Assessing Predictions of Violence: Being Accurate About Accuracy

Douglas Mossman

The prediction of violence occupies a prominent and controversial place in public mental health practice. Productive debate about the validity of violence predictions has been hampered by the use of methods for quantifying accuracy that do not control for base rates or biases in favor of certain outcomes. This article describes these problems and shows how receiver-operating characteristic analysis can be used to solve them. The article also reanalyzes 58 data sets from 44 published studies of violence prediction. Taken together, these data strongly suggest that mental health professionals’ violence predictions are substantially more accurate than chance. Short-term (1–7 day) clinical predictions seem no more accurate than long-term (>1 year) predictions. Past behavior alone appears to be a better long-term predictor of future behavior than clinical judgments and may also be a better indicator than cross-validated actuarial techniques.

The prediction of violence remains a prominent feature of public mental health practice (Lidz, Mulvey, & Gardner, 1993; Monahan, 1988; Monahan & Shah, 1989; Mulvey & Lidz, 1985; Poythress, 1990; Shah, 1978) despite a well-documented distrust of clinicians’ ability to make accurate predictions (Perlin, 1989). Scholars typically support this distrust by citing findings and subsequent critical appraisals of studies published in the 1970s and early 1980s, which suggested that clinicians’ predictions of future violence—particularly long-term predictions about outpatients—were usually incorrect (Ennis & Litwack, 1976; Faust & Ziskin, 1988; Monahan, 1978; Perlin, 1989). More recent studies have focused on predictions about subjects, followed for relatively short periods, whose actions can be closely observed (e.g., patients in hospitals). Monahan suggested that such studies represent “a second generation of thought on violence prediction” (Monahan, 1984, p. 11). They have been interpreted as indicating that clinicians’ judgments about dangerousness, particularly regarding issues related to civil commitment, have “a high degree of predictive validity” (McNeil & Binder, 1987, p. 197).

Investigators who want to evaluate the accuracy of violence predictions encounter several nettlesome methodologic issues (Monahan, 1981; Otto, 1992). They must decide which of an infinite variety of behaviors—including vehement gestures, verbal threats, damage to property, hospitalization, and drunk driving—“count” as examples of episodes of violence. They also must contend with the difficulties in ascertaining whether violent acts have occurred. For nonhospitalized subjects, arrest data often are used to tally violent acts, although arrest records document only a fraction of all violent actions actually committed, relatively overreport violence by individuals with significant psychopathology (Teplin, 1984), and relatively underreport crimes against younger victims and offenses that do not involve injury (Bureau of Justice Statistics, 1985). Although data gathered on hospitalized patients record a larger fraction of violent episodes than that which reaches the attention of law enforcement authorities (cf. Monahan, 1981, as cited with McNeil & Binder, 1987), such data may be affected by reporting biases, other patients’ psychopathology, or the ward milieu of a given hospital.

To be meaningful, results from studies of violence prediction should be reported with appropriate indices of accuracy—they should, that is, describe accuracy accurately. Unfortunately, many commonly used accuracy measures can lead to misunderstandings or misleading characterizations of clinicians’ ability to predict violence. In many studies, prediction accuracy is quantified as the fraction of correct predictions. However, if only 5% of the subjects are violent, a clinician who always predicts “no violence” would be right 95% of the time, whereas a clinician who errs on the side of caution and predicts violence 20% of the time can be correct about no more than 85% of the subjects. The fraction of correct predictions is one of many accuracy indices that fail to separate intrinsic ability to detect violent subjects from the effects of base rates and clinicians’ preferences about avoiding certain types of prediction errors.

This article recommends using receiver operating characteristic (ROC) analysis to evaluate attempts to detect or predict violence. ROC methods describe accuracy with indices of performance that are unaffected by base rates or by clinicians’ biases for or against Type I or Type II prediction errors. ROC methods “occupy a central or unifying position in the process of assessing and using diagnostic tools” in clinical medicine (Zweig & Campbell, 1993, p. 561) and have been used in criminological studies to evaluate predictions about juvenile delinquency (Fergusson, Fifield, & Slater, 1977) and prison furloughs (Serin & Lawson, 1987). The following section explains how ROC analysis can help investigators describe and evaluate violence predictions. The subsequent sections reevaluate previously published data and show how ROC methods can help in addressing several important questions: Are long-term predictions accurate? Are short-term predictions more accurate? What are the best methods for predicting violence?
Background: Quantifying Accuracy

As an illustration of procedures for quantifying clinicians’ ability to distinguish violent from nonviolent subjects, the following imaginary study is reviewed (cf. Otto, 1992, p. 108, n. 5).

Clinicians A. Good Short, Polly Spower, and Libby Tarian all evaluated patients presenting to an urban psychiatric emergency service. Polly Spower and Libby Tarian rated patients on a 5-point scale that reflected their opinions about each patient’s chance of acting violently during the week after evaluation. A rating of 1 indicated that the patient was judged very unlikely to be violent; a rating of 5 meant that the patient was judged very likely to be violent. The clinicians also made recommendations about which patients should be hospitalized because of their potential to act violently. A. Good Short also made recommendations about hospitalization, but did not assign ratings to patients.

The actual decisions about hospitalization were made by Kent B. Sued, who ignored the clinicians’ recommendations and released all patients. They were followed for 7 days by Tom Niscent, who observed all their acts over this time period. Tom Niscent reported patients’ behavior to Goldie Standard, who unambiguously determined which patients acted violently. This process resulted in a methodologically ideal (if ethically and legally questionable) study with data that were unaffected by the influence of treatment decisions on behavior, limited knowledge of actual behavior, or ambiguity in the definition of violence (Monahan, 1981; Monahan, 1988; Otto, 1992).

After several months, all three clinicians each had evaluated 1,000 patients, 10% of whom were violent. How should we assess the clinicians’ prediction accuracy?

One commonly used approach (see, e.g., Rofman, Askinazi, & Fant, 1980) assumes that clinicians’ recommendations about hospitalization or release represent judgments about the likelihood of future violence. If a clinician recommends hospitalization and the patient subsequently acts violently, the clinician’s recommendation is termed a true positive (TP) prediction. A false negative (FN) prediction is one in which the clinician did not recommend hospitalization for a patient who actually was violent. True negative (TN) and false positive (FP) predictions are defined similarly in the appendix. One can use a clinicians’ recommendations and the patients’ actual behavior to calculate what fraction of predictions were correct in light of subsequent events. One also can calculate the ratio of FP to TP predictions to find the odds that a prediction of violence was incorrect.

