Commentary: Assessing the Risk of Violence—Are "Accurate" Predictions Useful?

Douglas Mossman, MD

J Am Acad Psychiatry Law 28:272–81, 2000

In his 1999 Isaac Ray Award lecture (reprinted elsewhere in this issue),[1] Dr. Henry Steadman suggests that research over the next quarter century may yield "practical tools" for assessing the risk of violence in individuals with mental disorders. Despite "the various limitations of current knowledge," Dr. Steadman believes that recent studies justify "an optimism that would have been misplaced" two decades ago, when Professor John Monahan[2] authored these frequently quoted words about the accuracy of violence predictions:

> [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 were diagnosed as mentally ill* [pp. 47–49, emphasis in original].

In this article, I argue that past research on the accuracy of violence prediction deserves a more positive assessment than Monahan's words suggest, but that future research is unlikely to give clinicians and judicial decision-makers predictions instruments with much practical utility. To make this argument, I first attempt to reduce confusion about these issues by explaining how the accuracy of a test or detection system should be measured and what accuracy measurements mean. I then summarize evidence for the proposition that clinical judgments about future violence have better-than-chance accuracy. Next, I examine the practical import of currently available violence prediction methods (i.e., what these methods tell clinicians about the likelihood of violence in populations that they evaluate or treat). My discussion will reveal what seems like a paradox: even reasonably accurate assessment instruments may not have much practical value for clinicians who make decisions about violent patients. I then conclude with an exploration of this finding and its implications for clinical decision-making.

The Impact of Monahan's Monograph

Monahan's 1981 monograph[2] has had a lasting influence on courts' and legal scholars' perceptions about the ability of mental health professionals to gauge the potential for violence. The U.S. Supreme Court majority opinion in the *Barefoot v. Estelle*[3] decision recognized Monahan as the "leading thinker on th[e] issue" (463 U.S. at 901) of predicting violence, and Justice Blackmun's dissent in the case relied on Monahan's finding that "psychiatric testimony about a defendant's future dangerousness. . . is wrong two times out of three" (463 U.S. at 916). In *Heller v. Doe*,[4] a 1993 decision, the Supreme Court flatly declared, "Psychiatric predictions of future violent behavior by the mentally ill are inaccurate" (509 U.S. at 324). The 1998 edition of the leading treatise on mental disability law contrasts the possibility that predictions for short-term emergency hospitalization may be accurate with "the proven predictive failures as to *long-term* indeterminate future dangerousness" (p. 119, emphasis in original).[5]

Dr. Steadman's lecture is one of many recent instances showing that Monahan's 1981 description of

Dr. Mossman is Professor and Director, Division of Forensic Psychiatry, Wright State University School of Medicine, and Adjunct Professor, University of Dayton School of Law. Address correspondence to: Douglas Mossman, MD, WSU Dept. of Psychiatry, P. O. Box 927, Dayton, OH 45401-0927. E-mail: dmossman@pol.net

prediction accuracy also continues to influence mental health professionals and researchers who study psychiatric assessments of dangerousness. In her October 1998 Presidential Address to the American Academy of Psychiatry and the Law,[6] Dr. Renee Binder summarized her research as showing that "short-term predictions of violence risk are more accurate than has been reported in the literature about long-term predictions" (p. 197). That same month, Dr. Phillip Resnick, in a National Public Radio interview concerning the Massachusetts General Hospital's evaluation of boxer Mike Tyson,[7] commented that the evaluators were right to acknowledge "that no one can predict future behavior with accuracy."[7] The same NPR broadcast segment offered Dr. Margaret Hagen's view: "There's fifty years of research showing that you cannot accurately predict future violence on the basis of clinical judgment."[8] The November 1999 issue of *Psychiatric Services* contains a report by Hoptman and colleagues,[9] which asserts that "[a]ccuracy is low for long-term predictions" and that "accuracy is somewhat better for short-term predictions" (p. 1461).

These statements all are understandable interpretations of Monahan's "no more than one out of three predictions" assessment. Yet, as Monahan himself has recently noted,[10] predictions of violence, including long-term predictions based in clinical judgment, appear to have what I previously described as "a modest, better-than-chance level of accuracy" (Ref. 11, p. 790). This finding seems puzzling. How can predictions of violence be accurate when two-thirds of them are incorrect? We can answer this question by examining the accuracy of a familiar device that makes many wrong "predictions" but which we nonetheless think is very accurate: an airport's metal detector.

Airport Metal Detectors: Many "Wrong" but Accurate Predictions

Airports use metal detectors to determine whether prospective passengers are attempting to carry large metal weapons (e.g., firearms) on board. Most readers probably have never seen a metal detector identify a person who really is carrying a weapon, but almost all readers will have witnessed a detector make an "error" when its alarm is triggered by something innocuous (e.g., a cell phone or a belt buckle).

Suppose we refer to the alarm's sounding as a prediction that the passenger is carrying a weapon, and the alarm's not sounding as a prediction that the passenger is not carrying a weapon. A given passenger either is or is not carrying a weapon, and a detector's alarm either sounds or does not sound when a passenger walks through. Denote the possible events as follows:

W+: Passenger is carrying a weapon.
W−: Passenger is not carrying a weapon.
S+: Detector's alarm sounds.
S−: Detector's alarm does not sound.

A good starting point for thinking about the detector's accuracy is to describe its performance using the medical literature's familiar terms sensitivity and specificity.[12] The detector's sensitivity would be the probability that the detector's alarm sounds ("predicts" a weapon) when a passenger is actually carrying a weapon; symbolically, this probability is written $P(S+|W+)$. Specificity would be the probability that the detector does not sound (predicts no weapon) when a passenger is not carrying a weapon, or $P(S-|W-)$.

