The Inaccuracy of Clinical and Actuarial Predictions of Dangerous Behavior

authors masked

June 5, 2016

Abstract

The prediction of dangerous and/or violent behavior is important to the conduct of the United States criminal justice system when it makes decisions about restrictions of personal freedom such as preventive detention, forensic commitment, parole, and in some states such as Texas, when to permit an execution to proceed of an individual found guilty of a capital crime. This article discusses the prediction of dangerous behavior both through clinical judgment as well as actuarial assessment. The general conclusion drawn is that for both clinical and actuarial prediction of dangerous behavior, we are far from a level of accuracy that could justify routine use. To support this later negative assessment, two topic areas are emphasized: 1) the MacArthur Study of Mental Disorder and Violence, including the actuarial instrument developed as part of this project (the Classification of Violence Risk

(COVR)), along with all the data collected that helped develop the instrument; 2) the Supreme Court case of *Barefoot v. Estelle* (1983) and the American Psychiatric Association "friend of the court" brief on the (in)accuracy of clinical prediction for the commission of future violence. Although now over three decades old, *Barefoot v. Estelle* is still the controlling Supreme Court opinion regarding the prediction of future dangerous behavior and the imposition of the death penalty in states such as Texas; for example, see *Coble v. Texas* (2011) and the Supreme Court denial of *certiorari* in that case.

Keywords: clinical prediction, actuarial prediction, sensitivity, specificity, positive predictive value, clinical efficiency

An ability to predict and treat dangerous and/or violent behavior in criminal offenders is important to the administration of the criminal justice system in the United States. This prediction might be in the context of preventive detentions, parole decisions, forensic commitments, other legal forms of restriction on personal liberty, and in some states such as Texas, when to allow an execution to proceed of an individual found guilty of a capital offense. Behavioral prediction might rely on clinical judgment (usually through trained psychologists or other medically versed individuals) or by actuarial (statistical) assessments. In any case, concern should be on the accuracy of such predictions, and more pointedly, on the state of clinical and actuarial prediction of dangerous behavior. So, to pose the central question: are we at such a level of predictive accuracy that as a society we can justify the neces-

Table 1: A Generic 2×2 Contingency Table

	Outcome		
	A	$\bar{A}$	Totals
Prediction			
B	n_{BA}	$n_{B\bar{A}}$	n_B
$\bar{B}$	$n_{\bar{B}A}$	$n_{\bar{B}\bar{A}}$	$n_{\bar{B}}$
Totals	n_A	$n_{\bar{A}}$	n

sary false positives that would inappropriately restrict the personal liberty of those who would prove to be neither dangerous or violent, or that would lead to the execution of someone (otherwise serving a sentence of life without parole) who would not be dangerous or violent. Unfortunately, the conclusion reached here is that for both clinical and actuarial prediction of dangerous behavior, we are quite far from a level that could sanction routine use.[1]

Evidence about prediction accuracy can typically be presented in the form of a 2×2 contingency table defined by a cross-classification of individuals according to the events A and $\bar{A}$ (whether the person proved dangerous (A) or not ($\bar{A}$)); and B and $\bar{B}$ (whether the person was predicted to be dangerous (B) or not ($\bar{B}$)):

Prediction:

B (dangerous)

$\bar{B}$ (not dangerous)

Outcome (Post-Prediction):

A (dangerous)

$\bar{A}$ (not dangerous)

Table 1 gives a generic 2×2 contingency table presenting the available evidence on prediction accuracy; arbitrary cell frequencies are indicated using the appropriate subscript combinations of A and $\bar{A}$ and B and $\bar{B}$. There are five common statistics obtainable from the information given in such a 2×2 table that prove central to the evaluation of any diagnostic prediction system:

sensitivity = $P(B|A)$ = the conditional probability of predicting "dangerous" given that the person proves "dangerous" = n_{BA}/n_A;

specificity = $P(\bar{B}|\bar{A})$ = the conditional probability of predicting "not dangerous" given that the person proves "not dangerous" = $n_{\bar{B}\bar{A}}/n_{\bar{A}}$;

positive predictive value = $P(A|B)$ = the conditional probability that the person proves "dangerous" given the prediction of being "dangerous" = n_{BA}/n_B;

negative predictive value = $P(\bar{A}|\bar{B})$ = the conditional probability that the person proves "not dangerous" given the prediction of being "not dangerous" = $n_{\bar{B}\bar{A}}/n_{\bar{B}}$;

accuracy = the proportion of correct predictions = $(n_{BA} + n_{\bar{B}\bar{A}})/n$; this is just the sum of the two main diagonal frequencies in the 2×2 contingency table divided by the total sample size of n.

It is common to call $P(A)$ $(= n_A/n)$ a "base rate" (or "prior") for the *outcome* of being "dangerous" (A); $P(B)$ $(= n_B/n)$ may be called a "selection rate" for the *prediction* of being "dangerous" (B). The joint probability

of being predicted "dangerous" and actually being "dangerous" is denoted by $P(A \text{ and } B)$ $(= n_{AB}/n)$. Predictions and outcomes are considered "independent" whenever $P(A \text{ and } B)$ factors as $P(A)P(B)$ $(= (n_A/n)(n_B/n))$. Or equivalently, when $P(A|B) = P(A)$; that is, when the "positive predictive value," $P(A|B)$, is equal to the "base rate," $P(A)$. Generally, it is hoped that predictions are associated positively to outcomes and that $P(A|B) > P(A)$. The real question is how large the absolute difference should be between $P(A|B)$ and $P(A)$ to justify using the predictions in a criminal justice setting.

To legitimize the use of the term "probability" for the various expressions just defined, it is convenient to assume the operation of an "urn model." Here, there is a collection of n balls placed in a container; each ball is labeled A or $\bar{A}$, and also B or $\bar{B}$ according to the notationally self-evident table of frequencies just given. The sampling process considered is one of picking a ball blindly from the container, where the balls are assumed to be mixed thoroughly, and noting the occurrence of the events A or $\bar{A}$ and B or $\bar{B}$. Based on this physical idealization of such a selection process, it is intuitively reasonable to then assign probabilities according to the proportion of balls in the container satisfying the attendant conditions for sensitivity, the positive predictive value, and so on.