Imagine the clinicians’ discussion of their results, which are shown in Table 1. Polly Spower was right 54% of the time; A. Good Short, 86% of the time; and Libby Tarian, 89% of the time; however, Tarian and Short both recommended release of about half of the patients who became violent, whereas Spower only missed 7% of the violent patients. Polly Spower prefers her approach to clinical decision making because she feels that clinicians have a responsibility to protect the community, but she made almost five wrong predictions of violence for every correct one, whereas Libby Tarian’s FP:TP ratio was only 1.2. Tarian feels that her performance was superior because she places a high value on patients’ freedom. A. Good Short, whose FP:TP ratio was 3.3 and who was right almost as often as Libby Tarian, feels that his approach offers a good compromise.

All three clinicians performed significantly better than chance at sorting patients, but the correct fraction index and the FP:TP ratio obscure this. A clinician who had just recommended release for everybody would have had a correct fraction of 90%. If half of the patients had been violent, a clinician who randomly recommended admission for half the patients would have a FP:TP ratio of about 1, despite a performance that was no better than chance.

The appendix describes several accuracy indices that allow investigators to measure association in 2 X 2 contingency tables in ways that avoid confusing accuracy with the effects of base rates (Kraemer, 1985; Loeber & Dishion, 1983; Somoza & Mossman, 1990). In medical literature, sensitivity and specificity are frequently used to describe performance of diagnostic tests. If we interpret the clinicians’ recommendations as diagnostic tests of patients’ future violence, then sensitivity is the likelihood that a recommendation for hospitalization was made for a patient who is actually violent and specificity is the likelihood that a recommendation for release was made for a nonviolent patient.

Tables 1 and 2 show the sensitivities and specificities for the three clinicians. From these numbers alone, it is difficult to determine which (if any) predicted best or worst. Table 1 also shows each clinician’s relative improvement over chance (RIOC; Loeber & Dishion, 1983). The calculation of RIOC is explained in the appendix; it is one of several ways to summarize with a single index the degree to which values in a 2 x 2 table deviate from chance assignment (see Kraemer 1985, 1987). Because of the low base rate and her policy of erring on the side of caution, Polly Spower has a higher RIOC than Libby Tarian or A. Good Short, although she made more erroneous recommendations. Polly Spower, however, was no better at classifying than was Libby Tarian; indeed, Table 2 shows that Spower and Tarian performed identically. Spower simply preferred to minimize false-negative outcomes and avoided releasing violent patients, whereas Tarian preferred to minimize false-positive outcomes and recommended hospitalization only when she had a very strong suspicion that a patient would become violent.

These observations suggest that diagnostic accuracy should be measured with techniques that are not affected by base rates or clinicians’ preferences for certain outcomes (Swets, 1979). Ideally, accuracy should be described in a way that reflects the trade-offs between sensitivity and specificity and that is independent of a clinician’s actual cutoff or decision threshold. Table 2 allows one to calculate four sensitivity–specificity pairs using the divisions between the clinicians’ five rating categories as potential thresholds for hospitalization. At Libby Tarian’s strict threshold, hospitalization is recommended only for those patients rated >4. Sensitivity is 50%, and specificity is 93%. Put another way, the strictest threshold is associated with a violence detection rate, or a TP rate (TPR), of only 0.50, but the false alarm rate, or FP rate (FPR), is only 0.07. (Note that TPR = sensitivity and FPR = [1 – specificity].) At the second strictest threshold, hospitalization is recommended for patients rated >3, and those in categories 1, 2, and 3 are released. FPR increases to 0.16, and TPR increases to 0.69. We can calculate FPR and TPR for the other thresholds (see Table 2) in a similar way.

Signal detection theory (Swets & Pickett, 1982) provides a
### Table 1

#### Hospitalization Recommendations as Predictors of Future Violence

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Polly Spower</th>
<th>Libby Tarian</th>
<th>A. Good Short</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violent patient</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admit</td>
<td>93</td>
<td>50</td>
<td>55</td>
</tr>
<tr>
<td>Release</td>
<td>7</td>
<td>50</td>
<td>45</td>
</tr>
<tr>
<td>Nonviolent patient</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Admit</td>
<td>450</td>
<td>60</td>
<td>181</td>
</tr>
<tr>
<td>Release</td>
<td>450</td>
<td>840</td>
<td>719</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>0.93</td>
<td>0.50</td>
<td>0.55</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.50</td>
<td>0.93</td>
<td>0.80</td>
</tr>
<tr>
<td>Fraction correct</td>
<td>0.54</td>
<td>0.89</td>
<td>0.86</td>
</tr>
<tr>
<td>FP:TP ratio</td>
<td>4.8</td>
<td>1.2</td>
<td>3.3</td>
</tr>
<tr>
<td>Relative improvement over chance</td>
<td>0.85</td>
<td>0.44</td>
<td>0.41</td>
</tr>
<tr>
<td>Area under the ROC curve</td>
<td>0.856</td>
<td>0.856</td>
<td>0.752</td>
</tr>
</tbody>
</table>

**Note.** FP = false positive; TP = true positive; RIOC = relative improvement over chance; AUC = area under the ROC [receiver operating characteristic] curve.

A convenient way of depicting detection and false alarm rates at different decision thresholds. For each threshold, we can graph TPR as a function of FPR and thus plot each clinician's ROC. This term (originally derived from radar applications [Lusted, 1984]) suggests detection that is characterized by the threshold at which the receiver (e.g., a clinician) operates.

In Figure 1, the four thresholds of Polly Spower and Libby Tarian are represented by discrete points in the ROC square. The solid upper curve connecting these points was drawn using the "binormal assumption" of ROC analysis (Dorfman & Alf, 1969; Hanley, 1988; Metz, 1986; Mossman & Somoza, 1989; Somoza & Mossman, 1991; Swets, 1986), namely, observed frequency data arise from two latent distributions which, on some monotonic transformation of the decision scale, approximate two Gaussian (bell-shaped or "normal") distributions. This implies that the normal deviates, or Z-transforms, of the FPR and TPR can be fit to a straight line,

\[ Z_{TPR} = A + B \cdot Z_{FPR}, \]

where A is the populations' separation along the transformed decision scale in units of the standard deviation (SD) of the violent population, and B is the ratio of SD_{nonviolent} to SD_{violent}. For the upper curve, A = 1.5 and B = 1.0. A and B were calculated from the categorized data in Table 2 with maximum likelihood estimation software designed by A. Dorfman (Dorfman & Alf, 1969) and Metz (1986). (Somewhat less accurate estimates of A and B can be obtained by fitting Z_{TPR} and Z_{FPR} to a straight line using the least squares method.)