Suppose the detector has an adjustable dial, and technicians set the dial at a particular setting. Then, they test the detector by having many individuals (say, 10,000 persons chosen to represent typical passengers) pass through it twice. On their first walk through the detector, the individuals carry firearms; on their second pass through, the individuals are unarmed. The detector's alarm sounds 99.9 percent of the time that an armed person walks through and does not sound 90 percent of the time that an unarmed person walks through. The detector's sensitivity is $P(S+|W+) = .999$, and its specificity is $P(S-|W-) = .900$.

Although the metal detector is not perfect, these numbers clearly imply that it is a very accurate device. Yet one can portray the metal detector's performance in a way that makes it seem inaccurate. Suppose we evaluate accuracy by answering these questions:

(1) When the alarm sounds, how often is it a false alarm?

(2) In what fraction of cases does the detector give the correct answer?

(3) Does the metal detector do its job, which is to keep armed passengers off planes?

Suppose that only .1 percent of actual passengers try to carry a weapon on board, which we can represent as $P(W+) = .001$. After 1,000,000 real passengers passed through the metal detector, we might expect these detection results: passengers with weapons = P(W+) × 1,000,000 = 1,000; armed passengers detected = 1000 × P(S+|W+) = 999; passengers without weapons = 999,000; correctly identified unarmed passengers = P(S−|W−) × 999,000 = 899,991; false alarms = 999,000 − 899,991 = 99,999.

Let us now answer the three evaluation questions:

(1) The probability that an alarm is a false alarm is 99,999/(99,999 + 999) = .99. In other words, 99 percent of the predictions of weapons are wrong. Another way of putting this is that the ratio of false positive predictions to true positive predictions, FP:TP, is 99:1.

(2) The total number of correct predictions is 999 + 899,991 = 900,990, so the metal detector is right just over 90 percent of the time. But if this seems like good performance, consider the fact that if the alarm never sounded—if it always predicted W− (no weapon)—the metal detector would have been right 99.9 percent of the time. So, by using the metal detector, we get more incorrect predictions than we would get if we had not used the detector.

(3) If the one passenger who successfully concealed a weapon used it to hijack a plane, that event would make headlines. Media pundits would say that the detector had failed to do its job and that airport security was inadequate.

Suppose now that airport security personnel take the last criticism most seriously, and technicians adjust the metal detector setting to increase its sensitivity. Now, P(S+|W+) = .9999, but making this adjustment causes the specificity falls a bit, and P(S−|W−) = .80. Another 1,000,000 passengers pass through for whom P(W+), the probability of weapon carrying, remains .001. We expect these results: passengers with weapons = P(W+) × 1,000,000 = 1,000; armed passengers detected = 1000 × P(S+|W+) = 1,000; passengers without weapons = 999,000; correctly identified unarmed passengers = P(S−|W−) × 999,000 = 799,200; false alarms = 999,000−799,200 = 199,800.

We use these new results to re-answer the three evaluation questions:

(1) The probability that an alarm is a false alarm is 199,800/(199, 800 + 1,000) = .995. Now, 99.5 percent of the predictions of weapons are wrong, and FP:TP = 200:1.

(2) The number of correct predictions is 1,000 + 799,200 = 800,200. Now, the detector is right only 80 percent of the time.

(3) No passenger has successfully concealed a weapon; using this criterion, the detector's performance improved.

Four lessons emerge from this discussion. The first is that, by using poor indices of accuracy, it is possible to mislead oneself and conclude that a very accurate detection device is inaccurate. The problem with FP:TP and the correct fraction (CF) index is that they fail to factor out the "base rate" of the phenomenon being detected. This can be seen from looking at the formulae for these indices in the metal detector example:

$$FP{:}TP = \frac{[1-\mathrm{P}(W+)]\times[1-\mathrm{P}(S-|W-)]}{\mathrm{P}(W+)\times\mathrm{P}(S+|W+)} \qquad \text{(Eq. 1)}$$

$$CF =$$

$$\mathrm{P}(W+)\times\mathrm{P}(S+|W+)+[1-\mathrm{P}(W+)]\times\mathrm{P}(S-|W-)$$
$$= \mathrm{P}(W+)[\mathrm{P}(S+|W+)-\mathrm{P}(S-|W-)]+\mathrm{P}(S-|W-) \qquad \text{(Eq. 2)}$$

Inspection of Equation 1 reveals that FP:TP increases as the base rate, P(W+), decreases. Inspection of Equation 2 shows that whenever sensitivity, P(S+|W+), is greater than specificity, P(S−|W−), as was the case in the metal detector example, CF will decrease as P(W+) decreases. In the metal detector example, FP:TP was high and CF was low not because the detector was inaccurate but because the base rate of weapon carrying was low.

A second lesson is that erroneous predictions and even tragic consequences do not imply that a detection system "lacks" accuracy. A system can be both accurate and make errors. In fact, it may cloud the issue to talk about whether a detection method is accurate or not; what we should do, instead, is describe degrees of accuracy using indices that are not potentially misleading.

