Probabilistic Reasoning and Diagnostic Testing

For ye shall know the truth, and the truth shall set you free – Motto of the CIA (from John 8:31–32)

September 14, 2016

Topic Areas

1) Clinical Efficiency Revisted: Meehl and Rosen
Test Sensitivity and Specificity
The Positive Predictive Value and the Negative Predictive Value
"Betting the Base Rates"
The Test "Hit Rate"
Conditions for Clinical Efficiency: Meehl-Rosen; Dawes; Bokhari-Hubert

2) Diagnostic Test Evaluation
The ROC Curve
The Area Under the Curve (AUC)
Base Rate Independence for the AUC

3) Issues in Medical Screening; Other Screening Difficulties

Clinical Efficiency Revisited: Meehl and Rosen

We begin by (re)introducing a 2×2 contingency table cross-classifying n individuals by events A and $\bar{A}$ and B and $\bar{B}$ but now with terminology attuned to a diagnostic testing context.

The events B (positive) or $\bar{B}$ (negative) occur when the test says the person has "it" or doesn't have "it," respectively, whatever "it" may be.

The events A (positive) or $\bar{A}$ (negative) occur when the "state of nature" is such that the person has "it" or doesn't have "it," respectively.

The Generic 2×2 Contingency Table

		state of nature		
		A (pos)	$\bar{A}$ (neg)	row sums
test	B (pos)	n_{BA}	$n_{B\bar{A}}$	n_B
result	$\bar{B}$ (neg)	$n_{\bar{B}A}$	$n_{\bar{B}\bar{A}}$	$n_{\bar{B}}$
	column sums	n_A	$n_{\bar{A}}$	n

Using the urn model (that is, picking a person at random from the pool of size n) and conditionalizing on the state of nature (that is, knowing whether the state of nature for the selected individual is A (pos) or $\bar{A}$ (neg)), a number of common terms can be defined that are relevant to a diagnostic testing context:

Test Sensitivity and Specificity

		state of nature	
		A (pos)	$\bar{A}$ (neg)
test	B (pos)	$P(B\|A) = n_{BA}/n_A$ (sensitivity)	$P(B\|\bar{A}) = n_{B\bar{A}}/n_{\bar{A}}$ (false positive)
result	$\bar{B}$ (neg)	$P(\bar{B}\|A) = n_{\bar{B}A}/n_A$ (false negative)	$P(\bar{B}\|\bar{A}) = n_{\bar{B}\bar{A}}/n_{\bar{A}}$ (specificity)
column sums		$\frac{n_{BA}+n_{\bar{B}A}}{n_A} = 1.0$	$\frac{n_{B\bar{A}}+n_{\bar{B}\bar{A}}}{n_{\bar{A}}} = 1.0$

To give words to the two important concepts of test sensitivity and specificity, we have:

sensitivity $= P(B|A) =$ the probability that the test is positive if the person has "it";

specificity $= P(\bar{B}|\bar{A}) =$ the probability that the test is negative if the person doesn't have "it."

The Positive Predictive Value and the Negative Predictive Value

Using the urn model (that is, picking a person at random from the pool of size n) and conditionalizing on the diagnostic test result (that is, knowing whether the test result for the selected individual is B (pos) or $\bar{B}$ (neg)), a number of common terms can be defined that are relevant to a diagnostic testing context:

	state of nature	
	A (pos)	$\bar{A}$ (neg)
B (pos)	$P(A\|B) = n_{BA}/n_B$ (positive predictive value)	$P(\bar{A}\|B) = n_{B\bar{A}}/n_B$
$\bar{B}$ (neg)	$P(A\|\bar{B}) = n_{\bar{B}A}/n_{\bar{B}}$	$P(\bar{A}\|\bar{B}) = n_{\bar{B}\bar{A}}/n_{\bar{B}}$ (negative predictive) value)

Row Sums:

$B : \frac{n_{BA}+n_{B\bar{A}}}{n_B} = 1.0$

$\bar{B} : \frac{n_{\bar{B}A}+n_{\bar{B}\bar{A}}}{n_{\bar{B}}} = 1.0$

Again, to give words to the two important concepts of the positive and negative predictive values, we have:

positive predictive value $= P(A|B) =$ the probability that the person has "it" if the test says the person has "it";

negative predictive value $= P(\bar{A}|\bar{B}) =$ the probability that the person doesn't have "it" if the test says the person doesn't have "it."

Prediction According to the Base Rates (or "Betting the Base Rates")

Assuming that $P(A) \leq 1/2$ (this, by the way, can always be done without loss of any generality because the roles of A and $\bar{A}$ can be interchanged), prediction according to base rates would be to consistently say that a person doesn't have "it" (because $P(\bar{A}) \geq P(A)$).

