Psychological Medicine (2012), **42**, 895–898. © Cambridge University Press 2011
doi:10.1017/S003329171100208X

Risk assessment and receiver operating characteristic curves

G. Szmukler[1]*, B. Everitt[2] and M. Leese[1]

[1] *Health Service and Population Research Department, Institute of Psychiatry, King's College London, London, UK*
[2] *Department of Biostatistics, Institute of Psychiatry, King's College London, London, UK*

Risk assessment is now regarded as a necessary competence in psychiatry. The area under the curve (AUC) statistic of the receiver operating characteristic curve is increasingly offered as the main evidence for accuracy of risk assessment instruments. But, even a highly statistically significant AUC is of limited value in clinical practice.

Received 15 June 2011; Revised 5 September 2011; Accepted 7 September 2011; First published online 3 October 2011

Key words: AUC, risk assessment, ROC curve.

Introduction

While everyone will agree that risk assessment in mental health should be evidence based (Department of Health, 2007), there are problems with what constitutes the 'evidence' and the way it is reported. Our interest is primarily risk assessment in general adult mental health settings, but in order to elucidate the key issues we shall consider data from forensic settings as well. We aim to clarify for a non-specialist audience what the evidence means.

Assessing the quality of a risk assessment instrument

An influential paper by Mossman (1994) argued for the use of receiver operating characteristic (ROC) curves and the 'area under the curve' (AUC) statistic in the evaluation of risk assessment instruments. ROC curves and the AUC allow comparison of instruments and provide a single statistic concerning how well the instrument classifies patients into violent *versus* non-violent groups.

Most studies now rely on ROC curves and the statistical significance of the AUC as the evidence that a risk assessment instrument is a valid predictor of future violence. For instance, Douglas *et al.* (2008) in their overview of the Historical, Clinical, Risk Management-20 (HCR-20) risk assessment scheme cite the AUC as the key 'validity index' for the majority of studies reviewed. Recent studies of instruments (e.g. Harris *et al*. 2004; Gray *et al*. 2008; Snowden *et al*. 2009) rely on the statistical significance of AUCs.

To clarify the statistics underlying risk assessment, we examine the relationship between ROC curves and the accuracy of violence prediction for an individual who scores positively on a particular test. Research on ROC statistics and other approaches for assessing predictive accuracy are developing apace, but for our purposes a relatively simple approach will suffice. Developments are unlikely to affect accuracy of prediction in the individual case. For illustration, we shall use the results from a well-executed study by Snowden *et al*. (2007). They tested the predictive power of two risk assessment instruments in a large sample of mentally disordered offenders, following them for at least 2 years following discharge from hospital. We shall consider the Violence Risk Appraisal Guide (VRAG; Harris *et al.* 1993), an instrument with a substantial research base in forensic populations.

The first four columns in Table 1 show the numbers and percentages of patients reconvicted of a violent offence, for each score ('category') of the VRAG. Overall there were 26 cases of convictions for violence over 2 years out of 364 discharged patients (a base rate of 7.1%).

The AUC for the ROC curve is the proportion of ROC space lying under the curved line [see Fig. 1, which is the ROC curve we have derived for the Snowden *et al.* (2007) data]. An AUC of 0.5 would be expected due to chance; 1.0 would represent perfect classification. In the Snowden *et al.* (2007) study, the AUC was 0.77, statistically highly significant ($p < 0.001$). The AUC in fact represents the probability that a randomly selected violent patient scores higher on the VRAG than a randomly selected non-violent

* Address for correspondence: G. Szmukler, M.D., Professor of Psychiatry and Society, Health Service and Population Research Department, Institute of Psychiatry, King's College London, De Crespigny Park, London SE5 8AF, UK.
(Email: george.szmukler@kcl.ac.uk)

Table 1. *Relationships between each VRAG category if taken as a 'cut-off' and some key variables*[a]

VRAG category	Number violent	Total number in category	Percentage violent	Sensitivity	Specificity	Positive predictive value
2	0	17	0	1.00	0.01	0.07
3	0	44	0	1.00	0.06	0.08
4	3	84	3.6	1.00	0.19	0.09
5	2	77	2.6	0.88	0.43	0.11
6	9	80	11.3	0.81	0.63	0.14
7	8	44	18.2	0.46	0.86	0.20
8	3	12	25.0	0.15	0.97	0.25
9	1	4	25.0	0.04	0.99	0.25

VRAG, Violence Risk Appraisal Guide.

[a] Based on data from Snowden *et al.* (2007).