A third lesson comes from realizing that most airline passengers probably would not even consider being wrongfully identified as a weapon carrier to be a mistaken prediction; it is just a minor inconvenience that promotes safer air travel. If the consequences of false positive errors are trivial* and if false

* This was not always the case. When I took a plane flight in 1972, before metal detectors were in general use, security personnel identified me as fitting the "profile" of a hijacker and searched me before letting me board the plane. (The profile made a false-positive error; I wasn't carrying a weapon.) Were a full search necessary every time a prospective passenger triggered a metal detector's alarm, air travelers would

negative errors lead to serious problems, we should try to adjust a detection system so as to minimize the latter. When false positive and false negative errors both have important consequences (as frequently is the case in assessing violence), we have to consider both types of errors when thinking about how to calibrate the detection system.

The fourth lesson stems from the recognition that, like a metal detector, many diagnostic systems and prediction methods are "adjustable." Because its alarm threshold has many settings, a metal detector does not have a single level of sensitivity and specificity. In general, a prediction method should be evaluated and described in a way that characterizes the tradeoffs between sensitivity and specificity that occur as the "threshold" is adjusted throughout the range of possible values.

Receiver Operating Characteristic (ROC) Analysis and the Accuracy of Violence Predictions

In the mid-1990s, several writers[11, 13–15] recognized that adjustable thresholds are a feature of most violence prediction techniques, and that receiver operating characteristic analysis should therefore be used to describe the accuracy of violence prediction methods. ROC analysis allows investigators to characterize the tradeoffs between errors and correct identifications that arise from the intrinsic discrimination capacity of a detection method and to distinguish these features from the threshold or operating point used to make a decision.[16] ROC analyses typically utilize a ROC graph, which succinctly summarizes the results of a detection method as the threshold is moved throughout its range of possible values. Fig. 1 is an example of such a graph, based on a discriminant function for predicting violence described by Rice and Harris.[14] The graph plots the true positive rate (TPR = sensitivity) as a function of the false positive rate (FPR = 1 − specificity) and shows that as TPR increases, FPR increases too.

By making certain simple assumptions about the underlying shapes of the data distribution, one can fit the empirical results to a smooth ROC curve.[†] The better a test or detection system, the greater the area under the ROC curve (AUC) that describes the test's or system's performance. The AUC of a test or detection system has an important practical interpretation.[17] Applied to violence prediction, AUC equals the probability that the detection method would give a randomly selected, actually violent person a higher violence rating than a randomly selected, nonviolent person. A perfect violence detection method, one that always sorted violent and nonviolent persons correctly, would have an AUC of 1.0; a test that gave no information would have an AUC of .5 and would be described by the diagonal line in Fig. 1. For the ROC curve shown in Fig. 1, the AUC ± S.E. is .770 ± .020[‡], implying an accuracy level that is comparable to results from other studies in which discriminant functions were used to make long-range predictions of violence.[11]

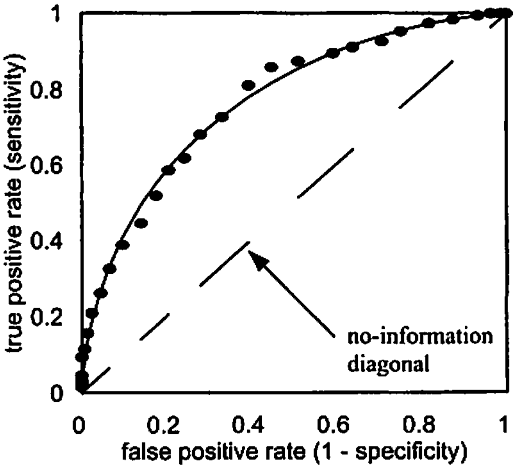

Figure 1. ROC curve (*smooth line*) fit to accuracy data (*individual points*) reported by Rice and Harris.[14] AUC = .770 ± .020. *Dashed diagonal line* represents the ROC for a test that provides no information.

Recognizing that ROC methods offer the best way to characterize accuracy helps us assess violence predictions for which investigators have reported results that permit calculation of TPR and FPR at only one cut-off point. Fig. 2 plots the results of the four studies[18–21] of clinical violence prediction reviewed in Monahan's 1981 monograph for which both sensitivity and specificity can be calculated.[§] In evaluating

have to arrive at airports several hours before boarding and making connections.

[†] A discussion of the binormal assumption used in ROC curve-fitting is found in Refs. 11 and 12. Briefly, the binormal assumption states that the points on a ROC curve can be summarized using the equation $Z_{TPR} = A + BZ_{FPR}$, where Z_{TPR} and Z_{FPR} are the normal deviates, or z-transforms, of TPR and FPR.

[‡] Rice and Harris report an AUC = .76 for these data. The small discrepancy arises because their article uses a trapezoidal AUC with only a few cut-offs rather than the area under a fitted binormal curve. AUCs calculated with the trapezoidal method can be expected to slightly underestimate the true area under a ROC curve. The fitted binormal ROC curve in Fig. 1 is described by the equation $Z_{TPR} = A + BZ_{FPR}$, where $A = 1.030$ and $B = .971$.