The probability of being correct in this prediction is $P(\bar{A})$ (which is greater than or equal to $1/2$).

Prediction according to the test would be to say the person has "it" if the test is positive, and doesn't have "it" if the test is negative.

The Test "Hit Rate" (or "Accuracy")

The probability of a correct diagnosis according to the test (called the "hit rate" or "accuracy") is:

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

$$(\frac{n_{BA}}{n_A})(\frac{n_A}{n}) + (\frac{n_{\bar{B}\bar{A}}}{n_{\bar{A}}})(\frac{n_{\bar{A}}}{n}) = \frac{n_{BA} + n_{\bar{B}\bar{A}}}{n} \ ,$$

which is just the sum of main diagonal frequencies in the 2×2 contingency table divided by the total sample size n.

A general condition can be given for when prediction by a test will be better than prediction by base rates (again, assuming that $P(A) \leq 1/2$).

It is for the accuracy to be strictly greater than $P(\bar{A})$:

$$P(B|A)P(A) + P(\bar{B}|\bar{A})P(\bar{A}) > P(\bar{A}).$$

Based on this first general condition, we give three equivalent conditions for clinical efficiency to hold that we attribute to Meehl and Rosen (1955), Dawes (1962), and Bokhari and Hubert (2015).

Meehl-Rosen condition: assuming that $P(A) \leq 1/2$, it is best to use the test (over base rates) if and only if

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

Dawes condition: assuming that $P(A) \leq 1/2$, it is better to use the test (over base rates) if and only if $P(\bar{A}|B) < 1/2$ (or, equivalently, when $P(A|B) > 1/2$; that is, when the positive predictive value is greater than $1/2$).

Bokhari-Hubert condition: assuming that $P(A) \leq 1/2$, it is better to use the test (over base rates) if and only if differential prediction holds between the row entries in the frequency table:

$n_{BA} > n_{B\bar{A}}$ but $n_{\bar{B}A} < n_{\bar{B}\bar{A}}$

In words, given the B (positive) row, the frequency of positive states of nature, n_{BA}, is greater than or equal to the frequency of negative states of nature, $n_{B\bar{A}}$; the opposite occurs within the $\bar{B}$ (negative) row.

The (Module 2) COVR Numerical Example Revisited

To give a numerical example of these conditions, the COVR 2×2 contingency table from Module 2 is used.

Recall that this table reports a cross-validation of an instrument for the diagnostic assessment of violence risk (B: positive (risk present); $\bar{B}$: negative (risk absent)) in relation to the occurrence of followup violence (A: positive (violence present); $\bar{A}$: negative (violence absent)):

		state of nature		
		A (pos)	$\bar{A}$ (neg)	row sums
prediction	B (pos)	19	36	55
	$\bar{B}$ (neg)	9	93	102
	column sums	28	129	157

To summarize what this table shows, we first note that 2 out of 3 predictions of "dangerous" are wrong ($.65 = 36/55$, to be precise); 1 out of 11 predictions of "not dangerous" are wrong ($.09 = 9/102$, to be precise).

The accuracy or "hit-rate" is .71 ($= (19 + 93)/157$).

If everyone was predicted to be "not dangerous", we would be correct 129 out of 157 times, the base rate for $\bar{A}$:
$P(\bar{A}) = 129/157 = .82$.

Because this is better than the accuracy of .71, all three conditions will fail for when the test would do better than the base rates:

Meehl-Rosen condition: for a specificity $= 93/129 = .72$, sensitivity $= 19/28 = .68$, and $P(A) = 28/157 = .18$,

$$P(A) \not> \frac{1 - \text{specificity}}{\text{sensitivity} + (1 - \text{specificity})}$$

$$.18 \not> \frac{1 - .72}{.68 + (1 - .72)} = .29$$

Dawes condition: the positive predictive value of $.35 = 19/55$ is not greater than $1/2$.

Bokhari-Hubert condition: there is no differential prediction because the row entries in the frequency table are ordered in the same direction.

Summary of the Dawes Condition: The False Positive Paradox

The Dawes condition described in the previous section shows the importance of clinical efficiency in the bottom-line justification for the use of a diagnostic instrument.

When you can do better with base rates than with a diagnostic test, the Dawes condition implies that the positive predictive value is less than $1/2$.

In other words, it is more likely that a person doesn't have "it" than they do, even though the test says the person has "it."

This anomalous circumstance has been called the "false positive paradox."

