Psychological Assessment
2015, Vol. 27, No. 3, 745–754

1040-3590/15/$12.00 http://dx.doi.org/10.1037/pas0000093

A New Condition for Assessing the Clinical Efficiency of a Diagnostic Test

Ehsan Bokhari and Lawrence Hubert
University of Illinois at Urbana-Champaign

When prediction using a diagnostic test outperforms simple prediction using base rates, the test is said to be "clinically efficient," a term first introduced into the literature by Meehl and Rosen (1955) in *Psychological Bulletin*. This article provides three equivalent conditions for determining the clinical efficiency of a diagnostic test: (a) Meehl-Rosen (Meehl & Rosen, 1955); (b) Dawes (Dawes, 1962); and (c) the Bokhari-Hubert condition, introduced here for the first time. Clinical efficiency is then generalized to situations where misclassification costs are considered unequal (for example, false negatives are more costly than false positives). As an illustration, the clinical efficiency of an actuarial device for predicting violent and dangerous behavior is examined that was developed as part of the MacArthur Violence Risk Assessment Study.

Keywords: clinical efficiency, diagnostic test evaluation, Meehl-Rosen condition, violence prediction

Suppose a diagnostic test is designed to determine whether a person has "it," whatever "it" may be. For example, when our interest is in predicting violence, the test should indicate whether the person will be violent in the future. Let B denote the event that the test is positive indicating the person has "it," and $\bar{B}$, the event that the test is negative indicating that the person does not have "it." Now, consider the events of whether a person truly has "it" or truly does not and denote these two events as A and $\bar{A}$, respectively. The events B and $\bar{B}$ will be called the *diagnostic test results* and the events A and $\bar{A}$, the *states of nature*.

Given the diagnostic test result and state of nature, a 2 × 2 contingency table can be constructed, as shown in Table 1. This table provides the frequencies for marginal events (for example, n_B is the number of people who test positive), or for joint events (for example, n_{BA} is the number of people who have "it" and test positive). In terms of violence prediction, n_B is the number predicted to be violent and n_{BA} is the number predicted to be and who are violent. The frequencies within the table have familiar names worth noting: n_{BA} is the number of *true positives*, $n_{B\bar{A}}$ is the number of *false positives*, $n_{\bar{B}A}$ is the number of *false negatives*, and $n_{\bar{B}\bar{A}}$ is the number of *true negatives*. Of particular importance are the marginal frequencies: n_A, representing the *base frequency* for those who have "it," and $n_{\bar{A}}$, the base frequency for those who do not. In addition, we may be interested in n_B and $n_{\bar{B}}$, the base frequencies for positive and negative diagnostic test outcomes, respectively; these are often called *selection frequencies*.

In addition to frequencies, various marginal, joint, and conditional probabilities can be defined. For example, $P(A) = \frac{n_A}{n}$; $P(A \cap B) = \frac{n_{BA}}{n}$; $P(A|B) = \frac{n_{BA}}{n_B}$; $P(B|A) = \frac{n_{BA}}{n_A}$; and so forth. These conditional probabilities are of general interest, and again it is worth noting their names. Conditionalizing on the state of nature, we have the following: $P(B|A) = \frac{n_{BA}}{n_A}$ is the sensitivity or true positive rate (TPR); $P(B|\bar{A}) = \frac{n_{B\bar{A}}}{n_{\bar{A}}}$ is the false positive rate (FPR); $P(\bar{B}|A) = \frac{n_{\bar{B}A}}{n_A}$ (= 1 – sensitivity) is the false negative rate (FNR); and $P(\bar{B}|\bar{A}) = \frac{n_{\bar{B}\bar{A}}}{n_{\bar{A}}}$ (= 1 – false positive rate) is the specificity or true negative rate (TNR). Conditionalizing on the diagnostic test result, $P(A|B) = \frac{n_{BA}}{n_B}$ is called the positive predictive value (PPV: the probability that the person has "it" given that the test is positive); $P(\bar{A}|\bar{B}) = \frac{n_{\bar{B}\bar{A}}}{n_{\bar{B}}}$ is the negative predictive value (NPV: the probability that the person does not have "it" given that the test is negative). The column marginal probabilities represent the base rates, or prior probabilities, for those who have "it" ($P(A)$) and those who do not ($P(\bar{A})$); the row marginal probabilities represent the selection ratios for those who are predicted to have "it" ($P(B)$) and those who are not ($P(\bar{B})$).

It is important to note the dependency among frequencies (and, consequently, probabilities). For instance, if we know the base and selection frequencies, the distribution of joint frequencies are subject to a single degree of freedom. As another example, given n_A and n_{BA}, $n_{\bar{B}A}$ is not free to vary. Similarly, given $P(\bar{B}|A)$, the laws

This article was published Online First April 20, 2015.

Ehsan Bokhari and Lawrence Hubert, Department of Psychology, University of Illinois at Urbana-Champaign.

Ehsan Bokhari is now at the Department of Statistics, University of Illinois at Urbana-Champaign.

Correspondence concerning this article should be addressed to Ehsan Bokhari, Department of Statistics, University of Illinois at Urbana-Champaign, Champaign, IL 61820. E-mail: bokhari1@illinois.edu

Table 1
A General 2 × 2 Contingency Table

	State of nature		
	A (positive)	$\bar{A}$ (negative)	Totals
Diagnostic test result			
B (positive)	n_{BA}	$n_{B\bar{A}}$	n_B
$\bar{B}$ (negative)	$n_{\bar{B}A}$	$n_{\bar{B}\bar{A}}$	$n_{\bar{B}}$
Totals	n_A	$n_{\bar{A}}$	n

of probability determine that $P(\bar{B}|A) = 1 - P(B|A)$. In other words, given the true and false positive rates, the true and false negative rates are redundant.