Clinical Prediction

The 2×2 contingency table given in Table 2 illustrates the typically poor prediction of dangerous behavior when based on clinical assessment. These data are from Kozol, Boucher, and Garofalo (1972), "The Diagnosis and Treatment of Dangerousness." Here, 2 out of 3 predictions of "dangerous" are wrong ($.65 = 32/49$ to be precise). We might note that this figure is the main source for the "error rate" reported so prominently in the landmark Supreme Court opinion(s) in *Barefoot v. Estelle* (1983). Also, about 1 out of 12 predictions of "not dangerous" are wrong ($.08 = 31/386$). Because the frequencies of actual "dangerous" behavior ($P(A) = 48/435 \approx .11$) and the prediction of "dangerous" behavior ($P(B) = 49/435 \approx .11$) are close in value, both the sensitivity and positive predictive values are about .35, and both the specificity and negative predictive values are about .92. This suggests that although the clinical prediction of being "not dangerous" may be pretty good (92% of such predictions are correct), the clinical prediction of "dangerous" is not (only 35% of such predictions are correct). Obviously, we do much better in predicting "not dangerous" than in predicting "dangerous," which will be the situation seen generally throughout this article.[2]

In his dissent opinion in the *Barefoot v. Estelle* case, Justice Blackmun quotes the American Psychiatric Association (APA) *amicus curiae* brief as follows: "[the] most that can be said about any individual is that a history of past violence increases the probability that future violence will occur."[3]

Table 2: A 2×2 Contingency Table for Predicting Dangerous Behavior By Clinical Assessment (Kozol et al. 1972)

	Outcome		
	A	$\bar{A}$	Totals
Prediction			
B	17	32	49
$\bar{B}$	31	355	386
Totals	48	387	435

Although "past violence" (B) might be associated with an increase in the probability of "future violence" (A), the error in that prediction can be very large, as it is here for the Kozol et al. data. Explicitly, the conditional probability of an outcome of "dangerous" given the prediction of "dangerous" ($P(A|B) = 17/49 = .35$) may be greater than the marginal probability of a future outcome of "dangerous" by itself ($P(A) = 48/435 = .11$), but this still implies that 2 out of 3 such predictions of "dangerous" are wrong. To us, the accuracy of behavioral prediction at this level is insufficient to justify any routine incarceration and/or death penalty policies based on these predictions; the same conclusion will hold for the actuarial prediction of "dangerous" discussed in a section to follow.

Clinical Efficiency

Some sixty years ago, Meehl and Rosen (1955) defined a notion of "clinical efficiency" to be when a diagnostic test is more accurate than just predicting using base rates (or alternatively worded, just "betting the base rates"). For

these Kozol et al. data, prediction by base rates would be to say everyone will be "not dangerous" because the number of people who are "not dangerous" (387) is larger than the number of people who are "dangerous" (48). Here, we would be correct in our predictions 89% of the time (.89 = 387/435). Based on clinical prediction, we would be correct a *smaller* percentage of the time (an accuracy of .86 = (17 + 355)/435). So, according to the Meehl and Rosen characterization, clinical prediction is *not* "clinically efficient" because one can do better by just predicting according to base rates.

A simple condition discussed at length in Bokhari and Hubert (2015) (which is originally attributed to Robyn Dawes (1962)), points to a minimal condition that a diagnostic "test" should probably satisfy (and which leads to prediction with the test being clinically efficient; that is, being better than prediction according to base rates. Explicitly, the condition is for the positive predictive value (PPV) to be greater than 1/2. If this minimal condition doesn't hold, as it doesn't here for the Kozol et al. data where the PPV has a value of .35, it will be more likely that a person is "not dangerous" than they are "dangerous" when the test actually predicts "dangerous." This is such an unusual situation that it has been referred to as the "false positive paradox," because false positive tests are more likely than true positive tests.

One possible (and maybe the only) justification for using a clinically inefficient test is to broaden the criterion being optimized by assigning unequal costs to the commission of false positives and false negatives. Thus, for a "test" to be "generalized clinically efficient," the total cost of using it must

be less than the total cost of just predicting by the base rates. For a discussion of how this generalization might be carried out and the restrictive conditions and bounds on how unequal costs can be assigned, see Bokhari and Hubert (2015).[4]

In commenting on the Kozol, et al. study, Monahan (1973) takes issue with the article's principal conclusion that "dangerousness can be reliably diagnosed and effectively treated" and notes that it "is, at best, misleading and is largely refuted by their own data." Mohahan concludes his critique with the following quotation from Wenk, Robison, and Smith (1972):

> Confidence in the ability to predict violence serves to legitimate intrusive types of social control. Our demonstration of the *futility* of such prediction should have consequences as great for the protection of individual liberty as a demonstration of the utility of violence prediction would have for the protection of society. (p. 402)

Actuarial Prediction

Paul Meehl in his iconic 1954 monograph, *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*, created quite a stir with his convincing demonstration that mechanical methods of data combination, such as multiple regression, outperform (expert) clinical prediction. The enormous amount of literature produced since the appearance of

this seminal contribution has uniformly supported this general observation; similarly, so have the extensions suggested for combining data in ways other than by multiple regression, for example, by much simpler unit weighting schemes, or those using other prior weights. It appears that individuals who are conversant in a field are better at selecting and coding information than they are at actually integrating it. Combining such selected information in a more mechanical manner will generally do better than the person choosing such information in the first place.[5]

The MacArthur Study of Mental Disorder and Violence

The MacArthur Research Network on Mental Health and the Law was created in 1988 by a major grant to the University of Virginia from the John D. and Catherine T. MacArthur Foundation. The avowed aim of the Network was to construct an empirical foundation for the next generation of mental health laws, assuring the rights and safety of individuals and society. New knowledge was to be developed about the relation between the law and mental health; new assessment tools were to be developed along with criteria for evaluating individuals and making decisions affecting their lives. The major product of the Network was the MacArthur Violence Risk Assessment Study; its principal findings were published in the very well-received 2001 book, *Rethinking Risk Assessment: The MacArthur Study of Mental Disorder and Violence* (John Monahan, et al., Oxford University Press). More importantly for us (and as a source of several illustrations given later), the

complete data set (on 939 individuals over 134 risk factors) has been made available on the web.[6]