[§] Concerning the results given in Ref. 20, Monahan reports that 7 percent of the group who were predicted to be nonviolent were subse-

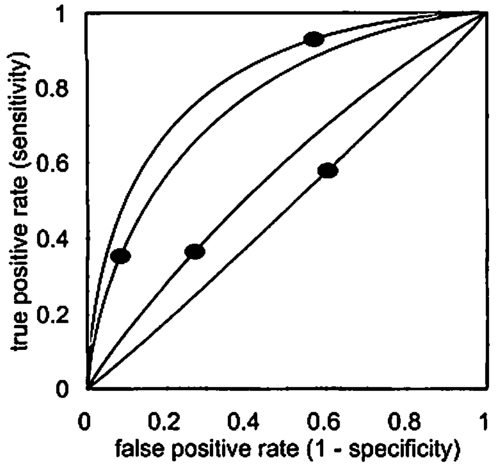

Figure 2. ROC curves drawn through single cut-offs. Values were derived from studies reviewed by Monahan.[2]

these studies, Monahan typically interpreted clinicians' yes–no recommendations about whether patients should be released from custody as though they were predictions of violence. Because these studies use yes–no recommendations, the clinicians' accuracy in each study yields only a single point in the ROC square. But of course, the clinicians could have made graded judgments about their patients' violence potential, which would be represented by multiple thresholds in a ROC square. Recognizing this, we can use reasonable assumptions[¶] to draw full ROC curves through each data point that we actually have, using the areas under each curve as indices of the clinicians' accuracy.

Table 1 lists the AUC ± S.E. for each long-term study, the average AUC ± S.E. for these studies, and for comparison, the average AUC ± S.E. for short- and medium-term violence predictions reviewed by Mossman.[11] (The average AUC and standard error for the long-term studies were calculated using a method described by Zhou[22]; the standard errors reported by Mossman[11] for the short- and medium-term averages have been recalculated using Zhou's method.) When we look at the individual results for the long-term studies, we see that they are quite heterogeneous. In one study,[18] accuracy was no better than chance, and in a second,[20] it was not significantly better than chance. In the two remaining studies,[19, 21] however, the accuracy of long-term clinical predictions was quite respectable. The weighted average accuracy for the four studies strongly suggests that clinicians' long-term predictions have better-than-chance accuracy. Moreover, long-term clinical predictions appear to be as accurate as short- and medium-term clinical predictions.

Table 1 Area Under the ROC Curve (AUC) and Standard Error (S.E.) for Short-, Medium-, and Long-Term Predictions of Violence

Data Source	AUC ± S.E.	p^a
Long-term[b] studies (from Monahan[2])		
Cocozza and Steadman[18]	0.483 ± 0.050	0.633
Kozol *et al.*[19]	0.763 ± 0.042	$<10^{-9}$
Steadman[20]	0.575 ± 0.050	0.067
Patuxent results[21]	0.822 ± 0.024	$<10^{-9}$
Average	0.666 ± 0.069	0.008
Average, medium-term[c] studies	0.719 ± 0.041	$<10^{-7}$
Average, short-term[d] studies	0.688 ± 0.033	$<10^{-8}$

[a] *p*, significance level.
[b] Follow-up period = 3–5 years.
[c] Follow-up period = 1–6 months.
[d] Follow-up period = 3–7 days.

To understand why this conclusion is different from the one reported by Monahan and other writers, it helps to take a close look at the results of the study by Kozol and colleagues[19] (cited also by Dr. Steadman[1] as Ref. 6). Kozol and colleagues found that of 49 subjects thought to be violent, only 35 percent had acted violently during a five-year follow-up period, which is consistent with the view that long-term predictions "are accurate in no more than one out of three" cases. Yet Kozol and colleagues also found that clinicians were correct concerning 92 percent of the 386 subjects whom they said would not be violent. The clinicians, in other words, were reasonably accurate; the FP:TP ratio reflects the fact that only 11 percent of the 435 subjects were violent during the follow-up period. This discussion illustrates the importance of recognizing that base rates affect the absolute numbers of prediction errors and of using accuracy indices that separate features of the detection process from the population's base rate.

Clinical and Actuarial Predictions

Most of the preceding discussion has focused on the accuracy of "clinical" assessments of violent risk, in which clinicians use their intuition, knowledge about the persons they are assessing, "gut instincts," and/or anything else they think may be relevant. By contrast, "actuarial" risk assessments typically re-

quently arrested and that two subgroups who were predicted to be violent had arrest rates of 39 and 46 percent. For this article's analyses, these last two rates were combined into a single rate of 42 percent.

[¶] The assumptions and their justifications are discussed in Ref. 11. In brief, one assumes that each binormal ROC curve is symmetric about the negative diagonal of the ROC square (i.e., the diagonal line running from (0, 1) to (1, 0)). This is equivalent to assuming that $B = 1$ in the equation $Z_{\mathrm{TPR}} = A + BZ_{\mathrm{FPR}}$.

quire clinicians to gather information about a (usually small) number of factors concerning the individuals they evaluate. The clinicians then categorize this information using some explicit scoring system and come up with a numerical value that summarizes the evaluees' risk of violence. The discriminant function evaluated by Rice and Harris[14] (the results of which were used to construct Fig. 1) is an example of an actuarial method for gauging violence potential.

Mental health professionals who are unfamiliar with studies comparing clinical judgments and actuarial methods may assume that the former are more accurate because they incorporate things like clinicians' experience, human pattern recognition abilities, and subtle nuances that are left out of simple formulae. The psychological literature consistently shows that the opposite is true, however; simple actuarial methods usually outperform clinical judgments in a variety of tasks,[23] including violence prediction.[11, 15, 24] Actuarial methods have advantages in addition to their superior performance. When used properly, they are systematic and impartial. They also have a "transparency" that clinical judgment lacks: whereas the reasoning behind clinical hunches is sometimes murky and ambiguous, actuarial judgments are based on data and an explicitly prescribed method of combining those data. This makes actuarial methods open to inspection, questioning, and when necessary, critique.