Clinical Efficiency

Base rates play an important role in prediction and decision making (Bar-Hillel, 1980; Kahneman & Tversky, 1973; Schwarz, Strack, Hilton, & Naderer, 1991; Faust & Nurcombe, 1989). The phrase "clinical efficiency" refers to prediction by a diagnostic test being better than prediction using just base rates (Meehl & Rosen, 1955). If $P(A) \leq \frac{1}{2}$, then prediction by base rates would be to say consistently that a person does not have "it" because then the probability of a correct prediction is $P(\bar{A}) \geq \frac{1}{2}$ (that is, the prediction is correct at least half the time). Similarly, if $P(A) \geq \frac{1}{2}$, prediction by base rates would be to always say that the person has "it." Prediction according to the diagnostic test is to say that the person has "it" when the test is positive and does not have "it" when the test is negative. To measure how "good" a diagnostic test is, consider the accuracy (or hit rate) of the test defined as

$$\begin{aligned} P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) &= \left(\frac{n_{BA}}{n_A}\right)\left(\frac{n_A}{n}\right) + \left(\frac{n_{\bar{B}\bar{A}}}{n_{\bar{A}}}\right)\left(\frac{n_{\bar{A}}}{n}\right) \\ &= \frac{n_{BA} + n_{\bar{B}\bar{A}}}{n}. \end{aligned}$$

This expression is just the sum of the main diagonal frequencies from a 2 × 2 contingency table (for example, see Table 1) divided by the total number of subjects, n.

Assuming $P(A) \leq \frac{1}{2}$, a general condition can be given for when prediction by a test will be better than prediction by base rates:

$$P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) > P(\bar{A}); \tag{1}$$

in words, when the base rate is at most $\frac{1}{2}$, the test should be used for prediction only if the accuracy of the test is greater than the proportion of the population not having "it."

Using the general condition presented in Equation 1, there are three important (and equivalent) conditions that can be derived for clinical efficiency. All three conditions are relevant to an attempt to predict an event having a typically low base rate by using a test possessing less than ideal sensitivity and specificity values; they characterize the circumstances when more accurate prediction would just be to use the larger base rate (that is, to say the person does not have "it") rather than to rely on the diagnostic test. These three equivalent conditions for base-rate prediction being superior to prediction from the test are attributed to Meehl and Rosen (1955); Dawes (1962), and Bokhari and Hubert; the introduction of this latter condition is the major justification for the current article.

Meehl-Rosen Condition

Assume $P(A) \leq \frac{1}{2}$. The Meehl-Rosen condition (Meehl & Rosen, 1955) states that it is best to use the test over base rates if and only if

$$P(A) > \frac{1 - P(\bar{B}|\bar{A})}{P(B|A) + (1 - P(\bar{B}|\bar{A}))}, \tag{2}$$

or in terms of specificity and sensitivity,

$$P(A) > \frac{1 - \text{specificity}}{\text{sensitivity} + (1 - \text{specificity})}.$$

Because $1 - P(\bar{B}|\bar{A}) = P(B|\bar{A})$, this condition implies that the test should be used for prediction over base rates if and only if the base-rate probability is larger than the ratio of the false positive rate to the sum of the true positive and false positive rates. The proof of the Meehl-Rosen condition can be found in Appendix A.

If $P(A) > \frac{1}{2}$, the Meehl-Rosen condition becomes

$$P(\bar{A}) > \frac{1 - P(B|A)}{P(\bar{B}|\bar{A}) + (1 - P(B|A))} = \frac{1 - \text{sensitivity}}{\text{specificity} + (1 - \text{sensitivity})},$$

and the proof is similar.

Dawes Condition

Assume $P(A) \leq \frac{1}{2}$. The Dawes condition (Dawes, 1962) states that it is best to use the test over base rates if and only if

$$P(\bar{A}|B) < \frac{1}{2}.$$

Equivalently, the Dawes condition can be written as $P(A|B) > \frac{1}{2}$, implying that prediction by the test is better than prediction by base rates if and only if the positive predictive value is greater than $\frac{1}{2}$ (for a proof, see Appendix B). If the positive predictive value is less than $\frac{1}{2}$ (that is, the Dawes condition fails to hold and it is better to just use base rates for prediction rather than the test), it is more likely that a person does not have "it" than they do even if the test says the person has "it." In other words, given a positive test result there is a higher probability that the person does not have "it" than they do. This has been called the "false positive paradox."

If $P(A) > \frac{1}{2}$, the Dawes condition becomes

$$P(\bar{A}|\bar{B}) > \frac{1}{2};$$

in words, when $P(A) > \frac{1}{2}$, the negative predictive value must be greater than $\frac{1}{2}$ for prediction by the test to outperform prediction by base rates. The proof is similar.

Bokhari-Hubert Condition

Assume $P(A) \leq \frac{1}{2}$. We claim that it is better to use the test over base rates if and only if differential prediction holds between the row entries in the contingency table: $n_{BA} > n_{B\bar{A}}$ and $n_{\bar{B}A} < n_{\bar{B}\bar{A}}$. The proof can be found in Appendix C. In words, when the number of true positives (n_{BA}) is greater than the number of false positives ($n_{B\bar{A}}$) and the number of true negatives ($n_{\bar{B}\bar{A}}$) is greater than the number of false negatives ($n_{\bar{B}A}$), prediction using the test is better than prediction by base rates.

This condition requires no probability calculations and can be seen directly in the contingency table—for base rates to be worse than the test, differential prediction must exist. All three conditions are equivalent; if the Bokhari-Hubert (BH) condition holds then the positive predictive value is greater than $\frac{1}{2}$ (because of the Dawes condition). In addition, the BH condition implies the negative predictive value is greater than $\frac{1}{2}$; thus, the BH condition is equivalent to both the positive and negative predictive values being greater than $\frac{1}{2}$. Unlike the Meehl-Rosen and Dawes conditions, when $P(A) > \frac{1}{2}$ the BH condition is exactly the same. Thus, if prediction by the test is better than prediction by base rates, the BH condition holds for any $P(A)$.