Summary of the Bokhari-Hubert Condition: Differential Prediction

For base rates to be worse than the test, the Bokhari-Hubert condition requires differential prediction to exist;

explicitly, within those predicted to be dangerous, the number who were dangerous (n_{BA}) must be greater than the number who were not dangerous ($n_{B\bar{A}}$);

conversely, within those predicted to be not dangerous, the number who were not dangerous ($n_{\bar{B}\bar{A}}$) must be greater than those who were dangerous ($n_{\bar{B}A}$).

Unequal Costs of Prediction Errors

One might conclude that it is ethically questionable to use a clinically inefficient test.

If you can't do better than just predicting with base rates, what is the point of using the diagnostic instrument in the first place.

The only mechanism that we know of that might justify the use of a clinically inefficient instrument would be to adopt severe unequal costs in the misclassification of individuals (that is, the cost of predicting "dangerous" when the "state of nature" is "not dangerous," and in predicting "not dangerous" when the "state of nature" is "dangerous").

But here we would soon have to acknowledge Sir William Blackstone's dictum (1765): "It is better that ten guilty escape than one innocent suffer."

Bokhari and Hubert (2015)

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

The Bokhari and Hubert paper (2015) (given at our web site where these slides are housed: bokhari_hubert.pdf) that discusses the three equivalent statements for clinical efficiency, also gives a generalized clinical efficiency condition that allows for the assignment of unequal costs to the false positives and false negatives.

Depending on how the costs of misclassification are assigned, a determination can be made as to when generalized clinical efficiency holds;

that is, when is the total costs of using a test less than the total costs obtained by just classifying through base rates?

Predicting Rare Events

When interests center on the prediction of a very infrequent event (such as the commission of suicide) and the cost of a false negative (releasing a suicidal patient) is greater than the cost of a false positive (detaining a non-suicidal patient), there still may be such a large number of false positives that implementing and acting on such a prediction system would be infeasible.

An older discussion of this conundrum is by Albert Rosen, "Detection of Suicidal Patients: An Example of Some Limitations in the Prediction of Infrequent Events," *Journal of Consulting Psychology* (*18*, 1954, 397–403).

[A particularly poignant change of example would be to replace "commission of suicide" with the phrase "commission of mass murder"]

Diagnostic Test Evaluation

There is a longer handout that we will use for figures and an extensive numerical example:

diagnostic_test_evaluation_example.pdf

The Receiver Operating Characteristic (ROC) curve of a diagnostic test is a plot of test sensitivity (the probability of a "true" positive) against 1.0 minus test specificity (the probability of a "false" positive).

As shown in Figure 1, when there is a single 2×2 contingency table, the ROC plot would be based on a single point.

In some cases, however, a diagnostic test might provide more than a simple dichotomy (for example, more than a value of 0 or 1, denoting a negative or a positive decision, respectively), and instead gives a numerical range (for example, integer scores from 0 to 20, as in the illustration in the handout on the Psychopathy Checklist, Screening Version (PCL:SV)).

The ROC Curve

In these latter cases, different possible "cutscores" might be used to reflect differing thresholds for a negative or a positive decision.

Figure 2 gives the ROC plot for the PCL:SV using three possible cutscores.

In general, the ROC curve is embedded in a box having unit-length sides.

It begins at the origin defined by a sensitivity of 0.0 and a specificity of 1.0, and ends at a sensitivity of 1.0 and a specificity of 0.0.

Along the way, the ROC curve goes through the various sensitivity and 1.0 − specificity values attached to the possible cutscores.

The Diagonal Lines of "No Discrimination"

The diagonals in both Figures 1 and 2 represent lines of "no discrimination" where sensitivity values are equal to 1.0 minus specificity values.

Restating, we have $P(B|A) = 1 - P(\bar{B}|\bar{A})$, and finally, $P(B|A) = P(B|\bar{A})$.

This last equivalence provides an interpretation for the "no discrimination" phrase: irrespective of the "state of nature" (A or $\bar{A}$), the probability of a "yes" prediction remains the same.

The Area Under the Curve (AUC)

For an ROC curve to represent a diagnostic test that is performing better than "chance," it has to lie above the "no discrimination" line where the probabilities of "true" positives exceed the probabilities of "false" positives (or equivalently, where sensitivities are greater than 1.0 minus the specificities).

The characteristic of good diagnostic tests is the degree to which the ROC curve "gets close to hugging" the left and top line of the unit-area box and where the sensitivities are much bigger than 1.0 minus specificities.

The most common summary measure of diagnostic test performance is the "area under the curve" (AUC), which ranges from an effective lower value of .5 (for the line of "no discrimination") to 1.0 for a perfect diagnostic test with sensitivity and specificity values both equal to 1.0.

So, as an operational comparison of diagnostic test performances, those with bigger AUCs are better.