Several factors—including an increasing concern about managing violence risk, sexual predator sentencing schemes, and investigators' awareness of the superiority of actuarial methods—have spurred the publication, over the past decade, of several actuarial tools[24–30] for assessing the risk of violence. Investigators are evaluating these instruments in diverse settings, and some apparently favorable findings are now being reported. For example, Douglas and colleagues[31] used the HCR-20 assessment method[30] as a violence predictor with patients discharged from civil psychiatric settings and followed for two years, and they found that this instrument had AUCs of .76-.80. Other studies of the HCR-20 that report on prediction of inpatient violence and violence by persons initially evaluated as outpatients have yielded AUCs that represented similar accuracy levels.[||]

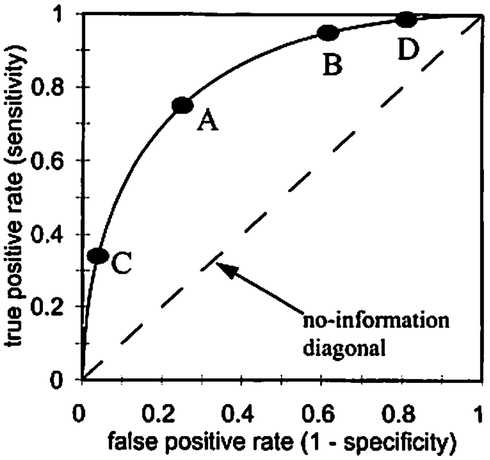

Figure 3. ROC curve for the hypothetical VPI; A, B, C, and D are possible operating points, or decision thresholds, for the instrument.

Are Modestly Accurate Predictions Useful?

Findings like the ones just presented imply that long-term predictions of violence can be accurate; they also support the assertion that actuarial methods of prediction probably are more accurate than unaided clinical judgment. Yet forensic clinicians should recognize that despite these findings, currently available risk assessment methods have very limited usefulness.

To understand this point, imagine that we have available a new violence prediction instrument, or "VPI," and that its AUC equals .83, making it an instrument with accuracy that is well above the average reported for cross-validated actuarial methods.[11] Assume also (for purposes of illustration) that the detection properties of the VPI are described by the ROC curve in Fig. 3, which includes the point in ROC space where FPR = .25 and TPR = .75.[**] Suppose, now, that we pick the VPI score *B* that corresponds to this point as our cut-off or operating point for the instrument. Scores greater than *B* will then represent a positive result (*R*+), and scores less than *B* will represent a negative result (*R*−). We use the VPI to evaluate a 320-member group of inpatients. Suppose, finally, that our experience tells us that one-fourth of the inpatients will engage in a seriously violent act during the follow-up period (a fairly typical violence rate in studies of inpatients),[32] and that we will use the *R*+ and *R*− results to dis-

[||] Although much of this research has not yet made its way into professional journals, it has been presented at meetings and is summarized in a Microsoft Word® document available at www.sfu.ca/psychology/groups/faculty/hart/violink.htm.

[**] One can show that a ROC curve for which AUC = .83 and which is symmetric about the negative diagonal through the ROC square will pass very near this point.

Table 2 Classification Results Using the Hypothetical VPI Instrument to Classify Patients' Risk of Violence

	Two-Way Classification			Three-Way Classification			
Behavior	High risk	Low risk	Totals	High risk	Unclassified	Low risk	Totals
Violent	60	20	80	27	49	4	80
Not violent	60	180	240	9	139	92	240
Totals	120	200	320	36	188	96	320

tinguish which patients are at high risk and low risk for violence.

The upper portion of Table 2 contains some simple Bayesian calculations showing that the 120 $R+$ patients comprise a high risk group whose members have a violence risk of 50 percent, and the 200 $R-$ patients are a low risk group whose members have a violence risk of about 10 percent. But how useful is this information? Clearly, we should be concerned about a group of patients half of whom will act violently. But should we ignore the potential risk posed by patients who have "only" a 1-in-10 chance of becoming violent? Most clinicians would respond, "Of course not!" In many clinical situations, moreover, mental health professionals would treat someone with a 10 percent risk of serious violence little differently from someone with a 50 percent risk; for both types of patient, clinicians usually would exercise high levels of concern in making follow-up plans and other treatment arrangements. Certainly, few clinicians who had to defend themselves in a *Tarasoff*-type lawsuit would want to tell jurors that they thought a patient's 10 percent risk of serious violence was not great enough to warrant careful efforts to prevent harm to others.

In an effort to make the VPI more helpful, one might explore whether it could be used to find subgroups or patients whose risk of violence was either high enough of low enough to justify levels or types of intervention different from those received by the "average" patient. Suppose we decide that low risk patients are those for whom the chance of acting violently is 4 percent, and high risk patients are those for whom the chance is 75 percent. We then choose cut-offs for the VPI such that patients with scores above a certain value (*C* in Fig. 3) will meet our "high risk" definition, and those below a certain value (*A* in Fig. 3) will meet our "low risk" definition. The lower portion of Table 2 describes the results of this process, which places 96 patients in the low risk group and 36 patients in the high risk group. Notice that 188 patients, or 59 percent of the entire group of 320, remain unclassified; because 49 (26%) of the unclassified patients act violently, their base rate of violence is virtually the same as the whole group's base rate. Looking at the results in the lower portion of Table 2, one can imagine skeptical interlocutors asking whether a 4 percent risk of serious violence is really low enough to ignore, and whether a 26 percent risk of serious violence is low enough to justify different treatment and precautions than might be imposed on the high risk group.[††]

