Law and Human Behavior, Vol. 18, No. 5, 1994

Comment

Further Comments on Portraying the Accuracy of Violence Predictions*

Douglas Mossman

Hart, Webster, and Menzies (1993) recently offered several recommendations for describing the accuracy of violence predictions using 2 × 2 contingency tables. This Comment describes some problems with their recommendations and suggests that researchers use receiver operating characteristic (ROC) analysis to quantify prediction accuracy.

In a recent Research Note, Hart, Webster, and Menzies (1993) recognize that most published research (e.g., Klassen & O'Connor, 1988; McNiel & Binder, 1987; Otto, 1992) on violence prediction has described accuracy using 2 × 2 contingency tables. These tables treat dangerousness assessments as binary (yes-or-no) predictions about the future, and portray the results of these predictions as true positive (TP), true negative (TN), false positive (FP), or false negative (FN) (see Table 1, top portion).

Hart and colleagues discuss several problems with describing clinical judgments about violence in this way, among which is the inconsistent use of statistical terminology in publications about violence prediction. They note that Monahan's landmark monograph (1981) uses "percent false positives" to designate the fraction of persons who were predicted to be violent but were not, whereas Otto's review of "second-generation" (post-1980) studies described prediction accuracy

* Address correspondence to Dr. Douglas Mossman, Wright State University, Department of Psychiatry, P. O. Box 927, Dayton, OH 45401-0927.

0147-7307/94/1000-0587$07.00/0

Table 1. Dependence of PPP, NPP, κ, and ϕ on Sequence and Prevalence

Prediction/test result	Actually violent (V+)	Actually not violent (V−)
Violent (T+)	True positives (TP)	False positives (FP)
Not violent (T−)	False negatives (FN)	True negatives (TN)

Test 1: TPR = 0.9, FPR = (1 − TNR) = 0.5
Test 2: TPR = 0.5, FPR = (1 − TNR) = 0.1
N = 1,000; Prevalence = 0.1

Sequence A: Test 1 selects patients to be evaluated by Test 2.

Test 1	V+	V−	Test 2	V+	V−
T+	90	450	T+	45	45
T−	10	450	T−	45	405
	PPP = 0.167			PPP = 0.500	
	NPP = 0.978			NPP = 0.900	
	κ = 0.135			κ = 0.399	
	ϕ = 0.241			ϕ = 0.400	

Sequence B: Test 2 selects patients to be evaluated by Test 1.

Test 2	V+	V−	Test 1	V+	V−
T+	50	90	T+	45	45
T−	50	810	T−	5	45
	PPP = 0.357			PPP = 0.500	
	NPP = 0.942			NPP = 0.900	
	κ = 0.340			κ = 0.340	
	ϕ = 0.346			ϕ = 0.400	

using false positive percentages that refer to the fraction of (ultimately) nonviolent persons who had been predicted to be violent. Although Hart and colleagues also caution researchers against using 2 × 2 tables to report prediction accuracy, they do not discuss in detail any other method of accuracy quantification. They feel that Monahan's way of describing 2 × 2 data "gives the information of greatest interest . . . , namely, the probability that a prediction of violence was incorrect" (p. 698). To reduce the likelihood of misunderstanding, Hart et al. suggest that investigators publish their raw data and calculate positive predictive power (PPP), negative predictive power (NPP), and an overall measure of accuracy such as κ or ϕ.

This Comment endorses Hart and colleagues' suggestion about publishing original data, but demurs to their suggestions concerning the best summary indices for 2 × 2 tables. Such tables indeed can be misleading, but for a different reason than Hart et al. give: 2 × 2 tables conflate intrinsic ability to detect future violence with the level of risk (or the "threshold," see Swets, 1992) that might prompt one to take action (e.g., to prevent violence).

