and false negative rates, particularly in tools aimed at any recidivism. I have also discussed the development and validation of a new scalable prediction tool, OxRec, which represents a methodological advance and provides a model for transparent reporting of such tools.

References

Abram, KM, Zwecker, NA, Welty, LJ, Hershfield, JA, Dulcan, MK and Teplin, LA (2015) 'Comorbidity and Continuity of Psychiatric Disorders in Youth after Detention: A Prospective Longitudinal Study' 72 *JAMA Psychiatry* 84.

Ægisdóttir, S, White, MJ, Spengler, PM, Maugherman, AS, Anderson, LA, Cook, RS, Nichols, CN, Lampropoulos, GK, Walker, BS and Cohen, G (2006) 'The Meta-analysis of Clinical Judgment Project: Fifty-Six Years of Accumulated Research on Clinical Versus Statistical Prediction' 34 *Counseling Psychologist* 341.

Barnoski, R and Aos, S (2003) 'Washington's Offender Accountability Act: An Analysis of the Department of Corrections' Risk Assessment' (Olympia, Washington State Institute for Public Policy).

Buchanan, A and Leese, M (2001) 'Detention of People with Dangerous Severe Personality Disorders: A Systematic Review' 358 *Lancet* 1955.

Coid, JW, Ullrich, S and Kallis, C (2013) 'Predicting Future Violence among Individuals with Psychopathy' 203 *British Journal of Psychiatry* 387.

Collins, GS, Ogundimu, EO and Altman, DG (2016) 'Sample Size Considerations for the External Validation of a Multivariable Prognostic Model: A Resampling Study' 35 *Statistics in Medicine* 214.

Desmarais, SL, Johnson, KL and Singh, JP (2016) 'Performance of Recidivism Risk Assessment Instruments in US Correctional Settings' 13 *Psychological Services* 206.

Douglas, KS and Reeves, KA (2010) *Historical-Clinical-Risk Management-20 (HCR-20) Violence Risk Assessment Scheme: Rationale, Application, and Empirical Overview* (Abingdon, Routledge).

Fazel, S, Chang, Z, Fanshawe, T, Långström, N, Lichtenstein, P, Larsson, H and Mallett, S (2016) 'Prediction of Violent Reoffending on Release from Prison: Derivation and External Validation of a Scalable Tool' 3 *Lancet Psychiatry* 535.

Fazel, S, Singh, JP, Doll, H and Grann, M (2012) 'Use of Risk Assessment Instruments to Predict Violence and Antisocial Behaviour in 73 Samples Involving 24,827 People: Systematic Review and Meta-analysis' 345 *British Medical Journal* e4692.

Fazel, S and Wolf, A (2018) 'Selecting a Risk Assessment Tool to Use in Practice: A 10-Point Guide' 21 *Evidence-Based Mental Health* 41.

Fazel, S, Wolf, S, Vasquez Martez, M and Fanshawe, T (2019) 'Prediction of violent reoffending in prisoners and individuals on probation: a Dutch validation study (OxRec)' *Scientific Reports* doi: 10.1038/s41598-018-37539-x.

Hamilton, Z, Neuilly, M-A, Lee, S and Barnoski, R (2015) 'Isolating Modeling Effects in Offender Risk Assessment' 11 *Journal of Experimental Criminology* 299.

Harris, GT, Rice, ME and Quinsey, VL (1993) 'Violent Recidivism of Mentally Disordered Offenders: The Development of a Statistical Prediction Instrument' 20 *Criminal Justice and Behavior* 315.

Howard, PD and Dixon, L (2011) 'Developing an Empirical Classification of Violent Offences for Use in the Prediction of Recidivism in England and Wales' 3 *Journal of Aggression, Conflict and Peace Research* 141.

——. (2012) 'The Construction and Validation of the OASys Violence Predictor: Advancing Violence Risk Assessment in the English and Welsh Correctional Services' 39 *Criminal Justice and Behavior* 287.

——. (2013) 'Identifying Change in the Likelihood of Violent Recidivism: Causal Dynamic Risk Factors in the OASys Violence Predictor' 37 *Law and Human Behavior* 163.

Imrey, PB and Dawid, AP (2015) 'A Commentary on Statistical Assessment of Violence Recidivism Risk' 2 *Statistics and Public Policy* 1.

Jeandarme, I, Edens, JF, Habets, P, Bruckers, L, Oei, K and Bogaerts, S (2017) 'PCL-R Field Validity in Prison and Hospital Settings' 41 *Law and Human Behavior* 29.

Khoury, MJ, Gwinn, M and Ioannidis, JP (2010) 'The Emergence of Translational Epidemiology: From Scientific Discovery to Population Health Impact' 172 *American Journal of Epidemiology* 517.

Lindhiem, O, Petersen, IT, Mentch, LK and Youngstrom, EA (2018) 'The Importance of Calibration in Clinical Psychology' *Assessment*, doi. org/10.1177/1073191117752055.

Mallett, S, Halligan, S, Thompson, M, Collins, GS and Altman, DG (2012) 'Interpreting Diagnostic Accuracy Studies for Patient Care' 345 *British Medical Journal* e3999.

Monahan, J and Skeem, JL (2016) 'Risk Assessment in Criminal Sentencing' 12 *Annual Review of Clinical Psychology* 489.

Royston, P and Sauerbrei, W (2008) *Multivariable Model-Building: A Pragmatic Approach to Regression Anaylsis Based on Fractional Polynomials for Modelling Continuous Variables* (Chichester, John Wiley & Sons).

Singh, JP (2013) 'Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer' 31 *Behavioral Sciences & the Law* 8.

Singh, JP, Desmarais, SL, Hurducas, C, Arbach-Lucioni, K, Condemarin, C, Dean, K, Doyle, M, Folino, JO, Godoy-Cervera, V, Grann, M, Ho, RMY, Large, MM, Nielsen, LH, Pham, TH, Rebocho, MF, Reeves, KA, Rettenberger, M, de Ruiter, C, Seewald, K and Otto, RK (2014) 'International Perspectives on the Practical Application of Violence Risk Assessment: A Global Survey of 44 Countries' 13 *International Journal of Forensic Mental Health* 193.

Singh, JP, Desmarais, SL and Van Dorn, RA (2013) 'Measurement of Predictive Validity in Violence Risk Assessment Studies: A Second-Order Systematic Review' 31 *Behavioral Sciences & the Law* 55.

Singh, JP, Grann, M and Fazel, S (2011) 'A Comparative Study of Violence Risk Assessment Tools: A Systematic Review and Metaregression Analysis of 68 Studies Involving 25,980 Participants' 31 *Clinical Psychology Review* 499.

——. (2013) 'Authorship Bias in Violence Risk Assessment? A Systematic Review and Meta-analysis' 8 *PloS One* e72484.

Webster, CD, Douglas, KS, Eaves, D and Hart, SD (1997) 'Assessing Risk of Violence to Others' in C Webster and M Jackson (eds), *Impulsivity: Theory, Assessment, and Treatment* (New York, Guilford Press).

Yang, M, Wong, SC and Coid, J (2010) 'The Efficacy of Violence Prediction: A Meta-analytic Comparison of Nine Risk Assessment Tools' 136 *Psychological Bulletin* 740.