The major analyses reported in *Rethinking Risk Assessment* are based on constructed classification trees; in effect, these are branching decision maps for using risk factors to assess the likelihood that a particular person will commit violence in the future. All analyses were carried out with an SPSS classification-tree program, called CHAID, now a rather antiquated algorithm (the use of this method without a modern means of cross-validation most likely led to the overfitting difficulties to be noted shortly). Moreover, these same classification tree analyses have been incorporated into a proprietary software product called the Classification of Violence Risk (COVR); it is available from the Florida-based company PAR (Psychological Assessment Resources). The program is to be used in law enforcement/mental health contexts to assess "dangerousness to others," a principal standard for inpatient or outpatient commitment or commitment to a forensic hospital.

One of the authors of the current article taught a graduate class entitled Advanced Multivariate Methods for the first time in the Fall of 2011, with a focus on recent classification and regression tree methods developed over the last several decades and implemented in the newer environments of MATLAB and R (but not in SPSS).[7] These advances include the use of "random forests," "bootstrap aggregation (bagging)," "boosting algorithms," "ensemble methods," and a number of techniques that avoid the dangers of overfitting and allow several different strategies of internal cross-validation.

To provide interesting projects for the class to present, a documented data set was obtained from the statistician on the original MacArthur Study; this was a transparent (re)packaging of the data already available on the web. This "cleaned-up" data set could provide a direct replication of the earlier SPSS analyses (with CHAID); but in addition and more importantly, all of the "cutting-edge" methods could now be applied that were unavailable when the original MacArthur study was completed in the late 1990s. At the end of the semester, five subgroups of the graduate students in the class reported on analyses they had done with the MacArthur data set and the prediction of the violence variable (each subgroup also had a different psychological test battery to focus on, for example, Brief Psychiatric Rating Scale, Novaco Anger Scale, Novaco Provocation Inventory, Big Five Personality Inventory, Psychopathy Checklist (Screening Version)). Every one of the talks essentially reported a "wash-out" when cross-validation and prediction was the emphasis as opposed to just constructing the classification structures in the first place. One could not do better than just predicting with base rates. This was a first indication that the prediction of "dangerous" was possibly not as advanced as the MacArthur Network might have us believe.[8]

A second major indication of a difficulty with prediction with the newer MacArthur assessment tools was given by the first cross-validation study done to justify the actuarial software COVR (mentioned earlier): "An Actuarial Model of Violence Risk Assessment for Persons with Mental Disorders" (John Monahan, et al., *Psychiatric Services*, 2005, *56*, 810–815). Portions of the

abstract for this article follow:

> Objectives: An actuarial model was developed in the MacArthur Violence Risk Assessment Study to predict violence in the community among patients who have recently been discharged from psychiatric facilities. This model, called the multiple iterative classification tree (ICT) model, showed considerable accuracy in predicting violence in the construction sample. The purpose of the study reported here was to determine the validity of the multiple ICT model in distinguishing between patients with high and low risk of violence in the community when applied to a new sample of individuals.
>
> Methods: Software incorporating the multiple ICT model was administered with independent samples of acutely hospitalized civil patients. Patients who were classified as having a high or a low risk of violence were followed in the community for 20 weeks after discharge. Violence included any battery with physical injury, use of a weapon, threats made with a weapon in hand, and sexual assault.
>
> Results: Expected rates of violence in the low- and high-risk groups were 1 percent and 64 percent, respectively. Observed rates of violence in the low- and high-risk groups were 9 percent and 35 percent, respectively, ... These findings may reflect the

"shrinkage" expected in moving from construction to validation samples.

Conclusions: The multiple ICT model may be helpful to clinicians who are faced with making decisions about discharge planning for acutely hospitalized civil patients.

John Monahan in his influential NIMH monograph, *The Clinical Prediction of Violent Behavior* (1977), observed that, even allowing for possible distortions in the research data, "it would be fair to conclude that the 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 thus had high base rates for it) and who were diagnosed as mentally ill." In other words, predictions that someone will be violent (and therefore subject to detention) will be wrong two out of three times. With such a dismal record of clinical prediction, there were high expectations that the MacArthur Network could produce a much better (actuarial) instrument in COVR. Unfortunately, that does not appear to be the case. The figures of 64% and 35% given in the abstract suggest two conclusions: in the training sample, the error in predicting dangerousness is 1 out of 3; whether this shows "considerable accuracy in predicting violence in the construction sample" is debatable, even assuming this inflated value is correct. The cross-validated proportion of 35% gives the error of being

Table 3: A 2×2 Contingency Table for Predicting Dangerous Behavior From the COVR Validation Study (Monahan et al. 2005)

	Outcome		
	A	$\bar{A}$	Totals
Prediction			
B	19	36	55
$\bar{B}$	9	93	102
Totals	28	129	157

wrong in the prediction of "dangerous" as 2 out of 3, the same value as for the clinical prediction data of Kozol, Boucher, and Garofalo (1972) discussed earlier. It is an understatement to then say: "These findings may reflect the "shrinkage" expected in moving from construction to validation samples." What it reflects is that actuarial prediction of violence is exactly as bad as clinical prediction. This may be one of the few examples in the behavioral science literature in which actuarial prediction doesn't do better than clinical prediction.

The complete 2×2 table from the COVR validation study is given in Table 3. As noted above, a prediction of "dangerous" is wrong 65% ($= 36/55$) of the time. A prediction of "not dangerous" is incorrect 9% ($= 9/102$) of the time (again, this is close to the 1 out of 12 incorrect predictions of "not dangerous" typically seen for purely clinical predictions). The accuracy is $(19 + 93)/157 = .71$). If everyone were predicted to be "not dangerous," we would be correct 129 out of 157 times, the base rate for $\bar{A}$: $P(\bar{A}) = 129/157 = .82$. Obviously, the accuracy of prediction using base rates (82%) is better than

for the COVR (71%), making the COVR not "clinically efficient" according to the Meehl and Rosen terminology.