ROC curves depict the clinicians' ability to differentiate violent from nonviolent subjects and also allow comparison of different systems or techniques for detecting violent patients. For example, the dashed curve in Figure 1 represents poorer discrimination than that represented by the solid curve because at any given FPR, TPR is greater for the solid curve than for the dashed curve. As a consequence, the area under the upper curve is larger than the area under the lower curve. Indeed, the area under the ROC curve (AUC) is a succinct and commonly used method for summarizing overall discriminating power. AUC equals the likelihood that a clinician would rate a randomly selected, actually violent person as more likely to be violent than a randomly selected, actually nonviolent person (Hanley & McNeil, 1982). Clinicians who could distinguish violent from nonviolent patients with nearly perfect accuracy would have ROC AUCs approaching 1.0, and those who performed no better than chance would obtain AUCs of 0.5. Z_{AUC}, the normal deviate of AUC, is related to the binormal ROC indices A and B as follows.

\[ Z_{AUC} = \frac{A}{\sqrt{1 + B^2}}. \]

Application of Metz and his colleagues' software to the rankings of Polly Spower and Libby Tarian shows that both clinicians achieved AUCs of 0.856. The lower curve in Figure 1 has an AUC of 0.752. It passes through the point in the ROC square representing the sensitivity and specificity for A. Good Short's hospitalization recommendations (see Table 1). This point falls below the upper
Figure 1. Receiver-operating characteristic curves for hypothetical studies of violence prediction accuracy, based on data shown in Table 2. Rectangles indicate clinicians' potential decision thresholds; solid curve depicts the discriminatory ability of imaginary clinicians Polly Spower (P.S.) and Libby Tarian (L.T.); dashed curve depicts accuracy inferred from recommendations by A. Good Short (A.G.S.) for hospitalization.

curve, implying that Short was not as good as Spower and Tarian at discriminating violent from nonviolent subjects. Although Short only made recommendations for hospitalization or release, it can be assumed reasonably that he could have given his patients ratings, one point along which would constitute his threshold for recommending hospitalization. An approximate ROC curve can be drawn for A. Good Short using a technique suggested by Weinstein (1980), in which a single sensitivity–specificity pair locates a ROC curve that is symmetric about the negative diagonal of the ROC square.

Method

The previous section's considerations suggest that published data about violence prediction could be reevaluated to determine whether ROC methods can help develop new perspectives on how accurate such predictions were. The reports evaluated in this article include studies that were familiar to me from prior research on the subject, supplemented by studies found through a search of the titles and abstracts in MEDLINE and PsycLIT databases from 1972 forward using the keyword conjunctions “predict + violence” and “predict + dangerous.” These studies contained either categorized data (i.e., data similar in form to that shown for Polly Spower and Libby Tarian in Table 2) or data that permitted calculation of a single sensitivity–specificity pair (i.e., data similar to that given for A. Good Short, representing a single point on a ROC curve). This procedure yielded 44 studies with data suitable for analysis, which I regard as a representative sample of the different types of violence prediction studies published over the past 2 decades. Although the sample is not intended to be exhaustive, I believe that it represents the majority of data on predicting violence obtainable from the past 2 decades' medical and psychological publications.

In many instances, published data required interpretation to be construed as indicative of prediction accuracy. The following comments illustrate typical procedures for evaluating data.

1. In their study of rates of arrest for violence among patients discharged from New York State Hospitals, Cocozza, Mellick, and Steadman (1978) showed that the number of previous arrests was correlated with the rate of postdischarge arrests. With some of their tabulated data, one can use arrest rates to calculate actual numbers of arrestees in each prior arrest category; these numbers can then be entered into Metz et al.'s ROCFIT software (Metz, 1986) to generate ROC indices of performance.

2. McNeil et al. (1988) developed a discriminant function that sorted patients into three categories of likelihood for violence while hospitalized; in another report (McNeil & Binder, 1991) they used clinicians' ratings to obtain similar rating categories. Both data sets allow one to calculate how many patients at each level did or did not engage in violence; this information can be used by the ROCFIT program to produce ROC indices of the categorization's accuracy in distinguishing violent patients.

3. The Michigan Department of Corrections (1978; see also Monahan, 1981) developed a method for categorizing parolees' likelihood of recidivism. Their data permitted estimation of actual numbers of parolees who fell into each likelihood category. Again, these numbers were suitable for entry into the ROCFIT software to obtain estimates of the department's prediction accuracy.

4. The 44 studies contain data on a large variety of approaches to predicting violence. In Table 3, the approaches are characterized as falling into one of four broad types. Clinical judgment studies contain data generated by clinicians' explicit predictions or their decisions. In several cases (e.g., Kozol, Boucher, & Garofalo, 1972; Steadman, 1977; Zeiss, Fenn, Tanke, & Yesavage, 1990), clinicians' recommendations about hospitalization or other types of disposition were interpreted as predictive acts; that is, as yes-or-no predictions about future violence or arrest that formed the basis for sensitivity–specificity descriptions of prediction accuracy. A clinician's reason for authorizing involuntary hospitalization—whether a patient was deemed dangerous to others—can be treated similarly (see, e.g., McNeil & Binder, 1987). Past behavior was presented in several studies in a manner that allowed it to be interpreted as a predictive device. For example, Steadman and Cocozza (1978) present data showing that past arrests for violent offenses were good predictors of subsequent arrests for violent offenses. Several studies presented results from rating systems or discriminant functions used to rank patients who acted violently. In some cases, accuracy data was presented for discriminant functions that simply were fitted retrospectively to data. In other studies, data were presented describing the accuracy of scales or functions that were prospectively validated on subjects distinct from the group used to derive the prediction formulae.