The accuracy of dangerousness assessments can be interpreted in much the same way that the accuracy of medical diagnostic technologies is described. If we treat such assessments as binary (yes-or-no) "tests" for future violence, then their accuracy can be summarized using the concepts of *sensitivity* and *specificity*

(Somoza & Mossman, 1990). A violence test's sensitivity, then, would be the probability that the test will be positive (T+) when administered to a violent (V+) person, and the test's specificity would be the probability that it will be negative (T−) when administered to a nonviolent (V−) person. Sensitivity equals the test's *true positive rate* (TPR) and specificity equals the *true negative rate* (TNR). A test's *false positive rate* (FPR) and *false negative rate* (FNR) are the likelihoods of misidentification of nonviolent and violent persons, respectively. From these definitions and data such as those shown in Table 1, we can show that TPR = TP ÷ (TP + FN) and FPR = (1 − TNR) = FP ÷ (TN + FP). Similarly, the *prevalence* (Pr) or base rate of violence is Pr = (TP + FN) ÷ (TP + FP + TN + FN).

As Hart and colleagues note, what clinicians often "really want to know" is PPP and NPP, the likelihoods that a positive or negative prediction is correct. Bayes's Theorem (Bayes, 1763; Mossman & Somoza, 1991) tells us that PPP and NPP are functions of the test properties and prevalence. PPP, NPP, κ and ϕ can be calculated from FPR (or TNR), TPR, and Pr as follows:

$$Q = \text{"level"} = TPR \cdot Pr + FPR \cdot (1 - Pr)$$

$$PPP = [TPR \cdot Pr] \div Q$$

$$NPP = [(1\text{-}FPR) \cdot (1 - Pr)] \div [1 - Q]$$

$$\kappa = \frac{CF - [Pr \cdot Q + (1 - Pr)] \cdot (1 - Q)]}{1 - [Pr \cdot Q + (1 - Pr) \cdot (1 - Q)]}$$

$$\phi = \sqrt{TPR \cdot PPP \cdot (1 - FPR) \cdot NPP} - \sqrt{(1 - TPR) \cdot (1 - PPP) \cdot FPR \cdot (1 - NPP)}$$

The problems with using PPP, NPP, κ and ϕ as indices of test performance are illustrated in the lower portions of Table 1, where two violence prediction techniques ("Test 1" and "Test 2") are used consecutively to identify violent persons. Assume that $N = 1{,}000$ and that $Pr = 0.1$. In Sequence A, Test 1 is used to select a subgroup who are then evaluated with Test 2; in Sequence B, Test 2 is used first. Because PPP, NPP, κ, and ϕ are prevalence-dependent, the values can be changed simply by varying the sequence of the tests. In Sequence A, the κ index makes Test 2 seem better than Test 1, but in Sequence B the tests' κ's are the same; the ϕ index rates Test 2 better than Test 1, but in Sequence B the reverse is true. Unless the base rate of violence will always be the same whenever a specific test is used, PPP, NPP, κ, and ϕ will *not* be the best indices for describing and comparing *intrinsic* properties of diagnostic tests.

If TPR and FPR (or TNR) are the best measures for summarizing the data in 2 × 2 tables, are 2 × 2 tables the best way of portraying the accuracy of dangerousness assessments? 2 × 2 tables can be misleading if they are taken to mean that there can be only two kinds of risk assessments (will be violent, won't be violent) or only two kinds of outcomes at followup (was violent, wasn't violent). Violence can and often should be conceptualized as a hierarchy of behaviors (see Morrison,

1992). But in many situations, clinicians want answers to questions about outcomes that appropriately are couched in binary terms: Will this NGRI patient act violently (or "get into any trouble") if conditionally released? Will this mentally disordered offender commit another violent felony? Will this emergency room evaluee, if sent home, injure a family member? Will this newly admitted patient hurt another patient or a staff member? Similarly, many situations (e.g., Mossman, in press a) call for yes-or-no *decisions* by clinicians: Should we recommend release? Should we send this patient home? Should we institute special precautions (e.g., seclusion)?