One can also imagine the difficulty in explaining why risk that falls below a certain point can be ignored or acted upon differently from risk that is just a bit higher.[34] To address this point, suppose we decided that a group of patients whose risk of serious violence equals that of the not-mentally-ill general public constitutes a population for whom no special measures are needed. Swanson and colleagues,[35] using responses from the Epidemiological Catchment Area survey, found that just over two percent of persons without a psychiatric diagnosis reported having committed a serious act of violence in the preceding year. By adjusting the VPI cut-off to point *D* in Fig. 3, one could identify a 47-member subgroup of patients of whom only one (2.1%) would be expected to act violently. Of course, this cut-off choice would leave 273 patients, more than 85 percent of the original 320, unclassified. Because 79 (29%) of these 273 patients would be expected to act violently, the VPI would give little information about the majority of patients beyond what one knew from their base rate alone. For most patients, therefore, the VPI would

[††] Recently, Steadman and colleagues[33] described how an iterative classification tree (ICT) could be used to assess 939 individuals in the MacArthur Violence Risk Assessment study. In Steadman and colleagues' three-way classification, the rate of violence in the low risk group ($n = 462$) was 3.9 percent; in the high risk group ($n = 257$), the rate was 43.6 percent; in the remaining 220 subjects, the rate was 20.9 percent. Steadman and colleagues calculated an AUC for the ICT of .82. This value almost certainly overestimates the ICT's true accuracy, however; the ICT was designed for these 939 subjects, and no cross-validation procedures were used to correct for over-optimism in the accuracy estimate.

not contribute anything to decisions about clinical management.

Having obtained these results, it is natural to wonder if a prediction method could conceivably help with clinical decision-making. The answer is yes, if the method were nearly infallible. For example, suppose the AUC for the VPI were .99 and that the ROC curve for this assessment instrument included the operating point where FPR = .05 and TPR = .95.[‡‡] The VPI would then sort patients into a 232-patient low risk subgroup, only 4 members of which (1.7%) would act violently, and an 88-patient high risk subgroup of which 76 members (86%) would act violently. Put differently, members of the high risk subgroup would have a violence risk that was more than 50 times that of the low risk subgroup, and most clinicians, I suspect, would feel comfortable with planning very different sorts of treatment for these persons. Notice, however, that even this superb prediction tool would miscategorize 5 percent of patients. For one percent of the patients, moreover, miscategorization would result in a failure to identify and take appropriate steps to prevent harm by one violent individual.

Conclusions

If asked what we will be doing 24 hours from now, many of us could give a short-term prediction that would be both confident and specific (e.g., "Seeing my patient, Mr. Jones"), and usually these sorts of short-term predictions turn out to be fairly accurate. By contrast, if we were asked what we will be doing 24 months from now, most of us would give long-term predictions that would be general and hedged (e.g., "I'll probably be seeing patients"), and we would not be at all surprised if we were wrong. We have similar levels of confidence in our short- and long-term predictions about other people. The sorts of "reason-giving explanation"[36] that we use to explain persons' behavior seem reliable only over short periods of time, because the specific motives, beliefs, or desires that we usually invoke to explain and rationalize a person's actions ("he took a drink of water because he wanted to quench his thirst") are operative for relatively short time periods.

Most of us, mental health professionals included, typically think and talk about violent actions using ordinary language explanatory paradigms exemplified by the sentence, "Jones hit Smith because Jones thought Smith made a threat." That is, we think and talk about violent actions just as we do other actions; because we regard them as emanating from specific beliefs and desires, we usually talk about violent actions using reason-giving explanations. Consequently, we might expect that our ability to say who will and will not be violent would share the limitations of our everyday language explanatory schemata. We might expect ourselves to do reasonably well at assessing a patient's violence risk for the next few days, during which time our knowledge about his present emotional state and his currently operative beliefs and desires would be relevant to his specific actions. As time elapsed and the patient's emotional state changed, the specific information obtained in a clinical assessment would be less and less pertinent, and we would expect our ability to predict his behavior would deteriorate. As Dix[37] put it several years ago, "Intuition suggests that psychiatrists' predictive ability is substantially greater when it is called into play concerning the short-term risk posed by persons whose assaultive tendencies are related to symptoms of identifiable serious mental illnesses" (p. 256).

Our current scientific understanding of violent behavior offers clinicians another view of violent behavior, however. From this perspective, an individual's use of violence reflects his relatively static sociodemographic characteristics, enduring behavioral patterns (e.g., his likelihood of seeking and remaining in outpatient treatment), and long-term likelihood of being in certain mental states (e.g., mistrustful or intoxicated). These traits make it more likely in general that one will act violently, whatever one's specific current situation might be.[35, 38, 39]

Psychiatric impairments alter one's interpretations of events, one's ability to resolve conflicts, and one's relationships with family and friends.[40] These effects tend to be chronic features of mental disorders, which may be why mental illness has a chronic, small, but statistically detectable effect on a person's violence risk. Similarly, several other personal characteristics that are statistically associated with violence—including sex, age, level of education, poverty, propensity to become intoxicated, likelihood of not adhering to treatment recommendations, and reaction to stressors—can provide information that helps make reasonable statements about an individual's long-term violence risk, because these charac-

[‡‡] A ROC curve for which AUC = .99 and which is symmetric about the negative diagonal through the ROC square will include this point.

teristics also are long-term features of an individual's physical and psychological make-up. It should not be surprising, then, that clinicians' intuitive judgments about individuals' long-term violence risk have better-than-chance accuracy. We should expect that, by using simple actuarial prediction tools that focus one's attention on known risk factors, clinicians could have considerable success in sorting patients into subgroups that, over extended periods of time, have larger and smaller proportions of individuals who become violent.