Diagnostic Test Evaluation Generally

In the assessment of how well a diagnostic test performs, the statistics of sensitivity, specificity, accuracy, and the positive and negative predictive values (the PPV and the NPV) obviously play central roles. For example, by simply noting whether the PPV is greater than 1/2, an immediate determination as to "clinical efficiency" can be made and whether the test outperforms simple base rate prediction. Judging by the literature on the prediction of dangerous and/or violent behavior, the preferred measure of prediction accuracy now seems to be the "area under the curve" (the AUC). The curve being referred to here is the Receiver Operating Characteristic (ROC) curve of a diagnostic test, which is a plot of test sensitivity (the probability of a "true" positive) against 1.0 minus test specificity (the probability of a "false" positive) over different possible "cutscores" that might be used to reflect differing thresholds for a negative or a positive decision. When there is a single 2×2 contingency table (that is, a single cutscore), the ROC plot would be based on a single point, and the AUC is the simple average of sensitivity and specificity.

For the COVR validation study, we have a sensitivity of .68 (= 19/28) and a specificity of .72 (= 93/129), giving an AUC of .70, which is a number that appears to be about the norm for actuarial instruments considered useful

in the prediction of dangerous behavior. Note that this is in conjunction with a PPV of .35 (= 19/55) and an NPV of .91 (= 93/102); so, 2 out of 3 predictions of "dangerous" are incorrect, and the COVR is not "clinically efficient." But again it might be emphasized that we do much better in predicting "not dangerous" than in predicting "dangerous."

Although the AUC may be the "currently in vogue" means to assess predictive accuracy, a sole reliance on it is flawed for several reasons. First, whenever the base rate for the condition being assessed is relatively low (as it is for "dangerous" behavior), the AUC is not a good measure for conveying the adequacy of the actual predictions made from a diagnostic test. The AUC only evaluates the test itself and not how the test performs when used on specific populations with differing base rates for the presence or absence of the condition being assessed.

The use of the AUC as a measure of diagnostic value can be misleading in assessing conditions with unequal base rates, such as being "dangerous." This misinformation can be further compounded when AUC measures become the basic data subjected to a meta-analysis. Our general suggestion is to rely on some function of the positive and negative predictive values to evaluate a diagnostic test. These measures incorporate both specificity and sensitivity as well as the base rates in the sample for the presence or absence of the condition under study.

In contrast to some incorrect understandings in the literature about the invariance of specificity and sensitivity across samples, sizable subgroup vari-

ation can be present in the sensitivity and specificity values for a diagnostic test; this is called "spectrum bias" and is discussed thoroughly by Ransohoff and Feinstein (1978). Also, sensitivities and specificities are subject to a variety of other biases that have been known for some time (for example, see Begg, 1971). In short, because ROC measures are generally *not* invariant across subgroups, however formed, we do not agree with the sentiment expressed in the otherwise informative review article by John A. Swets, Robyn M. Dawes, and John Monahan, "Psychological Science Can Improve Diagnostic Decisions," *Psychological Science in the Public Interest* (2000, *1*, 1–26). We quote:

These two probabilities [sensitivity and specificity] are independent of the prior probabilities (by virtue of using the priors in the denominators of their defining ratios). The significance of this fact is that ROC measures do not depend on the proportions of positive and negative instances in any test sample, and hence, generalize across samples made up of different proportions. All other existing measures of accuracy vary with the test sample's proportions and are specific to the proportions of the sample from which they are taken.

A particularly pointed critique of the sole reliance on specificity and sensitivity (and thus on the AUC) is given in an article by Karel Moons and Frank Harrell (*Academic Radiology*, *10*, 2003, 670–672), entitled "Sensitivity and Specificity Should Be De-emphasized in Diagnostic Accuracy Studies." We give several telling paragraphs from this article below:

... a single test's sensitivity and specificity are of limited value to clinical practice, for several reasons. The first reason is obvious. They are reverse probabilities, with no direct diagnostic meaning. They reflect the probability that a particular test result is positive or negative given the presence (sensitivity) or absence (specificity) of the disease. In practice, of course, patients do not enter a physician's examining room asking about their probability of having a particular test result given that they have or do not have a particular disease; rather, they ask about their probability of having a particular disease given the test result. The predictive value of test results reflects this probability of disease, which might better be called "posttest probability."

It is well known that posttest probabilities depend on disease prevalence and therefore vary across populations and across subgroups within a particular population, whereas sensitivity and specificity do not depend on the prevalence of the disease. Accordingly, the latter are commonly considered characteristics or constants of a test. Unfortunately, it is often not realized that this is a misconception.

Various studies in the past have empirically shown that sensitivity, specificity, and likelihood ratio vary not only across different populations but also across different subgroups within particular populations.

...

Since sensitivity and specificity have no direct diagnostic meaning and vary across patient populations and subgroups within populations, as do posttest probabilities, there is no advantage for researchers in pursuing esti-

mates of a test's sensitivity and specificity rather than posttest probabilities. As the latter directly reflect and serve the aim of diagnostic practice, researchers instead should focus on and report the prevalence (probability) of a disease given a test's result – or even better, the prevalence of a disease given combinations of test results.

Finally, because sensitivity and specificity are calculated from frequencies present in a 2×2 contingency table, it is always best to remember the possible operation of Berkson's bias (or fallacy)—the relationship that may be present between two dichotomous variables in one population may change dramatically for a selected sample based on some other variable or condition, for example, hospitalization, being a volunteer, age, economic status, and so on.

Several Examples Using Data From the MacArthur Study

We give three more examples using data from the MacArthur Study that involve variables generally thought to be "good" predictors of "dangerous" behavior. Two of the variables are "prior arrest" and "prior violence," and are considered here as diagnostic "tests" in their own right. The third uses data from an explicit diagnostic instrument, the Psychopathy Checklist, Screening Version (PCL:SV).