Independence of Base Rates for the AUC

Figure 1 helps show the independence of base rates for the AUC;

the AUC is simply the average of sensitivity and specificity when only one cutscore is considered, and neither sensitivity or specificity is a function of base rates:

A = (1 - sens)(1 - spec)
B = (1/2)(1 - spec)(sens)
C = (1/2)(1 - sens)(spec)
AUC = 1.0 - (A + B + C) = (1/2)(sensitivity + specificity)

We can also see explicitly how different normalizations (using base rates) are used in calculating an AUC or accuracy:

$$P(B|A) = n_{BA}/n_A = \text{sensitivity}$$
$$P(\bar{B}|\bar{A}) = n_{\bar{B}\bar{A}}/n_{\bar{A}} = \text{specificity}$$

$$\text{AUC} = ((n_{BA}/n_A) + (n_{\bar{B}\bar{A}})/n_{\bar{A}})/2$$

$$\text{accuracy} = (n_{BA} + n_{\bar{B}\bar{A}})/n$$

Note that only when $n_A = n_{\bar{A}}$ (that is, when the base rates are equal), are accuracy and the AUC identical.

In instances of unequal base rates (such as in the prediction of "dangerous behavior"), the AUC can be a very poor measure of diagnostic test usage in a particular sample.

The Wilcoxon Test Statistic Interpretation of the AUC

As developed in detail by Hanley and McNeil (1982), it is possible to calculate numerically the AUC for an ROC curve that is constructed for multiple cutscores by first computing a well-known two-sample Wilcoxon test statistic.

An numerical example is given for this in the handout for the PCL:SV data; the AUC turns out to be .73.

Summary Comments About Using the AUC to Evaluate Diagnostic Tests

The answer we have for the general question of "how should a diagnostic test be evaluated?" is in contrast to current widespread practice.

Whenever the base rate for the condition being assessed is relatively low (for example, for "dangerous" behavior), the area under the ROC curve (AUC) is not necessarily a good measure for conveying the adequacy of the actual predictions made from a diagnostic test.

The AUC does not incorporate information about base rates.

It only evaluates the test itself and not how the test actually performs when used on a specific population with differing base rates for the presence or absence of the condition being assessed.

What About the PPV and the NPV?

The use of AUC as a measure of diagnostic value can be very misleading in assessing conditions with unequal base rates, such as being "dangerous."

This misinformation is 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.

What Should Be a Minimal Condition That a Diagnostic Test Should Have?

A simple condition given earlier (and attributed to Robyn Dawes) points to a minimal condition that a diagnostic test should probably satisfy (and which leads to prediction with the test being better than just prediction according to base rates): the positive predictive value must be greater than 1/2.

If this minimal condition does not hold, it will be more likely that a person doesn't have "it" than they do, even where the test says the person has "it."

As noted earlier, this situation is so unusual that it has been referred to as the "false positive paradox."

Issues in Medical Screening

It might be an obvious statement to make, but in our individual dealings with doctors and the medical establishment generally, it is important for all to understand the positive predictive values (PPVs) for whatever screening tests we now seem to be constantly subjected to, and thus, the number, (1 – PPV), referring to the false positives;

that is, if a patient tests positive, what is the probability that "it" is not actually present.

It is a simple task to plot PPV against $P(A)$ from 0 to 1 for any given pair of sensitivity and specificity values. Such a plot can show dramatically the need for highly reliable tests in the presence of low base rate values for $P(A)$ to attain even mediocre PPV values.

A Personal Favorite: Prostate Screening

Besides a better understanding of how PPVs are determined, there is a need to recognize that even when a true positive exists, not every disease needs to be treated.

In prostate cancer screening, for example, the worst danger is one of overdiagnosis and overtreatment, leading to more harm than good (see, for example, Gina Kolata, "Studies Show Prostate Test Save Few Lives," *New York Times*, March 19, 2009).

When I informed my doctor that I no longer would give blood for a PSA screening test, she agreed completely – the only reason such tests were done routinely was to practice "defensive medicine" on behalf of their clinics, and to prevent possible lawsuits arising from such screening tests not being administered routinely.

In other words, clinics get sued for underdiagnosis but not for overdiagnosis and overtreatment.

Other Screening Difficulties Discussed in the Readings

Premarital screenings for HIV – (a new "Wasserman" test, but not for syphilis)

Prenatal screening for Down's syndrome

The Screening efforts of the Transportation Security Administration (TSA)

Colon cancer screening through sigmoidoscopy or colonoscopy

Computer tomography scans for lung cancer (CT scans)

Life-line Screening: A Final Cautionary Example

www.lifelinescreening.com