Refining the assessment of the sensitivity and specificity of diagnostic tests, with applications to prostate cancer screening and non-small cell lung cancer staging

Vance W. Berger[1]
and Lisa Semanick[2]

Key words: neoplasms, diagnostic techniques and procedures, early diagnosis, neoplasm staging, sensitivity and specificity, predictive value of tests.

[1] United States of America, National Cancer Institute, Division of Cancer Prevention, Biometry Research Group, Bethesda, Maryland, United States, and University of Maryland at Baltimore County, Baltimore County, Maryland, United States. Send correspondence to: Vance W. Berger, Biometry Research Group, National Cancer Institute, Executive Plaza North, Suite 3131, 6130 Executive Boulevard, MSC 7354, Bethesda, Maryland 20892-7354, United States of America; telephone: (301) 435-5303; fax: (301) 402-0816; e-mail: vb78c@nih.gov

[2] The Johns Hopkins University, Baltimore, Maryland, United States of America.

Diagnostic tests are often used to detect or stage a disease as well as to determine a course of subsequent treatment. For example, elevated prostate-specific antigen (PSA) levels or abnormal digital rectal examination (DRE) findings are indicators of the potential for prostate cancer (1), and they are often followed by biopsy. Likewise, computed tomography (CT) and positron-emission tomography (PET) can be used for staging non-small cell lung cancer, that is, to detect mediastinal lymph node metastases (2). Conventional statistical analysis of the detection properties of any given diagnostic test is based on a 2x2 table that cross-classifies subjects by presence or absence of cancer (according to some gold standard, often biopsy) and a positive or negative test result. Yet not all tumors are the same. Size, stage, severity, or some other factors may render some cancers, but not others, potentially lethal. Also, some cancers, though not yet symptomatic, may have metastasized to the point that treatment would not result in much benefit. As such, some tumors are more in need of being treated, and therefore of being detected, than are others.

It has been recognized that two tests, even with identical sensitivities and specificities, may still be distinguishable by the type of tumors they tend to detect or miss. Therefore, it has been suggested that the standard 2x2 table, which allows no provision for this consideration, be modified by splitting the true-positive group into two categories, based on the size of the tumor detected (3). However, it is preferable to split both the true-positive group and the false-negative group. Doing this requires a measurement of amenability to treatment that is available even for subjects with negative results. It is not the purpose of this article to identify the ideal measures of amenability to treatment, but rather to illustrate how such measures would be used to refine the assessment of a diagnostic test. To this end, we use DRE status as a measure of amenability to treatment when evaluating PSA for prostate cancer, and we use CT status as a measure of amenability to treatment when evaluating PET for non-small cell lung cancer.

THE CONCEPT OF AMENABILITY TO TREATMENT

The standard definition of sensitivity is the number of true-positive tests divided by the total number of cancers, that is, P{positive finding | cancer}.

TABLE 1. General structure of a refined table to evaluate a screening test

	Cancer severity				
	Severe	Moderate	Mild	None	Total
Diagnostic test+	x_1	x_2	x_3	x_4	$x_1+x_2+x_3+x_4$
Diagnostic test–	$n_1–x_1$	$n_2–x_2$	$n_3–x_3$	$n_4–x_4$	$n_1+n_2+n_3+n_4–x_1–x_2-x_3-x_4$
Total	n_1	n_2	n_3	n_4	$n_1+n_2+n_3+n_4$

Specificity is generally defined as the number of true-negative tests divided by the total number of non-cancers, that is, P{negative finding | no cancer}. A rational utility function would be an increasing function of both the sensitivity and the specificity (4), so one would want each of these quantities to be large.

The definitions of sensitivity and specificity are unambiguous in the usual context of a binary classification of the true status (cancer or no cancer), so there are only two columns, and a 2x2 table results. But when weighing the risks and benefits of any treatment modality for any disease, it will often turn out that some patients benefit greatly, some less, some not at all, and some may even be harmed by treatment. Hence, the key issue in evaluating a diagnostic test is the extent to which it provides benefits. With this is mind, we propose that a diagnostic test be evaluated based on not only its ability to detect tumors but also its ability to zero in on those tumors that are most amenable to treatment. False-positives should count against the test in proportion to the harm incurred, and true-positives should count for the test in proportion to the benefit accrued.