First, adopting prior arrest as a diagnostic "test": dangerous—one or more prior arrests (B); not dangerous—no prior arrests ($\bar{B}$), we have the contingency table in Table 4. Here, 3 out of 4 predictions of "dangerous"

Table 4: A 2×2 Contingency Table for Predicting Dangerous Behavior From Prior Arrest (Data From the MacArthur Risk Assessment Study)

	Outcome		
	A	$\bar{A}$	totals
Prediction			
B	103	294	397
$\bar{B}$	39	354	393
totals	142	648	790

are wrong (.74 = 294/397); 1 out of 10 predictions of "not dangerous" are wrong (.10 = 39/393). The accuracy of the test is (103 + 354)/790 = .58, and the correctness of prediction by base rates is 648/790 = .82; thus, "prior arrest" is not a clinically efficient "test." Although "prior arrest" is not clinically efficient, this does not imply that it is therefore independent of being "dangerous." Or to say this differently, the positive predictive value ($P(A|B) = 103/397 = .26$) is not equal to the marginal prior probability ($P(A) = 142/790 = .18$) as strict independence would require; however, the magnitude of $P(A|B)$ in relation to $P(A)$ (a difference of only .08) is not really large enough to justify the use of the "test" as a prediction system for "dangerous" behavior.

Second, using prior violence as a diagnostic "test": dangerous—prior violence (B); not dangerous—no prior violence ($\bar{B}$), we obtain Table 5. In this case, 7 out of 10 predictions of "dangerous" are wrong (.69 = 106/154); 1 out of 6 predictions of "not dangerous" are wrong (.16 = 128/785). The accuracy of the test, (48 + 657)/939 = .75, is less than the the correctness

Table 5: A 2×2 Contingency Table for Predicting Dangerous Behavior From Prior Violence (Data From the MacArthur Risk Assessment Study)

	Outcome		
	A	$\bar{A}$	Totals
Prediction			
B	48	106	154
$\bar{B}$	128	657	785
Totals	176	763	939

of prediction by base rates: $763/939 = .81$; thus, "prior violence" is not a clinically efficient "test" either. Just as for "prior arrest," "prior violence" is most likely not independent of being "dangerous" (that is, $P(A|B) = 48/154 = .31 > P(A) = 176/939 = .19$). Although the difference of .12 between $P(A|B)$ and $P(A)$ is a slight increase over that for the "prior arrest" variable, it is still not really large enough to form a defensible prediction system.

The Psychopathy Checklist, Screening Version (PCL:SV) is supposedly the single "best" variable for the prediction of violence based on the data from the MacArthur Risk Assessment Study. It consists of twelve items, each scored 0, 1, or 2 during the course of a structured interview. The items are identified below by short labels:

1) Superficial; 2) Grandiose; 3) Deceitful; 4) Lacks Remorse; 5) Lacks Empathy; 6) Doesn't Accept Responsibility; 7) Impulsive; 8) Poor Behavioral Controls; 9) Lacks Goals; 10) Irresponsible; 11) Adolescent Antisocial Behavior; 12) Adult Antisocial Behavior

Table 6: A 2×2 Contingency Table for Predicting Dangerous Behavior From the Psychopathy Check List: Screening Version (Data From the MacArthur Risk Assessment Study)

	Outcome		
	A	$\bar{A}$	Totals
Prediction			
B	29	26	55
$\bar{B}$	130	675	805
Totals	159	701	860

The total score on the PCL:SV ranges from 0 to 24, with higher scores supposedly more predictive of being dangerous and/or violent.

Table 6 is based on PCL:SV data from the MacArthur Risk Assessment Study, using a cutscore of 18; that is, when above or at the cutscore, predict "violence"; when below the cutscore, predict "nonviolence." The basic statistics for the various diagnostic test results at that cutscore are given below:

accuracy: $(29 + 675)/860 = 704/860 = .819 \approx .82$ (which is slightly better than using base rates)

base rate: $701/860 = .815 \approx .82$

sensitivity: $29/159 = .18$

specificity: $675/701 = .96$

AUC $= (.18 + .96)/2 = .57$

positive predictive value: $29/55 = .53$

negative predictive value: $675/805 = .84$

Because the PPV is .53 (and slightly greater than 1/2), the PCL:SV is clin-

ically efficient at a cutscore of 18; but it is barely so and only in the third decimal place: the accuracy is .819 which is slightly larger than the base rate of .815.

Across the various 2×2 contingency tables that have been given, the same general pattern is present in the PPV and NPV values. There are high NPV values suggesting an enhanced ability to predict "not dangerous"; but values of PPV at about one-half or less that indicate a poor ability to predict "dangerous" behavior; these "tests" are generally not clinically efficient and show the "false positive paradox." This is an unfortunate conclusion for the use of such predictions in the criminal justice contexts because it is just these later predictions of "dangerous" behavior that need to be highly accurate to justify the limitation on personal liberty and freedom, and in some states the imposition of the death penalty.

Searching for Accurate Predictions of Dangerous Behavior: Is it a Fool's Errand?

Given the data presented earlier in this article (that is, from Kozol et al. 1972, the MacArthur Study, and the COVR cross-validation), and also from a number of meta-analytic studies such as the one cited in the first endnote, it might be obvious to state that the accurate prediction of dangerous behavior is difficult, whether done actuarially or clinically.[9] The general reasons for this difficulty are not hard to find: the behavior being predicted has a rather

low base rate, and the clinical and/or actuarial mechanisms available to make these predictions are too fallible to justify routine use in a criminal justice context. We will attempt to expand on this general conclusion in this section with the use of Bayes' theorem defined in the context of a 2×2 contingency table containing the frequencies of "prediction" and "outcome."

