Behavioral Sciences and the Law
Behav. Sci. Law (2013)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/bsl.2050

Evaluating Risk Assessments Using Receiver Operating Characteristic Analysis: Rationale, Advantages, Insights, and Limitations

Douglas Mossman, M.D.*

The last two decades have witnessed major changes in the way that mental health professionals assess, describe, and think about persons' risk for future violence. Psychiatrists and psychologists have gone from believing that they could not predict violence to feeling certain they can assess violence risk with well-above-chance accuracy. Receiver operating characteristic (ROC) analysis has played a central role in changing this view. This article reviews the key concepts underlying ROC methods, the meaning of the area under the ROC curve (AUC), the relationship between AUC and effect size *d*, and what these two indices tell us about evaluations of violence risk.

The area under the ROC curve and *d* provide succinct but incomplete descriptions of discrimination capacity. These indices do not provide details about sensitivity–specificity trade-offs; they do not tell us how to balance false-positive and false-negative errors; and they do not determine whether a diagnostic system is accurate enough to make practically useful distinctions between violent and non-violent subject groups. Justifying choices or clinical practices requires a contextual investigation of outcomes, a process that takes us beyond simply knowing global indices of accuracy. **Copyright © 2013 John Wiley & Sons, Ltd.**

INTRODUCTION

The last two decades have witnessed major changes in the methods that mental health professionals use to assess, describe, and think about patients' risk for future violence. When the *Tarasoff* decisions of the 1970s established clinicians' potential liability for their patients' harmful acts, experts on violence prediction believed that mental health professionals could not say anything useful about whether their patients would harm others. The following often-quoted statements capture this viewpoint:

- "[P]sychiatrists have absolutely no expertise in predicting dangerous behavior— indeed, they may be *less* accurate predictors than laymen—and . . . they usually err by overpredicting violence" (Ennis & Litwack, 1974, pp. 734–735).

*Correspondence to: Dr. Douglas Mossman, UC Department of Psychiatry, 260 Stetson Street, Suite 3200, Cincinnati, Ohio 45219-0559. E-mail: douglas.mossman@uc.edu

Copyright © 2013 John Wiley & Sons, Ltd.

- "[T]he 'best' clinical research currently in existence indicates that psychiatrists and psychologists are accurate in no more than one out of three predictions of violent behavior over a several-year period among institutionalized populations that had both committed violence in the past . . . and who were diagnosed as mentally ill" (Monahan, 1981, pp. 47–49).
- "In general, mental health professionals . . . are more likely to be wrong than right when they predict legally relevant behavior. When predicting violence, dangerousness, and suicide, they are far more likely to be wrong than right" (Morse, 1978, p. 600)

In an influential 1984 article, John Monahan (1984) suggested that mental health professionals' dismal performance in violence prediction studies might stem from methodological problems, such as fuzzy definitions of violence, erratic ascertainment of whether violence occurred, or the long study periods (usually years) covered by the studies. He thought that a "second generation" (p. 10) of studies–studies that looked at predictions over shorter prediction periods, used better prediction techniques, and were more careful at identifying and quantifying violence–might show that mental health professionals had at least some ability to gauge violence risk.

By the late 1990s, however, experts on violence prediction held very different views about what clinicians might accomplish, and they began to think in terms of "assessing risk" rather than making "violence predictions." In 1997, for example, Monahan noted that, in contrast to what had seemed true a decade earlier, clinicians could "distinguish violent from non-violent patients with a modest, better-than-chance level of accuracy" (Monahan, 1997, p. 317; see also Monahan, 2002, pp. 110–111).

Monahan's new conclusion resulted not so much from new empirical findings (though important studies with new findings had appeared in print) as from re-evaluations of previously available data. These re-evaluations applied receiver operating characteristic (ROC) analytic methods to problems involving violence prediction. ROC analysis recognizes that diagnosticians have varying levels of confidence about whether an either/or event will occur and that proper descriptions of detection accuracy must reflect these varying levels. The conceptual approach underlying ROC analysis has been used since the 1970s to evaluate radiological technologies and diagnoses (Goodenough, Rossmann, & Lusted, 1974; McNeil et al., 1975), and applications of ROC analysis began appearing in psychiatric publications in the late 1980s (Erdman et al., 1987; Mossman & Somoza, 1989; Murphy et al., 1987).

This article reviews key concepts underlying ROC analytic methods and their contribution to understanding what happens when clinicians or investigators evaluate violence risk. The next section explains the "motivation" for using ROC methods: a recognition that judgments about the occurrence of a future event usually fall along an implicit continuum from very low to very high probability, which means that statistical descriptions must reflect this key feature of predictions. The article then provides an intuitive description of the ROC accuracy indices and their relationship to other accuracy indices that readers are likely to encounter in publications on violent prediction. Subsequent sections describe limitations in what these indices tell us about the practical value and application of detection methods and prediction systems.

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

UNDERSTANDING THE MATHEMATICS OF VIOLENCE PREDICTION[1]

Predicting Violence: Studies from the 1970s and 1980s

Violent actions vary in their intensity and in the frequency with which they occur. Yet in most studies that examine predictors and predictions of future violence, investigators treat violence as a binary phenomenon. That is, once investigators define or operationalize a "violent act" for purposes of a study, they consider whether a subject acted violently or not and how well this yes-or-no outcome was predicted. For example, the MacArthur Violence Risk Assessment Study (Steadman et al., 1998) defined violence to others as acts that caused physical injury, sexual assaults, assaults that involved using a weapon, or threats made with a weapon in hand. The MacArthur investigators classified each subject as either violent or not, whether the subject committed one, two, or several such acts of violence towards others during a post-hospitalization follow-up period.

In most studies of violence prediction published before the mid-1990s, judgments about prediction accuracy assumed that assessments about future events had binary, will-or-will-not-occur forms as well. Table 1 explains the meaning of several indices used in describing the accuracy of binary predictions.

Two assumptions underlie the terms defined in Table 1: (i) violence either occurs (V+) or does not (V–); (ii) adopting medical terminology, a "test"–here, the prediction about whether violence will occur or not–is either "yes" (positive, T+) or "no" (negative, T–).

If a clinician predicts that a person will act violently and the person later does commit violence, that prediction is a "true positive" (TP). If a clinician predicts that a person will act violently but the person does not, the prediction is a "false positive" (FP). Similarly, a prediction of non-violence can turn out to be a "true negative" (TN) or a "false negative" (FN). In Table 1, TP, FP, FN, and TN designate the numbers of persons in each of these four categories. From TP, FP, FN, and TN, one can calculate all the accuracy indices listed underneath the 2×2 contingency matrix.

