Module 4: 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)

Abstract: Two main questions are discussed that relate to diagnostic testing. First, when does prediction using simple base rate information outperform prediction with an actual diagnostic test?; and second, how should the performance of a diagnostic test be evaluated in general? Module 2 on the (un)reliability of clinical and actuarial prediction introduced the Meehl and Rosen (1955) notion of "clinical efficiency," which is a phrase applied to a diagnostic test when it outperforms base rate predictions. In the first section to follow, three equivalent conditions are given for when "clinical efficiency" holds; these conditions are attributed to Meehl and Rosen (1955), Dawes (1962), and Bokhari and Hubert (2015). The second main section of this module introduces the Receiver Operating Characteristic (ROC) curve, and contrasts the use of a common measure of test performance, the "area under the curve" (AUC), with possibly more appropriate performance measures that take base rates into consideration. A final section of the module discusses several issues that must be faced when implementing screening programs: the evidence for the (in)effectiveness of cancer screening for breast (through mammography) and prostate (through the prostate-specific antigen (PSA) test); premarital screening debacles; prenatal screening; the cost of screening versus effectiveness; the ineffectiveness of airport behavioral detection programs implemented by the Transportation

Security Administration (TSA); informed consent and screening; the social pressure to screen.

Contents

1 Clinical Efficiency

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:

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

As in the introductory Module 1, a physical "urn" model is tacitly assumed that will generate a probability distribution according to the frequency distribution just given. There are n such balls in the urn with each ball labeled B or $\bar{B}$ and A or $\bar{A}$. There are n_{BA} balls with the labels B and A; $n_{B\bar{A}}$ balls with the labels B and $\bar{A}$; $n_{\bar{B}A}$ balls with the labels $\bar{B}$ and A; $n_{\bar{B}\bar{A}}$ balls with the labels $\bar{B}$ and $\bar{A}$. When a single ball is chosen from the urn "at random" and the two labels observed, a number of different event probabilities (and conditional probabilities) can be defined. For example, $P(B) = n_B/n$; $P(A) = n_A/n$; $P(A \text{ and } B) = n_{BA}/n$; $P(A|B) = n_{BA}/n_B$; and so on.

Using the urn model and conditionalizing on the state of nature, a number of common terms can be defined that are relevant to a diagnostic testing context:

		state of nature	
		A (pos)	$\bar{A}$ (neg)
diagnostic	B (pos)	$P(B\|A) = n_{BA}/n_A$ (sensitivity)	$P(B\|\bar{A}) = n_{B\bar{A}}/n_{\bar{A}}$ (false positive)
test 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."

Using the urn model and conditionalizing on the diagnostic test results, several additional terms relevant to a diagnostic testing context can be defined::

		state of nature		
		A (pos)	$\bar{A}$ (neg)	row sums
diagnostic	B (pos)	$P(A\|B) = n_{BA}/n_B$ (positive predictive value)	$P(\bar{A}\|B) = n_{B\bar{A}}/n_B$	$\frac{n_{BA}+n_{B\bar{A}}}{n_B} = 1.0$
test result	$\bar{B}$ (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)	$\frac{n_{\bar{B}A}+n_{\bar{B}\bar{A}}}{n_{\bar{B}}} = 1.0$

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."

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. Thus, 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). This last reference provides a formal proof of equivalence.

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.

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 (positive)	$\bar{A}$ (negative)	row sums
prediction	B (positive)	19	36	55
	$\bar{B}$ (negative)	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 (= (10 + 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.

1.1 Measuring the Degree of Clinical Efficiency

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."

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 ($n_{\bar{B}A}$).

As a way of assessing the degree of clinical efficiency, the Goodman-Kruskal (λ) Index of Prediction Association can be adopted. The lambda coefficient 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 naive prediction based just on marginal column frequencies. For the 2×2 contingency table of frequencies, it is defined as:

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

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. In other words, no differential prediction of a column event is made based on knowledge of what particular row an object belongs to. A non-zero $\lambda_{\text{column|row}}$ is just another way of specifying the Bokhari-Hubert differential prediction condition. The upper limit for $\lambda_{\text{column|row}}$ is 1.0, which corresponds to perfect prediction with the diagnostic test, and where test accuracy is 1.0.

To justify $\lambda_{\text{column|row}}$ as an index of clinical efficiency through a "proportional reduction in error measure," suppose the Bokhari-Hubert condition holds for the 2×2 contingency table and assume that $P(A) \leq 1/2$. Now, consider a ball picked randomly from the urn, and that we are asked to predict the "state of nature" in the absence of any information about the diagnostic test result; we would predict $\bar{A}$ (negative) and be wrong with probability $n_A/n = P(A)$. If asked to predict the "state of nature" but were told there is a diagnostic test result of B (positive) for this randomly selected ball, we would predict A (positive) and be wrong $n_{B\bar{A}}/n_B = P(\bar{A}|B)$. If the test result is $\bar{B}$ (negative), we would predict $\bar{A}$ (negative) and be wrong with probability $n_{\bar{B}A}/n_B = P(A|\bar{B})$. Thus, incorporating the probability of picking a ball from B or $\bar{B}$, the probability of error when given the diagnostic test result must be $P(\bar{A}|B)P(B)+P(A|\bar{B})P(\bar{B})$. Recalling that the probability of error when not knowing the diagnostic test result is $P(A)$, consider the proportional reduction in error measure defined by

$$\frac{P(A) - [P(\bar{A}|B)P(B) + P(A|\bar{B})P(\bar{B})]}{P(A)} \; .$$

After some simple algebra, this reduces to $\lambda_{\text{column|row}}$.

It might be noted in passing that significance testing in a 2×2 table with the usual chi-squared test of association tells us nothing about differential prediction. For example, the chi-squared test could show a significant relation between the A and $\bar{A}$, and the B and $\bar{B}$ events, but if $\lambda_{\text{column|row}}$ is zero, there is no differential prediction, and therefore base rates will outperform the use of a diagnostic test. More generally, when attempting to predict an event having a low base rate (for example, "dangerous") by using a "test" possessing less than ideal sensitivity and specificity values, it is common to be more accurate in prediction merely by using the larger base rate (for example, "not dangerous") rather than the diagnostic test.

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").[1]

The Bokhari and Hubert paper (2015) that discusses the three equivalent statements for clinical efficiency, also gives a generalized clinical efficiency condition (a generalized Bokhari-Hubert condition [GBH]) that allows for the assignment of unequal costs to the false

[1]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."