5. Several studies presented multiple sets of data regarding prediction accuracy. In some studies, the same subjects' behavior was evaluated using different definitions of violence (e.g.,
Table 3
Characteristics of Selected Studies of Violence Prediction

<table>
<thead>
<tr>
<th>Study</th>
<th>N</th>
<th>VF</th>
<th>Time period</th>
<th>Type</th>
<th>D/C</th>
<th>AUC ± SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black &amp; Spinks (1985)</td>
<td>125</td>
<td>0.488</td>
<td>5 years</td>
<td>r</td>
<td>c</td>
<td>0.947 ± 0.023</td>
</tr>
<tr>
<td>Blomhoff, Sein, &amp; Fris (1990)</td>
<td>59</td>
<td>0.424</td>
<td>38 days (avg)</td>
<td>b</td>
<td>o</td>
<td>0.738 ± 0.068</td>
</tr>
<tr>
<td>Cocozza &amp; Steadman (1974)</td>
<td>98</td>
<td>0.204</td>
<td>28.5 months (avg)</td>
<td>r</td>
<td>r</td>
<td>0.889 ± 0.050</td>
</tr>
<tr>
<td>Cocozza &amp; Steadman (1976)</td>
<td>257</td>
<td>0.148</td>
<td>3 years</td>
<td>j</td>
<td>c</td>
<td>0.483 ± 0.050</td>
</tr>
<tr>
<td>Cocozza, Melick, &amp; Steadman (1978), #1</td>
<td>1,920</td>
<td>0.009</td>
<td>19 months</td>
<td>b</td>
<td>c</td>
<td>0.847 ± 0.058</td>
</tr>
<tr>
<td>Cocozza et al. (1978), #2</td>
<td>1,938</td>
<td>0.017</td>
<td>15 months</td>
<td>b</td>
<td>c</td>
<td>0.933 ± 0.039</td>
</tr>
<tr>
<td>Convit, Jaeger, Lin, Meisner, &amp; Volavka (1988), #1</td>
<td>51</td>
<td>0.608</td>
<td>3 months</td>
<td>r</td>
<td>o</td>
<td>0.759 ± 0.069</td>
</tr>
<tr>
<td>Convit et al. (1988), #2</td>
<td>79</td>
<td>0.251</td>
<td>3 months</td>
<td>v</td>
<td>o</td>
<td>0.684 ± 0.074</td>
</tr>
<tr>
<td>Copas &amp; Whiteley (1976), #1</td>
<td>104</td>
<td>0.423</td>
<td>2–3 years</td>
<td>r</td>
<td>h,c</td>
<td>0.809 ± 0.044</td>
</tr>
<tr>
<td>Copas &amp; Whiteley (1976), #2</td>
<td>87</td>
<td>0.471</td>
<td>3 years</td>
<td>v</td>
<td>c</td>
<td>0.664 ± 0.061</td>
</tr>
<tr>
<td>Farrington (1989), #1</td>
<td>411</td>
<td>0.122</td>
<td>23 years</td>
<td>b</td>
<td>c</td>
<td>0.803 ± 0.039</td>
</tr>
<tr>
<td>Farrington (1989), #2</td>
<td>411</td>
<td>0.122</td>
<td>19 years</td>
<td>b</td>
<td>c</td>
<td>0.904 ± 0.029</td>
</tr>
<tr>
<td>Farrington (1989), #3</td>
<td>411</td>
<td>0.122</td>
<td>15 years</td>
<td>b</td>
<td>c</td>
<td>0.885 ± 0.031</td>
</tr>
<tr>
<td>Hedlund, Sletten, Altman, &amp; Evenson (1973)</td>
<td>2,762</td>
<td>0.097</td>
<td>1 month (avg)</td>
<td>v</td>
<td>o</td>
<td>0.757 ± 0.018</td>
</tr>
<tr>
<td>Hoffman &amp; Beck (1985)</td>
<td>1,806</td>
<td>0.262</td>
<td>5 years</td>
<td>v</td>
<td>c</td>
<td>0.695 ± 0.014</td>
</tr>
<tr>
<td>Holland, Holt, &amp; Beckett (1982), #1</td>
<td>198</td>
<td>0.111</td>
<td>32 months</td>
<td>b</td>
<td>c</td>
<td>0.536 ± 0.062</td>
</tr>
<tr>
<td>Holland et al. (1982), #2</td>
<td>198</td>
<td>0.111</td>
<td>32 months</td>
<td>b</td>
<td>c</td>
<td>0.614 ± 0.060</td>
</tr>
<tr>
<td>Janofsky, Spears, &amp; Neubauer (1988), #1</td>
<td>47</td>
<td>0.149</td>
<td>7 days</td>
<td>j</td>
<td>o</td>
<td>0.737 ± 0.114</td>
</tr>
<tr>
<td>Janofsky et al. (1988), #2</td>
<td>47</td>
<td>0.149</td>
<td>7 days</td>
<td>b</td>
<td>o</td>
<td>0.896 ± 0.082</td>
</tr>
<tr>
<td>Jones, Beideman, &amp; Fowler (1981)</td>
<td>141</td>
<td>0.340</td>
<td>2 years</td>
<td>r</td>
<td>c</td>
<td>0.851 ± 0.037</td>
</tr>
<tr>
<td>Kandel, Brennan, Mednick, &amp; Michelson</td>
<td>265</td>
<td>0.072</td>
<td>9 years</td>
<td>v</td>
<td>c</td>
<td>0.672 ± 0.070</td>
</tr>
<tr>
<td>Kerk (1989)</td>
<td>68</td>
<td>0.176</td>
<td>3 days</td>
<td>j</td>
<td>o</td>
<td>0.684 ± 0.091</td>
</tr>
<tr>
<td>Klasson &amp; O'Connor (1988a), #1</td>
<td>109</td>
<td>0.321</td>
<td>3 months</td>
<td>h,c</td>
<td>0.943 ± 0.028</td>
<td></td>
</tr>
<tr>
<td>Klasson &amp; O'Connor (1988a), #2</td>
<td>127</td>
<td>0.276</td>
<td>3 months</td>
<td>r</td>
<td>c</td>
<td>0.976 ± 0.019</td>
</tr>
<tr>
<td>Klasson &amp; O'Connor (1988b)</td>
<td>239</td>
<td>0.192</td>
<td>6 months</td>
<td>r</td>
<td>h,c</td>