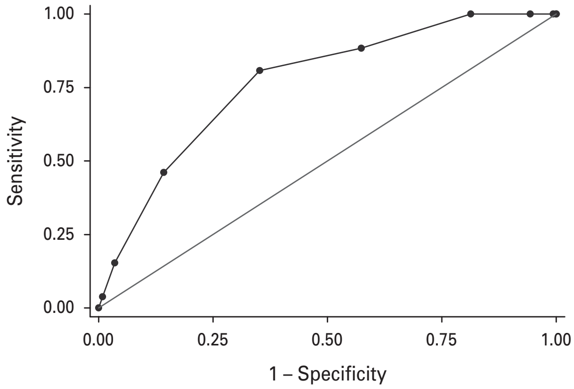

Fig. 1. Receiver operating characteristic (ROC) curve based on data from Snowden *et al.* (2007). Area under ROC curve = 0.7688.

patient, in this case, in nearly eight out of 10 instances. The original paper did not provide confidence limits for the AUC, but the 95% confidence interval (CI) using Stata 11 software (StataCorp LP, USA) indicates that it lies between 0.68 and 0.85.

The ROC and its relationship to an individual case

The ROC curve represents a series of 'sensitivity' *versus* 'specificity' values. So, for example, we can read off from the ROC curve that when the 'sensitivity' was 0.46, the 'specificity' was 0.86 (note that the ROC curve is plotted against (1 – specificity).

'Sensitivity', $a/(a+c)$, is a characteristic of the group of patients who turned out to be violent. 'Specificity', $d/(b+d)$, is a characteristic of the group of patients who turned out to be non-violent (see Table 2). If, for example, we were to take a score of 7 or above as the VRAG cut-off point, the sensitivity would be 0.46 (see Fig. 1 and Table 1). This indicates that approximately five out of 10 persons who turn out to be violent will have scored 7 or above. 'Specificity', in this case 0.86 at a cut-off of 7 or above, refers to the probability that a non-violent person will score as negative on the test. But it is important to note that this means that approximately one out of 10 will score as positive. 'False positives' – the bugbear of predictive tests for uncommon events – emerge through the 'specificity' of the test. Because there were so many more non-violent patients than violent ones in the population of interest – 13 times as many – the number of 'false positives' at a cut-off of 7 or above was four times the number of 'true positives'. This corresponds to a positive predictive value (PPV) of 0.2; that is, the test at this cut-off would have been wrong eight times out of 10.

Table 2. *Grid for calculation of sensitivity, specificity and positive predictive value*

	Turns out to be violent	Turns out not violent
Test predicts violence	*a*	*b*
Test predicts non-violence	*c*	*d*

The PPV is crucial in assessing the meaning of a positive test result for a particular patient in a particular setting. It is the proportion of cases predicted by the test to be violent who turn out to be violent – $a/(a+b)$ in the 2×2 table (Table 2). Table 1 shows the PPV for this sample at each VRAG category if it were

https://doi.org/10.1017/S003329171100208X Published online by Cambridge University Press

to be taken as the cut-off. As illustrated above, using a score of 7 or above as the cut-off, the PPV was 0.2.

Table 1 shows key characteristics of the data that are obscured by presentations of results in terms only of AUCs. In addition to the PPV at each possible cut-off, the table also shows the 'trade off' between 'sensitivity' and 'specificity' – or between 'true positives' and 'false positives'. The 'cost' of the VRAG getting one out of four predictions correct (PPV 0.25 for cut-off at VRAG category 8), is missing 22/26 of the violent cases.

Perhaps we would fare better if we used VRAG categories rather than cut-offs? Missing from the original paper are estimates of the uncertainties in the estimated proportions of patients in VRAG categories who were violent. For example, the 95% CI for the estimated proportion of patients scoring from 7 to 9, inclusive, was 0.11 to 0.33. If one were to argue that risk should be assessed according to groups, it is clear that the precision based on what appears to be a reasonably large study like this one is poor. Furthermore, the well-documented effect of 'shrinkage' is likely to make the situation worse when predicted probabilities are applied in new situations (for example, different samples). The accuracy of allocation of subjects to VRAG categories is yet another source of error. Clearly a very much larger number of subjects from the population of these patients would be required for achieving reasonable accuracy.

The value of a test such as the VRAG is further crucially influenced by the 'base rate' of violence in the population of interest. The ROC is, in theory at least, independent of the base rate of violence. In Snowden *et al*. (2007), the base rate was 7.1%. If the base rate were, say 15%, it can readily be shown that the PPV, using category 7 as a cut-off, would be 0.37. At a base rate of 5% it would be 0.15; and at 1%, it would be 0.03 – that is, wrong in 97 cases out of 100.

