1 Diagnostic Test Evaluation

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 to follow on the Psychopathy Checklist, Screening Version (PCL:SV)). 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 discussed below using three possible cutscores.

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

Figure 1: An ROC curve for a diagnostic test having just one cutscore.

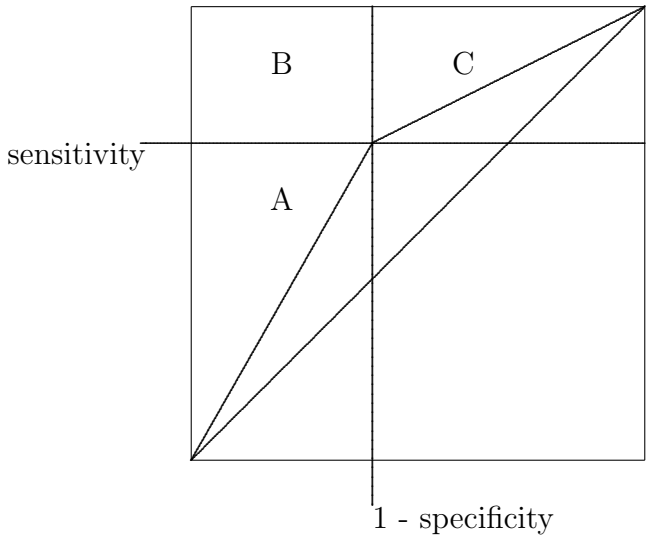
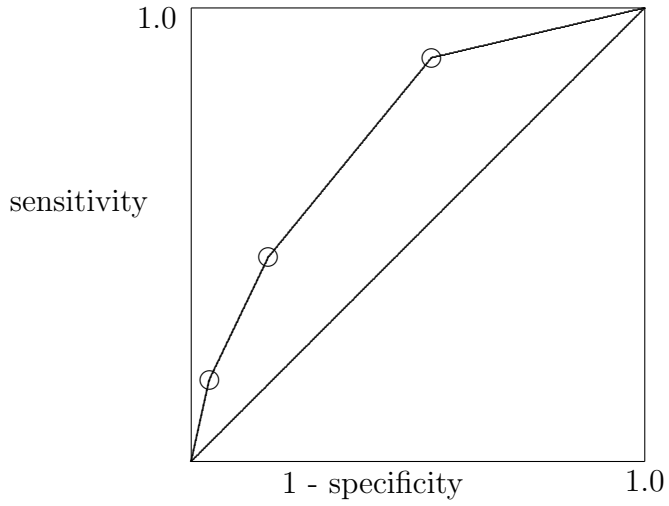


Figure 2: An ROC curve for the PCL:SV having three cutscores.



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.

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.

1.1 An Example Using the Psychopathy Checklist, Screening Version (PCL:SV): Data From the MacArthur Risk Assessment Study

The Psychopathy Checklist, Screening Version (PCL:SV) is the single best variable for the prediction of violence based on the data from the MacArthur Risk Assessment Study. It consists of twelve items, with each item being 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

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

Based on the MacArthur Risk Assessment Study data of Table 1, the three cutscores of 6, 12, and 18 were used to predict violence at followup (that is, when above or at a specific cutscore, predict "violence"; when below the cutscore, predict "nonviolence"). The basic statistics for the various diagnostic test results are given below:

Table 1: Data from the MacArthur Risk Assessment Study on the Psychopathy Checklist, Screening Version.

PCL-SV	block	violence at followup		block	totals
Score	yes	yes	no	no	
0		0	34		34
1		1	45		46
2		1	54		55
3		6	48		54
4	18	1	57	328	58
5		4	41		45
6		5	49		54
7		8	51		59
8		10	57		67
9		13	38		51
10	69	9	40	254	49
11		16	31		47
12		13	37		50
13		12	19		31
14		9	14		23
15		7	26		33
16	43	3	13	93	16
17		7	10		17
18		5	11		16
19		10	10		20
20		5	6		11
21		4	1		5
22	29	5	5	26	10
23		0	2		2
24		5	2		7
totals		159	701		860

Cutscore of 6:

		violence		
		Yes (A)	No ($\bar{A}$)	row sums
prediction	Yes (B)	141	373	414
	No ($\bar{B}$)	18	328	446
	column sums	159	701	860

accuracy: $(141 + 328)/860 = .55$
base rate: $(373 + 328)/860 = 701/860 = .815 \approx .82$
sensitivity: $141/159 = .89$
specificity: $328/701 = .47$
positive predictive value: $141/414 = .34$
negative predictive value: $328/446 = .74$

Cutscore of 12:

		violence		
		Yes (A)	No ($\bar{A}$)	row sums
prediction	Yes (B)	72	119	191
	No ($\bar{B}$)	87	582	669
	column sums	159	701	860

accuracy: $(72 + 582)/860 = .76$
base rate: $701/860 = .815 \approx .82$
sensitivity: $72/159 = .45$
specificity: $582/701 = .83$
positive predictive value: $72/191 = .38$
negative predictive value: $582/669 = .87$

Cutscore of 18:

		violence		
		Yes (A)	No ($\bar{A}$)	row sums
prediction	Yes (B)	29	26	55
	No ($\bar{B}$)	130	675	805
	column sums	159	701	860

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: $675/701 = .96$

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

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

As noted earlier, a common measure of diagnostic adequacy is the area under the ROC curve (or AUC). Figure 2 gives the ROC plot for the PCL:SV data based on the following sensitivity and 1.0 − specificity values:

cutscore	sensitivity	specificity	1 - specificity
6	.89	.47	.53
12	.45	.83	.17
18	.18	.96	.04

The AUC in this case has a value of .73, as computed in the section to follow. Only the cutpoint of 18 gives a better accuracy than using base rates, and even here, the accuracy is only minimally better than with the use of base rates: $704/860 = .819 > 701/860 = .815$. Also, the area under the ROC curve is not necessarily a good measure of clinical efficiency because it does not incorporate base rates. It is only a function of the test itself and not of its use on a sample of individuals.