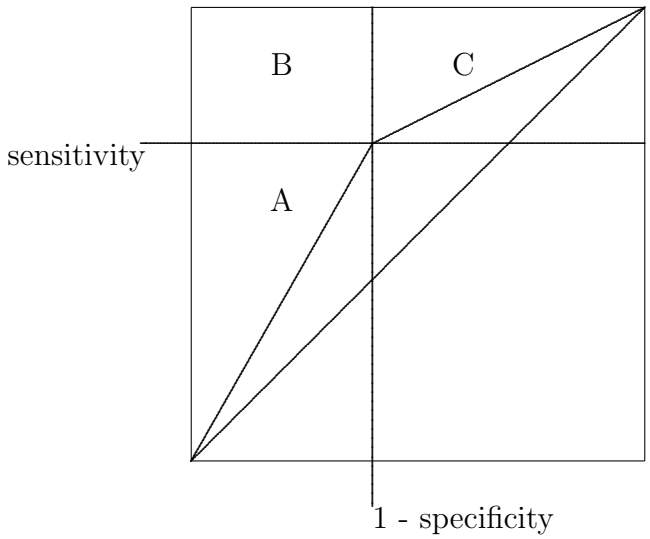
positives and false negatives. Depending on how the costs of misclassification are assigned, a determination can be made as to when generalized clinical efficiency holds; that is, when is the total costs of using a test less than the total costs obtained by just classifying through base rates? Further, depending on the specific data available in the 2×2 contingency table (such as that for the COVR instrument given earlier in this section), statements such as the following can be made based on explicit bounds given in Bokhari and Hubert (2015): for generalized clinical efficiency to hold, false negatives (releasing a dangerous person) cannot be considered more than 10.3 times more costly than false positiveS (detaining a non-dangerous person); also, one needs to have false negatives be more than twice as costly as false positives. So, in summary, false negatives must be at least twice as costly as false positives but no more than about ten times as costly.

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).

2 Diagnostic Test Evaluation

The Receiver Operating Characteristic (ROC) curve of a diagnostic test is a plot of test sensitivity (the probability of a "true" posi-

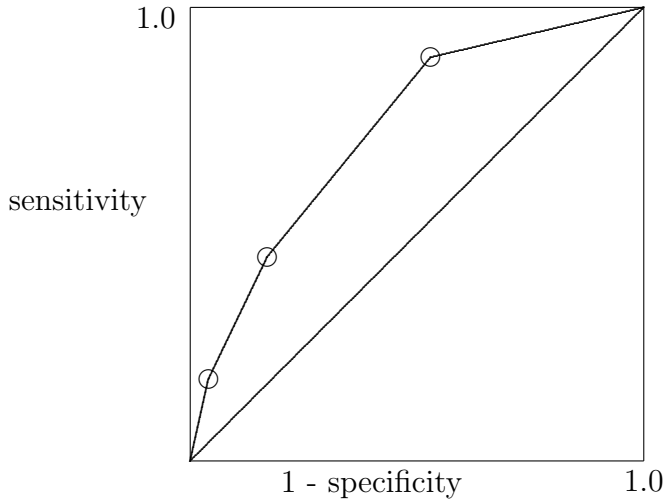
Figure 1: An ROC curve for a diagnostic test having just one cutscore.



tive) 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 $1.0 -$

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



specificity values attached to the possible cutscores. 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.

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.

2.1 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:

Cutscore of 6:

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

		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.

2.2 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 $(m_{12}m_{21} + (1/2)m_{22}m_{21} + (1/2)m_{11}m_{12})$.

2.3 A Modest Proposal for Evaluating a Diagnostic Test When Different Cutscores Can Be Set

One suggestion for evaluating a diagnostic test when different cutscores are possible is to set a cutscore so that the proportion of positive predictions is "close" to the prior probability of a positive "state of nature" — and to then look at the consistency of subject classifications by A and B and by $\bar{A}$ and $\bar{B}$. To give an example, we use the PCL:SV data and a cutscore of 13:

		violence		
		Yes (A)	No ($\bar{A}$)	row sums
prediction	Yes (B)	60	100	160
	No ($\bar{B}$)	99	601	700
	column sums	159	701	860

accuracy: $(60 + 601)/860 = .77$
base rate: $701/860 = .815 \approx .82$
sensitivity: $60/159 = .38$
specificity: $601/701 = .86$
positive predictive value: $60/160 = .38$
negative predictive value: $601/700 = .86$

Here, $P(A \cap B | A \cup B)$ = the proportion of positive classifications (by A or B) that are consistent = $60/(60 + 100 + 99) = 60/259 = .23$; so, only 1/4 of the time are the positive classifications consistent; $P(\bar{A} \cap \bar{B} | \bar{A} \cup \bar{B})$ = the proportion of negative classifications (by $\bar{A}$ or $\bar{B}$) that are consistent = $601/(601 + 100 + 99) = 601/800 = .75$; so, 3/4 of the time the negative classifications are consistent.[2]

[2]These two types of consistency index just presented may be of particular value when

Note that from Bayes' theorem, we have the two statements:

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

and

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

If $P(A) = P(B)$ (and thus, $P(\bar{A}) = P(\bar{B})$), $P(A|B) = P(B|A)$ and $P(\bar{A}|\bar{B}) = P(\bar{B}|\bar{A})$. Or, in words, the positive predictive value is equal to the sensitivity, and the negative predictive value is equal to the specificity. This is seen numerically in the example given above where $P(A)$ and $P(B)$ are very close (that is, $P(A) = .185$; $P(B) = .186$).

Possibly the use of these measures will eliminate the terminological confusion about what a "false positive" means; one usual interpretation is 1 - specificity (which does not take base rates into account): the probability that the test is positive given that the person doesn't have "it"; the other is 1 - the negative predictive value (which does take base rates into account): the probability that the person has "it" given that the test is negative. Also, for a "false negative," the usual interpretation is 1 - sensitivity (which does not take base rates into account): the probability that the test is negative given that the person has "it"; the other is 1 - positive predictive value (which does take base rates into account): the probability that the person doesn't have "it" given that the test is positive. By equating $P(A)$

two distinct diagnostic tests are to be compared. Here, no explicit "state of nature" pair of events (A and $\bar{A}$) would be available, but one of the diagnostic tests would serve the same purpose.

and $P(B)$, the confusions about the meaning of a "false positive" and a "false negative" can be finessed because different interpretations can be given as to what is "false" and what is "positive" or "negative."

Because of the equivalence of sensitivity and the positive predictive value and of specificity and the negative predictive value when the base rates $P(A)$ and $P(B)$ are equal, another measure of diagnostic accuracy but one that does take base rates into account would be the simple average of the positive and negative predictive values. This would correspond to an AUC measure for the single cutpoint that equalizes the base rates $P(A)$ and $P(B)$; that AUC measure would be, as usual, the simple average of specificity and sensitivity.

3 Summary Comments

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.

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.

A simple condition given in an earlier section of this module (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."

As an another measure of diagnostic accuracy we might consider a weighted function of the positive and negative predictive values, such as the simple proportion of correct decisions. When the positive and negative predictive values are each weighted by the probabilities that the diagnostic test is positive or negative, and these values then summed, the simple measure of accuracy (defined as the proportion of correct decisions) is obtained.