Relationship to measures of association. The Goodman-Kruskal lambda coefficient (Goodman & Kruskal, 1954) is a proportional-reduction-in-error measure for predicting a column event (A or $\bar{A}$) from knowledge of a row event (B or $\bar{B}$) over a naïve prediction based solely on marginal column frequencies (n_A and $n_{\bar{A}}$). Thus, the Goodman-Kruskal lambda coefficient can be considered a measure of association between the diagnostic test result and the state of nature. For the 2 × 2 contingency table (for example, Table 1), lambda is defined as:

$$\lambda_{\text{column|row}} = \frac{\max\left\{n_{BA}, n_{B\bar{A}}\right\} + \max\left\{n_{\bar{B}A}, n_{\bar{B}\bar{A}}\right\} - \max\left\{n_A, n_{\bar{A}}\right\}}{n - \max\left\{n_A, n_{\bar{A}}\right\}}.$$

If $\lambda_{\text{column|row}}$ is zero, the maximum of the column marginal frequencies is the same as the sum of the maximum frequencies within rows; therefore, no differential prediction of a column event is made based on knowledge of what particular row an object belongs to. A nonzero $\lambda_{\text{column|row}}$ is an alternative way of specifying the BH differential prediction condition. If $\lambda_{\text{column|row}} = 0$, $\max\{n_{BA},n_{B\bar{A}}\} + \max\{n_{\bar{B}A},n_{\bar{B}\bar{A}}\} = \max\{n_A,n_{\bar{A}}\}$. If $\max\{n_A,n_{\bar{A}}\} = n_A = n_{BA} + n_{\bar{B}A}$, $n_{BA} \geq n_{B\bar{A}}$ and $n_{\bar{B}A} \geq n_{\bar{B}\bar{A}}$, and the condition fails to hold. Similarly, if $\max\{n_A,n_{\bar{A}}\} = n_{\bar{A}} = n_{\bar{B}\bar{A}} + n_{B\bar{A}}$, $n_{B\bar{A}} \geq n_{BA}$ and $n_{\bar{B}\bar{A}} \geq n_{B\bar{A}}$, and again the condition fails to hold.

An alternative and more popular test of association is based on Pearson's chi-squared statistic (Pearson, 1900). Although this test can be used for significance testing in a 2 × 2 contingency table, it says nothing about differential prediction. For instance, this test may show a significant relation between the state of nature (A and $\bar{A}$) and the diagnostic test results (B and $\bar{B}$), but when $\lambda_{\text{column|row}}$ is zero, there is no differential prediction and the use of base rates will outperform the diagnostic test.

Relationship to odds ratio and relative risk. Odds ratios, or relative odds, are another way of measuring association. The odds of an event is defined as the ratio of the probability that a person has "it" to the probability that a person does not have "it," given a specific diagnostic test result:

$$O_B = \frac{P(A|B)}{P(\bar{A}|B)} = \frac{\dfrac{n_{BA}}{n_B}}{\dfrac{n_{B\bar{A}}}{n_B}} = \frac{n_{BA}}{n_{B\bar{A}}}$$

and

$$O_{\bar{B}} = \frac{P(A|\bar{B})}{P(\bar{A}|\bar{B})} = \frac{\dfrac{n_{\bar{B}A}}{n_{\bar{B}}}}{\dfrac{n_{\bar{B}\bar{A}}}{n_{\bar{B}}}} = \frac{n_{\bar{B}A}}{n_{\bar{B}\bar{A}}}.$$

The first term, O_B, gives the odds of a person having "it" when they test positive for having "it"; the second term, $O_{\bar{B}}$, gives the odds that a person has "it" when they do not test positive for "it." In Bayesian terms, the odds can be thought of as posterior odds, given the test result; the prior odds are $\frac{P(A)}{P(\bar{A})}$. The odds ratio is simply the ratio of the two odds, $OR = \frac{O_B}{O_{\bar{B}}}$. Thus, the odds ratio compares which group (B vs. $\bar{B}$) is more likely to have "it." If the BH condition holds, then $n_{BA} > n_{B\bar{A}} \Leftrightarrow O_B > 1$ and $n_{\bar{B}\bar{A}} > n_{\bar{B}A} \Leftrightarrow O_{\bar{B}} < 1$. This means that if the person tests positive for "it," the odds are greater than not that they do have "it"; if the person tests negative for "it," the odds are greater that they do not have "it" than they do. If $O_B > 1$ and $O_{\bar{B}} < 1$, then $OR > 1$. Therefore, the BH condition implies that the odds ratio is greater than one; thus, the odds someone has "it" is greater in the group that tests positive for "it." Of course, none of the entries in the denominators can be zero, but when the BH condition holds, only $n_{B\bar{A}}$ has any possibility of being equal to 0.

Relative risk is the ratio of the probability that a person has "it" given they tested positive for having "it" to the probability that a person has "it" given that they did not test positive for "it." The relative risk is defined as

$$RR = \frac{P(A|B)}{P(A|\bar{B})} = \frac{\dfrac{n_{BA}}{n_B}}{\dfrac{n_{\bar{B}A}}{n_{\bar{B}}}} = \frac{n_{BA}n_{\bar{B}}}{n_{\bar{B}A}n_B}.$$

This ratio is greater than one if and only if $n_{BA}n_{\bar{B}} > n_{\bar{B}A}n_B$. If the BH condition holds, it can be shown (see Appendix D for proof) that the relative risk is necessarily greater than one. For this implication to work, $n_{\bar{B}A} > 0$. In summary, if $n_{\bar{B}A},n_{B\bar{A}} > 0$, the BH condition holds if and only if $O_B > 1$ and $O_{\bar{B}} < 1$; the BH condition also implies $OR > 1$ and $RR > 1$.