An Imaginary Study of Violence Prediction

To understand typical applications of binary accuracy indices, imagine the following dilemma faced by Dr. Jones, the superintendent of the imaginary, 600-bed Farblundget State Psychiatric Hospital (FSPH). The governor's office has informed Dr. Jones that, due to budgetary cutbacks, FSPH must downsize by quickly releasing one-third of its in patients, making sure not to release any "dangerous" individuals.

In response, Dr. Jones and the FPSH clinical staff set about to identify those patients who, they hope, have the best chance of not doing anything violent if they leave the hospital. A week later, the FSPH staff members have identified 197 patients who, they believe, have the lowest chances of doing violence if released to the community. Then, FSPH learns that the upcoming budget crisis is even worse than Dr. Jones's superiors previously thought; FSPH must close, and all 600 of its patients will be released.

On hearing this news, social scientists realize that they now have (to quote Professor Monahan's description of research on the consequences of patients transferred after

[1] Portions of this section are adapted from Mossman (2006).

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

Table 1. Chief methods of characterizing the accuracy of binary violence predictions

	Patients' Actual behavior		
Predictions:	Violent (V+)	Not violent (V–)	*Sums*:
"Will be violent" (T+)	TP = true positives	FP = false positives	TP + FP
"Will not be violent" (T–)	FN = false negatives	TN = true negatives	FN + TN
Sums:	TP + FN	FP + TN	

Base rate = BR = $\frac{TP+FN}{TP+FP+FN+TN}$

Sensitivity = true positive rate (tpr) = $P(\mathrm{T}+|\mathrm{V}+) = \frac{TP}{TP+FN}$

Specificity = true negative rate (tnr) = $P(\mathrm{T}-|\mathrm{V}-) = \frac{TN}{FP+TN}$

False positive rate (fpr) = 1 – tnr = $P(\mathrm{T}+|\mathrm{V}-) = \frac{FP}{FP+TN}$

Correct fraction = CF = $\frac{TP+TN}{TP+FP+FN+TN} = \boldsymbol{tpr} \times \boldsymbol{BR} + \boldsymbol{tnr} \times (1 - \boldsymbol{BR})$

Percent correct = $\frac{TP+TN}{TP+FP+FN+TN} \times 100\%$

Positive predictive value = PPV = $P(\mathrm{V}+|\mathrm{T}+) = \frac{TP}{TP+FP}$

Negative predictive value = $NPV = P(\mathrm{V}-|\mathrm{T}-) = \frac{TN}{FN+TN}$

Note that:

$$\boldsymbol{PPV} = P(\boldsymbol{V}+|\boldsymbol{T}+) = \frac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FP}} = \frac{\dfrac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FP}+\boldsymbol{FN}+\boldsymbol{TN}}}{\dfrac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FP}+\boldsymbol{FN}+\boldsymbol{TN}} + \dfrac{\boldsymbol{FP}}{\boldsymbol{TP}+\boldsymbol{FP}+\boldsymbol{FN}+\boldsymbol{TN}}}$$

$$= \frac{\dfrac{\boldsymbol{TP}+\boldsymbol{FN}}{\boldsymbol{TP}+\boldsymbol{FP}+\boldsymbol{FN}+\boldsymbol{TN}} \times \dfrac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FN}}}{\left(\dfrac{\boldsymbol{TP}+\boldsymbol{FN}}{\boldsymbol{TP}+\boldsymbol{FP}+\boldsymbol{FN}+\boldsymbol{TN}} \times \dfrac{\boldsymbol{TP}}{\boldsymbol{TP}+\boldsymbol{FN}}\right) + \left(\dfrac{\boldsymbol{FP}+\boldsymbol{TN}}{\boldsymbol{TP}+\boldsymbol{FP}+\boldsymbol{FN}+\boldsymbol{TN}} \times \dfrac{\boldsymbol{FP}}{\boldsymbol{FP}+\boldsymbol{TN}}\right)}$$

$$= \frac{\boldsymbol{BR} \times \boldsymbol{tpr}}{(\boldsymbol{BR} \times \boldsymbol{tpr}) + ([1 - \boldsymbol{BR}] \times \boldsymbol{fpr})} = \frac{P(\boldsymbol{V}+)P(\boldsymbol{T}+|\boldsymbol{V}+)}{P(\boldsymbol{V}+)P(\boldsymbol{T}+|\boldsymbol{V}+) + [1 - P(\boldsymbol{V}+)]P(\boldsymbol{T}+|\boldsymbol{V}-)} = \frac{P(\boldsymbol{V}+)P(\boldsymbol{T}+|\boldsymbol{V}+)}{P(\boldsymbol{T}+)},$$

which is a statement of Bayes' theorem.

Baxtrom v. Herold (1966)) "an excellent opportunity for naturalistic research" (Monahan, 1981, p. 46) to find out how accurately clinicians predict violence. The scientists decide to monitor the former FSPH patients for the first year after their release by meeting periodically with them and their families and by checking their police records to see whether they commit any violent acts. In this study, any confirmable report of striking, physically fighting with, or doing physical harm to another person will identify a former patient as having been "violent" (cf. Steadman et al., 1998). For the purposes of the study, investigators will consider the 197 patients whom the clinicians would have released under their original instructions as "predicted not violent"; the other 403 patients are "predicted violent."

Twelve months later, 150 former FSPH patients–one-quarter of those released–have committed violent acts, and more details about this result appear in Figure 1. Suppose one asks, "What is the likelihood that a patient whom the clinicians predicted violent actually became violent?" Answering this question calls for a calculation of the positive predictive value (PPV in Table 1) of the FSPH clinicians' "predictions," and the answer is 134 out of 403 (33%) cases, or "one-third of the time." This implies that two-thirds of the FSPH clinicians' "predictions" of violence were wrong.