Just saying that a measure is "good" because it is independent of base rates doesn't make it "good" for the use to which it is being put (or, in the jargon of computer science, a "bug" doesn't suddenly become a "feature" by bald face assertion). As an example from the MacArthur data given in Module 2 on the cross-validation of an

actuarial model of violence risk assessment, the AUC would be given as the simple average of sensitivity and specificity (AUC = (.68 + .72)/2 = .70). This number tells us precious little of importance in how the diagnostic test is doing with the cross-validation sample. The (very low) accuracy or "hit-rate" measure is .71, which is worse than just using the base rate (.82) and predicting that everyone will be "not dangerous." Using the test, 2 out of 3 predictions of dangerousness are wrong; 1 out of 11 predictions of "not dangerous" are wrong. It is morally questionable to have one's liberty jeopardized by an assessment of being " dangerous" that is wrong 2 out of 3 times (or, in some Texas cases, one's life, such as in Barefoot v. Estelle (1983) discussed at length in Module 2).

In contrast to some incorrect understandings in the literature about the invariance of specificity and sensitivity across samples, sizable subgroup variation 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 estimates 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 operation of Berkson's 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, and so on.

4 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, $(1 - \text{PPV})$, 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.

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 the case of another personal favorite of ours, prostate cancer screening, its low accuracy makes mammograms look good, where 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). Armed with this information, we no longer give blood for a PSA screening test. When we so informed our doctors as to our wishes, they 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.[3]

[3]We list several additional items that are relevant to screening: an article by Sandra G. Boodman for the *AARP Bulletin* (Januaary 1, 2010) summarizes well what its title offers: "Experts Debate the Risks and Benefits of Cancer Screening." A cautionary example of breast cancer screening that tries to use dismal specificity and sensitivity values for detecting the HER2 protein, is by Gina Kolata, "Cancer Fight: Unclear Tests for New Drug," *New York Times*, April 19, 2010). The reasons behind proposing cancer screening guidelines and the contemporary emphasis on evidence-based medicine is discussed by Gina Kolata in "Behind Cancer Guidelines, Quest for Data" (*New York Times*, November 22, 2009). Other articles that involve screening discuss how a fallible test for ovarian cancer (based on the CA-125 protein) might be improved using a particular algorithm to monitor CA-125 fluctuations more precisely (Tom Randall, *Bloomberg Businessweek*, May 21, 2010, "Blood Test for Early Ovarian Cancer May Be Recommended for All"); three items by Gina Kolata concern food allergies (or nonallergies, as the case may be) and a promising screening test for Alzheimer's: "Doubt Is Cast on Many Reports of Food Allergies" (*New York Times*, May 11, 2010); and "I Can't Eat That. I'm Allergic" (*New York Times*, May 15, 2010); "Promise Seen for Detection of Alzheimer's" (*New York Times*, June 23, 2010); a final item to mention discusses a promising alternative to mammogram screening: "Breast Screening Tool Finds Many Missed Cancers" (Janet Raloff, *ScienceNews*, July 1, 2010).

A good way to conclude this discussion of issues involving (cancer) screening is to refer the reader to three items from the *New York Times*: an OpEd article ("The Great Prostate Mistake," March 9, 2010) by Richard J. Ablin, a recent piece by Gina Kolata summarizing a large longitudinal randomized controlled Canadian study on the value of mammograms ("Vast Study Casts Doubt On Value of Mammograms"; February 11, 2014), and a second article by Gina Kolata on the severe overdiagnosis of thyroid cancer in South Korea.

Dr. Ablin is a research professor of immunobiology and pathology at the University of Arizona College of Medicine, and President of the Robert Benjamin Ablin Foundation for Cancer Research. Most importantly for our purposes, he is the individual who in 1970 discovered the PSA test for detecting prostate cancer; his perspective on the issues is therefore unique:[4]

[4]To show the ubiquity of screening appeals, we reproduce a solicitation letter to LH from Life Line Screening suggesting that for only $139, he could get four unnecessary screenings right in Champaign, Illinois, at the Temple Baptist Church:

Dear Lawrence,

Temple Baptist Church in Champaign may not be the location that you typically think of for administering lifesaving screenings. However, on Tuesday, September 22, 2009, the nation's leader in community-based preventive health screenings will be coming to your neighborhood.

Over 5 million people have participated in Life Line Screening's ultrasound screenings that can determine your risk for stroke caused by carotid artery diseases, abdominal aortic aneurysms and other vascular diseases. Cardiovascular disease is the #1 killer in the United States of both men and women—and a leading cause of permanent disability.

Please read the enclosed information about these painless lifesaving screenings. A package of four painless Stroke, Vascular Disease & Heart Rhythm screenings costs only $139. Socks and shoes are the only clothes that will be removed and your screenings will be completed in a little more than an hour.

You may think that your physician would order these screenings if they were necessary. However, insurance companies typically will not pay for screenings unless there are symptoms. Unfortunately, 4 out of 5 people that suffer a stroke have no apparent symptoms or warning signs. That is why having a Life Line Screening is so important to keep you and

I never dreamed that my discovery four decades ago would lead to such a profit-driven public health disaster. The medical community must confront reality and stop the inappropriate use of P.S.A. screening. Doing so would save billions of dollars and rescue millions of men from unnecessary, debilitating treatments.

Several excerpts are provided below from the Gina Kolata article on the Canadian mammogram study:

One of the largest and most meticulous studies of mammography ever done, involving 90,000 women and lasting a quarter-century, has added powerful new doubts about the value of the screening test for women of any age.

It found that the death rates from breast cancer and from all causes were the same in women who got mammograms and those who did not. And the screening had harms: One in five cancers found with mammography and treated was not a threat to the woman's health and did not need treatment such as chemotherapy, surgery or radiation.

The study, published Tuesday in *The British Medical Journal*, is one of the few rigorous evaluations of mammograms conducted in the modern era of more effective breast cancer treatments. It randomly assigned Canadian

---

your loved ones healthy and independent.

"These screenings can help you avoid the terrible consequences of stroke and other vascular diseases. I've seen firsthand what the devastating effects of stroke, abdominal aortic aneurysms and other vascular diseases can have on people and I feel it is important that everyone be made aware of how easily they can be avoided through simple, painless screenings." — Andrew Monganaro, MD, FACS, FACC (Board Certified Cardiothoracic and Vascular Surgeon)

I encourage you to talk to your physician about Life Line Screening. I am confident that he or she will agree with the hundreds of hospitals that have partnered with us and suggest that you participate in this health event.

We are coming to Champaign for one day only and appointments are limited, so call 1-800-395-1801 now.

Wishing you the best of health,
Karen R. Law, RDMS, RDCS, RVT
Director of Clinical Operations

women to have regular mammograms and breast exams by trained nurses or to have breast exams alone.

Researchers sought to determine whether there was any advantage to finding breast cancers when they were too small to feel. The answer is no, the researchers report.

...

Dr. Kalager, whose editorial accompanying the study was titled "Too Much Mammography," compared mammography to prostate-specific antigen screening for prostate cancer, using data from pooled analyses of clinical trials. It turned out that the two screening tests were almost identical in their overdiagnosis rate and had almost the same slight reduction in breast or prostate deaths.

"I was very surprised," Dr. Kalager said. She had assumed that the evidence for mammography must be stronger since most countries support mammography screening and most discourage PSA screening.