Relationship to diagnostic likelihood ratios. The positive diagnostic likelihood ratios can be used to assess the performance of a diagnostic test. A positive diagnostic likelihood ratio, DLR_B, provides the likelihood that a positive test (indicating that a person has "it") occurs in an individual who truly does have "it" than one who does not. Similarly, a negative diagnostic likelihood ratio, $DLR_{\bar{B}}$, indicates the likelihood that a negative test (indicating a person does not have "it") occurs in an individual who truly does have "it" than one who does not. The diagnostic likelihood ratios are defined as follows:

$$DLR_B = \frac{P(B|A)}{P(B|\bar{A})} = \frac{\dfrac{n_{BA}}{n_A}}{\dfrac{n_{B\bar{A}}}{n_{\bar{A}}}} = \frac{n_{BA}}{n_{B\bar{A}}}\left(\frac{n_{\bar{A}}}{n_A}\right),$$

$$DLR_{\bar{B}} = \frac{P(\bar{B}|A)}{P(\bar{B}|\bar{A})} = \frac{\dfrac{n_{\bar{B}A}}{n_A}}{\dfrac{n_{\bar{B}\bar{A}}}{n_{\bar{A}}}} = \frac{n_{\bar{B}A}}{n_{\bar{B}\bar{A}}}\left(\frac{n_{\bar{A}}}{n_A}\right).$$

Ideally, a diagnostic test has $DLR_B > 1$ and $DLR_{\bar{B}} < 1$. If the BH condition holds, then we know $n_{BA} > n_{B\bar{A}}$; $DLR_B > 1$ if $\dfrac{n_{BA}}{n_{B\bar{A}}} > \dfrac{n_A}{n_{\bar{A}}}$, which is always true if $P(A) \leq \frac{1}{2}$. Similarly, if the BH condition holds, $n_{BA} > n_{B\bar{A}}$ and $DLR_{\bar{B}} < 1$ if $\frac{n_{\bar{B}A}}{n_{\bar{B}\bar{A}}} < \frac{n_A}{n_{\bar{A}}}$, which is always true if $P(A) \geq \frac{1}{2}$. Thus, if the BH condition holds, at least one of the two ideal diagnostic likelihood ratio conditions also holds.

And what about when the BH condition is not met? If the condition fails, it is either because (a) $n_{BA} \leq n_{B\bar{A}}$, or (b) $n_{\bar{B}\bar{A}} \leq n_{\bar{B}A}$, or both (a) and (b) are true. If (a) is true and $P(A) \geq \frac{1}{2}$, then $DLR_B \leq 1$. Similarly, if (b) is true and $P(A) < \frac{1}{2}$, then $DLR_{\bar{B}} \geq 1$. The other two situations (where (a) is true but $P(A) < \frac{1}{2}$ or (b) is true but $P(A) > \frac{1}{2}$) will depend on the data.

The BH Condition and Unequal Costs

The discussion up to this point considers only the special case of misclassification costs being equal; that is, the costs of false positives and false negatives are weighted the same. Attention is now given to the case where costs are not considered equal and decisions should be made so the overall cost is minimized.

Suppose that the cost of a false negative is k times more costly than a false positive, where $k > 0$. In determining whether the test outperforms base-rate prediction where costs are considered unequal, clinical efficiency requires the lowest-cost option. Rather than requiring $P(A) \leq \frac{1}{2}$, or equivalently $P(A) \leq P(\bar{A})$, consider the case when $C(P(A)) \geq C(P(\bar{A}))$, where $C(\cdot)$ is a cost function. In words this states that the cost of predicting that everyone has "it" is more than the cost of predicting no one has "it." The costs can be defined as $C(P(A)) = n_{\bar{A}}$ (the number of false positives when predicting everyone to have "it") and $C(P(\bar{A})) = kn_A$ (the number of false negatives, weighted by k, when predicting no one to have "it"). Thus, if $C(P(A)) \geq C(P(\bar{A}))$, then $n_{\bar{A}} \geq kn_A$. Note that when $k = 1$, $n_{\bar{A}} \geq n_A$, which is equivalent to $P(A) \leq P(\bar{A})$.

Assume $C(P(A)) \geq C(P(\bar{A}))$; that is, assume that the cost of prediction using the base rates is lowest when predicting no one to have "it." Prediction using the test is better than prediction using the base rates if and only if $C(\text{Test}) < C(P(\bar{A}))$, where the cost of the test is defined as $C(\text{Test}) = n_{B\bar{A}} + kn_{\bar{B}A}$ (the number of false positives plus k times the number of false negatives). This will be referred to as "generalized clinical efficiency." When unequal costs are assumed, generalized clinical efficiency can be characterized by a generalized Bokhari-Hubert (GBH) condition: a test outperforms base rate prediction if and only if $n_{BA} > \frac{1}{k}n_{B\bar{A}}$ and $n_{\bar{B}\bar{A}} > kn_{\bar{B}A}$. This holds true when $C(P(\bar{A})) \geq C(P(A))$ as well (see Appendix E for proof).

Note that when $k = 1$ (that is, costs are equal) the original BH condition holds, as desired. When $k > 1$ (so that false negatives are considered more costly than false positives), $\frac{1}{k} < 1$, and the requirement in the first row is weaker (that is, more easily satisfied) than when costs are equal because the number of true positives only needs to be larger than $\frac{1}{k}$ times the number of false positives; for the second row, the requirement is stronger (that is, less easily satisfied) than when costs are equal because the number of true negatives now needs to be larger than k times the number of false negatives. Similarly, when $0 < k < 1$ (so that false positives are considered more costly than false negatives), the requirement for differential prediction is stronger in the first row but weaker in the second.

One interesting consequence of the above results is when the frequencies are known but the costs are not. If the GBH condition is not met, one is able to determine upper and lower bounds for k so that the GBH condition is satisfied; the first requirement of the GBH condition provides a lower bound, the second requirement provides an upper bound. To see this, simply solve for k. The first part of the condition states that $n_{BA} > \frac{1}{k}n_{B\bar{A}}$; thus, the lower bound for k is $\frac{n_{B\bar{A}}}{n_{BA}}$. The second part states that $n_{\bar{B}\bar{A}} > kn_{\bar{B}A}$ so the upper bound is $k < \frac{n_{\bar{B}\bar{A}}}{n_{\bar{B}A}}$. Thus, for a test to satisfy the GBH condition, the costs of a false negative to a false positive must be bounded as follows:

$$\frac{n_{B\bar{A}}}{n_{BA}} < k < \frac{n_{\bar{B}\bar{A}}}{n_{\bar{B}A}}.$$

In words, the bounds state that the cost of a false negative relative to a false positive cannot be less than the ratio of false positives to true positives and cannot be more than the ratio of true negatives to false negatives. If k is not within the bounds, then the GBH condition fails and the test is not generalized clinically efficient.