Yet the standard definition of sensitivity awards full credit to a test that detects a tumor, regardless of the characteristics of that tumor. A tumor that remains asymptomatic but that has already spread to the point that treatment can be only palliative may, if detected, result in less clinical benefit than a tumor that could still be excised. Likewise, a propensity for overdetection of indolent tumors that, had they not been detected, might not have resulted in harm is a bad property of a diagnostic test (5). Therefore, the gold standard should not represent cancer that is overdiagnosed (6). Consider, for example, our Table 1, which presents the general structure of a refined table to evaluate a screening test. Our Table 1 differs from the standard 2x2 table in that our table has four columns for cancer severity according to the gold standard: three positive severity levels, and a negative finding. Therefore, our Table 1 is a 2x4 table, instead of the usual 2x2 table for binary tests and binary gold standards or the 4x2 table (for example) used for receiver operating characteristic (ROC) curve methods (4) when the test provides a classification with four categories. For example, in an article by Cooner et al. (7), that article's Table 1 shows a data set in which the PSA is classified into three categories. For the purposes of this paper we consider only binary classifications of any diagnostic test.

Our Table 1 presents all relevant information, and so can be considered to be the gold standard by which any other subset of information may be judged. The interpretation of our Table 1 depends on the definitions of the columns in terms of the need for (or amenability to) treatment. There may be no benefit in detecting severe cancers or mild cancers. That is because severe cancers have already progressed to the point that treatment is futile, and mild cancers, if left untreated, would not result in any harm. It would also be possible to define the cancer-severity columns in a monotonic fashion, so that the benefit in detecting a cancer would increase with the severity of that cancer. Such distinctions could not be drawn if all the cancers were grouped together.

Nothing is lost in switching from the usual 2x2 table to our Table 1's 2x4 format. And, one could reconstruct the usual 2x2 table by selecting the appropriate cutpoint and dichotomizing our Table 1 (8). Yet something is certainly gained with our 2x4 format. Analogous to the 4x2 table used, for example, for ROC curve methods when the test provides a classification with four categories, the richer classification in our Table 1 also allows for multiple cutpoints, each defining a combination of sensitivity and specificity. To illustrate the advantages of our Table 1 format, in the next two sections we consider prostate cancer and non-small cell lung cancer, respectively. In each case, we present a data set that will allow for a 2x3 table, and we proceed based on monotonic ordering. That is, we consider the more severe tumors to be the ones most in need of being detected, but we then reverse this assumption and consider the most severe tumors to be beyond hope. In such a case, there would be less benefit in detecting these than in detecting the moderate ones.

THE EXAMPLE OF PROSTATE CANCER

The slow growth of most prostate cancers makes the need for treatment hard to define, but

patient characteristics, such as age, may play a role. When evaluating PSA as a screening test for prostate cancer, for this article we considered only the DRE status. We did that because of the availability of a data set in which all subjects had been screened with both PSA and DRE, and had been biopsied independently of the results of the screening tests. We extracted the data from a data set that had been created by Cooner and that was later presented by Baker (9) in an evaluation of the performance of combinations of prostate-cancer markers. Instead of evaluating the predictive ability of the pair of screening tests used in combination, we refined the evaluation of one screening tool by considering the status of the other. The subset of subjects who had both screening tests and a biopsy may not be representative of the entire sample. In turn, the entire sample itself may not be representative of the target population of subjects who would conceivably be screened. Our purpose is not to make a statement about the Cooner data set per se, or even about PSA or DRE, but rather to illustrate a novel, more informative presentation of the data, complete with any sampling biases it may (or may not) contain.

We used our Table 2 to evaluate PSA, with its rows for PSA+ and PSA–, and columns for DRE+ cancer, DRE– cancer, and no cancer. The order in which the PSA and DRE tests are administered could influence the results, especially if the DRE is administered first. One could use counterbalancing techniques (10) to ensure that any carryover effect is balanced. However, it might just be simpler to administer the PSA first for all the subjects in the study.

DRE+ tumors are probably more aggressive than DRE– tumors. For example, Table 4 in the Cooner et al. article (7) shows a data set in which disease was confined in 42 of 65 (65%) of the DRE– tumors, but in only 94 of 177 (53%) of the DRE+ tumors. The trend in Table 5 of the Cooner et al. article (7) is reversed, so that the DRE– tumors appear to be more severe. Nevertheless, in our Table 2, the second column ("DRE– cancer," that is, cancer, but not sufficient to cause an abnormal DRE finding) is intermediate between the first column ("DRE+ cancer") and the third column ("No cancer").