Consider the prediction of "dangerous" (B) or "not dangerous" ($\bar{B}$), and the outcome of "dangerous" (A) or "not dangerous" ($\bar{A}$). Two forms of Bayes' theorem are of interest; first, a compact form:

$$P(A|B) = P(B|A)\left(\frac{P(A)}{P(B)}\right)$$

In words, the positive predictive value (PPV), $P(A|B)$, is the sensitivity, $P(B|A)$, times the ratio of the prior base rate for the outcome, $P(A)$, to the selection rate for the prediction, $P(B)$. Second, there is a more expansive rewriting using an alternative expression for $P(B)$ [$= P(B|A)P(A) + (1 - P(\bar{B}|\bar{A})(1 - P(A))$]:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + (1 - P(\bar{B}|\bar{A})(1 - P(A))}$$

In the discussion below, we will use some convenient shorthand notation: "sens" for "sensitivity," "spec" for "specificity," "prior" for $P(A)$, "select" for $P(B)$, "PPV" for the "positive predictive value," and "NPV" for the "negative predictive value." Based on these abbreviations, the second longer form of Bayes' theorem gets rewritten as PPV = sens $\times$ prior / (sens $\times$ prior

+ (1-spec) $\times$ (1-prior)).

The minimal condition that we believe is necessary for a test to have justifiable value in the criminal justice setting, is for it to be clinically efficient and thus outperform simple base rate prediction (which would amount to predicting consistently the higher base rate condition of "not dangerous"). We will assume therefore that the costs of a false positive and a false negative are equal; this is a reasonable beginning assumption given the broad definitions of "violence" typically used, which are not restricted to instances of physical injury and/or death. For a mechanism to extend this discussion to unequal costs, see Bokhari and Hubert (2015),

Given the minimal condition of clinical efficiency for a diagnostic mechanism to be of value, there is also the simple equivalence of the PPV being greater than 1/2. In other words, when the PPV is greater than 1/2, the "false positive paradox" is avoided and it is more likely that the person proves "dangerous" than "not dangerous" when the test actually predicts "dangerous." It is noted that the PPV is less than 1/2 for all the data sets given earlier except when the cutscore of 18 is used for the PCL:SV; here, the PPV is .53, with the test accuracy of .819 minimally outperforming the base rate prediction accuracy of .815 by .004.

Based on the minimal condition of the PPV being greater than 1/2 and using the first compact form of Bayes' theorem, the condition can be restated as

$$P(B|A)\left(\frac{P(A)}{P(B)}\right) \geq \frac{1}{2}$$

or as, sens $\times$ (prior/select) $\geq 1/2$. Thus, the larger the selection proportion is in relation to the prior proportion, the less the PPV. In fact, when the selection proportion is more than twice the prior base rate, the PPV condition *must* fail since the sensitivity is always less than or equal to 1.0. The diagnostic "test" of "prior arrest" exemplifies this kind of mandatory failure of the PPV condition.

Judging from the literature on actuarial instruments for risk assessment (for example COVR (Classification of Violence Risk), VRAG (Violence Risk Appraisal Guide), HCR-20 (Historical Clinical Risk-20), among others), reported values for the AUC are about .75 and lower; reported levels of the prior proportion of "dangerous" are about .20 (as it is in the MacArthur Study). Using the second form of Bayes' theorem, a simple condition can be derived for when the PPV is greater than or equal to 1/2: sens $\times$ prior $\geq$ (1-spec) $\times$ (1-prior). As shown below, there are several ways this last result can be used to demonstrate the difficulty of obtaining a clinically efficient diagnostic test for assessing a low base rate behavior (such as being "dangerous") with the usual levels of fallibility present in commonly available risk assessment instruments.[10]

1) If we assume for convenience that sens = spec, then it must be true that sens $\geq$ (1-prior) and spec $\geq$ (1-prior). Remembering that the AUC for one cutscore is simply the average of the sensitivity and specificity, a prior of .20 and sensitivity and specificity values of .75 would lead to a failure of the PPV condition. In other words, for a commonly seen base rate for "dangerous"

and for a typical AUC value seen in practice, the various instruments are not clinically efficient.

2) Again, assuming a prior of .20 and an AUC of .75 (so the average of sensitivity and specificity is .75), for the PPV to be greater than or equal to 1/2, the specificity would need to be greater than or equal to 5/6; the sensitivity would need to be less than or equal to 2/3 (but greater than or equal to 1/2 since the specificity cannot be larger than 1.0). In other words, in the presence of a low base rate condition, clinical efficiency requires a *very high* specificity value along with particular upper and lower bounds on the sensitivity value.

So, to answer the question posed in this section's heading: for the generally low base rates present for "violence" or "dangerous" and the level of fallibility present in the available instruments in terms of sensitivity and specificity, the search for an accurate risk assessment instrument that might be used routinely in criminal justice contexts, may indeed be a fool's errand.

Some Concluding Comments

The American Psychological Association (APA) in 2011 filed a "friend of the court brief" with the U.S. Supreme Court in the case of *Billy Wayne Coble v. State of Texas*. The brief asked the Supreme Court to grant Mr. Coble's petition for *certiorari* (that is, for review of a lower court decision) based partly on the "unreliability" of testimony as to the risk of "future

dangerousness" given by a forensic psychiatrist, Dr. Richard Coons. The Supreme Court denied *certiorari* in June of 2011, citing *Barefoot v. Estelle* as justification for the denial. Billy Wayne Coble is currently an inmate on the Texas death row, waiting for his execution date to be set.[11]

The APA brief makes three assertions about predicting future dangerousness, two of which we would affirm and one with which we would disagree and instead assert that the facts would suggest otherwise. The two we would affirm:

Unstructured clinical testimony like that at issue is not based on science and should not be relied upon to establish future dangerousness.

Unstructured clinical risk-assessment testimony is unduly persuasive to juries.

The one that we would question:

In contrast to Dr. Coon's unstructured approach, structured risk-assessment methods are scientifically based and can reliably inform assessments of future dangerousness in a variety of contexts.