Meehl-Rosen and Dawes Conditions

The Meehl-Rosen condition does not appear to be generalizable when costs are unequal. The Dawes condition can be, however; using the GBH condition, it can be shown that the generalizable Dawes condition is

$$P(\bar{A}|B) < 1 - \frac{1}{k+1},$$

when $C(P(A)) \geq C(P(\bar{A}))$. The condition is best stated in terms of the positive predictive value: $P(A|B) > \frac{1}{k+1}$. For $k = 1$, this is the original Dawes condition.

The more general Dawes condition given above implies a test is generalized clinically efficient when $C(P(A)) \geq C(P(\bar{A}))$, but not when $C(P(A)) \leq C(P(\bar{A}))$; that is, the condition may be met when $C(P(A)) \leq C(P(\bar{A}))$ but it is not necessarily true that the test is generalized clinically efficient. Thus, one needs to know k even though the costs of false positives and false negatives will not always be known; the GBH condition holds regardless of the relationship between $C(P(A))$ and $C(P(\bar{A}))$ (that is, it is independent of whether

$C(P(A))$ is greater than or less than $C(P(\bar{A}))$), making it more usable in practice because when k is unknown, bounds for k to make the test generalized clinically efficient can be constructed.

The Dawes condition for unequal costs only characterizes generalized clinical efficiency when $C(P(A)) \geq C(P(\bar{A}))$. Because the condition was derived from the GBH condition it can be extended so that it is no longer dependent on the relationship between $C(P(A))$ and $C(P(\bar{A}))$ (that is, it holds when $C(P(A)) \leq C(P(\bar{A}))$ as well). What may be referred to as the generalized Dawes condition can be stated as $P(A|B) > \frac{1}{k+1}$ and $P(\bar{A}|\bar{B}) > \frac{1}{1+\frac{1}{k}}$. For example, if the cost of false negatives is weighted as three times the cost of false positives then the generalized Dawes condition states that the positive predictive value needs to be greater than $\frac{1}{4}$ and the negative predictive value needs to be greater than $\frac{3}{4}$ to have the test be generalized clinically efficient (that is, the cost of the test is less than that of prediction using the base rates). The fact that these two lower bounds sum to one is not a coincidence—this will always be the case. The generalized Dawes condition characterizes a test that is generalized clinically efficient regardless of the relationship between the costs using base-rate predictions, $C(P(A))$ and $C(P(\bar{A}))$.

Given that the generalized Dawes condition holds for all k, bounds can be constructed for k in terms of the positive and negative predictive values as follows:

$$\frac{1}{P(A|B)} - 1 < k < \frac{P(\bar{A}|\bar{B})}{1 - P(\bar{A}|\bar{B})}.$$

(Some authors have referred to the first term, the inverse of the positive predictive value, as "the number needed to detain" [Buchanan, 2008]). For example, if the positive predictive value is .25 and the negative predictive value is .9, then the cost of a false negative relative to a false positive would need to be between three and nine for the test to be generalized clinically efficient. Figure 1 gives an illustration of the lower bounds of the PPV and NPV for different values of k. The shaded region are the values the PPV and NPV can take for generalized clinical efficiency to be met, across different values of k.

Predicting Violence and Dangerousness

Predicting violent and dangerous behavior continues to be a heavily debated topic; the importance of base rates in predicting violent behavior has been discussed elsewhere (Doren, 1998; Vrieze & Grove, 2008; Wollert, 2006), although not all agree (for example, see Harris & Rice, 2007). We begin with a numerical example of predicting violence using an actuarial model developed in the MacArthur Violence Risk Assessment Study called the Classification of Violence Risk (COVR; Monahan et al., 2001). The COVR was designed for the diagnostic assessment of violence risk (event B, risk present; event $\bar{B}$, risk absent) in relation to the occurrence of follow-up violence (event A, violence present; event $\bar{A}$, violence absent) among persons with mental illnesses. Table 2 displays the results from the construction sample (that is, the sample used to construct the COVR instrument). Those who are classified as having a "high" or "very-high" risk of violence are predicted to be violent (that is, risk is considered present); those who are classified as "low" or "very-low" are not (risk absent). The base rate for violence in the sample is $P(A) = \frac{128}{756} = .17 < \frac{1}{2}$.

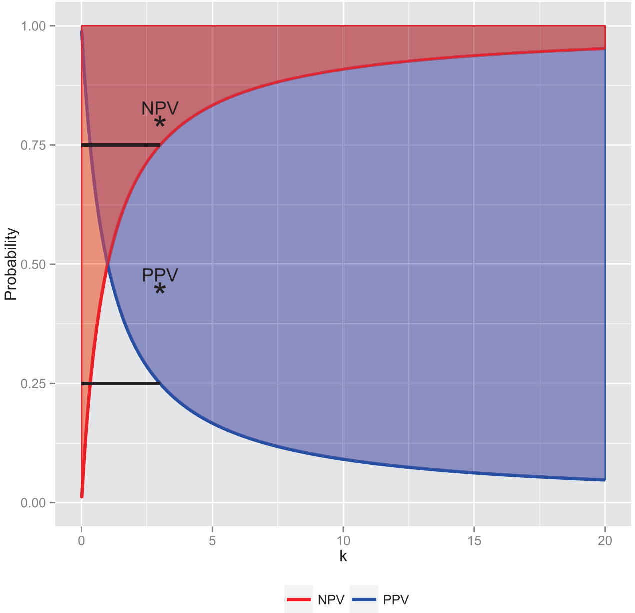

Figure 1. The figure gives a visual demonstration of the lower bounds on the positive predictive value (PPV) and negative predictive value (NPV). The decreasing curve represents the lower bound on the PPV; the shaded region above it gives the values of the PPV that satisfy the first part of the generalized Dawes condition across k. The increasing curve is for the NPV; the shaded region above it gives the values of the NPV that satisfy the second part of the generalized Dawes condition across k. When both the PPV and NPV fall in the shaded region(s) for a given k (such as for the stars given in the figure representing a test with a PPV of .45 and a NPV of .80 for the case when $k = 3$; the horizontal lines represent the lower bounds at $k = 3$), the test satisfies the generalized Dawes condition and is therefore generalized clinically efficient. The two curves intersect at $k = 1$, where both the PPV and NPV lower bounds are equal to one half. See the online article for the color version of this figure.