Suppose one next asks, "In what fraction of cases were clinicians' predictions correct?" This calls for a calculation of the correct fraction (CF in Table 1). The data in Figure 1 suggest that FSPH clinicians were "right" about the 134 actually violent former patients

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

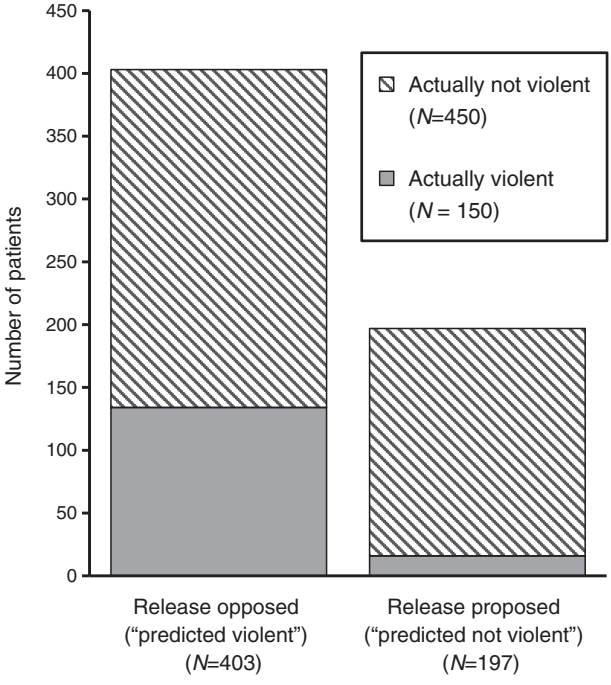

Figure 1. Results of an imaginary study on violence prediction.

whose release they opposed (TP = 134) and the 181 actually non-violent former patients whose release they recommended (TN = 181). The FSPH clinicians thus correctly categorized 315 (just over half) of the 600 patients, and CF = 0.525. Notice, though, that evaluating accuracy via the CF statistic would lead us to conclude that the FSPH clinicians could have done better at predicting by simply recommending that all patients be released—then, the clinicians would have been correct 75% of the time. Here, then, is a fictional example of a commonly observed phenomenon: when the base rate (BR) of a phenomenon is low, one often can make better predictions by simply guessing that the phenomenon will not occur than by trying to figure out whether the phenomenon will occur (Faust & Ahern, 2011).

Some readers may object at this point, saying, "The FSPH clinicians were not trying to maximize the PPV or CF of yes-or-no decisions: they were trying to avoid releasing 'dangerous' patients under an order to downsize the hospital census. How well would they have done?"

As it turns out, the scientists conducting this "violence prediction" study have access to more information about the FSPH clinicians' decision-making, and examination of this information shows that the clinicians had implemented a sound policy of release planning. In making their judgments about patients' potential post-release aggressiveness, FSPH clinicians classified patients as falling into five "dangerousness" categories: 1 = "well below average"; 2 = "below average"; 3 = "average"; 4 = "well above average"; 5 = "well above average"

Figure 2 shows the numbers of actually violent and non-violent patients who fell into each risk category. The clinicians' mandate not to release patients who were "dangerous" led them to plan to release only those patients who had below-average levels of risk. Their choice of this decision threshold (the dashed, vertical line in Figure 2) led to a discharge policy that was highly sensitive to violence. Numerically, this sensitive-and-cautious

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

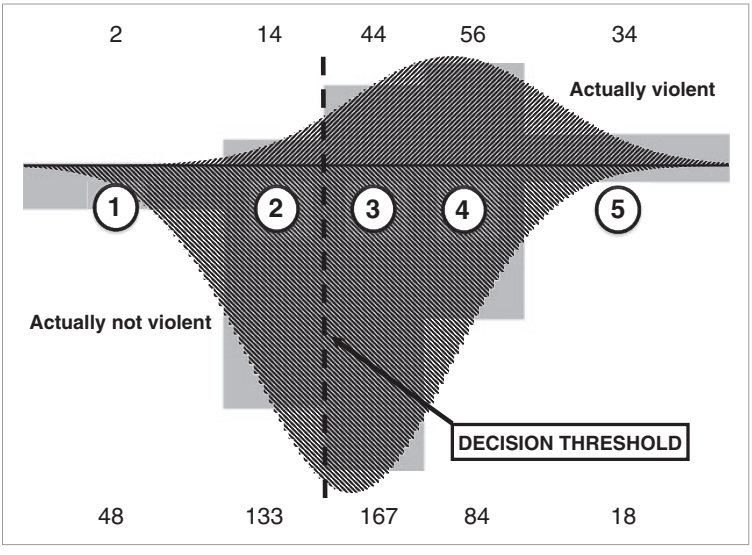

Figure 2. Numbers of actually violent and non-violent patients who fell into each of five risk categories in the imaginary study.

policy is expressed in the finding that nearly 90% of the violent patients would not have been released had FSPH downsized but not closed, though only 40% of the non-violent patients would have returned to the community. Rather than simply betting the base rate, Dr. Jones and his staff at FSPH responded to their mandate to avoid releasing "dangerous" patients. They set a release-decision threshold at a point that reflected an (implicit) policy judgment that making a "false negative" error (release of a patient who would become violent) was much worse than making a "false positive" error (retention of a non-violent patient).

ROC Analysis

In the 1990s, studies of violence prediction began evaluating the accuracy of violence predictions using statistical methods that separate effects of base rates and decision thresholds from intrinsic detection capabilities (Mossman, 1994a, 1994b; Gardner et al., 1996; Rice & Harris, 1995). The chief statistical tool for accomplishing this was ROC analysis. The first article to apply ROC analysis to violence prediction showed that, in contrast to what courts and mental health publications had believed, clinicians could "distinguish violent from non-violent patients with a modest, better-than-chance level of accuracy" (Mossman, 1994a, p. 790) and that short-term predictions covering several days were no more accurate than predictions that covered a year or more. By the middle of the 21st century's first decade, ROC indices had become investigators' standard tools for describing instruments that assess the risk of future violence and the recidivism potential of sex offenders.

Figure 2 depicts the FSPH clinicians' five-category ratings of patients as histograms, with rectangular regions extending downward to represent the non-violent patients and upward for the violent patients. Superimposed on the histograms are two Gaussian ("normal" or "bell-shaped") distributions. Figure 3 contains a ROC curve with some features added to illustrate the relationship between the categories in Figure 2 and the area

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

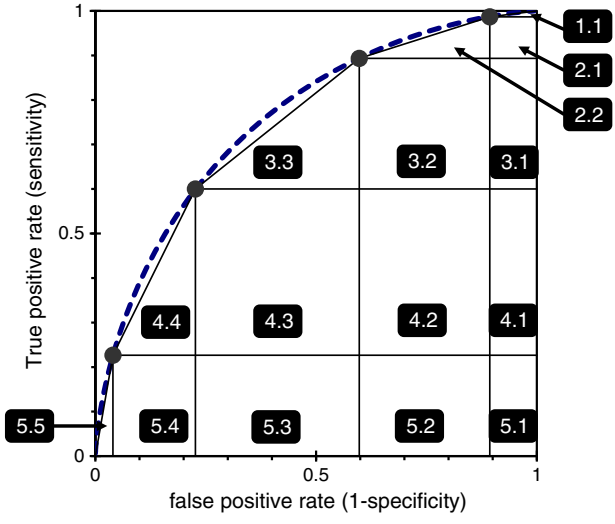

Figure 3. Receiver operating characteristic curve describing the accuracy of the Farblundget State Psychiatric Hospital FSPH clinicians with features that illustrate the relationship between risk categories in Figure 2 and the trapezoidal area under the curve.

under the curve in Figure 3. Looking at Figure 2, one can see that the five risk categories used by the FSPH clinicians created four potential decision thresholds (in addition to keeping everyone or discharging everyone) for making release recommendations. The threshold that the clinicians would have used put them close to the desired goal of reducing the hospital population by one-third, but it also meant that they set their apparent sensitivity for detecting violent patients at 134/150 = 0.89 and their specificity at 181/450 = 0.40. Had they proposed discharging only patients whose risk was judged to be "well below average," the sensitivity would have increased to 148/150 = 0.99, but the specificity would have fallen to 48/450 = 0.11. Other decision thresholds would have yielded other values for specificity and sensitivity.