This article has reviewed published data indicating that clinical judgments about long-term violence risk have better-than-chance accuracy and that the accuracy of such judgments is similar to the accuracy of clinical judgments about short- and medium-term violence risk. Recently published findings strongly suggest that actuarial methods probably can help clinicians do better than what their unaided clinical judgment would tell them about a person's intermediate- and long-term risk of violence. Despite these findings, however, currently available prediction techniques frequently may not help clinicians make decisions about patient management. This is not because these violence prediction techniques are inaccurate, but because they are not accurate enough to sort patients into subgroups with meaningfully different levels of risk.

If violence prediction techniques are not accurate enough to make practical differences in clinical management, this does not mean that mental health professionals cannot do things to reduce violence. For example, considerable recent evidence suggests that nonadherence to medication and (especially) substance abuse are risk factors for violent behavior during the months after hospital discharge,[41, 42] and that friends and family members are the persons most likely to be the targets of violence.[42–44] Evidence also suggests that outpatient commitment and close community follow-up may improve outpatient outcomes and continuation in treatment.[45, 46] It thus seems reasonable to suppose that educating families and intensively following former inpatients after discharge (perhaps using outpatient commitment or intensive case management[47, 48] to improve adherence to community treatment) might be effective ways for mental health professionals to reduce violence.

It is important, however, to recognize that these measures are desirable and should be undertaken anyway, for the sake of patients and members of their social network. Whether or not such interventions reduce violence, these measures are beneficial because they ultimately enhance patients' autonomy. Patients need and deserve these treatments because they are good treatments; administering these treatments can be fully justified on therapeutic, nonutilitarian grounds alone. If a clinician has a well-founded belief that a patient needs and deserves certain treatments, that belief alone should motivate the clinician and justify making arrangements for the patient to get those treatments. Under such circumstances, the impact of treatment on the patient's violence potential should be of relatively minor importance in a clinician's decision-making, if it is important at all. Sound clinical interventions may be socially useful because they reduce violence potential in patients who can be identified as high risk, but violence reduction should be a side effect of, rather than a justification for, those interventions.

References

1. Steadman HJ: From dangerousness to risk assessment of community violence: taking stock at the turn of the century. J Am Acad Psychiatry Law 28:265–71, 2000
2. Monahan J: The Clinical Prediction of Violent Behavior. Rockville, MD: National Institute of Mental Health, 1981
3. Barefoot v. Estelle, 463 U.S. 880 (1983)
4. Heller v. Doe, 509 U.S. 312 (1993)
5. Perlin ML: Mental Disability Law—Civil and Criminal (ed 2). Charlottesville, VA: Lexis Law Publishing, 1998
6. Binder RL: Are the mentally ill dangerous? J Am Acad Psychiatry Law 27:189–201, 1999
7. Schouten R, Schmahmann JD, Medoff D, Deters TJ, Henderson D, Jordan BD: Independent medical evaluation of Michael Gerard Tyson, September 30, 1998 (the report to the Nevada Boxing Commission, prepared by the Law and Psychiatry Service at Massachusetts General Hospital, was made available to the public despite Dr. Schouten's objection), found at http://www.lvrj.com/lvrj home/sports/packages/tyson/report(site visited May 22, 2000)
8. Siegel R: Mike Tyson, of Sound Mind? National Public Radio, "All Things Considered," Oct 14, 1998 (Robert Siegel's interviews of Drs. Resnick and Hagen may be heard at http://www.npr.org/ramfiles/atc/19981014.atc.09.ram (site visited May 22, 2000))
9. Hoptman MJ, Yates KF, Patalinjug MB, Wack RC, Convit A: Clinical prediction of assaultive behavior among male psychiatric patients at a maximum-security forensic facility. Psychiatr Serv 50:1461–6, 1999
10. Monahan J: Clinical and actuarial predictions of violence, in Modern Scientific Evidence: The Law and Science of Expert Testimony (vol 1). Edited by Faigman D, Kaye D, Saks M, Sanders J. St. Paul, MN: West Publishing Co., pp 300–18, 1997
11. Mossman D: Assessing predictions of violence: being accurate about accuracy. J Consult Clin Psychol 62:783–92, 1994
12. Somoza E, Mossman D: Introduction to neuropsychiatric decision making: binary diagnostic tests. J Neuropsychiatry Clin Neurosci 2:297–300, 1990