What is at issue is the term "reliably," which we would interpret as meaning "accurately." Given the empirical literature on structured risk assessment instruments and with AUC values around .75 and prior base rates of about .20, positive predictive values would generally be less than 1/2 (as we have noted earlier); that is, in the prediction of "dangerous" we would more likely to be wrong than right—the "false positive paradox" arises. To us, this is not a reliable means to "inform assessment of future dangerousness in a variety

of contexts."

In choosing between clinical and actuarial prediction strategies, preference should be tilted toward actuarial strategies, if only to avoid the capricious judgment of individuals such as Coons who offer no scientifically justified bases at all for their predictions. But that is not then to say that the various actuarial mechanisms available are at a level of accuracy that could automatically justify their routine use in imposing severe restrictions on personal liberty (or for allowing executions to proceed, as in Texas). One problem is in the ambiguity as to what constitutes "dangerous" behavior and how encompassing the definition is (for example, verbal threats made with some makeshift weapon in hand but with no attendant physical injury or touching, are routinely considered to be "violent"). Another difficulty is in the "number needed to detain" (NND) to prevent one occurrence of "dangerous" behavior. As a definition of NND, we simply take the reciprocal of the PPV. For example, given the Kozol, et al. data, where the PPV is .35, about three people ($3 \approx 1/.35$) would need to be detained (indefinitely?) to have an expectation of preventing one act of "violence," however that term might be defined. Possibly, it is best to remember Sir William Blackstone's adage from the *Commentaries on the Laws of England* (1765): "It is better that ten guilty escape than one innocent suffer."

As a final point, when assertions are made as to the "reliability" (or better, the "accuracy") of risk assessment, those assertions should be more than just noting that the conditional probability of "dangerous" given a

positive prediction of "dangerous" (that is, the PPV) is greater than the prior probability of "dangerous." An assertion of "reliability" must be made based on the size of the PPV (and of the NPV, as well), with the minimal condition on the PPV being greater than 1/2, although the closer to 1.0, the better. Unfortunately, consideration of the PPV is often hidden by the predominant use of the AUC measure. The empirical evaluation of a risk assessment instrument in practice should be grounded in its performance in the various groups to which it will be applied. Mere assertions of "reliability" based on the typical values seen for the AUC are not enough.

References

Begg, C. B. (1987) Bias in the assessment of diagnostic tests. *Statistics in Medicine*, *6*, 411–423.

Bokhari, E. and Hubert, L. (2015) A new condition for assessing the clinical efficiency of a diagnostic test. *Psychological Assessment*, *27*, 745–754.

Bokhari, E. and Hubert, L. (in press) The lack of cross-validation can lead to inflated results and spurious conclusions: A re-analysis of the MacArthur Violence Risk Assessment Study. *Journal of Classification*, in press.

Dawes, R. M. (1962) A note on base rates and psychometric efficiency. *Journal of Consulting Psychology*, *26*, 422–424.

Dawes, R. M. (2005) The ethical implications of Paul Meehl's work on comparing clinical versus actuarial prediction methods. *Journal of Clinical Psychology*,

61, 1245–1255.

Kozol, H. L., Boucher, R. J. and Garofalo, R. (1972) The diagnosis and treatment of dangerousness. *Crime and Delinquency*, *18*, 371–392.

Meehl, P. (1954) *Clinical Versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis, Minnesota: University of Minnesota Press.

Meehl, P. and Rosen, A. (1955) Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin*, *52*, 194–215.

Monahan, J. (1973) Dangerous offenders: A critique of Kozol et al. *Crime and Delinquency*, **19**, 418–420.

Monahan, J. (1977) *The Clinical Prediction of Violent Behavior*. Lanham, Maryland: Jason Aronson, Inc.

Monahan, J., Steadman, H., et al. (2001) *Rethinking Risk Assessment: The MacArthur Study of Mental Disorder and Violence*. New York: Oxford University Press.

Monahan, J., Steadman, H., et al. (2005) An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric Services*, *56*, 810–815.

Moons, K. and Harrell, F. (2003) Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Academic Radiology*, *10*, 670–672.

Perlin, M. L. (2013) *Mental Disability and the Death Penalty*. Lanham, Maryland: Rowman & Littlefield.

Ransohoff, D. F. and Feinstein, R. R. (1978) Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*, *299*, 926–930.

Snowden, R. J., Gray, N. S., Taylor, J. and MacCulloch, M. J. (2007) Actuarial prediction of violent recidivism in mentally disordered offenders. *Psychological Medicine*, *37*, 1539–1549.

Swets, J. A., Dawes, R. M. and Monahan, J. (2000) Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26.

Szmulker, G., Everitt, B. and Leese, M. (2012) Risk assessment and receiver operating characteristic curves. *Psychological Medicine*, *42*, 895–898.

Vrieze, S. I. and Grove, W. (2008) Predicting sex offender recidivism. I. Correcting for item overselection and accuracy overestimation in scale development. II. Sampling error-induced attenuation of predictive validity over base rate information. *Law and Human Behavior*, *32*, 266–278.

Wenk, E. A., Robinson, J. O. and Smith, G. W. (1972) Can violence be predicted? *Crime and Delinquency*, *18*, 393–402.

Notes

[1]There is a large meta-analytic literature on the use of risk assessment instruments that would stongly support this rather disappointing overall conclusion. One of the most comprehensive such studies appeared in the Open Access *British Medical Journal (BJM)* on July 24, 2012, by Seena Fazel, Jay P. Singh, Helen Doll, and Martin Grann ("Use of Risk Assessment Instruments to Predict Violence and Antisocial Behaviour in 73 Samples Involving 24,827 People: Systematic Review and Meta-analysis"). The overall conclusion of this meta-analysis is stated as follows:

Although risk assessment tools are widely used in clinical and criminal justice settings, their predictive accuracy varies depending on how they are used. They seem to identify low risk individuals with high levels of accuracy, but their use as sole determinants of detention, sentencing, and release is not supported by the current evidence. Further research is needed to examine their contribution to treatment and management.