Figure 3 contains a ROC graph, which depicts the accuracy of the FSPH clinicians as a set of trade-offs between sensitivity and specificity as a detection method's decision threshold is moved throughout the entire range of possibilities. It is customary to construct a ROC graph by plotting the true positive rate of a detection method (tpr, equal to sensitivity) as a function of the false positive rate of the detection method (fpr = 1 – specificity). Returning for a moment to Table 1, one sees that unlike PPV and CF, tpr and fpr are not functions of the base rate, which means that ROC graphs depict features of diagnostic systems independently of the base rate.

The large circles in Figure 3 represent the four cut-offs that would be formed by placing decision thresholds at the four boundaries between the FSPH clinicians' five categories of risk shown in Figure 2. The smooth ROC curve in Figure 3 (the dashed line) links these four cut-offs under the "binormal assumption" of ROC analysis (Dorfman & Alf, 1969; Metz, Herman, & Shen, 1998; Somoza & Mossman, 1991), which is illustrated using the bell-shaped curves in Figure 2. In the context of violence prediction, the binormal assumption suggests that discrimination capacity can be succinctly summarized by assuming that the violence prediction method partially separates violent and non-violent individuals along a continuous, latent decision axis. The violent and non-violent

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

populations thus form two overlapping, "normal" distributions with different means and standard deviations (SDs; Mossman, 1994a).

For our imaginary FSPH study, the binormal assumption implies that we believe the clinicians had the ability to rate each patient's violence risk along a continuum of risk–or, at least, along a latent decision axis with many gradations in risk levels. We can regard the clinicians' choice to classify patients into five categories of risk and the boundaries (thresholds) between those categories as somewhat arbitrary: the clinicians might have used (say) three or seven risk categories, and if they had done so, the number and location of the boundaries (thresholds) between categories would have had different locations along the underlying decision axis.

Receiver operating characteristic analysis gives investigators several ways of summarizing prediction accuracy, but here, I focus on three commonly used indices. The first index, the area under the curve (AUC), is a simple summary of overall accuracy (Hanley & McNeil, 1982). AUC answers the question, "If we randomly select one patient from the violent group and one patient from the non-violent group, what is the probability that the FSPH clinicians would have assigned a higher probability of violence to the actually violent patient?" We can understand the calculation of AUC via an unusual feature of Figure 3, the rectangular and triangular partitions labeled with boxes such as 4.3. Returning to Figure 2 for a moment, one sees that (for example) the probability of randomly selecting a violent patient who belongs to category 4 is 56/150 = 0.373 and the probability of randomly selecting a non-violent patient who belongs to category 3 is 167/450 = 0.371. Therefore, the joint probability of both events is $\left(\frac{56}{150}\right) \times \left(\frac{167}{450}\right) = 0.139$. In Figure 3, this probability calculation corresponds to the square partition labeled 4.3. This partition falls entirely within the total AUC, because all violent patients rated "4" have higher ratings than all the non-violent patients rated "3" and therefore, all such random pairings of patients would lead to correct assignments of patients to either the violent or non-violent group. In cases where both the violent and non-violent members of a random pair have the same rating, clinicians could flip a coin to designate one member of the pair as "predicted violent," and they would be correct half the time. Thus, for example, the probability of selecting a violent and non-violent patient who are both rated "4" and correctly classifying them by guessing is $\frac{1}{2} \times \left(\frac{56}{150}\right) \times \left(\frac{84}{450}\right) = 0.0697$, and Figure 3 depicts this calculation graphically using the triangular partition labeled 4.4. The total AUC equals the proportion of all possible random pairs of violent and non-violent patients who would be classified correctly using the five-category ratings, or the sum of the area partitions shown in Figure 3. More formally, one calculates this "trapezoidal" area under the ROC curve, AUC_T, as follows:

$$\begin{aligned} \boldsymbol{AUC_T} = & \sum_{k=1}^{K-1} \left[P(\text{rating} = k|V-) \sum_{l=k+1}^{K} P(\text{rating} = l|V+) \right] \\ & + \frac{1}{2} \sum_{k=1}^{K} P(\text{rating} = k|V-) P(\text{rating} = k|V+) \end{aligned} \tag{1}$$

where $k = \{1, 2, \ldots, K\}$ are the rating categories, $l = k + 1$ is the category that is one greater than category k, and $P(\text{rating} = k \mid V[i])$ means "the probability of a k rating, given violence status $V[i]$, $i = \{-, +\}$."

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

From the previous paragraph, one sees that a perfect violence prediction method—one that always would give a randomly chosen violent person a higher rating than a randomly chosen non-violent person—would have an AUC of 1.0. A violence prediction method that gave no information about future behavior–that is, a method that did no better than a coin toss at distinguishing a violent person from a non-violent person—would have an AUC of 0.5. Calculated using equation (1), $AUC_T = 0.743$, a result that is typical of findings from recent studies of violence prediction methods (Douglas & Reeves, 2010; Singh, Grann, & Fazel, 2011; Yang, Wong, & Coid, 2010).