13. Mossman D: Further comments on portraying the accuracy of violence predictions. Law Hum Behav 18:587–93, 1994
14. Rice M, Harris G: Violent recidivism: assessing predictive validity. J Consult Clin Psychol 63:737–48, 1995
15. Gardner W, Lidz CW, Mulvey EP, Shaw EC: Clinical versus actuarial predictions of violence in patients with mental illness. J Consult Clin Psychol 64:602–9, 1996
16. Mossman D, Somoza E: ROC curves, test accuracy, and the description of diagnostic tests. J Neuropsychiatry Clin Neurosci 3:330–3, 1991
17. Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143: 29–36, 1982
18. Cocozza JJ, Steadman HJ: The failure of psychiatric predictions of dangerousness: clear and convincing evidence. Rutgers L Rev 29: 1084–1101, 1976
19. Kozol H, Boucher R, Garofalo R: The diagnosis and treatment of dangerousness. Crime Delinquency 18:371–92, 1972
20. Steadman HJ: A new look at recidivism among Patuxent inmates. Bull Am Acad Psychiatry Law 5:200–9, 1977
21. State of Maryland: Maryland's defective deliquency statute—a progress report (unpublished manuscript; also cited in Ref. 1). Annapolis, MD: Department of Public Safety and Correctional Services, 1973
22. Zhou XH: Empirical Bayes combinations of estimated areas under TOC curves using estimating equations. Med Decis Making 16:24–8, 1996
23. Dawes RM, Faust D, Meehl PE: Clinical versus actuarial judgment. Science 243:1668–74, 1989
24. Quinsey VL, Harris GT, Rice ME, Cormier, CA: Violent Offenders: Appraising and Managing Risk. Washington, DC: American Psychological Association, 1998
25. Boer DP, Hart SD, Kropp PR, Webster CD: The Manual for the Sexual Violence Risk–20. Burnaby, British Columbia, Canada: Simon Fraser University, 1997
26. Hanson RK: The Development of a Brief Actuarial Risk Scale for Sexual Offense Recidivism. Ottawa, Ontario: Department of the Solicitor General of Canada, Public Works and Government Services Canada, cat. no. JS4–1/1997-4E, 1997
27. Hare RD: The Revised Psychopathy Checklist. Toronto: Multi-Health Systems, 1991
28. Hart SD, Cox D, Hare RD: Manual for the Screening Version of the Hare Psychopathy Checklist—Revised (PCL:SV). Toronto: Multi-Health Systems, 1995
29. Webster CD, Harris GT, Rice M, Cormier C, Quinsey V: The Violence Prediction Scheme. Toronto: Centre of Criminology, University of Toronto, 1994
30. Webster CD, Douglas KS, Eaves E, Hart SD: HCR-20: Assessing Risk for Violence (ver 2). Vancouver: Mental Health, Law, and Policy Institute, Simon Fraser University, 1997
31. Douglas KS, Ogloff JRP, Nicholls TL, Grant I: Assessing risk for violence among psychiatric patients: the HCR-20 risk assessment scheme and the Psychopathy Checklist: screening version. J Consult Clin Psychol 61:917–30, 1999
32. Borum R: Improving the clinical practice of violence risk assessment: technology, guidelines, and training. Am Psychol 51:945–56, 1996
33. Steadman HJ, Silver E, Monahan J, *et al*: A classification tree approach to the development of actuarial violence risk assessment tools. Law Hum Behav 24:83–100, 2000
34. Mossman D, Hart KJ: How bad is civil commitment?—a study of attitudes toward violence and involuntary hospitalization. Bull Am Acad Psychiatry Law 21:181–94, 1993
35. Swanson JW, Holzer CE III, Ganju VK, Jono RT: Violence and psychiatric disorder in the community: evidence from the epidemiologic catchment area surveys. Hosp Community Psychiatry 41:761–70, 1990
36. Morse SJ: Craziness and criminal responsibility. Behav Sci Law 17:147–64, 1999
37. Dix GE: A legal perspective on dangerousness: current status. Psychiatr Ann 13:243–56, 1983.
38. Swartz MS, Swanson JW, Hiday VA, Borum R, Wagner R, Burns BJ: Taking the wrong drugs: the role of substance abuse and medication noncompliance in violence among severely mentally ill individuals. Soc Psychiatry Psychiatr Epidemiol 33(Suppl 1): S75–S80, 1998
39. Swanson J, Borum R, Swartz M, Hiday V: Violent behavior preceding hospitalization among persons with severe mental illness. Law Hum Behav 23:185–204, 1999
40. Swanson J, Swartz M, Estroff S, Borum R, Wagner R, Hiday V: Psychiatric impairment, social contact, and violent behavior: evidence from a study of outpatient-committed persons with severe mental disorder. Soc Psychiatry Psychiatr Epidemiol 33(Suppl 1):S86–94, 1998
41. Swartz MS, Swanson JW, Hiday VA, Borum R, Wagner HR, Burns BJ: Violence and severe mental illness: the effects of substance abuse and nonadherence to medication. Am J Psychiatry 155:226–31, 1998
42. Steadman HJ, Mulvey EP, Monahan J, *et al*: Violence by people discharged from acute psychiatric inpatient facilities and by others in the same neighborhoods. Arch Gen Psychiatry 55:393–401, 1998
43. Tardiff K, Marzuk PM, Leon AC, Portera L: A prospective study of violence by psychiatric patients after hospital discharge. Psychiatr Serv 48:678–81, 1997
44. Estroff SE, Swanson JW, Lachicotte WS, Swartz M, Bolduc M: Risk reconsidered: targets of violence in the social networks of people with serious psychiatric disorders. Soc Psychiatry Psychiatr Epidemiol 33(Suppl 1):S95–S101, 1998
45. Hiday VA, Scheid-Cook TL: A follow-up of chronic patients committed to outpatient treatment. Hosp Community Psychiatry 40:52–9, 1989
46. Swartz MS, Swanson JW, Wagner HR, Burns BJ, Hiday VA, Borum R: Can involuntary outpatient commitment reduce hospital recidivism?—findings from a randomized trial with severely mentally ill individuals. Am J Psychiatry 156:1968–75, 1999
47. Mueser KT, Bond GR, Drake RE, Resnick SG: Models of community care for severe mental illness: a review of research on case management. Schizophr Bull 24:37–74, 1998
48. Issakidis C, Sanderson K, Teesson M, Johnston S, Buhrich N: Intensive case management in Australia: a randomized controlled trial. Acta Psychiatr Scand 99:360–7, 1999