<td>0.906 ± 0.030</td>
</tr>
<tr>
<td>Klasson &amp; O'Connor (1989)</td>
<td>265</td>
<td>0.253</td>
<td>3 months</td>
<td>h,c</td>
<td>0.763 ± 0.037</td>
<td></td>
</tr>
<tr>
<td>Kozol, Boucher, &amp; Garofalo (1972)</td>
<td>435</td>
<td>0.110</td>
<td>5 years</td>
<td>j</td>
<td>c</td>
<td>0.763 ± 0.042</td>
</tr>
<tr>
<td>Levinson &amp; Ramsay (1979)</td>
<td>53</td>
<td>0.245</td>
<td>3 months</td>
<td>j</td>
<td>r</td>
<td>0.565 ± 0.094</td>
</tr>
<tr>
<td>Lidz, Mulvey, &amp; Gardner (1993)</td>
<td>730</td>
<td>0.359</td>
<td>6 months</td>
<td>j</td>
<td>o</td>
<td>0.661 ± 0.022</td>
</tr>
<tr>
<td>McNeil &amp; Binder (1987)</td>
<td>101</td>
<td>0.168</td>
<td>3 days</td>
<td>j</td>
<td>o</td>
<td>0.752 ± 0.072</td>
</tr>
<tr>
<td>McNeil &amp; Binder (1991)</td>
<td>149</td>
<td>0.174</td>
<td>7 days</td>
<td>j</td>
<td>o</td>
<td>0.713 ± 0.063</td>
</tr>
<tr>
<td>McNeil, Binder, &amp; Greenfield (1988), #1</td>
<td>238</td>
<td>0.181</td>
<td>3 days</td>
<td>b</td>
<td>o</td>
<td>0.682 ± 0.044</td>
</tr>
<tr>
<td>McNeil et al. (1988), #2</td>
<td>238</td>
<td>0.181</td>
<td>3 days</td>
<td>r</td>
<td>o</td>
<td>0.612 ± 0.062</td>
</tr>
<tr>
<td>Menzies, Walker, &amp; Sepejik (1985)</td>
<td>203</td>
<td>0.286</td>
<td>24 months</td>
<td>r</td>
<td>r</td>
<td>0.670 ± 0.047</td>
</tr>
<tr>
<td>Michigan Department of Corrections (1978)</td>
<td>1,100</td>
<td>0.102</td>
<td>14 months</td>
<td>v</td>
<td>c</td>
<td>0.735 ± 0.026</td>
</tr>
<tr>
<td>Mullen &amp; Reininger (1982)</td>
<td>165</td>
<td>0.091</td>
<td>2 years (avg)</td>
<td>j</td>
<td>c</td>
<td>0.577 ± 0.080</td>
</tr>
<tr>
<td>Panek &amp; Wagner (1989)</td>
<td>36</td>
<td>0.667</td>
<td>5 years</td>
<td>v</td>
<td>o</td>
<td>0.877 ± 0.057</td>
</tr>
<tr>
<td>Panek, Wagner, &amp; Suen (1979)</td>
<td>78</td>
<td>0.513</td>
<td>3 months</td>
<td>v</td>
<td>o</td>
<td>0.725 ± 0.057</td>
</tr>
<tr>
<td>Payne, McCabe, &amp; Walker (1974)</td>
<td>334</td>
<td>0.353</td>
<td>2 years</td>
<td>r</td>
<td>c</td>
<td>0.721 ± 0.030</td>
</tr>
<tr>
<td>Phillips &amp; Naar (1983)</td>
<td>63</td>
<td>0.476</td>
<td>Several weeks/months</td>
<td>j</td>
<td>o</td>
<td>0.795 ± 0.058</td>
</tr>
<tr>
<td>Roffman, Askinsazi, &amp; Fant (1980)</td>
<td>118</td>
<td>0.492</td>
<td>45 days</td>
<td>o</td>
<td>o</td>
<td>0.802 ± 0.041</td>
</tr>
<tr>
<td>Selby (1984)</td>
<td>100</td>
<td>0.500</td>
<td>Several years</td>
<td>r</td>
<td>o</td>
<td>0.961 ± 0.018</td>
</tr>
<tr>
<td>Sepejik, Menzies, Webster, &amp; Jensen (1983), #1</td>
<td>360</td>
<td>0.467</td>
<td>2 years</td>
<td>b</td>
<td>r</td>
<td>0.591 ± 0.030</td>
</tr>
<tr>
<td>Sepejik et al. (1983), #2</td>
<td>364</td>
<td>0.464</td>
<td>2 years</td>
<td>j</td>
<td>r</td>
<td>0.636 ± 0.029</td>
</tr>
<tr>
<td>Steadman &amp; Cocozza (1978)</td>
<td>166</td>
<td>0.143</td>
<td>2–3 years</td>
<td>b</td>
<td>c</td>
<td>0.755 ± 0.061</td>
</tr>
<tr>
<td>Steadman &amp; Morrissey (1982), #1</td>
<td>257</td>
<td>0.391</td>
<td>Several months-years</td>
<td>r</td>
<td>o</td>
<td>0.633 ± 0.036</td>
</tr>
<tr>
<td>Steadman &amp; Morrissey (1982), #2</td>
<td>282</td>
<td>0.301</td>
<td>Several months-years</td>
<td>v</td>
<td>o</td>
<td>0.566 ± 0.038</td>
</tr>
<tr>
<td>Steadman &amp; Morrissey #3</td>
<td>250</td>
<td>0.075</td>
<td>Several months-years</td>
<td>v</td>
<td>o</td>
<td>0.591 ± 0.094</td>
</tr>
<tr>
<td>Steadman (1977)</td>
<td>152</td>
<td>0.342</td>
<td>3 years</td>
<td>j</td>
<td>c</td>
<td>0.575 ± 0.050</td>
</tr>
<tr>
<td>Swett &amp; Hartz (1984), #1</td>
<td>199</td>
<td>0.211</td>
<td>23 days (median)</td>
<td>v</td>
<td>o</td>
<td>0.641 ± 0.048</td>
</tr>
<tr>
<td>Swett &amp; Hartz (1984), #2</td>
<td>201</td>
<td>0.224</td>
<td>23 days (median)</td>
<td>v</td>
<td>o</td>
<td>0.646 ± 0.052</td>
</tr>
<tr>
<td>Viskunen, De Jong, Bartko, Goodwin, &amp; Linnoila (1989)</td>
<td>57</td>
<td>0.228</td>
<td>36 ± 18 months</td>
<td>v</td>
<td>c</td>
<td>0.864 ± 0.052</td>
</tr>
</tbody>
</table>