Receiver operating characteristic (ROC) analysis offers a means for portraying the accuracy of violence assessments so that clinicians can perceive the relationship between level of risk and decision choice. Although some tests for predicting violence truly are binary (e.g., sex), most prediction variables have more than two values (e.g., number of previous arrests, income, years of education), and the results of most violence assessment methods (or actuarial methods or clinical judgments) thus reflect levels of confidence about the outcome. Typically, at the highest levels of confidence, only a few of the actually violent persons will be correctly identified (the test sensitivity will be low), but very few nonviolent persons will be misidentified (specificity will be high), and the ratio of true positive to false positive predictions will be high. As lower levels of confidence are included among those for which a test will be interpreted as positive, sensitivity increases but specificity decreases. ROC analysis helps investigators summarize these trade-offs between sensitivity and specificity which are the defining features of most violence detection methods. Two specific published examples will help demonstrate how ROC methods describe accuracy and elucidate key features of violence predictions.

Table 2A describes the performance of a discriminant function that uses prerelease data on mentally ill offenders to sort those who committed violent offenses from those who did not during a multiyear period following release from confinement (Harris, Rice, & Quinsey, 1993). Table 2A groups offenders according to their discriminant function score (DFS), with those in the first bin having the highest scores (i.e., highest likelihood of reoffending) and those in the sixth bin having the lowest scores. These data are summarized in a succinct, pictorial form in Figure 1, where the upper *ROC curve* plots TPR as a function of FPR. Both Table 2A and Figure 1 show that as the threshold (i.e., the DFS) is lowered, more violent recidivists are detected (TPR increases), but more nonrecidivists are misidentified (FPR increases, too).

The divisions between the bins represent five potential thresholds for making decisions about release of the offenders should this instrument be used to predict future behavior. These thresholds are marked along the upper curve in Figure 1. Once a particular threshold is chosen, the discriminant function's performance could be described using a 2×2 contingency table. But plotting the data graphically helps us realize that the discriminant function has many possible thresholds, and that the 2×2 table for the particular DFS used to make a decision (e.g., "below you go, above you stay") does not tell the whole story about the test's intrinsic performance.

Table 2. Examples of Violence Prediction Data with Multiple Thresholds

A. Discriminant Function—Violent Recidivists (after Harris et al., 1993, Table 4)

Category	1	2	3	4	5	6
Recidivists (N = 191)	78	37	23	28	15	10
Nonrecidivists (N = 427)	50	44	38	86	102	107
Likelihood ratio[a]	3.5	1.9	1.4	0.73	0.33	0.21
True positive rate	0.41	0.60	0.72	0.87	0.95	1.0
False positive rate	0.12	0.22	0.31	0.51	0.75	1.0

Binormal ROC indices: A = 1.04, B = 1.04, AUC = 0.765 ± 0.021

B. Nurses' Judgments: Attacks by Inpatients (after McNiel & Binder, 1991, Table 1)

Category	High	Moderate	Low
Physical attacks (N = 26)	6	12	8
No physical attack (N = 123)	9	38	76
Likelihood ratio[a]	3.2	1.5	0.50
True positive rate	0.23	0.69	1.0
False positive rate	0.073	0.38	1.0

Binormal ROC indices: A = 0.824, B = 1.07, AUC = 0.713 ± 0.063

[a] The "stratum-specific likelihood ratio" for category i, $SSLR_i$ (Pierce & Cornell, 1993) is $SSLR_i = p(T_i|V+) \div p(T_i|V-)$, where "$p(x|y)$" means "the probability of x, given y." $SSLR_i$ is the ratio of the probability of being in category T_i, given that one is violent, to the probability of being in category T_i, given that one is not violent. Notice that the *odds* of being violent given that one falls into category T_i, or $O(V+|T_i)$, are given by $SSLR_i \times [Pr \div (1 - Pr)]$. The odds prior to testing of being violent $O(V+) = [Pr \div (1 - Pr)]$. We can thus state that

$$\text{Posterior odds} = O(V+|T_i) = SSLR_i \times \text{prior odds} = SSLR_i \times O(V+),$$

which is the original formulation of Bayes's Theorem (1763).