Finally, and as noted above, a recent example of a medical screening fiasco and the resulting overdiagnoses and overtreatments, involves thyroid cancer, and the detection of tiny and harmless tumors that are better left undisturbed. The situation is particularly serious in South Korea, as pointed out by the excerpts given below from an article by Gina Kolata ("Study Warns Against Overdiagnosis of Thyroid Cancer," *New York Times*, November 5, 2014):

To the shock of many cancer experts, the most common cancer in South Korea is not lung or breast or colon or prostate. It is now thyroid cancer, whose incidence has increased fifteenfold in the past two decades. "A tsunami of thyroid cancer," as one researcher puts it.

Similar upward trends for thyroid cancer are found in the United States and Europe, although not to the same degree. The thyroid cancer rate in the United States has more than doubled since 1994.

Cancer experts agree that the reason for the situation in South Korea and elsewhere is not a real increase in the disease. Instead, it is down to screening,

which is finding tiny and harmless tumors that are better left undisturbed, but that are being treated aggressively.

South Koreans embraced screening about 15 years ago when the government started a national program for a variety of cancers – breast, cervix, colon, stomach and liver. Doctors and hospitals often included ultrasound scans for thyroid cancer for an additional fee of $30 to $50.

Since South Korea adopted widespread cancer screening in 1999, thyroid cancer has become the most diagnosed cancer in the country. But if this early detection were saving lives, the already-low death rate from thyroid cancer should have fallen, not remained steady.

In the United States and Europe, where there are no formal, widespread screening programs for thyroid cancer, scans for other conditions, like ultrasound exams of the carotid artery in the neck or CT scans of the chest, are finding tiny thyroid tumors.

Although more and more small thyroid cancers are being found, however, the death rate has remained rock steady, and low. If early detection were saving lives, death rates should have come down.

That pattern – more cancers detected and treated but no change in the death rate – tells researchers that many of the cancers they are finding and treating were not dangerous. It is a phenomenon that researchers call overdiagnosis, finding cancers that did not need treatment because they were growing very slowly or not at all. Left alone, they would probably never cause problems. Overdiagnosis is difficult to combat. Pathologists cannot tell which small tumors are dangerous, and most people hear the word "cancer" and do not want to take a chance. They want the cancer gone.

But cancer experts said the situation in South Korea should be a message to the rest of the world about the serious consequences that large-scale screening of healthy people can have.

"It's a warning to us in the U.S. that we need to be very careful in our advocacy of screening," said Dr. Otis W. Brawley, chief medical officer at the American Cancer Society. "We need to be very specific about where we have good data that it saves lives."

Colon cancer screening wins Dr. Brawley's unqualified endorsement. Breast cancer screening saves lives, he said, and he advocates doing it, but he said it

could also result in overdiagnosis. Even lung cancer screening can be susceptible to overdiagnosis, with as many as 18 percent of patients treated when they did not need to be, Dr. Brawley said.

The soaring increase in thyroid cancers in South Korea is documented in a paper published on Thursday in the New England Journal of Medicine. The authors report not only that the number of diagnoses escalated as screening became popular, but also that the newly detected cancers were almost all very tiny ones. These tiny cancers, called papillary thyroid cancers, are the most common kind and are the sort typically found with screening. They are known to be the least aggressive.

The epidemic was not caused by an environmental toxin or infectious agent, said Dr. H. Gilbert Welch of Dartmouth, an author of the paper. "An epidemic of real disease would be expected to produce a dramatic rise in the number of deaths from disease," he said. "Instead we see an epidemic of diagnosis, a dramatic rise in diagnosis and no change in death."

Cancer experts stress that some thyroid cancers are deadly – usually they are the larger ones. And, they say, if a person notices symptoms like a lump on the neck or hoarseness, they should not be ignored.

"But there is a real difference between not ignoring something obvious and telling the population to try really hard to find something wrong," Dr. Welch said.

Thyroid cancer tends to be particularly indolent. On autopsy, as many as a third of people have tiny thyroid cancers that went undetected in their lifetime. Once a cancer is found, though, treatment is onerous and involves removing the thyroid. Patients must then take thyroid hormones for the rest of their lives. For some, Dr. Brawley said, the replacement hormones are not completely effective, and they end up with chronically low thyroid hormone levels, feeling depressed and sluggish as a result.

In a small percentage of those having thyroid surgery, surgeons accidentally damage the nearby vocal cords – that happened to the 2 percent of South Korean patients who ended up with vocal cord paralysis. Or they damage the parathyroid glands, tiny yellow glands just behind the thyroid that control calcium levels in the body. When the parathyroids are damaged, as happened in 11 percent of patients in South Korea, patients get hypoparathyroidism, a

difficult condition to treat.

In South Korea, some doctors, including Dr. Hyeong Sik Ahn of the College of Medicine at Korea University in Seoul, the first author of the new paper, have called for thyroid cancer screening to be banned. But their calls were mostly ignored, Dr. Ahn explained in an email. "Most thyroid doctors, especially surgeons, deny or minimize harms."

Thyroid experts in the United States are calling for restraint in diagnosing and treating tiny tumors. A few places, like Memorial Sloan-Kettering Cancer Center in Manhattan, offer patients with small tumors the option of simply waiting and having regular scans to see if the tumor grows. But few patients have joined the program.

"Once we have made a diagnosis of cancer it is difficult to say, 'Don't do anything,'" said Dr. Ashok R. Shaha, a thyroid cancer surgeon at Memorial Sloan-Kettering who is concerned about the zeal to diagnose and treat tiny tumors. Doctors as well as patients can be wary, he said. "In the U.S. we have a fear that if we miss a cancer the patient will sue."

Dr. R. Michael Tuttle, who runs the wait-and-see program at Memorial-Sloan Kettering, said the best way to encourage observation of very low-risk thyroid cancer instead of aggressive treatment was to "stop the diagnosis." That means, he said, "decrease screening and decrease F.N.A.," meaning fine needle aspiration, which is used to examine thyroid lumps noticed coincidentally.

And the lesson from South Korea should be heeded, said Dr. Barnett S. Kramer, director of the division of cancer prevention at the National Cancer Institute.

"The message for so long is that early detection is always good for you," he said. But this stark tale of screening gone wrong "should acutely raise awareness of the consequences of acting on the intuition that all screening must be of benefit and all diagnoses at an early stage are of benefit."