Other Accuracy Indices

Calculating the effect size of the prediction method is another way of summarizing overall accuracy of a prediction method. Looking at the bell-shaped distributions in Figure 2, we see that risk ratings of the violent patients tend to place them to the right compared with the non-violent patients; in this case, it turns out that the rightward displacement equals 1 SD. Put another way, the effect of the FSPH clinicians' assessment is to separate the distributions of violent and non-violent patients by 1 SD, or more simply, the effect size equals 1. Often, studies in behavioral science refer to the effect size as "Cohen's *d*," acknowledging a frequently cited text that uses this statistic (Cohen, 1988). Formally, *d* is the standardized mean distance between the ratings of the members of the violent and non-violent patients, that is,

$$d = \frac{\bar{x}_{V+} - \bar{x}_{V-}}{s_{pooled}} \tag{2a}$$

where $\bar{x}_{V+}$ and $\bar{x}_{V-}$ are the mean ratings of the V + and V − populations and s_{pooled} is the pooled SD:

$$s_{pooled} = \sqrt{\frac{(n_{V+} - 1)s_{V+}^2 + (n_{V-} - 1)s_{V-}^2}{n_{V+} + n_{V-}}} \tag{2b}$$

In equation (2b), $n_{\text{V+}}$ and $n_{\text{V}-}$ are the numbers of violent and non-violent individuals, and s_{V+}^2 and s_{V-}^2 are the variances of the population ratings.

A third way of describing accuracy utilizes the binormal assumption and refers to the locations of the normal distributions along the latent decision axis. If we assign the mean and SD of the non-violent population the values of 0 and 1, respectively, we can express the properties of this violence detection system using the following linear equation:

$$Z_{TPR} = A + BZ_{FPR} \tag{3}$$

In equation (3), Z_{TPR} and Z_{FPR} are the normal deviates, or *z*-transforms, of the true and false positive rates, respectively; *A* equals the distance between the means of the violent and non-violent populations (measured in units of the SD of the non-violent population) and *B* equals the ratio of the SDs of the non-violent and violent populations (*i.e.*, $\frac{SD_{V-}}{SD_{V+}}$). When $B = 1$, *A* equals the effect size.

Values of AUC calculated using the binormal assumption incorporate the idea mentioned earlier that numbers of rating categories and locations of cut-off points are somewhat arbitrary; (fpr, tpr) values inferred from counting subjects in rating categories are best

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

considered estimates of possible thresholds along a continuous decision axis. One computes the binormal area under the ROC curve, AUC_Z, as

$$\mathbf{AUC}_z = \Phi\left(\frac{\mathrm{A}}{\sqrt{1+\mathrm{B}^2}}\right) \tag{4}$$

where $\Phi(\bullet)$ is the cumulative normal distribution function (cndf). Maximum likelihood estimates of the binormal ROC indices AUC_z, A, and B and corresponding SDs can be obtained using free, downloadable software developed by Metz and colleagues (available at http://metz-roc.uchicago.edu/MetzROC/software, explained in Metz & Kronman, 1980), and this article's author has written code for the free software program WinBUGS (Lunn et al., 2009) that provides Bayesian estimates of binormal ROC indices and their credible intervals. Applying the latter method to the FSPH data yields these index estimates and their 95% intervals: $A = 0.98$ [0.75–1.21], $B = 0.99$ [0.82–1.17], $AUC_Z = 0.76$ [0.71–0.80].

One can interconvert Cohen's d and AUC_T by considering one other statistic, the common language effect size (CLES, McGraw & Wong, 1992). Under most circumstances, CLES is roughly equal to AUC_Z, and AUC_Z is just a bit larger than AUC_T. By setting B equal 1, we can adapt equation (4) to produce the following relationship:

$$d = A \geq \sqrt{2}\left(\Phi^{-1}\boldsymbol{AUC_T}\right) \tag{5}$$

where $\Phi^{-1}(\bullet)$ is the inverse of the cndf, or $z_{(\bullet)}$. When B is very different from unity, one can use the binormal indices to compute an effect size d_a as a multiple of the perpendicular distance between the line described by equation (1) and the origin of an ROC curve plotted in normal deviate space:

$$d_{\boldsymbol{a}} = \sqrt{2}\left(\Phi^{-1}\boldsymbol{AUC}_Z\right) = \sqrt{2}\frac{A}{\sqrt{1+B^2}} \tag{6}$$

Another interpretation of d_a is that it equals the difference between the populations' two means relative to the root mean square of SD_{V+} and SD_{V-} (Simpson & Fitter, 1973). More details on these interrelationships appear in Rice and Harris (2005) and Ruscio (2008).

USING ACCURACY INDICES: SOME EXAMPLES

The just-described indices allow several methods of describing and evaluating risk assessment instruments. Here, we shall consider three such methods: comparisons of diagnostic methods or detection systems; efforts to set decision thresholds; and attempts to classify effect sizes as "small," "medium," or "large."

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

Comparing Risk Assessment "Technologies"

Along with the application of ROC methods to violence prediction, the 1990s also witnessed a turn toward a different type of violence prediction "technology." Before 1990, studies of psychiatric violence prediction usually evaluated how well unstructured clinical judgments by mental health professionals predicted patients' future violence. Here, "clinical judgment" refers to an evaluation process in which professionals mentally combines whatever data they think is relevant "in their heads" to assess persons' risks of becoming violent. Clinical judgments about dangerousness might incorporate aspects of the professionals' knowledge, personal experience, "gut" feelings and other intuitions, and whatever other information about the situation that seems relevant to the problem. This process is called "clinical" because it mimics how physicians arrive at judgments about their patients' diagnoses and treatments: doctors interview and examine patients, think about what is probably going wrong, and then suggest what patients should do and prescribe treatments.

This clinical process differs greatly from what insurance company actuaries do when judging, say, an individual's life expectancy for the purposes of setting an insurance premium. Actuaries often use simple formulae, tables, algorithms, or other pre-specified ways to combine information and arrive at predictions or risk estimates. For at least two decades, psychologists have used the term "actuarial judgment" to describe prediction processes that resemble what insurance actuaries do when they evaluate the risk of future events (Dawes, Faust, & Meehl, 1989). The essence of actuarial judgment is that it generates predictions or risk estimates from empirically based relationships between specific types of data and the events to be predicted. Because actuarial approaches to judging risk use fixed, predetermined formulae or algorithms, they are sometimes called "statistical," "mechanical," or "formal" methods for making judgments (Grove & Meehl, 1996; Grove et al., 2000).

The 1990s marked the first time that psychologists published studies of how well various actuarial "technologies" or "instruments" assessed the risk of future violence (e.g., Gardner et al., 1996; Rice & Harris, 1995). Typically, using an actuarial risk assessment instrument (ARAI) requires a clinician to gather information about 10–20 factors concerning the individuals undergoing evaluation. The clinician then scores that information about each factor using some pre-specified method, such as an instruction manual or equation. This process generates a numerical value or rating category that describes the evaluees' risk of violence.