The smooth curve drawn through the five thresholds utilizes the "binormal assumption" of ROC curve fitting (Hanley, 1988), which derives from the empirical finding that when (FPR, TPR) pairs are plotted as normal deviates or "Z-transforms," the pairs tend to fall along a straight line (Swets, 1986). A test's performance *throughout its entire range of thresholds* can thus be summarized used the relationship $Z_{TPR} = A + BZ_{FPR}$, where A is the intercept and B is the slope of the ROC plotted in normal deviate space. Another ROC performance index, the *area under the ROC curve* (AUC), provides a single, global estimate of overall performance. AUC equals the likelihood that the DFS of one nonviolent subject drawn at random from the nonviolent population would be lower that the DFS of a randomly selected violent subject (Hanley & McNiel, 1982).

Although clinicians often are *asked* whether someone will be violent or not, and although they often must make yes-or-no judgment-based *decisions* about patients, they can (and probably should, see Grisso & Appelbaum, 1992) actually categorize patients into more than two anticipated likelihoods or levels of risk for violence. Clinicians can be asked to give probabilistic ratings for subjects, and the accuracy of clinicians' ratings can be treated as though the clinicians were reporting the results of "mental" or implicit discriminant functions. McNiel and Binder (1991), for example, asked physicians and nurses to estimate the likelihood that newly admitted patients would act violently within one week of admission. The clinicians estimated probabilities on an eleven-point scale (0%, 10%, . . . , 100%);

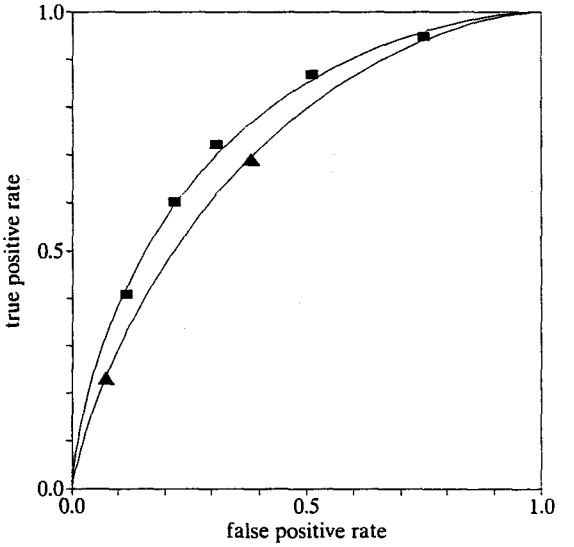

Fig. 1. ROC curve describing performance of two methods for predicting violence, based on data of Harris et al. (1993) (upper curve, squares) and McNiel and Binder (1991) (lower curve, triangles).

the authors grouped these rankings into three categories ("low" = 0–33%, "moderate" = 34%–66%, and "high" = 67%–100%) for analysis and publication. An example is shown in Table 2B, and the ROC curve corresponding to this data is shown in Figure 1.

By comparing either the curves in Figure 1 or the ROC indices in Tables 2A and 2B, we can see that the two methods—which use different time frames, criteria for violence, and study populations—have very similar accuracies (and in fact, their accuracies are typical of violence assessments; see Mossman, in press [b]). Their AUCs, for example, do not differ significantly ($z = 0.783$, $p = 0.57$ [two-sided]). Had we compared the methods in any other way, we might not have realized this. More to the point, the binormal ROC indices tell us a great deal more than any of the 2 × 2 summary indices. In fact, ROC indices even tell us more than the tabulated data, since they allow for estimation of test performance at any threshold and can be used for a host of statistical comparisons of test performance (e.g., McNiel & Hanley, 1984; Metz, 1986; McClish, 1989; Somoza & Mossman, 1992).

In a concluding footnote, Hart et al. remind readers that their discussion deals simply with the calculation of statistical indices, which is only one of many methodological problems that make it difficult to interpret results of violence prediction studies. ROC analysis will not solve many methodologic problems (e.g., ascertainment of "what actually happens" in natural settings or the "open texture" of the term "violence"; see Otto, 1992; Mossman, 1994 in press b). It will help investigators and clinicians better understand the nature of violence risk assessment, however, and will prevent them from conflating intrinsic discrimination accuracy, prevalence, and choice of decision threshold when violence predictions are evaluated.