Before we leave the topic of medical screening completely, there are several additional issues having possible ethical and probabilistic implications that should at least be raised, if only briefly:[5]

[5]Besides miscarriages of justice that result from confusions involving probabilities, others

Premarital screening: From the early part of the 20th century, it has been standard practice for states to require a test for syphilis before a marriage license was issued. The rationale for this requirement was so the disease was not passed on to a newborn in the birth canal, with the typical result of blindness, or to an unaffected partner. Besides requiring a test for syphilis, many states in the late 1980s considered mandatory HIV evaluations before marriage licenses were issued. Illinois passed such a law in 1987 that took effect on January 1, 1988, and continued through August of 1989. It was a public health disaster. In the first six months after enactment, the number of marriage licenses issued in Illinois dropped by 22.5%; and of the 70,846 licenses issued during this period, only eight applicants tested positive with a cost of $312,000 per seropositive identified individual. Even for the eight

have suffered because of failures to clearly understand the fallibility of diagnostic testing. Probably the most famous example of this is the disappearance of Azaria Chamberlain, a nine-week-old Australian baby who disappeared on the night of August 17, 1980, while on a camping trip to Ayers Rock. The parents, Lindy and Michael Chamberlain, contended that Azaria had been taken from their tent by a dingo. After several inquests, some broadcast live on Australian television, Lindy Chamberlain was tried and convicted of murder, and sentenced to life imprisonment. A later chance finding of a piece of Azaria's clothing in an area with many dingo lairs, lead to Lindy Chamberlain's release from prison and eventual exoneration of all charges.

The conviction of Lindy Chamberlain for the alleged cutting of Azaria's throat in the front seat of the family car rested on evidence of fetal hemoglobin stains on the seat. Fetal hemoglobin is present in infants who are six months old or younger—Azaria Chamberlain was only nine weeks old when she disappeared. As it happens, the diagnostic test for fetal hemoglobin is very unreliable, and many other organic compounds can produce similar results, such as nose mucus and chocolate milkshakes, both of which were present in the vehicle (in other words, the specificity of the test was terrible). It was also shown that a "sound deadener" sprayed on the car during its production produced almost identical results for the fetal hemoglobin test.

The Chamberlain case was the most publicized in Australian history (and on a par with the O.J. Simpson trial in the United States). Because most of the evidence against Lindy Chamberlain was later rejected, it is a good illustration of how media hype and bias can distort a trial.

identified as positive, the number of false positives was unknown; the more definitive follow-up Western blot test was not available at that time. This particular episode was the most expensive public health initiative ever for Illinois; the understated conclusion from this experience is that mandatory premarital testing is not a cost-effective method for the control of human immunodeficiency virus infection. For a further discussion of the Illinois experience in mandatory HIV premarital testing, see Turnock and Kelly (1989).

Prenatal screening: The area of prenatal screening inevitably raises ethical issues. Some screening could be labeled quickly as unethical, for example, when selective abortions occur as the result of an ultrasound to determine the sex of a fetus. In other cases, the issues are murkier.[6] For instance, in screening for Down's syndrome because of a mother's age, acting solely on the use of noninvasive biomedical markers with poor selectivity and sensitivity values is questionable; the further screening with more invasive methods, such as amniocentesis, may be justifiable even when considering an accompanying one to two percent chance of the invasive test inducing a miscarriage. At least in the case of screening for Down's syndrome, these trade-offs between invasive screening and the risk of spontaneous miscarriage may no longer exist given a new noninvasive DNA blood test announced in the *British Medical Journal* in January 2011, "Noninvasive Prenatal Assessment of Trisomy 21 by Multiplexed Maternal

[6]There is also the fear that increasingly sophisticated prenatal genetic testing will enable people to engineer "designer babies," where parents screen for specific traits and not for birth defects per se. The question about perfection in babies being an entitlement is basically an ethical one; should otherwise healthy fetuses be aborted if they do not conform to parental wishes? To an extent, some of this selection is done indirectly and crudely already when choices are made from a sperm bank according to desired donor characteristics.

Plasma DNA Sequencing: Large Scale Validity Study." The article abstract follows:

Objectives: To validate the clinical efficacy and practical feasibility of massively parallel maternal plasma DNA sequencing to screen for fetal trisomy 21 among high risk pregnancies clinically indicated for amniocentesis or chorionic villus sampling.

Design: Diagnostic accuracy validated against full karyotyping, using prospectively collected or archived maternal plasma samples.

Setting: Prenatal diagnostic units in Hong Kong, United Kingdom, and the Netherlands.

Participants: 753 pregnant women at high risk for fetal trisomy 21 who underwent definitive diagnosis by full karyotyping, of whom 86 had a fetus with trisomy 21.

Intervention: Multiplexed massively parallel sequencing of DNA molecules in maternal plasma according to two protocols with different levels of sample throughput: 2-plex and 8-plex sequencing.

Main outcome measures: Proportion of DNA molecules that originated from chromosome 21. A trisomy 21 fetus was diagnosed when the z-score for the proportion of chromosome 21 DNA molecules was greater than 3. Diagnostic sensitivity, specificity, positive predictive value, and negative predictive value were calculated for trisomy 21 detection.

Results: Results were available from 753 pregnancies with the 8-plex sequencing protocol and from 314 pregnancies with the 2-plex protocol. The performance of the 2-plex protocol was superior to that of the 8-plex protocol. With the 2-plex protocol, trisomy 21 fetuses were detected at 100% sensitivity and 97.9% specificity, which resulted in a positive predictive value of 96.6% and negative predictive value of 100%. The 8-plex protocol detected 79.1% of the trisomy 21 fetuses and 98.9% specificity, giving a positive predictive value of 91.9% and negative predictive value of 96.9%.

Conclusion: Multiplexed maternal plasma DNA sequencing analysis could be used to rule out fetal trisomy 21 among high risk pregnancies. If referrals for amniocentesis or chorionic villus sampling were based on the sequencing test results, about 98% of the invasive diagnostic procedures could be avoided.

Costs of screening: All screening procedures have costs attached, if only for the laboratory fees associated with carrying out the diagnostic test. When implemented on a more widespread public health basis, however, screenings may soon become cost-prohibitive for the results obtained. The short-lived premarital HIV screening in Illinois is one example, but new diagnostic screening methods seem to be reported routinely in the medical literature. These then get picked up in the more popular media, possibly with some recommendation for further broad implementation. A societal reluctance to engage in such a process may soon elicit a label of "medical rationing" (possibly, with some further allusion to socialized medicine, or what one can expect under "Obama-care").[7]

One recent example of a hugely expensive but (mostly) futile screening effort is by the Transportation Security Administration (TSA) and its airport passenger screening program. We give excerpts from three reports that appeared in the *New York Times* in 2013 and 2014:

"Report Says TSA Screening Is Not Objective" (Michael S. Schmidt, June 4, 2013) –

The Transportation Security Administration has little evidence that an airport passenger screening program, which some employees believe is a magnet for racial profiling and has cost taxpayers nearly one billion dollars, screens passengers objectively, according to a report by the inspector general for the Homeland Security Department.

[7]One possible mechanism that may be a viable strategy for keeping the cost of screenings under some control is through a clever use of statistics. Depending on what is being assessed (for example, in blood, soil, air), it may be possible to test a "pooled" sample; only when that sample turns out to be "positive" would the individual tests on each of the constituents need to be carried out.

The T.S.A.'s "behavioral detection program" is supposed to rely on security officers who pull aside passengers who exhibit what are considered telltale signs of terrorists for additional screening and questioning. It is illegal to screen passengers because of their nationality, race, ethnicity or religion.