Thus the statistical significance of the AUC of the ROC alone offers little help when it comes to a particular patient, yet this is the statistic that is now most relied upon in the risk assessment literature.

ROCs and AUCs in other studies

How good are the AUCs generated by risk assessment instruments? The VRAG has an average AUC of 0.72 over more than 30 independent replications (Harris & Rice, 2007). A review by Douglas *et al*. (2008) of the HCR-20, another commonly used instrument, if scored, indicates a range of AUCs from 0.48 to 0.84, and, like the VRAG, AUC is commonly around 0.7. The most comprehensive attempt to produce a risk assessment instrument, tested on almost 1000 general psychiatric patients, that by the MacArthur Foundation, produced on its first iteration an AUC of 0.80. However, on replication with an independent sample, the AUC shrunk to 0.63 (with some later re-adjustments in the categorisation of violence, revised to 0.70; Monahan *et al*. 2005). Thus the study of Snowden *et al*. (2007) with its ROC of 0.77 is at the high end, but the limitations of the test in practice, as we have shown, are major.

Implications for practice and policy

In general psychiatric practice, base rates for serious violence are a magnitude smaller than in forensic care, perhaps 1% per annum for patients with a psychosis. False-positive rates are extremely high even with the best instruments. When the potential costs to false positives are great (for example, coercive interventions), the situation becomes serious. The problem – also dealt with in relation to offenders, in complex statistical terms by Cook & Michie (2010) – is the inherent limitation of population-derived, probabilistic data to predict individual behaviour. Our analysis above provides a concrete example of how this arises.

There is much more to the notion of risk than the probability of damage. The perception of risk (which includes factors such as the level of perceived control over the hazard, or the dread it engenders), its social amplification, as well as the degree of 'moral outrage' should the hazard occur (due to the belief that someone has failed in their duty to prevent it) are clearly important. A demand, impossible to meet, is being placed on clinicians to predict who will be dangerous. There are duties, highly challenging in this case, on researchers and practitioners to engage with the public to improve their understanding of what is possible and impossible, and why.

Declaration of Interest

None.

References

Cooke DJ, Michie C (2010). Limitations of diagnostic precision and predictive utility in the individual case: a challenge for forensic practice. *Law and Human Behavior* **34**, 259–274.

Department of Health (2007). *Best Practice in Managing Risk: Principles and Evidence for Best Practice in the Assessment and Management of Risk to Self and Others in Mental Health Services*. Department of Health: London.

Douglas KS, Guy LS, Weir J (2008). HCR-20 Violence risk assessment scheme: overview and annotated bibliography

https://doi.org/10.1017/S003329171100208X Published online by Cambridge University Press

(current up to October 3, 2006) (http://www.violence-risk.com/hcr20annotated.pdf). Accessed 28 May 2008.

Gray NS, Taylor J, Snowden RJ (2008). Predicting violent reconvictions using the HCR-20. *British Journal of Psychiatry* **192**, 384–387.

Harris GT, Rice ME (2007). Characterizing the value of actuarial violence risk assessments. *Criminal Justice and Behavior* **34**, 1638–1658.

Harris GT, Rice ME, Camilleri JA (2004). Applying a forensic actuarial assessment (the Violence Risk Appraisal Guide) to nonforensic patients. *Journal of Interpersonal Violence* **19**, 1063–1074.

Harris GT, Rice ME, Quinsey VL (1993). Violent recidivism of mentally disordered offenders: the development of a statistical prediction instrument. *Criminal Justice and Behavior* **20**, 315–335.

Monahan J, Steadman HJ, Robbins PC, Appelbaum P, Banks S, Grisso T, Heilbrun K, Mulvey EP, Roth L, Silver E (2005). An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric Services* **56**, 810–815.

Mossman D (1994). Assessing predictions of violence: being accurate about accuracy. *Journal of Consulting and Clinical Psychology* **62**, 783–792.

Snowden RJ, Gray NS, Taylor J, Fitzgerald S (2009). Assessing risk of future violence among forensic psychiatric inpatients with the Classification of Violence Risk (COVR). *Psychiatric Services* **60**, 1522–1526.

Snowden RJ, Gray NS, Taylor J, MacCulloch MJ (2007). Actuarial prediction of violent recidivism in mentally disordered offenders. *Psychological Medicine* **37**, 1539–1549.

https://doi.org/10.1017/S003329171100208X Published online by Cambridge University Press