[2]The Supreme Court decision in *Barefoot v. Estelle* (1983) concerns the prediction of "dangerous" behavior. The Court holding in this landmark case was as follows:

There is no merit to petitioner's argument that psychiatrists, individually and as a group, are incompetent to predict with an acceptable degree of reliability that a particular criminal will commit other crimes in the future, and so represent a danger to the community. In effect, the Court held that no matter what the data might show, and for both clinical and actuarial prediction, such predictions of future crime can be made at an acceptable level to be of value in the criminal justice system (and in the Texas context, to permit an execution to proceed, as it did for Thomas Barefoot). Up to the present, the *Barefoot v. Estelle* opinion is controlling whenever issues of behavioral prediction of dangerous or violent behavior come before the court.

[3]A listing of the Blackmun dissent is available at

`http://cda.psych.uiuc.edu/barefoot_majority_opinion_blackmun_dissent.pdf`

The APA brief is at

`http://cda.psych.uiuc.edu/barefoot_apa_brief.pdf`

[4]It is of interest to note that some individuals with vested interests in clinical and/or actuarial prediction strategies, raise the unequal cost issue rather gratuitously whenever commenting on a critique of their preferred prediction method. No justifiable assignment of such unequal costs is ever actually made that would justify an otherwise clinically inefficient prediction strategy. We quote Vrieze and Grove (2008) below; they were confronted by such a reviewer for their critique of an actuarial instrument for predicting sex offender recidivism:

It is true, as pointed out by one reviewer, that a proper decision analysis would consider not only the proportion of erroneous classifications but also the costs associated with these errors—distinguishing positive from negative classification errors as we do not. However and unfortunately, there is no general agreement among decision makers about the relative importance of false positive and false negative prediction errors. Therefore, we followed custom in the statistical discrimination literature and assumed zero-one loss: zero relative cost if no prediction error occurs, versus a cost of one if either a false negative or a false positive prediction error occurs. Absent a legislative clarification of how important is mistakenly committing a low-risk sex offender, as compared to letting a high risk sex offender live at large in the community, we are loath to impose our private evaluations of relative importance. (pp. 271–272)

[5]A 2005 article by Robyn Dawes in the *Journal of Clinical Psychology* (*61*, 1245–1255) has the intriguing title "The Ethical Implications of Paul Meehl's Work on Comparing Clinical Versus Actuarial Prediction Methods." Dawes' main point is that given the overwhelming evidence we now have, it is unethical to use clinical judgment in preference to the use of statistical prediction rules. We quote from the abstract:

> Whenever statistical prediction rules ... are available for making a relevant prediction, they should be used in preference to intuition. ... Providing service that assumes that clinicians "can do better" simply based on self-

confidence or plausibility in the absence of evidence that they can actually do so is simply unethical. (p. 1245)

[6]See `http://www.macarthur.virginia.edu/read_me_file.html`

[7]We note that another author of this current article was, at the time, a graduate student in this same class.

[8]For the data used by the multivariate class, see

`http://cda.psych.uiuc.edu/statistical_learning_course/MacOnline/`

An "in press" paper by Bokhari and Hubert, to appear in the *Journal of Classification*, provides a detailed re-analysis of these same data, with an emphasis on the newer methodologies for cross-validation ("The lack of cross-validation can lead to inflated results and spurious conclusions: A re-analysis of the MacArthur Violence Risk Assessment Study.")

[9]The meta-analytic study referenced in the first endnote gives summary data on the use of various risk assessment instruments for violent offending (VO) (thirty samples) and for sexual offending (SO) (twenty samples). The median positive and negative predictive values (PPV and NPV) and the interquartile ranges (IQR) are given below:

	Median	IQR
PPV:VO	.41	.27–.60
NPV:VO	.91	.81–.95
PPV:SO	.23	.09–.41
NPV:SO	.93	.82–.98

It is clear that most uses of the risk assessment instruments are not clinically efficient; moreover, the pattern holds up of being able to predict "not dangerous" much better than "dangerous."

[10]To give an example of cross-validating a risk assessment instrument other than the COVR, we consider the more well-known VRAG, and show its performance in the usual 2×2 contingency table format for one well-run cross-validation study (see Szmukler, et al. (2012) for another reanalysis of these same data). The table below is from Snowden,

et al. (2007) on a sample of 364 mentally disordered offenders who were followed for two years after discharge from the hospital. (A VRAG score of seven or more (on a nine-point scale) was used to predict the presence of "violence.")

	Outcome		
	A (violent)	$\bar{A}$ (not violent)	Totals
Prediction			
B (violent)	12	48	60
$\bar{B}$ (not violent)	14	290	304
Totals	26	338	364

sensitivity = 12/26 = .46;

specificity = 290/338 = .86;

base rate for "violent" = 26/364 = .07;

accuracy = 302/364 = .83;

base-rate prediction success = 338/364 = .93;

positive predictive value (PPV) = 12/60 = .20;

negative predictive value (NPV) = 290/304 = .95.

So, again, there is a good ability to predict the lack of "violence" (with a NPV of .95), coupled with a dismal performance in predicting the presence of "violence" (with a PPV of .20). Explicitly, the PPV of .20 indicates that 4 out of 5 predictions of "violence" are incorrect. Obviously, the VRAG for this sample is not clinically efficient since the calculated PPV is less than 1/2; also, and as noted earlier using the compact form of Bayes' theorem, because the selection proportion is more than twice the prior base rate, the PPV condition must automatically fail.

[11]For a recent and thorough review of the literature on the prediction of dangerous or violent behavior as it relates to the death penalty, see Michael L. Perlin, *Mental Disability and the Death Penalty: The Shame of the States* (2013; Rowman & Littlefield); Chapter 3 is particularly relevant: "Future Dangerousness and the Death Penalty." A good resource

generally for material on the prediction of dangerous behavior and related forensic matters is the Texas Defender Service (`www.texasdefender.org`), and the publications it has freely available at its web site:

A State of Denial: Texas Justice and the Death Penalty (2000)

Deadly Speculation: Misleading Texas Capital Juries with False Predictions of Future Dangerousness (2004)

Minimizing Risk: A Blueprint for Death Penalty Reform in Texas (2005)