Note. Webster et al. (1984) = Webster, Sepejik, Menzies, Slomen, Jensen, & Butler (1984). VF = fraction of subjects who were violent. D/C = definition or criterion for deeming subject violent: c = committed crime; h = hospitalized; o = observations in hospital or other facility; r = review of reports or records. AUC = area under the ROC (receiver-operating characteristic) curve. Avg = average.

* Types of prediction are classified as follows: j = clinical judgment; b = past behavior; r = discriminant function, retrospective fit; v = discriminant function, prospectively validated.
threats, actual attacks, or both), or several different prediction
techniques (e.g., clinicians' impressions, preadmission behav-
ior, past history, or discriminant functions). Some studies con-
tained predictions on more than one group of patients (e.g., a
retrospective fit of a discriminant function and a validation test
of that function's discriminating power [see, e.g., McNeil et al.,
1988]). For several studies, therefore, one could calculate sev-
eral AUCs to describe accuracy for different definitions or pre-
diction techniques. Although it may be misleading to choose
any one estimate as representative for a multiple-result study, it
also could be misleading to count multiple-result studies more
than once. As this article's intent is illustrative rather than rig-
gorously meta-analytical, those cases where it seemed instructive
to present more than one data set per study are represented by
numbered data sets, the characteristics of which are summa-
rized in Table 3.

6. Some reports (Black & Spinks, 1985; Cocozza et al.,
1978; Copas & Whiteley, 1976; McNeil & Binder, 1991; McNeil
et al., 1988; Michigan Department of Corrections, 1978; Payne,
McCabe, & Walker, 1974; Virkkonen, DeJong, Bartko, Good-
win, & Linnola, 1989) contained data that represented mul-
tiple points in a ROC square and that could be analyzed
directly by Metz and his colleagues' software to obtain estimates
of AUC ± SE. In the majority of studies, however, the data al-
lowed calculation of only a single specificity–specificity pair
representing only one point in the ROC square. There is no
to ignore the majority of published information about vio-
ience prediction accuracy. For this article, therefore, when data
allowed calculation of only one specificity–specificity pair, the
AUC was estimated by assuming that the single pair located
ROC curves that were symmetric about the negative diagonal,
(i.e., that B = 1 in Equation 1).² The standard errors of AUC for
these single-point ROC curves were calculated using the
method described by Hanley and McNeil (1982).

Unlike the imaginary study described in the previous section,
no real publications used the services of omniscient colleagues
to determine whether violent acts actually occurred. Investiga-
tors instead used a variety of definitions or criteria for deeming
someone violent, such as information from arrest records or
hospital incident reports. Despite investigators' best efforts,
these criteria cannot eliminate statistical uncertainty from ac-
curacy quantification. It was impossible to determine what
effect this uncertainty had on accuracy indices, and this article's
analysis therefore assumes (as did the authors of the original
articles) that this uncertainty had a minimal or random impact
on the accuracy data. Readers who are interested in more so-
plicated statistical approaches to this issue may consult

Recently, McClish (1992) described a method for obtaining a
weighted average of ROC areas taken from different studies, and
a chi-square test to determine whether the average represents a
common AUC for the group of AUCs. McClish's techniques
were used to obtain average AUCs for several subgroupings of
the data sets.

Results

Table 3 contains the results from reevaluating 58 data sets
from 44 published studies of violence prediction. These studies
involved over 16,000 psychiatric patients, indictees, and parol-
ees. Table 3 summarizes each study's population size, observation
period, prediction type, fraction of subjects who were violent,
criterion for determining who was violent, and AUC. The ap-
proximate 95% confidence interval for each data set's AUC may
be obtained by evaluating AUC ± 1.96 (SE) using the rightmost
column; when the confidence interval lies entirely above 0.5
(the AUC associated with a nondiscriminating test), one may
assume that predictions had better-than-chance accuracy.

Because Table 3 uses a single scale to characterize accuracy,
it is tempting to combine and compare AUCs from various data
sets. Table 4 lists weighted averages for several different sub-
groups of studies, but readers should recognize that these aver-
ages are at best provisional. The averages in Table 4 were calcu-
lated from Table 3's representative, but not comprehensive,
sampling of the literature. For some studies where data were
analyzed several different ways, not all the findings are included,
and the data subsets shown in Table 3 are merely intended to be
illustrative. The data selected reflect results obtained by using a
very broad variety of study subjects, population sizes, clinical
situations, observation periods, base rates, and definitional cri-
teria for violence. Indeed, the studies' heterogeneity suggests
that a simple conclusion about the accuracy of clinicians' vio-
ience predictions cannot be rendered, and any conclusion using
the data in Tables 3 and 4 should be advanced tentatively. The
following observations are advanced with all these caveats in
mind, in the belief that they offer a useful (if imperfect) view of
clinical violence prediction that can, at the very least, stimulate
readers to reconsider how data on prediction accuracy can best
be formulated and interpreted.

1. For 47 of 58 data sets (81%), prediction accuracy was
significantly better than chance—in some cases, far better. The
median AUC for all 58 data sets is 0.73 and the weighted average
AUC ± SE is 0.7777 ± 0.0048, well above the 0.5 level indicat-
ing no-better-than-chance discrimination.

2. Those studies listed in Table 3 that were published after
1986 are most reflective of the methodological concerns voiced
in Monahan's call for a "second generation" of research on vio-
lence prediction (Monahan, 1984). The accuracies in these
more recent reports can be compared with those found in ear-
lier (pre-1986) studies. Twenty-seven of 36 data sets (75%) from
"first-generation" studies deal with long-term predictions cov-
ering a year or more, but 15 of 22 data sets (68%) from second-
generation studies deal with predictions covering less than a
year. Ten data sets from the older studies report no-better-than-
chance discrimination, but the more recent studies report bet-
ter-than-chance discrimination in all but one case. Overall, the
average AUC from the newer studies is clearly higher than the
average for the older studies, but this may only mean that au-
thors are less inclined to submit (or journals less inclined to
publish) results from negative studies. Moreover, both of these

² One could substitute other values for B if one knew these were more
appropriate. Letting B = 1 is a reasonable approach for the data evalu-
ated in this article because this is roughly the average value of B ob-
tained through reevaluation of those data on violence prediction that
were reported in a manner that allowed conventional ROC curve fitting;
that is, where authors published data from which one could calculate
two or more possible decision thresholds.
subgroups are highly heterogeneous, and better comparisons may come from examining smaller strata sharing common AUCs.