Because $n_{BA} = 105 > 60 = n_{B\bar{A}}$ and $n_{\bar{B}\bar{A}} = 568 > 23 = n_{\bar{B}A}$, the BH condition is satisfied implying the COVR outperforms base-rate prediction. The other equivalent conditions can be verified but require some calculations. The sensitivity of the test is $\frac{105}{128} = .82$, the specificity is $\frac{568}{628} = .90$, and the Meehl-Rosen condition is satisfied: $\frac{(1-.9)}{.82+(1-.9)} = .10 < .17 = P(A)$. The positive predictive value is $\frac{105}{165} = .64 > .5$, so the Dawes condition is satisfied. Finally, because the accuracy of the test is $\frac{105+568}{756} = .89$ and this is greater than $1 - P(A) = .83$, the general condition for clinical efficiency is satisfied as well.

Next, data from a Monahan et al. (2005) study attempting to validate the COVR model on new data are considered. Table 3 displays the results in the form of a 2 × 2 contingency table. The base rate for violence in this sample is $\frac{28}{157} = .18 < \frac{1}{2}$.

The model correctly predicts violence in approximately one third of the patients ($\frac{19}{55} = .35$); the model also correctly predicts nonviolence in about 10 of every 11 patients ($\frac{93}{102} = .91$). Overall, the model correctly diagnosed three of every four patients; the accuracy of their test was $\frac{(19+93)}{157} = .71$. Because $P(A) \leq \frac{1}{2}$, prediction by base rates would be to say that all patients will not commit violence. In doing so, one would be

Table 2
A 2 × 2 Contingency Table for Predicting Risk of Violence in the Classification of Violence Risk Construction Sample

	State of nature		
	A (Violence present)	$\bar{A}$ (Violence absent)	Totals
Prediction			
B (Risk present)	105	60	165
$\bar{B}$ (Risk absent)	23	568	591
Totals	128	628	756

Note. Compare Table 6.7 in Monahan et al. (2001, p. 125).

correct 82% of the time for this sample ($P(\bar{A}) = \frac{129}{157} = .82$). Because prediction using base rates is better than prediction using the test, all three of the conditions fail, as demonstrated below.

The specificity is $\frac{93}{129} = .72$, the sensitivity of the test is $\frac{19}{28} = .68$, and the base rate for violence is $P(A) = .18$. Verifying the Meehl-Rosen condition, we see

$$.18 = P(A) \ngtr \frac{1 - \text{specificity}}{\text{sensitivity} + (1 - \text{specificity})} = \frac{1 - .72}{.68 + (1 - .72)} = .29,$$

so the condition fails to hold. The PPV of the test is PPV $= \frac{19}{55} = .35 < \frac{1}{2}$, so the Dawes condition fails to hold. Finally, because $n_{\bar{B}\bar{A}} = 93 > 9 = n_{\bar{B}A}$ but $n_{BA} = 19 < 36 = n_{B\bar{A}}$, the BH condition also fails. Another easy way to detect the failure of this latter condition is to note there is no differential prediction because the row entries in the contingency table (see Table 3) are ordered in the same direction. This demonstrates the ease with which the BH condition can be verified, relative to the other conditions.

Given unequal costs, the GBH condition is $n_{BA} > \frac{1}{k} n_{B\bar{A}}$ and $n_{\bar{B}\bar{A}} > k n_{\bar{B}A}$, and an upper and lower bound for k can be constructed so that a test is clinically efficient: $\frac{n_{B\bar{A}}}{n_{BA}} < k < \frac{n_{\bar{B}\bar{A}}}{n_{\bar{B}A}}$. Using the data from Table 3, the second requirement of the GBH condition states that the upper bound for k is $\frac{n_{\bar{B}\bar{A}}}{n_{\bar{B}A}} = \frac{31}{3} \approx 10.3$; that is, false negatives cannot be considered more than 10.3 times more costly than false positives or the GBH condition fails in the second row. For the first requirement, the lower bound for k is $\frac{n_{B\bar{A}}}{n_{BA}} = \frac{36}{19} \approx 1.9$. (Note the result is the same if the PPV and NPV are use to construct the bounds.) In other words, one would need to consider false negatives to be almost twice as costly as false positives for the COVR to be clinically efficient in the validation sample.

The authors also provide “revised estimates” of their sample, reclassifying participants from both groups by “using a slightly more inclusive operational definition of violence” (Monahan et al., 2005, p. 814). In their revised estimate their accuracy is better (it is exactly equal to $P(\bar{A})$), but all three conditions again fail to hold.

Conclusion

The BH condition presented in this article provides a simple condition for determining whether prediction from a diagnostic test outperforms prediction by base rates; this condition is equivalent to those presented by Meehl and Rosen (1955) and Dawes (1962). Unlike these latter two conditions, the BH pattern is unchanged with respect to base-rate probability. The BH condition also has several interpretive relationships with measures of association, such as the Goodman-Kruskal lambda and the odds ratios.

The simplicity of the BH condition relies on the use of 2 × 2 contingency tables; this in turn leads to a question as to why researchers typically fail to present data in this simple form. Besides its use in assessing the BH condition directly, 2 × 2 contingency tables provide all the information needed to determine the quality of a diagnostic test at a given cut score.

This article has also demonstrated that a popular instrument for predicting violence outperforms base-rate prediction in the construction sample but fails to do so in a validation sample; this is an example of the expected shrinkage in the accuracy of a diagnostic measure when used on new data. Unfortunately, the COVR’s failure to outperform base-rate prediction appears to be the norm among violence prediction measures. In a meta-analysis of 73 samples, Fazel, Singh, Doll, and Grann (2012) determined that the median positive predictive value among the measures examined was .41, suggesting that the instruments failed to satisfy the BH condition in over half of the studies if costs are considered equal.