Social science publications on ARAIs now number in the hundreds (if not thousands), and they usually quantify accuracy using ROC indices. Using these indices, one can compare diverse prediction methods, including methods that rely on clinical or actuarial judgment. Studies of actuarial prediction methods have consistently contained results that imply well-above-chance levels of predictive accuracy, but so do clinical judgment studies when evaluated with ROC methods (Mossman, 1994a, 2000). However, AUC values from clinical judgment studies are generally smaller than AUC values from actuarial studies. Few studies have directly compared actuarial and clinical judgments about the same population of evaluees, but most scholars and social scientists believe that actuarial judgments are superior. Adding confidence to this judgment are findings from a vast body of research that has shown the superiority of actuarial judgments in many other types of predictions (Grove et al., 2000).

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

Less clear is whether actuarial judgment is better than another approach to risk assessment, structured professional judgment (SPJ). In implementing SPJ, clinicians begin their evaluations using scales containing factors with theoretical or empirically demonstrated relationships to the outcome of interest (e.g., violent offending), just as they do in making actuarial judgments; indeed, many SPJ scales have been evaluated statistically as though their numerical results represented a form of actuarial assessment (e.g., Douglas et al., 1999). In SPJ, however, the scales and their numerical results serve as starting points from which clinicians anchor their judgments about potential outcomes, judiciously taking into account factors not included on the scale (e.g., a patient's recent threat) when considering risk of violence. Proponents believe that SPJ is superior to actuarial judgment because it allows consideration of relevant factors not included in ARAIs (Sreenivasan et al., 2010). Also, SPJ instruments typically include features or factors that can guide decisions about management (i.e., reduction) of risk (Hart & Logan, 2011), which is more helpful to clinicians than merely knowing someone's risk level.

The conceptualization of accuracy that underlies ROC analytic methods is central to recent meta-analytic studies that have compared accuracy of various risk assessment strategies. For example, a meta-analysis of measures for assessing risk of recidivism by sex offenders used effect sizes (which, as the previous section explained, are directly related to AUCs) to evaluate whether evaluators would achieve superior accuracy by adjusting risk ratings or by simple summing items. In one set of analyses, the authors concluded that differences in accuracy "were not large enough to be meaningful"; in another set, "the adjusted scores showed lower predictive accuracy than did the unadjusted actuarial scores" (Hanson & Morton-Bourgon, 2009, p. 7). Another meta-analysis that examined 68 studies of SPJ instruments and ARAIs found median AUCs that ranged between 0.66 and 0.78; the authors "found no evidence that, compared with SCJ tools, actuarial instruments produced better levels of predictive validity" (Singh et al., 2011, p. 510).

Selecting Thresholds

Actuarial risk assessment instruments and SPJ instruments are imperfect methods of distinguishing those individuals who will be violent from those who will not. As I have noted earlier, recent studies of actuarial violence prediction typically report AUCs of 0.65–0.80. This implies that, though the ratings or scores of violent persons are, on average, higher than those of non-violent persons (so that the probability of violence increases as the score increases), the score distributions of violent and non-violent individuals overlap considerably. This means that users of ARAIs or SPJ tools usually cannot make decisions based on the outcome of a violence risk assessment alone; users also must make judgments about what level of risk should trigger what course of action.

In the case of the fictional FSPH clinicians, the decision threshold was effectively set by their directive to release one-third of their patients, which was very close to the proportion of patients who fell into the two lowest-risk rating categories. In many legal contexts where ARAIs are used to assess the recidivism risks of sex offenders and their eligibility for commitments as "sexually violent predators," statutes limit commitment only to those offenders who are "likely to engage in repeat acts of sexual violence" (K.S.A. § 59-29a02, 2009). At least in theory, this could mean that some specific probability p of reoffending would justify commitment, so that only those

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

offenders whose ARAI scores implied a risk greater than *p* would be eligible for commitment.

In most circumstances, however, the level of risk that justifies taking a particular action is far from clear. In fact, studies described by Mossman (2006) suggest that for civil commitment aimed at averting violence, people disagree enormously (over five orders of magnitude) about balancing the trade-offs between wrongfully committing a non-violent person and releasing a person who is not violent. This huge range of disagreement about balancing false-positive and false-negative errors implies a big range of possible cut-offs for any risk assessment instrument currently or likely to become available. This may help to explain Swets, Dawes, and Monahan's (2000, p. 22) observation:

> Assessing benefits and costs can be problematic; publicizing them can leave the decision maker vulnerable to criticism. How many safe people should be hospitalized as "dangerous" to prevent discharging one patient who turns out to be violent? No court has ever answered that question with a number. Judges are notoriously reluctant to set decision threshold that depend on overt cost-benefit consideration, as are many other professionals and officials.

Effect Sizes: Does One Rule Fit All?

In his introduction to the use of effect sizes in power analyses, Cohen (1988) suggests a rule of thumb:

> If the investigator thinks that the effect ... is small, he might posit a **d** value such as .2 or .3. If he anticipates it to be large, he might posit **d** as .8 or 1.0. If he expects it to be medium (or simply seeks to straddle the fence on the issue), he might select some such value as **d** = .5 (p. 25) [boldface type in original].

Cohen proposes these categories "*as a convention* [italics in original]," noting that offering these definitions

> is fraught with dangers: The definitions are arbitrary, such qualitative concepts as "large" are sometimes understood as absolute, sometimes as relative; and thus they run a risk of being misunderstood (p. 12) ...

> This risk is nevertheless accepted in the belief that more is to be gained than lost by supplying a common conventional frame of reference which is recommended for use *only when no better basis for estimating the ES* [effect size] *is available.* (p. 25, italics added)

As the context makes clear, Cohen intends that the "small-medium-large" characterization of *d* be used as an guesstimation tool to plan the sample sizes for an experiment intended to detect the presence of an effect (say, of a treatment intervention), not as a way to judge the accuracy of detection systems. Many investigators have done just this, however:

- In discussing accuracy statistics, Dolan and Doyle (2000) state, "In general, Cohen's $d > 0.50$ or ROC-AUCs > 0.75 are considered large effect sizes" (pp. 303–304).
- Douglas and Reeves (2010) state, "Although there are no formal categories, AUC values of approximately .65 to .70 may be considered moderate to large, and approximately .70 and above may be considered large" (p. 165).

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

- In discussing the relationship between *d* and AUC, Rice and Harris (2005) comment, "Cohen stated that the values of *d* for small, medium, and large effects, respectively, are .2, .5, and .8" (p. 617). Though Rice and Harris note that Cohen "provided this rule of thumb tentatively" (p. 618), they later say that *d* values from ARAI studies had exceeded 0.8, levels that Cohen had characterized as "about as high as they come" in applied psychology (Rice & Harris, 2005 p. 619, quoting Cohen, 1988, p. 81).
- Anderson and Hanson (2010) state, "Static-99 is as accurate as any of the other available measures for the prediction of sexual recidivism . . . By conventional standards [referring to Rice & Harris, 2005], the effect size is medium to large" (p. 259).