According to the report, the T.S.A. has not assessed the effectiveness of the program, which has 2,800 employees and does not have a comprehensive training program. The T.S.A. cannot "show that the program is cost-effective, or reasonably justify the program's expansion," the report said.

As a result of the T.S.A.'s ineffective oversight of the program, it "cannot ensure that passengers at U.S. airports are screened objectively," the report said.

...

In August, *The Times* reported that more than 30 officers at Logan International Airport in Boston had said that the program was being used to profile passengers like Hispanics traveling to Florida or blacks wearing baseball caps backward.

The officers said that such passengers were being profiled by the officers in response to demands from managers who believed that stopping and questioning them would turn up drugs, outstanding arrest warrants or immigration problems.

The managers wanted to generate arrests so they could justify the program, the officers said, adding that officers who made arrests were more likely to be promoted. The Homeland Security Department said then that its inspector general was investigating the matter, although the coming report does not address the program at Logan Airport.

In a written statement, Representative Bennie Thompson, Democrat of Mississippi, the ranking member on the House Homeland Security Committee, said that the report "deals yet another blow to T.S.A.'s efforts to implement a behavioral detection screening program."

Mr. Thompson added that he would be offering an amendment to the Homeland Security appropriations bill this week that would "prevent any more taxpayer dollars from being spent on this failed and misguided effort."

"At Airports, A Misplaced Faith in Body Language" (John Tier-

ney, March 23, 2012) –

Like the rest of us, airport security screeners like to think they can read body language. The Transportation Security Administration has spent some $1 billion training thousands of "behavior detection officers" to look for facial expressions and other nonverbal clues that would identify terrorists.

But critics say there's no evidence that these efforts have stopped a single terrorist or accomplished much beyond inconveniencing tens of thousands of passengers a year. The T.S.A. seems to have fallen for a classic form of self-deception: the belief that you can read liars' minds by watching their bodies.

Most people think liars give themselves away by averting their eyes or making nervous gestures, and many law-enforcement officers have been trained to look for specific tics, like gazing upward in a certain manner. But in scientific experiments, people do a lousy job of spotting liars. Law-enforcement officers and other presumed experts are not consistently better at it than ordinary people even though they're more confident in their abilities.

"Theres an illusion of insight that comes from looking at a person's body," says Nicholas Epley, a professor of behavioral science at the University of Chicago. "Body language speaks to us, but only in whispers."

...

"The common-sense notion that liars betray themselves through body language appears to be little more than a cultural fiction," says Maria Hartwig, a psychologist at John Jay College of Criminal Justice in New York City. Researchers have found that the best clues to deceit are verbal – liars tend to be less forthcoming and tell less compelling stories – but even these differences are usually too subtle to be discerned reliably.

One technique that has been taught to law-enforcement officers is to watch the upward eye movements of people as they talk. This is based on a theory from believers in "neuro-linguistic programming" that people tend to glance upward to their right when lying, and upward to the left when telling the truth.

But this theory didn't hold up when it was tested by a team of British and North American psychologists. They found no pattern in the upward

eye movements of liars and truth tellers, whether they were observed in the laboratory or during real-life news conferences. The researchers also found that people who were trained to look for these eye movements did not do any better than a control group at detecting liars.

"Behavior Detection Isn't Paying Off" (The Editorial Board, April 6, 2014) –

A multiyear experiment in behavior detection is only worsening the Transportation Security Administration's reputation for wastefulness. Since 2007, the T.S.A. has trained officers to identify high-risk passengers on the basis of mostly nonverbal signs, like fidgeting or sweating, which may indicate stress or fear. The total price tag: nearly $1 billion.

In theory we're all for the T.S.A. devoting resources to human intelligence, but this particular investment does not appear to be paying off.

As John Tierney wrote in *The Times* on March 25, the T.S.A. "seems to have fallen for a classic form of self-deception: the belief that you can read liars' minds by watching their bodies." He cited experiments showing that people are terrible at spotting liars. One survey of more than 200 studies found that "people correctly identified liars only 47 percent of the time, less than chance."

The T.S.A.'s behavior-detection officers are no better. The Government Accountability Office told Congress in November that T.S.A. employees could not reliably single out dangerous passengers and that the program was ineffective.

In its review of 49 airports in 2011 and 2012, the G.A.O. calculated that behavior-detection officers designated passengers for additional screening on 61,000 occasions. From that group, 8,700, or 14 percent, were referred to law enforcement. Only 4 percent of the 8,700, or 0.6 percent of the total, were arrested – none for suspected terrorism. (The T.S.A. said the Federal Air Marshal Service earmarked certain cases for further investigation, but could not provide the G.A.O. with details.) The G.A.O. attributed these poor results to a general "absence of scientifically validated evidence" for training T.S.A. employees in the dark art of behavior detection, and urged Congress to limit future funding.

The union representing T.S.A. officers has defended the program, which costs roughly $200 million a year, arguing that an "imperfect deterrent to terrorist attacks is better than no deterrent at all." But behavior detection is far from the country's only shield, and "imperfect" is an understatement. Congress should take the G.A.O.'s advice.

Besides initial screening costs and those involved in dealing with follow-up procedures for all the false positives identified, there may also be costs involved in the particular choice among alternatives for a diagnostic procedure. If one strategy has demonstrable advantages but increased costs over another, based on an evidence-based assessment it still may be cost-effective to choose the higher-priced alternative. But if the evidence does not document such an advantage, it would seem fiscally prudent in controlling the increasing societal health-care costs to not choose the more expensive option as the default, irrespective of what professional pressure groups may want and who would profit the most from the specific choices made. A case in point is the use of colonoscopy in preference to sigmoidoscopy. We quote from a short letter to the editor of the *New York Times* by John Abramson (February 22, 2011) entitled "The Price of Colonoscopy":

Colon cancer screening with colonoscopy—viewing the entire colon—has almost completely replaced more limited sigmoidoscopy, which costs as little as one-tenth as much. Yet studies have repeatedly failed to show that colonoscopy reduces the risk of death from colon cancer more effectively than sigmoidoscopy.

A recent example of a breakthrough in medical screening for lung cancer that may end up being very cost-ineffective was reported in a *News of the Week* article by Eliot Marshall, appearing in *Sci-*

ence (2010), entitled "The Promise and Pitfalls of a Cancer Breakthrough." It reviews the results of a $250 million study sponsored by the National Cancer Institute (NCI) named the National Lung Screening Trial (NLST). The diagnostic test evaluated was a three-dimensional low-dose helical computed tomography (CT) scan of an individual's lung. Although Harold Varmus commented that he saw "a potential for saving many lives," others saw some of the possible downsides of widespread CT screening, including costs. For example, note the comments from the NCI Deputy Director, Douglas Lowy (we quote from the *Science* news item):[8]

[8]Continued from the main text:

In NLST (National Lung Screening Trial), about 25% of those screened with CT got a positive result requiring followup. Some researchers have seen higher rates. Radiologist Stephen Swensen of the Mayo Clinic in Rochester, Minnesota, says that a nonrandomized study he led in 2005 gave positive results for 69% of the screens. One difference between the Mayo and NLST studies, Swensen says, is that Mayo tracked nodules as small as 1 to 3 millimeters whereas NLST, which began in 2002, cut off positive findings below 4 mm.