When an event that is being predicted is rare (so the base rate is small), clinical efficiency can be extremely difficult to meet. An example of such a rare event is suicide; Rosen (1954) noted that “[t]he low incidence of suicide is in itself a major limitation in the development of an effective suicide predictor, for in any attempt at prediction of infrequent behavior, a large number of false positives are obtained,” implying that the test is not likely be clinically efficient. For a test used to predict a rare event, it may be that the only way for it to be clinically efficient in the generalized sense is to weight false negatives as far more costly than false positives, and which arguably is reasonable for an event like suicide. Given the large number of false positives, however, Rosen (1954) noted that “it would be impractical to treat as suicidal the prodigious number of misclassified cases.” Even if the unequal costs are justifiable, it may not be practical to implement.

In assessing and predicting violence, it is often not the case that the costs of false negatives and false positives are equally weighted. As one reviewer suggested, “a child should not go to a parent who has a roughly one chance in three of becoming violent, even though the parent has a higher likelihood of being nonviolent than violent,” suggesting that false negatives are more costly than false positives. The authors agree with the reviewer but note that there are other situations where the opposite may be true and false positives are considered more costly than false negatives. An instance of this may be determining whether an individual accused of

Table 3
A 2 × 2 Contingency Table for Predicting Risk of Violence in a Classification of Violence Risk Validation Sample (Monahan et al., 2005)

	State of nature		
	A (Violence present)	$\bar{A}$ (Violence absent)	Totals
Prediction			
B (Risk present)	19	36	55
$\bar{B}$ (Risk absent)	9	93	102
Totals	28	129	157

committing a crime should be indefinitely detained because he or she is deemed a threat to society. Using a device in this context that is correct in only one out of three positive predictions seems ethically questionable. In both the situations presented, it may be appropriate to use an actuarial tool such as the COVR, but depending on the costs of misclassification, the decision made as a result of the actuarial device may be seen as inappropriate. Costs should be decided a priori and justified both ethically and legally.

References

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44,* 211–233.

Buchanan, A. (2008). Risk of violence by psychiatric patients: Beyond the actuarial versus clinical assessment debate. *Psychiatric Services, 59,* 184–190.

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

Doren, D. M. (1998). Recidivism base rates, predictions of sex offender recidivism, and the "sexual predator" commitment laws. *Behavioral Sciences and the Law, 16,* 97–114.

Faust, D., & Nurcombe, B. (1989). Improving the accuracy of clinical judgment. *Psychiatry: Interpersonal and Biological Processes, 52,* 197–208.

Fazel, S., Singh, J. P., Doll, H., & Grann, M. (2012). Use of risk assessment instruments to predict violence and antisocial behaviour in 73 samples involving 24,827 people: Systematic review and meta-analysis. *British Medical Journal, 345,* e4692.

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of American Statistical Association, 49,* 732–762.

Harris, G. T., & Rice, M. E. (2007). Characterizing the value of actuarial violence risk assessments. *Criminal Justice and Behavior, 34,* 1638–1658.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80,* 237–251.

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

Monahan, J., Steadman, H. J., Robbins, P. C., Appelbaum, P. S., Banks, S., Grisso, T., . . . Silver, E. (2005). An actuarial model of violence risk assessment for persons with mental disorders. *Psychiatric Services, 56,* 810–815.

Monahan, J., Steadman, H. J., Silver, E., Appelbaum, P. S., Robbins, P. C., Mulvey, E. P., . . . Banks, S. (2001). *Rethinking risk assessment: The MacArthur study of mental disorder and violence*. New York, NY: Oxford University Press.

Pearson, K. (1900). On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series, 5,* 157–175.

Rosen, A. (1954). Detection of suicidal patients: An example of some limitations in the prediction of infrequent events. *Journal of Consulting Psychology, 18,* 397.

Schwarz, N., Strack, F., Hilton, D., & Naderer, G. (1991). Base rates, representativeness, and the logic of conversation: The contextual relevance of "irrelevant" information. *Social Cognition, 9,* 67–84.

Vrieze, S. I., & Grove, W. M. (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.

Wollert, R. (2006). Low base rates limit expert certainty when current actuarials are used to identify sexually violent predators: An application of Bayes's theorem. *Psychology, Public Policy, and Law, 12,* 56–85.

(*Appendices follow*)

Appendix A

The Meehl-Rosen Condition: Proof

$$\begin{aligned}
P(A) &> \frac{1 - P(\bar{B}|\bar{A})}{P(B|A) + (1 - P(\bar{B}|\bar{A}))} \Leftrightarrow \\
P(A)[P(B|A) + (1 - P(\bar{B}|\bar{A}))] &> 1 - P(\bar{B}|\bar{A}) \Leftrightarrow \\
P(B|A)P(A) + P(A) - P(\bar{B}|\bar{A})P(A) &> 1 - P(\bar{B}|\bar{A}) \Leftrightarrow \\
P(B|A)P(A) - P(\bar{B}|\bar{A})P(A) + P(\bar{B}|\bar{A}) &> 1 - P(A) \Leftrightarrow \\
P(B|A)P(A) + P(\bar{B}|\bar{A})(1 - P(A)) &> P(\bar{A}) \Leftrightarrow \\
P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) &> P(\bar{A}).
\end{aligned}$$

Appendix B

The Dawes Condition: Proof

Before proving the Dawes condition, note that $P(A|B)P(B) = P(A \cap B) = P(B|A)P(A)$ and $P(A|B)P(B) + P(A|\bar{B})P(\bar{B}) = P(A)$. These two equalities are used in the proof.

$$\begin{aligned}
P(\bar{A}|B) &< \frac{1}{2} \Leftrightarrow \\
2P(\bar{A}|B) &< 1 \Leftrightarrow \\
P(\bar{A}|B) + P(\bar{A}|B) &< 1 \Leftrightarrow \\
P(\bar{A}|B) &< 1 - P(\bar{A}|B) \Leftrightarrow \\
P(\bar{A}|B) &< P(A|B) \Leftrightarrow \\
P(\bar{A}|B)P(B) &< P(A|B)P(B) \Leftrightarrow \\
P(\bar{A}|B)P(B) + P(\bar{A}|\bar{B})P(\bar{B}) &< P(A|B)P(B) + P(\bar{A}|\bar{B})P(\bar{B}) \Leftrightarrow \\
P(\bar{A}) &< P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A})
\end{aligned}$$