The problem with the "small-medium-large" categorization is that it may not correlate well with practical usefulness. To see why, we consider two examples.

The Gambler

An American roulette wheel has 38 slots, 16 of which are red, 16 are black, and two are green. A roulette player doubles his money every time he bets red or black correctly. A gambler develops a method for predicting when the roulette ball will land in a red slot. The method has a sensitivity of 0.58 and a specificity of 0.58, so its effect size is just over 0.4. The gambler decides he will place $100 bets for 4 hours each night, betting red when the prediction is "red" and black when the prediction is "not red." On a typical evening, the gambler can make 30 bets an hour.

Does the gambler's prediction method have a small, medium, or large effect? To answer this question, we calculate what winnings the gambler can expect each night. First, consider 3,800 spins of the roulette wheel: we expect 1,600 "red" outcomes, 1,600 "blue" outcomes, and 200 "green" outcomes. Now, the gambler can expect to win on 58% of the "red" and "black" outcomes and to lose on all the "green" outcomes, so overall he expects to win on 2,088 (55%) of the 3,800 spins. Next, consider a 4-hour, 120-spin gambling session. The gambler expects to win 55% of the time—that is, on average, he will make 66 winning and 54 losing bets each gambling session, giving him an expectation of making $1,200 a night. Our gambler may not get rich at this rate, but most readers would regard his expected winnings as a "large" sum for 4 hours of work (often with free snacks and drinks!).

In patient Violence

Barzman et al. (2011) recently described preliminary results for their Brief Rating of Aggression by Children and Adolescents (BRACHA), an ARAI for aggressive behavior by children and adolescents who are about to undergo psychiatric hospitalization. In one analysis, Barzman et al. found that their ARAI achieved an AUC of 0.82 in predicting violence toward others, which (using equation 5) implies an effect size $d_a = 1.3$. Suppose that this finding held up in a cross-validation study. Would the effect be large enough to make a difference in management?

Barzman et al. noted that in their study sample, the 6-day base rate for violence toward others was 15%. Using the cut-off that produces maximum diagnostic information (MDI) creates an operating point such that (fpr, tpr) = (0.33, 0.82). Using this operating point to dichotomize patients into higher- and lower-risk subgroups, and based on the formulae for PPV and NPV in Table 1, a clinician could separate patients into two groups in which the expected rates of violence were 31% versus 3.6%–an odds ratio of 12. But would

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

knowing that a patient fell into one or the other subgroup make a difference in the patient's clinical management? Put another way: should a clinician or hospital ignore a patient's violence potential and take no prevention or protection measures if the patient "only" has a 3.6% chance of assaulting someone else in the next week? Probably not.

An ARAI with a very large effect size might create a difference that mattered. For example, an ARAI with an effect size $d_a = 2.5$ (AUC = 0.96) would produce MDI at the operating point (fpr, tpr) = (0.058, 0.824), capable of separate a population with a violence base rate of 0.15 into subgroups of patients with quite high (72%) and quite low (1.2%) risks of violence. Of course, no ARAI achieves this sort of accuracy, and none is likely to do so. Moreover, on a 24-bed unit filled with low-risk patients whose average length of stay was 10 days, one would expect 11 violent patients among its 876 admissions each year—a number high enough that staff members would still take significant safety precautions and would need training to deal with the problem.

CONCLUSIONS

Receiver operating characteristic analysis was recognized two decades ago as "a fundamental tool in clinical medicine" (Zweig & Campbell, 1993). Reporting area under the ROC curve is the method of summarizing accuracy that has become "most common in the risk assessment field" (Douglas & Reeves, 2010, p. 164). For practitioners and other clinical "consumers" of scientific information, ROC graphs provide easily apprehended depictions of diagnostic performance. ROC methods offer evaluators several statistical approaches toward conceptualizing and characterizing the discrimination capabilities of diagnostic and detection systems. Several ROC accuracy indices provide succinct ways of conveying features of diagnostic performance. AUC and the effect size *d* are particularly popular because they offer global descriptions of discrimination power, have intuitive appeal, allow comparisons of different methods, and are useful summaries in meta-analytic studies.

The AUC and *d* indices have limitations, however. They do not tell us about the sensitivity–specificity trade-offs that are the crucial features of any imperfect detection system. Nor do they tell us anything about how to balance the false-positive and false-negative errors that occur when an imperfect prediction method is applied. Finally, AUC and *d* do not tell whether, in practical terms, the discrimination capacity of a diagnostic system is large enough to permit useful distinctions between violent and non-violent subject groups. Justifying choices or clinical practices requires a contextual investigation of outcomes, a process that takes us beyond simply stating global indices of accuracy. Deciding whether a diagnostic or prediction method is adequate for any particular application requires a careful consideration of the consequences of cut-off choices, the benefits of making correct judgments, and the types and the costs of errors.

REFERENCES

Anderson, D., & Hanson, R. K. (2010). Static-99: An actuarial tool to assess risk of sexual and violent recidivism among sexual offenders. In Otto R. K., & Douglas K. S. (Eds.), *Handbook of Violence Risk Assessment* (pp. 251–267). New York: Routledge/Taylor & Francis Group.

Barzman, D., Brackenbury, L., Sonnier, L., Schnell, B., Cassedy, A., Salisbury, Sorter M; & Mossman, D. (2011). Brief Rating of Aggression by Children and Adolescents (BRACHA): Development of a tool to assess risk of inpatients' aggressive behavior. *The Journal of the American Academy of Psychiatry and the Law*, *39*, 170–179.