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:

$P(B|A) = n_{BA}/n_A =$ sensitivity
$P(\bar{B}|\bar{A}) = n_{\bar{B}\bar{A}}/n_{\bar{A}} =$ specificity
AUC $= ((n_{BA}/n_A) + (n_{\bar{B}\bar{A}})/n_{\bar{A}})/2$
accuracy $= (n_{BA} + n_{\bar{B}\bar{A}})/n$ $(= P(A|B)P(B) + P(\bar{A}|\bar{B})P(\bar{B}))$

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, the AUC can be a poor measure of diagnostic test usage in a particular sample. We will come back to this issue shortly and suggest several alternative measures to the AUC that do take base rates into consideration when evaluating the use of diagnostic tests in populations where one of the base rates may be small, such as in the prediction of "dangerous" behavior.

1.2 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. Given two groups of individuals each with a score on some test, the Wilcoxon test statistic can be interpreted as follows: choose a pair of individuals at random (and with replacement) from the two groups (labeled A and $\bar{A}$, say, in anticipation of usage to follow), and assess whether the group A score is greater than the group $\bar{A}$ score. If this process is continued and the proportion of group A scores greater than those from group $\bar{A}$ is computed, this later value will converge to the proportion of all possible pairs constructed from the groups A and $\bar{A}$ in which the value for the A group member is greater than or equal to that for the $\bar{A}$ group member. In particular, we ask for the probability that in a randomly selected pair of people, where one committed violence and the other did not, the psychopathy score for the person committing violence is greater than that for the person not committing violence. This is the same as the two-sample Wilcoxon statistic (with a caveat that we will need to have a way of dealing with ties); it is also an interpretation for the AUC.

What follows is an example of the Wilcoxon test statistic calculation that relates directly back to the PCL:SV results of Table 1 and the computation of the AUC for Figure 2. Specifically, we compute the Wilcoxon statistic for a variable with four ordinal levels (I, II, III, and IV, with the IV level being the highest, as it is in the PCL:SV example):

	Violence Present	
	Yes (A)	No ($\bar{A}$)
I	m_{11}	m_{12}
II	m_{21}	m_{22}
III	m_{31}	m_{32}
IV	m_{41}	m_{42}
totals	n_A	$n_{\bar{A}}$

There is a total of $n_A n_{\bar{A}}$ pairs that can be formed from groups A and $\bar{A}$. The number of pairs for which the group A score is strictly greater than the group $\bar{A}$ score is:

$$\{m_{12}(m_{21} + m_{31} + m_{41})\}+$$
$$\{m_{22}(m_{31} + m_{41})\}+$$
$$\{m_{32}(m_{41})\}$$

The number of pairs for which there is a tie on the ordinal variable is:

$$(m_{11}m_{12}) + (m_{21}m_{22}) + (m_{31}m_{32}) + (m_{41}m_{42})$$

By convention, the Wilcoxon test statistic is the number of "strictly greater" pairs plus one-half of the "tied" pairs, all divided by the total number of pairs:

$$[\{m_{12}(m_{21} + m_{31} + m_{41}) + (1/2)(m_{11}m_{12})\}+$$
$$\{m_{22}(m_{31} + m_{41}) + (1/2)(m_{21}m_{22})\}+$$
$$\{m_{32}(m_{41}) + (1/2)(m_{31}m_{32})\} + \{(1/2)(m_{41}m_{42})\}]/[n_A n_{\bar{A}}]$$

For the PCL:SV results of Table 1:

	Violence Present		
	Yes (A)	No ($\bar{A}$)	row totals
I	18	328	346
II	69	254	323
III	43	93	136
IV	29	26	55
column totals	159	701	860

the Wilcoxon test statistic = 81,701.5/111,459.0 = .73 = AUC.

Using only the cutscore of 18:

	Violence Present	
	Yes(A)	No($\bar{A}$)
(No) (I + II + III)	130	675
(Yes) (IV)	29	26
column totals	159	701

the Wilcoxon statistic =

$$[(675)(29) + (1/2)(675)(130) + (1/2)(26)(29)]/[(159)(701)] = .57 ;$$

here, the AUC is merely defined by the average of sensitivity and specificity: $(.18 + .96)/2 = .57$

The relation just shown numerically can also be given in the notation used for the general Wilcoxon test:

sensitivity = m_{21}/n_A

specificity = $m_{12}/n_{\bar{A}}$

So, the average of sensitivity and specificity $((1/2)((m_{21}/n_A)+(m_{12}/n_{\bar{A}})))$ is equal to (after some algebra) the Wilcoxon statistic $(m_{12}m_{21} + (1/2)m_{22}m_{21} + (1/2)m_{11}m_{12})$.