3. Four types of predictive acts or techniques are evaluated in these studies, and Table 4 suggests that the four techniques are not equally accurate. Not surprisingly, the average accuracy of retrospectively fitted equations are higher than for the other three methods. For four of the studies of retrospective fits, the discriminant equations were retested in samples different from those used to derive the equations (Convit, Jaeger, Lin, Meisner, & Volavka, 1988; Copas & Whiteley, 1976; Klassen & O'Connor, 1988a, 1989; Steadman & Morrissey, 1982). In these "validation fits," which are a better way of gauging the discriminant functions' true performance, a substantial shrinkage (Copas, 1985) in prediction accuracy was the consistent result. The average AUC of discriminant functions (or "actuarial methods" [Faust & Ziskin, 1988]) evaluated using cross-validation groups (0.7130 ± 0.0085) is greater than the average for clinical predictions (0.6718 ± 0.0115, z = 2.881, p = .0040 [two-sided]). Examination of Table 4 shows that these differences are explained by the relative accuracies of long-term (≥1 year) predictions of each type; the average accuracy of validation-fit discriminant functions covering less than 1 year is comparable to the average for clinical predictions.

4. The accuracy achieved by A. Good Short (AUC = 0.752) equals the best AUC shown in Table 3 for clinical predictions covering a week or less (McNeil & Binder, 1987). From Table 4, one can see that the average accuracies for clinical predictions covering periods of 1 year or longer (n = 7, AUC = 0.6448 ± 0.0169), short (1–7 day) time periods (n = 6, AUC = 0.6877 ± 0.0333), and intermediate (0.5–12 month) time periods (n = 10, AUC = 0.6974 ± 0.0179) reveal no significant differences related to the length of time that the prediction governed. To reduce methodological heterogeneity, one can restrict comparisons to those studies published since 1986 that assessed clinical predictions over short and long observation periods. Table 4 shows an average accuracy for short-term clinical predictions (i.e., those reported by Janofsky, Spears, and Neubauer [1988], Kirk [1989], and McNeil & Binder [1987, 1991]) that is not different from the accuracy of the long-term predictions described by Zeiss et al. (1990).

5. I noted earlier that investigators who gauge accuracy using the correct fraction index or the FP:TP ratio may mistake the effect of a higher base rate for superior accuracy. A close look at some studies' original data provides illustrations of this point. If, for example, one compares the FP:TP ratio for Kozol et al. (1972) with that for McNeil and Binder (1987, #1), one finds that, for the former study, FP:TP = 1.9, whereas for the latter, FP:TP = 0.38. Two of three predictions of violence were incorrect in Kozol et al.'s study, but, as Table 3 shows, the accuracy of predictions was similar to that found by McNeil and Binder (1987).

6. Tables 3 and 4 support using past behavior to predict violence. Whether the prediction covered several days or several months, past behavior was a better-than-chance predictor of future behavior. The data of Janofsky et al. (1988) suggest that past behavior alone (#2) may have been a better indicator of violence potential than were clinicians' judgments (#1). (The significance of the difference shown in Table 3 cannot be evaluated because the intertest correlation of subjects could not be ascertained from the data published.) The average accuracy of predictions based on past behavior is higher than the average for clinical judgments (z = 6.49, p < 10⁻⁶ [two-sided])

### Table 4: Average ROC Areas for Selected Subgroups of Violence Predictions

<table>
<thead>
<tr>
<th>Group of predictions</th>
<th>N</th>
<th>Avg AUC ± SE*</th>
<th>χ²_homog</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>All data sets</td>
<td>58</td>
<td>0.7777 ± 0.0048</td>
<td>923.6</td>
<td>&lt;10⁻⁶</td>
</tr>
<tr>
<td>Pre-1986 sets</td>
<td>36</td>
<td>0.7492 ± 0.0059</td>
<td>590.0</td>
<td>&lt;10⁻⁶</td>
</tr>
<tr>
<td>Post-1986 sets</td>
<td>22</td>
<td>0.8358 ± 0.0084</td>
<td>244.6</td>
<td>0.044</td>
</tr>
<tr>
<td>Short-term sets</td>
<td>9</td>
<td>0.6925 ± 0.0234</td>
<td>15.91</td>
<td>0.0169</td>
</tr>
<tr>
<td>Medium-term sets</td>
<td>14</td>
<td>0.8240 ± 0.0096</td>
<td>236.2</td>
<td>&lt;10⁻⁶</td>
</tr>
<tr>
<td>Long-term sets</td>
<td>35</td>
<td>0.7664 ± 0.0057</td>
<td>616.1</td>
<td>&lt;10⁻⁶</td>
</tr>
<tr>
<td>Clinical judgments</td>
<td>17</td>
<td>0.6718 ± 0.0115</td>
<td>69.26</td>
<td>&lt;10⁻⁶</td>
</tr>
<tr>
<td>Behavior-based predictions</td>
<td>13</td>
<td>0.7797 ± 0.0120</td>
<td>152.1</td>
<td>&lt;10⁻⁶</td>
</tr>
<tr>
<td>DFs, *retrospective fit</td>
<td>14</td>
<td>0.8885 ± 0.0081</td>
<td>205.9</td>
<td>&lt;10⁻⁶</td>
</tr>
<tr>
<td>DFs, *validation fit</td>
<td>14</td>
<td>0.7130 ± 0.0085</td>
<td>59.94</td>
<td>&lt;10⁻⁶</td>
</tr>
<tr>
<td>Pre-1986 clinical short-term</td>
<td>2</td>
<td>0.6036 ± 0.0618</td>
<td>0.0263</td>
<td>0.87</td>
</tr>
<tr>
<td>Post-1986 clinical short-term</td>
<td>4</td>
<td>0.7222 ± 0.0395</td>
<td>0.5392</td>
<td>0.91</td>
</tr>
<tr>
<td>All clinical short-term</td>
<td>6</td>
<td>0.6877 ± 0.0333</td>
<td>4.391</td>
<td>0.50</td>
</tr>
<tr>
<td>All clinical medium-term</td>
<td>4</td>
<td>0.6974 ± 0.0179</td>
<td>0.6974</td>
<td>0.87</td>
</tr>
<tr>
<td>All clinical long-term</td>
<td>7</td>
<td>0.6448 ± 0.0169</td>
<td>37.97</td>
<td>2 × 10⁻⁶</td>
</tr>
<tr>
<td>Post-1986 behavior-based short-term</td>
<td>2</td>
<td>0.7303 ± 0.0387</td>
<td>7.312</td>
<td>0.0066</td>
</tr>
<tr>
<td>All behavior-based long-term</td>
<td>10</td>
<td>0.7868 ± 0.0129</td>
<td>141.0</td>
<td>&lt;10⁻⁶</td>
</tr>
<tr>
<td>Post-1986 DF*-validation medium-term</td>
<td>2</td>
<td>0.7477 ± 0.0329</td>
<td>1.216</td>
<td>0.27</td>
</tr>
<tr>
<td>All DF*-validation medium-term</td>
<td>5</td>
<td>0.7030 ± 0.0222</td>
<td>8.208</td>
<td>0.083</td>
</tr>
<tr>
<td>Pre-1986 DF*-validation long-term</td>
<td>7</td>
<td>0.7110 ± 0.0094</td>
<td>39.46</td>
<td>10⁻⁶</td>
</tr>
<tr>
<td>Post-1986 DF*-validation long-term</td>
<td>2</td>
<td>0.7956 ± 0.0442</td>
<td>6.486</td>
<td>0.011</td>
</tr>
<tr>
<td>All DF*-validation long-term</td>
<td>9</td>
<td>0.7147 ± 0.0092</td>
<td>51.39</td>
<td>&lt;10⁻⁶</td>
</tr>
</tbody>
</table>