REFERENCES

Bayes, T. (1763). An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London, 53,* 370–375.

Grisso, T., & Appelbaum, P. S. (1992). Is it unethical to offer predictions of future violence? *Law and Human Behavior, 16,* 621–634.

Hanley, J. A. (1988). The robustness of the "binormal" assumptions used in fitting ROC curves. *Medical Decision Making, 8,* 197–203.

Hanley, J. A., & McNiel, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29–36.

Harris, G. T., Rice, M. E., & Quinsey, V. L. (1993). Violent recidivism of mentally disordered offenders: The development of a statistical prediction instrument. *Criminal Justice and Behavior, 20,* 315–335.

Hart, S. J., Webster, C. D., & Menzies, R. J. (1993). A note on portraying the accuracy of violence predictions. *Law and Human Behavior, 17,* 695–700.

Klassen, D., & O'Connor, W. A. (1988). A prospective study of predictors of violence in adult male mental health admissions. *Law and Human Behavior, 12,* 143–158.

McClish, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making, 9,* 190–195.

McNiel, B. J., Keller, E., & Adelstein, S. J. (1975). Primer on certain elements of medical decision making. *New England Journal of Medicine, 293,* 211–215.

McNiel, B. J., & Hanley, J. A. (1984). Statistical approaches to the analysis of receiver operating characteristic (ROC) curves. *Medical Decision Making, 4,* 137–150.

McNiel, D. E., & Binder, R. L. (1987). Predictive validity of judgments of dangerousness in emergency civil commitment. *American Journal of Psychiatry, 144,* 197–200.

McNiel, D. E., & Binder, R. L. (1991). Clinical assessment of risk of violence among psychiatric inpatients. *American Journal of Psychiatry, 148,* 1317–1321.

Metz, C. E. (1986). Statistical analysis of ROC data in evaluating diagnostic performance. In D. E. Herbert & R. H. Myers (Eds.), *Multiple regression analysis: Applications in the health sciences* (pp. 365–384). Washington, DC: American Institute of Physics. (ROC analysis software is available from Charles E. Metz, Ph.D., Department of Radiology, University of Chicago, Chicago, IL 60637.)

Monahan, J. (1981). *The clinical prediction of violent behavior.* DHHS Publication ADM 81-92. Rockville, MD: National Institute of Mental Health.

Morrison, E. F. (1992). A hierarchy of aggressive and violent behaviors among psychiatric inpatients. *Hospital and Community Psychiatry, 43,* 505–506.

Mossman, D. (in press a). Dangerousness decisions: An essay on the mathematics of clinical violence predictions and involuntary hospitalization. *University of Chicago Law School Roundtable.*

Mossman, D. (in press b). Assessing predictions of violence: Being accurate about accuracy. *Journal of Consulting and Clinical Psychology.*

Mossman, D., & Somoza, E. (1991). Neuropsychiatric decision making: The role of prevalence in diagnostic testing. *Journal of Neuropsychiatry and Clinical Neurosciences, 3,* 84–88.

Otto, R. K. (1992). Prediction of dangerous behavior: A review and analysis of "second-generation" research. *Forensic Reports, 5,* 103–134.

Pierce, J. C., & Cornell, R. G. (1993). Integrating stratum-specific likelihood ratios with the analysis of ROC curves. *Medical Decision Making, 13,* 141–151.

Somoza, E., & Mossman, D. (1990). Introduction to neuropsychiatric decision-making: Binary diagnostic tests. *Journal of Neuropsychiatry and Clinical Neurosciences, 2,* 297–300.

Somoza, E., & Mossman, D. (1992). Comparing and optimizing diagnostic tests: an information-theoretical approach. *Medical Decision Making, 12,* 179–188.

Swets, J. A. (1986). Forms of empirical ROCs in discrimination and diagnostic tasks: implications for theory and measurement of performance. *Psychological Bulletin, 99,* 181–198.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47,* 522–532.