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

Baxtrom v. Herold. (1966). 383 U.S. 107.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences 2nd ed.* Hillsdale, NJ: Lawrence Erlbaum Associates.
Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668–1674.
Dolan, M., & Doyle, M. (2000). Violence risk prediction. Clinical and actuarial measures and the role of the Psychopathy Checklist. *The British Journal of Psychiatry*, *177*, 303–311.
Dorfman, D. D., & Alf, E., Jr. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data, *Journal of Mathematical Psychology*, *6*, 487–496.
Douglas, K. S., & Reeves, K. A. (2010). Historical-Clinical-Risk Management-20 (HCR-20) violence risk assessment scheme: Rationale, application, and empirical overview. In Otto R. K., & Douglas K. S. (Eds.), *Handbook of violence risk assessment* (pp. 147–185). New York, NY: Routledge/Taylor & Francis.
Douglas, K. S., Ogloff, J. R. P., Nicholls, T. L., & Grant, I. (1999). Assessing risk for violence among psychiatric patients: The HCR-20 risk assessment scheme and the Psychopathy Checklist: Screening Version. *Journal of Consulting and Clinical Psychology*, *67*, 917–930.
Ennis, B. J., & Litwack, T. R. (1974). Psychiatry and the presumption of expertise: Flipping coins in the courtroom. *California Law Review*, *62*, 693–752.
Erdman, H. P., Greist, J. H., Gustafson, D. H., Taves, J. E., & Klein, M. H. (1987). Suicide risk prediction by computer interview: a prospective study. *The Journal of Clinical Psychiatry*, *48*, 464–467.
Faust, D., & Ahern, D. C. (2011). Clinical judgment and prediction. In Faust D. (Ed.), *Coping with psychiatric and psychological testimony* 6th ed. (pp. 147–208). New York: Oxford University Press.
Gardner, W., Lidz, C. W., Mulvey, E. P., & Shaw, E. C. (1996). A comparison of actuarial methods for identifying repetitively violent patients with mental illnesses. *Law and Human Behavior*, *20*, 35–48.
Goodenough, D. J., Rossmann, K., & Lusted, L. B. (1974). Radiographic applications of receiver operating characteristic (ROC) analysis. *Radiology*, *110*, 89–95.
Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law*, *2*, 293–323.
Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*, 19–30.
Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36.
Hanson, R. K., & Morton-Bourgon, K. E. (2009). The accuracy of recidivism risk assessments for sexual offenders: A meta-analysis of 118 prediction studies. *Psychological Assessment*, *21*, 1–21.
Hart, S. D., & Logan, C. (2011). Formulation of violence risk using evidence- based assessments: the structured professional judgment approach. In Sturmey P., & McMurran M. (Eds.), *Forensic case formulation* (pp. 83–106). Chichester, West Sussex, UK: John Wiley & Sons.
Kans. Stat. Annot. § 59-29a02. (2009).
Lunn, D., Spiegelhalter, D., Thomas, A., & Best, N. (2009). The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*, *28*, 3049–3067.
McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. *Psychological Bulletin*, *111*, 361–365.
McNeil, B. J., Varady, P. D., Burrows, B. A., & Adelstein, S. J. (1975). Measures of clinical efficacy—cost-effectiveness calculations in the diagnosis and treatment of hypertensive renovascular disease. *The New England Journal of Medicine*, *293*, 216–221.
Metz, C. E., & Kronman, H. B. (1980). Statistical significance tests for binormal ROC curves. *Journal of Mathematical Psychology 22*, 218–243.
Metz, C. E., Herman, B. A., & Shen, J. H. (1998). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, *17*, 1033–53.
Monahan, J. (1981). *The Clinical Prediction of Violent Behavior.* Rockville, MD: National Institute of Mental Health.
Monahan, J. (1984). The prediction of violent behavior: toward a second generation of theory and policy. *The American Journal of Psychiatry*, *141*, 1–11.
Monahan, J. (1997). Clinical and actuarial predictions of violence. In Faigman D., Kaye D., Saks M., & Sanders J. (Eds.), *Modern scientific evidence: The law of science and expert testimony* (Vol. 1 pp. 300–318). St. Paul, MN: West Group.
Monahan, J. (2002). Clinical and actuarial predictions of violence. B. Scientific status. In Faigman D. L., Kaye D. H., Saks M. J., & Sanders J. (Eds.), *Science and the law: Social and behavioral sciences issues* (pp. 90–112). St. Paul, MN: West Group.
Morse, S. J. (1978). Crazy behavior, morals, and science: An analysis of mental health law. *California Law Review*, *51*, 527–654.
Mossman, D. (1994a) Assessing predictions of violence: being accurate about accuracy. *Journal of Consulting and Clinical Psychology*, *62*, 783–792.
Mossman, D. (1994b). Further comments on portraying the accuracy of violence prediction. *Law and Human Behavior*, *18*, 587–593.
Mossman, D. (2000). Assessing the risk of violence – are "accurate" predictions useful? *The Journal of the American Academy of Psychiatry and the Law*, *28*, 272–281.

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl

Mossman, D. (2006). Critique of pure risk assessment or, Kant meets *Tarasoff*. *University of Cincinnati Law Review*, *75*, 523–609.
Mossman, D., & Somoza, E. (1989). Maximizing diagnostic information from the dexamethasone suppression test: an approach to criterion selection using receiver operating characteristic analysis. *Archives of General Psychiatry*, *44*, 653–660.
Murphy, J. M., Berwick, D. M., Weinstein, M. C., Borus, J. F., Budman, S. H., & Klerman, G. L. (1987). Performance of screening and diagnostic tests. Application of receiver operating characteristic analysis. *Archives of General Psychiatry*, *44*, 550–555.
Rice, M. E., & Harris, G. T. (1995). Violent recidivism: Assessing predictive validity. *Journal of Consulting and Clinical Psychology*, *63*, 737–748.
Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's *d*, and *r*. *Law and Human Behavior*, *29*, 615–620.
Ruscio, J. (2008). A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods*, *13*, 19–30.
Simpson, A. J., & Fitter, M. J. (1973). What is the best index of detectability? *Psychology Bulletin*, *80*, 481–488.
Singh, J. P., Grann, M., & Fazel, S. (2011). A comparative study of violence risk assessment tools: a systematic review and metaregression analysis of 68 studies involving 25,980 participants. *Clinical Psychology Review*, *31*, 499–513.
Somoza, E., & Mossman, D. (1991). ROC curves and the binormal assumption. *The Journal of Neuropsychiatry and Clinical Neurosciences*, *3*, 436–439.
Sreenivasan, S., Weinberger, L. E., Frances, A., & Cusworth-Walker, S. (2010). Alice in actuarial land: through the looking glass of changing Static-99 norms. *The Journal of the American Academy of Psychiatry and the Law*, *38*, 400–406.
Steadman, H., Mulvey, E., Monahan, J., Robbins, P., Appelbaum, P., Grisso, T., . . . Silver, E. (1998). Violence by people discharged from acute psychiatric inpatient facilities and by others in the same neighborhoods. *Archives of General Psychiatry*, *55*, 393–401.
Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Sciences in the Public Interest (Supplement to Psychological Science)*, *1*, 1–26.
Yang, M., Wong, S. C. P., & Coid, J. (2010). The efficacy of violence prediction: A meta-analytic comparison of nine risk assessment tools. *Psychological Bulletin*, *136*, 740–767.
Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*, 561–577.

Copyright © 2013 John Wiley & Sons, Ltd.

DOI: 10.1002/bsl