One negative consequence of CT screening, Lowy said at the teleconference, is that it triggers follow-up scans, each of which increases radiation exposure. Even low-dose CT scans deliver a "significantly greater" exposure than conventional chest x-rays, said Lowy, noting that, "It remains to be determined how, or if, the radiation doses from screening ... may have increased the risks for cancer during the remaining lifetime" of those screened. Clinical followup may also include biopsy and surgery, Lowy said, "potentially risky procedures that can cause a host of complications."

G. Scott Gazelle, a radiologist and director of the Institute for Technology Assessment at Massachusetts General Hospital in Boston, has been analyzing the likely impacts of lung cancer screening for a decade. He agrees that people are going to demand it—and that "there are going to be a huge number of false positives." He was not surprised at NLST's finding of a lifesaving benefit of 20%. His group's prediction of mortality reduction through CT scans, based on "micromodeling" of actual cancers and data from previous studies, was 18% to 25%, right on target. But Gazelle says this analysis, now under review, still suggests that a national program of CT screening for lung cancer "would not be cost effective." Indeed, the costs seem likely to be three to four times those of breast cancer screening, with similar benefits.

Advocates of screening, in contrast, see the NLST results as vindicating a campaign to put advanced computer technology to work on lung cancer. The detailed images of early

Lowy, also speaking at the teleconference, ticked off some "disadvantages" of CT screening. One is cost. The price of a scan, estimated at about $300 to $500 per screening, is the least of it. Big expenses ensue, Lowy said, from the high ratio of people who get positive test results but do not have lung cancer. Even if you focus strictly on those with the highest risk—this trial screened smokers and ex-smokers who had used a pack of cigarettes a day for 30 years—"20% to 50%" of the CT scans "will show abnormalities" according to recent studies, said Lowy. According to NCI, about 96% to 98% are false positives. (p. 900)

Besides controlling health-case expenditures by considering the cost-effectiveness of tests, there are other choices involved in who should get screened and at what age. In an article by Gina Kolata in the *New York Times* (April 11, 2011), "Screening Prostates at Any Age," a study is discussed that found men 80 to 85 years old are being screened (using the PSA test) as often as men 30 years younger. Both the American Cancer Society and the American Urological Society discourage screenings for men whose life expectancy is ten years or less; prostate cancer is typically so slow-growing that it would take that long for any benefits of screening to appear. In addition, the United States Preventative Services Task Force recommends that screening should stop at 75. Given the observations we made about prostate screening in the previous section and the OpEd article by Richard Ablin, it appears we have an instance, not

tumors in CT scans are "exquisite," says James Mulshine, vice president for research at Rush University Medical Center in Chicago, Illinois, and an adviser to the pro-screening advocacy group, the Lung Cancer Alliance in Washington, D.C. He thinks it should be straightforward to reduce the number of biopsies and surgeries resulting from false positives by monitoring small tumors for a time before intervening. There are 45 million smokers in the United States who might benefit from CT screening, says Mulshine. He asks: Do we provide it, or "Do we tell them, 'Tough luck'?"

of practicing "evidence-based medicine," but a more likely one of "(Medicare) greed-induced medicine."

Informed consent and screening: Before participation in a screening program, patients must give informed consent, with an emphasize on the word "informed." Thus, the various diagnostic properties of the test should be clearly communicated, possibly with the use of Gigerenzer's "natural frequencies"; the risk of "false positives" must be clearly understood, as well as the risks associated with any follow-up invasive procedures. All efforts must be made to avoid the type of cautionary tale reported in Gigerenzer et al. 2007: at a conference on AIDS held in 1987, the former senator from Florida, Lawton Childs, reported that of twenty-two (obviously misinformed about false positives) blood donors in Florida who had been notified they had tested HIV-positive, seven committed suicide.

To inform patients properly about screening risks and benefits, the medical professionals doing the informing must be knowledgeable themselves. Unfortunately, as pointed out in detail by Gigerenzer et al. 2007, there is now ample evidence that many in the medical sciences are profoundly confused. An excellent model for the type of informed dialogue that should be possible is given by John Lee in a short "sounding board" article in the *New England Journal of Medicine* (1993, *328*, 438–440), "Screening and Informed Consent." This particular article is concerned with mammograms for detecting breast cancer but the model can be easily extended to other diagnostic situations where informed consent is required. Finally, to show that the type of exemplar dialogue that Lee models is not now widespread, we refer the reader to an editorial by Gerd Gigerenzer

in *Maturitas* (2010, *67*, 5–6) entitled "Women's Perception of the Benefit of Breast Cancer Screening." The gist of the evidence given in the editorial should be clear from its concluding two sentences: "Misleading women, whether intentionally or unintentionally, about the benefit of mammography screening is a serious issue. All of those in the business of informing women about screening should recall that medical systems are for patients, not the other way around" (p. 6).

The (social) pressure to screen: Irrespective of the evidence for the value for a diagnostic screen, there are usually strong social pressures for us to engage in this behavior. These urgings may comes from medical associations devoted to lobbying some topic, from private groups formed to advocate for some position, or from our own doctors and clinics not wishing to be sued for underdiagnosis. The decision to partake or not in some screening process, should depend on the data-driven evidence of its value, or on the other side, of the potential for harm. On the other hand, there are many instances where the evidence is present for the value of some ongoing screening procedure. One of the current authors (LH) takes several medications, all to control surrogate endpoints (or test levels), with the promise of keeping one in a reasonable healthy state. Eye drops are used to control eye pressure (and to forestall glaucoma); lisinopril and amlodipine to keep blood pressure under control (and prevent heart attacks); and a statin to keep cholesterol levels down (and again, to avoid heart problems).

In addition to contending with social pressures to screen wherever those pressures may come from, there is now what seems to be a

never-ending stream of media reports about new screening devices to consider or updated guidelines to follow about who should be screened, when to screen, and how often. There is now, for example, the possibility of genomic scans for a variety of mutations that might increase the risk of breast or ovarian cancer, or of the use of newer three-dimensional and hopefully more sensitive mammography. For the later we give several paragraphs from a Denise Grady article from the *New York Times* (June, 24, 2014), entitled "3-D Mammography Test Appears to Improve Breast Cancer Detection Rate":

> Adding a newer test to digital mammograms can increase the detection rate for breast cancer and decrease nerve-racking false alarms, in which suspicious findings lead women to get extra scans that turn out normal, a study found.
>
> Millions of women will get the newer test, tomosynthesis, this year. The procedure is nearly identical to a routine mammogram, except that in mammography the machine is stationary, while in tomosynthesis it moves around the breast. Sometimes called 3-D mammography, the test takes many X-rays at different angles to create a three-dimensional image of the breast. It was approved in the United States in 2011.
>
> The verdict is still out on the long-term worth of this new technology. The new results are promising but not definitive, according to experts not associated with the study, published Tuesday in The Journal of the American Medical Association. Tomosynthesis has not been around long enough to determine whether it saves lives or misses tumors.
>
> Even so, more and more mammography centers are buying the equipment, which is far more costly than a standard mammography unit, and marketing the test to patients as a more sensitive and accurate type of screening. It has come on the scene at a time when the value of breast cancer screening and the rising costs of health care are increasingly debated.