*Note. ROC = receiver-operating characteristic. Avg = average. AUC = area under the ROC curve.

*a Weighted average and standard error calculated using method of McClish (1992). \( \chi^2_{\text{homog}} \) = test for homogeneity of ROC areas; \( df = N - 1; p = \) likelihood that average AUC is a common AUC. DF = discriminant function.
higher than the average for cross-validated discriminant functions \( z = 4.54, p = 5 \times 10^{-6} \) [two-sided].

7. Table 3 contains no study that directly compared mental health professionals’ prediction accuracy with the accuracy of nonclinicians. McNeil and Binder (1991) present data on physicians’ and nurses’ accuracy, but the significance of any differences in performance cannot be evaluated because the correlation in predictions was not published. Tables 3 and 4 permit mental health professionals to claim, with reasonable confidence, that they usually do better than chance in discriminating violent from nonviolent subjects, no matter what the time frame. But the studies in Tables 3 and 4 do not allow mental health professionals to claim any special expertise in predicting violence. Tables 3 and 4 suggest that a nonclinician furnished with knowledge of past behavior may outperform a mental health professional relying solely on information garnered from a clinical interview.

Conclusions and Implications

Mental health professionals are frequently placed in situations where they are called on to anticipate future violence, and clarity about the nature of violence prediction is therefore a matter of great practical importance. Clinicians are apt to think about violence predictions as yes-or-no judgments, because their decisions often involve taking or not taking a particular course of action (e.g., releasing or not releasing a patient, warning or not warning a third party). However, investigators studying clinicians’ ability to discriminate violent from nonviolent patients should recognize that accuracy in making this discrimination is only one factor that contributes to clinicians’ decisions based on anticipation of future violence. The threshold at which a rational decision maker takes action also reflects information about the base rate of violence and about the relative risks and benefits associated with various clinical outcomes (Mossman & Somoza, 1992).

ROC analysis lets investigators evaluate the accuracy of violence predictions using statistical tools that are unaffected by underlying base rates or biases favoring certain prediction outcomes. This article shows how prediction accuracy can be conceptualized and studied separately from other factors that influence clinical decisions. ROC analysis produces indices of performance that offer distinct advantages over the indices used in most published studies of violence prediction accuracy, and future reports of prediction accuracy should use ROC methods and indices for quantifying results.

Clinicians should not forget that what ROC analysis quantifies—the trade-off between sensitivity and specificity—is a fundamental feature of their ability to anticipate violence. This means that clinicians (and the general public) should realize that a fraction of the decisions based on assessments of potential future violence will inevitably be mistaken. Because clinicians cannot avoid making mistakes, they have to choose what kind of mistakes they prefer to make. When, for example, clinicians must decide whether to hospitalize someone involuntarily, the proportion of FP and FN mistakes they make will depend on how they feel about the consequences of wrongfully hospitalizing someone who is not violent and the consequences of releasing someone who is.

This article shows readers how ROC analysis can be used to succinctly convey accuracy data from an idealized hypothetical study, and it offers some tentative assessments of violence prediction accuracy using statistical methods that avoids the pitfalls of previously used descriptions of accuracy. This article’s reevaluation of representative data from the past 2 decades suggests that clinicians are able to distinguish violent from nonviolent patients with a modest, better-than-chance level of accuracy. The data evaluated here do not suggest that short-term violence predictions are more accurate than long-term predictions, and these data do not allow mental health professionals to claim any special ability to discriminate violent patients from those who will not be violent.

The accuracy of violence assessment has important implications for clinical practice and public policy. Future investigators, armed with appropriate statistical methods for evaluating and thinking about their data, should be able to provide information about prediction accuracy that clarifies, rather than muddles, clinical and legal decision making.

References


under a receiver operating characteristic (ROC) curve. *Radiology.* 143, 29-36.


**Appendix**

**Definitions of Some Terms Used to Describe Prediction Accuracy**

<table>
<thead>
<tr>
<th>Clinician's recommendation about hospitalization</th>
<th>Actual behavior</th>
<th>Admit patient</th>
<th>Release patient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Violent</td>
<td>True positive (TP)</td>
<td>False negative (FN)</td>
<td></td>
</tr>
<tr>
<td>Not violent</td>
<td>False positive (FP)</td>
<td>True negative (TN)</td>
<td></td>
</tr>
</tbody>
</table>

Base rate (BR) = (TP + FN)/(TP + FP + FN + TN)

Correct fraction (CF) = (TP + TN)/(TP + FP + FN + TN)

FP:TP ratio = FP/TP

Selection ratio (SR) = (TP + FP)/(TP + FP + FN + TN)

True positive rate (TPR) = Sensitivity = TP/(TP + FN)

True negative rate (TNR) = Specificity = TN/(TN + FP)

False positive rate (FPR) = (1 - Specificity) = FP/(FP + TN)

Risk ratio = TPR/FPR

Odds ratio = (TP · TN)/(FP · FN) or 

\[
\text{Relative improvement over chance (RIOC)} = \frac{\text{CF} - (\text{BR} \cdot \text{SR}) + (1 - \text{BR})(1 - \text{SR})}{1 - (\text{SR} - \text{BR})} - \frac{((\text{BR} \cdot \text{SR}) + (1 - \text{BR})(1 - \text{SR}))}{(\text{BR} \cdot \text{SR}) + (1 - \text{BR})(1 - \text{SR})}
\]

Received May 11, 1993
Revision received September 14, 1993
Accepted October 12, 1993