(*Appendices continue*)

Appendix C

The Bokhari-Hubert Condition: Proof

We begin with the general condition:

$$\begin{aligned}
P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) &> P(\bar{A}) \Leftrightarrow \\
\frac{n_{BA} + n_{\bar{B}\bar{A}}}{n} &> \frac{n_{\bar{A}}}{n} \Leftrightarrow \\
n_{BA} + n_{\bar{B}\bar{A}} &> n_{\bar{A}} \Leftrightarrow \\
n_{BA} + n_{\bar{B}\bar{A}} &> n_{B\bar{A}} + n_{\bar{B}\bar{A}} \Leftrightarrow \\
n_{BA} &> n_{B\bar{A}}.
\end{aligned}$$

Thus, $P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) > P(\bar{A}) \Leftrightarrow n_{BA} > n_{B\bar{A}}$. Now,

$$\begin{aligned}
n_{BA} &> n_{B\bar{A}} \Leftrightarrow \\
n_{\bar{B}A} + n_{BA} + n_{\bar{B}\bar{A}} &> n_{\bar{B}A} + n_{B\bar{A}} + n_{\bar{B}\bar{A}} \Leftrightarrow \\
n_{\bar{B}\bar{A}} &> n_{\bar{B}A} + n_{B\bar{A}} + n_{\bar{B}\bar{A}} - \ (n_{\bar{B}A} + n_{BA}) \Leftrightarrow \\
n_{\bar{B}\bar{A}} &> n_{\bar{B}A} + n_{\bar{A}} - n_{A}.
\end{aligned}$$

Because (by assumption) $n_{\bar{A}} > n_A$, we have $n_{\bar{A}} - n_A > 0$, and therefore, $n_{\bar{B}\bar{A}} > n_{\bar{B}A}$, as desired.

Appendix D

The Bokhari-Hubert Condition and Relative Risk: Proof

We assume the Bokhari-Hubert condition holds.

$$\begin{aligned}
n_{BA} > n_{B\bar{A}} \text{ and } n_{\bar{B}\bar{A}} > n_{\bar{B}A} &\Rightarrow n_{BA}n_{\bar{B}\bar{A}} > n_{B\bar{A}}n_{\bar{B}A} \Rightarrow \\
n_{BA}n_{\bar{B}\bar{A}} + n_{BA}n_{\bar{B}A} &> n_{B\bar{A}}n_{\bar{B}A} + n_{BA}n_{\bar{B}A} \Rightarrow \\
n_{BA}\,(n_{\bar{B}\bar{A}} + n_{\bar{B}A}) &> n_{\bar{B}A}\,(n_{B\bar{A}} + n_{BA}) \Rightarrow \\
n_{BA}n_{\bar{B}} &> n_{\bar{B}A}n_{B} \Rightarrow \\
\frac{n_{BA}n_{\bar{B}}}{n_{\bar{B}A}n_{B}} &> 1 \Rightarrow \\
RR &> 1
\end{aligned}$$

(*Appendices continue*)

Appendix E

The Generalized Bokhari-Hubert Condition and Unequal Costs: Proof

We begin with the general condition of clinical efficiency under unequal costs where false negatives ($n_{\bar{B}A}$) are considered k times more costly than false positives ($n_{B\bar{A}}$), for some $k > 0$. Assume that the number of negative cases is greater than or equal to k times the number of positive cases ($n_{\bar{A}} \geq kn_A$) so that $C(P(A)) \geq C(P(\bar{A}))$.

$$\begin{aligned}
C(P(\bar{A})) &> C(\text{Test}) \Leftrightarrow \\
kn_A &> n_{B\bar{A}} + kn_{\bar{B}A} \Leftrightarrow \\
k(n_{BA} + n_{\bar{B}A}) &> n_{B\bar{A}} + kn_{\bar{B}A} \Leftrightarrow \\
kn_{BA} &> n_{B\bar{A}} \Leftrightarrow \\
n_{BA} &> \frac{1}{k}n_{B\bar{A}}.
\end{aligned}$$

Now,

$$\begin{aligned}
kn_{BA} &> n_{B\bar{A}} \Leftrightarrow \\
n_{\bar{B}\bar{A}} + kn_{BA} + kn_{\bar{B}A} &> n_{\bar{B}\bar{A}} + n_{B\bar{A}} + kn_{\bar{B}A} \Leftrightarrow \\
n_{\bar{B}\bar{A}} &> n_{\bar{A}} - kn_A + kn_{\bar{B}A}.
\end{aligned}$$

Because $n_{\bar{A}} - kn_A \geq 0$ by assumption, it holds that $n_{\bar{B}\bar{A}} > kn_{\bar{B}A}$.

Now consider the case when $kn_A \geq n_{\bar{A}}$.

$$\begin{aligned}
C(P(A)) &> C(\text{Test}) \Leftrightarrow \\
n_{\bar{A}} &> n_{B\bar{A}} + kn_{\bar{B}A} \Leftrightarrow \\
n_{B\bar{A}} + n_{\bar{B}\bar{A}} &> n_{B\bar{A}} + kn_{\bar{B}A} \Leftrightarrow \\
n_{\bar{B}\bar{A}} &> kn_{\bar{B}A}.
\end{aligned}$$

In addition,

$$\begin{aligned}
n_{\bar{B}\bar{A}} &> kn_{\bar{B}A} \Leftrightarrow \\
kn_{BA} + n_{\bar{B}\bar{A}} + n_{B\bar{A}} &> kn_{BA} + kn_{\bar{B}A} + n_{B\bar{A}} \Leftrightarrow \\
kn_{BA} &> kn_A - n_{\bar{A}} + n_{B\bar{A}}.
\end{aligned}$$

Because, by assumption, $kn_A - n_{\bar{A}} \geq 0$, it holds that $kn_{BA} > n_{B\bar{A}}$, or equivalently, $n_{BA} > \frac{1}{k}n_{B\bar{A}}$. Thus, the GBH condition for unequal costs remains the same regardless of the relationship between $C(P(A))$ and $C(P(\bar{A}))$.

Received August 2, 2014
Revision received December 23, 2014
Accepted December 23, 2014 ■