A variety of medically-related agencies issue guidelines periodically that concern general health practice. Unfortunately, some of these

may be conflicting depending on the agencies involved and who they represent. As a good controversial case in point, there is the ongoing debate about the wisdom of annual pelvic exams for women. An editorial given below from the *New York Times* (authored by "The Editorial Board"; July 2, 2014), and entitled "The Dispute Over Annual Pelvic Exams," illustrates well the type of confusion that might be present among "dueling" recommendations:

> Two major medical groups have taken opposing positions on whether healthy, low-risk women with no symptoms should have an annual pelvic exam. The American College of Physicians, the largest organization of physicians who practice internal medicine, strongly advised against the exams, which many women find distasteful or painful. The American College of Obstetricians and Gynecologists, the leading group of specialists providing health care for women, immediately reiterated its support for yearly pelvic exams for asymptomatic adult women.
>
> The exams at issue are not the Pap smears used to detect cervical cancers. Those are still recommended although there is disagreement on how often they should be done. The new dispute involves the "bimanual examination," in which a doctor inserts two gloved fingers into a woman's vagina and presses down on her abdomen with the other hand to check from both sides the shape and size of her uterus, ovaries and fallopian tubes. It also involves procedures that use a speculum to open the vagina for examination.
>
> Oddly enough, both professional groups agree there is no credible scientific evidence that the annual pelvic examinations save lives. They simply disagree over whether that lack of evidence matters much.
>
> The College of Physicians thinks it does. In a review of published scientific studies from 1946 through January 2014, it found no evidence that the pelvic exams provide any benefit in asymptomatic, nonpregnant adult women and significant evidence of harm, such as unnecessary surgeries, fear, anxiety and pain. The exams drive some women to avoid the doctors and can be traumatic for rape victims. The physicians organization estimated the annual cost of the exams at $2.6 billion. Unnecessary follow-up tests drive the cost even

higher.

By contrast, the gynecologists group argues that the "clinical experiences" of gynecologists, while not "evidence-based," demonstrate that annual pelvic exams are useful in detecting problems like incontinence and sexual dysfunction and in establishing a dialogue with patients about a wide range of health issues.

In recent years, medical groups and researchers have issued changing and sometimes conflicting recommendations on how often women should get a routine mammogram, how often to get pap smears, and now, whether to get an annual pelvic exam. Women will need to make their own judgments about procedures that many of them, and their doctors, may have used for years as a matter of standard practice.

The decision to institute or encourage widespread diagnostic screening should be based on evidence that shows effectiveness in relation to all the costs incurred. Part of the national discussion in the United States of evidence-based medical decision making is now taking place for the common screening targets of cervical, prostate, and breast cancer. Until recently it was considered an inappropriate question to ask whether it might be best if we didn't screen and identify a nonlethal cancer, and thus avoid debilitating and unnecessary treatment. A recent survey article by Gina Kolata makes these points well: "Considering When it Might Be Best Not to Know About Cancer" (*New York Times*, October 29, 2011). The United Kingdom is somewhat more advanced than the United States with respect to guidelines when screening programs should be implemented. The British National Health Service has issued useful "appraisal criteria" to guide the adoption of a screening program. The appendix to follow reproduces these criteria.

4.1 Appendix: U.K. National Screening Committee Programme Appraisal Criteria

Criteria for appraising the viability, effectiveness and appropriateness of a screening programme —

Ideally all the following criteria should be met before screening for a condition is initiated:

The Condition:

1. The condition should be an important health problem.

2. The epidemiology and natural history of the condition, including development from latent to declared disease, should be adequately understood and there should be a detectable risk factor, disease marker, latent period or early symptomatic stage.

3. All the cost-effective primary prevention interventions should have been implemented as far as practicable.

4. If the carriers of a mutation are identified as a result of screening, the natural history of people with this status should be understood, including the psychological implications.

The Test:

5. There should be a simple, safe, precise and validated screening test.

6. The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.

7. The test should be acceptable to the population.

8. There should be an agreed policy on the further diagnostic investigation of individuals with a positive test result and on the choices available to those individuals.

9. If the test is for mutations, the criteria used to select the subset of mutations to be covered by screening, if all possible mutations are not being tested, should be clearly set out.

The Treatment:

10. There should be an effective treatment or intervention for patients identified through early detection, with evidence of early treatment leading to better outcomes than late treatment.

11. There should be agreed evidence-based policies covering which individ-

uals should be offered treatment and the appropriate treatment to be offered.

12. Clinical management of the condition and patient outcomes should be optimised in all health care providers prior to participation in a screening programme.

The Screening Programme:

13. There should be evidence from high quality Randomised Controlled Trials that the screening programme is effective in reducing mortality or morbidity. Where screening is aimed solely at providing information to allow the person being screened to make an "informed choice" (e.g., Down's syndrome, cystic fibrosis carrier screening), there must be evidence from high quality trials that the test accurately measures risk. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened.

14. There should be evidence that the complete screening programme (test, diagnostic procedures, treatment/intervention) is clinically, socially and ethically acceptable to health professionals and the public.

15. The benefit from the screening programme should outweigh the physical and psychological harm (caused by the test, diagnostic procedures and treatment).

16. The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (i.e., value for money). Assessment against this criteria should have regard to evidence from cost benefit and/or cost effectiveness analyses and have regard to the effective use of available resources.

17. All other options for managing the condition should have been considered (e.g., improving treatment, providing other services), to ensure that no more cost effective intervention could be introduced or current interventions increased within the resources available.

18. There should be a plan for managing and monitoring the screening programme and an agreed set of quality assurance standards.

19. Adequate staffing and facilities for testing, diagnosis, treatment and programme management should be available prior to the commencement of the screening programme.

20. Evidence-based information, explaining the consequences of testing, investigation and treatment, should be made available to potential participants to assist them in making an informed choice.

21. Public pressure for widening the eligibility criteria for reducing the screening interval, and for increasing the sensitivity of the testing process, should be anticipated. Decisions about these parameters should be scientifically justifiable to the public.

22. If screening is for a mutation, the programme should be acceptable to people identified as carriers and to other family members.

References

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

[2] Bokhari, E., & Hubert, L. (2015). A new condition for assessing the clinical efficiency of a diagnostic test. *Psychological Assessment*, *xx*, xxx–xxx.

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

[4] Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96.

[5] Hanley, J. A. & McNeil, B. J. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) curve. *Radiology*, *143*, 29–36.

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

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

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

[9] Turnock, B. J., & Kelly, C. J. (1989). Mandatory premarital testing for human immunodeficiency virus: The Illinois experience. *Journal of the American Medical Association*, *261*, 3415–3418.