Various contingencies can be considered. The futility assumption is that DRE+ tumors are less amenable to treatment, given the higher likelihood of extracapsular disease that responds poorly to surgery or radiation. Conversely, the monotonicity assumption is that these DRE+ tumors are the ones most in need of treatment, given their severity.

It is not our purpose to address this important issue. Instead, our purpose is to clarify the role played by some measure of amenability to treatment in assessing a diagnostic test. Hence, the truth or falsity of either the futility assumption or the monotonicity assumption is somewhat tangential to our argument. As we illustrate in this paper our approach would prove useful given the truth of either assumption. Our display is useful under either scenario, that is, either futility or monotonicity, and we do not address which one is true.

TABLE 2. Prostate-specific antigen (PSA) assessment, accounting for the digital rectal examination (DRE) status of cancers detected and missed

	DRE+ cancer	DRE– cancer	No cancer	Total
PSA+	152	42	192	386
PSA–	23	12	322	357
Total	175	54	514	743

TABLE 3. Assessment of prostate-specific antigen (PSA) based on splitting off the "No cancer" column

	Cancer (DRE+ or DRE–)[a]	No cancer	Total
PSA+	194	192	386
PSA–	35	322	357
Total	229	514	743

[a] DRE = digital rectal examination.

We proceed first under the monotonicity assumption, and the sensitivity and specificity are both defined in terms of 2x2 tables. Our Table 3 presents the data based on combining the first two columns, that is, the DRE+ and DRE– columns. In that way, any cancer counts as positive, regardless of its severity or its need for treatment. The sensitivity and specificity for our Table 3 are 194/(194 + 35) = 0.847 and 322/(192 + 322) = 0.626, respectively. Clearly, the tumors in the first column need to be treated (per the monotonicity assumption), and those in the last column do not, but partial credit might be appropriate for the middle column. If so, then combining the middle column with the last one, per the strict definition of the sensitivity and specificity, might not be appropriate. It is also possible to instead combine the middle category with the first one, and separate the last column from them.

Our Table 4 presents the data based on combining the last two columns (DRE– cancer and no cancer) from Table 2. In that way, only cancer sufficient to cause an abnormal DRE counts as positive. The sensitivity and specificity for this table are 152/(152 + 23) = 0.869 and 334/(234 + 334) = 0.588, respectively. Neither binary classification (that is, merging the middle column with one or the other of the extreme columns) tells the whole story. Similarly, neither measure of sensitivity (or of specificity) suffices on its own. In fact, with more than

TABLE 4. Assessment of prostate-specific antigen (PSA) based on splitting off the digital-rectal-examination-positive (DRE+) column

	DRE+ cancer	No cancer or DRE– cancer	Total
PSA+	152	234	386
PSA–	23	334	357
Total	175	568	743

three classifications, a dichotomization tells even less of the story.

When 2x2 tables are required to summarize an inherently ordered categorical structure, such as we have here for the ordered categorical variable "need for detection," one approach is to present the entire set of these 2x2 tables, rather than arbitrarily selecting only one of them. This set of all the 2x2 tables is known as the Lancaster decomposition (11), and it is relevant whenever there are ordered categories. It also lends itself to ROC analyses (12). Similarly, we propose that when evaluating a test in which the disease can be classified into several categories, all relevant measures of sensitivity and specificity be presented, along with the complete table (such as our Table 2). In fact, our Table 2 provides all the information available in both our Table 3 and our Table 4, and the pair of sensitivities is more informative than either one is by itself, and likewise the pair of specificities is more informative than either one is by itself.

We now reverse the monotonicity assumption underlying the ordering structure inherent in our Table 2. That is, we now assume that DRE+ tumors are less amenable to treatment than DRE– tumors are. If this is the case, then the columns of our Table 2 would need to be permuted, with DRE– tumors now coming first, and DRE+ tumors being in the middle. That is because now detection of a DRE+ tumor would confer benefit that is intermediate, that is, more than the benefit from mistakenly calling normal tissue cancerous but less than the benefit from detecting a DRE– tumor. Our Table 2 would now be replaced with our Table 5. Combining the first two columns (DRE– tumors and DRE+ tumors) results again in Table 3, with sensitivity and specificity of 194/(194 + 35) = 0.847 and 322/(192 + 322) = 0.626, respectively. However, one could also combine the last two columns in order to obtain our Table 6, in which only DRE– tumors count as those needing detection. Now the sensitivity and specificity are 42/(42 + 12) = 0.777 and 345/(345 + 344) = 0.501, respectively. The roles of PSA and DRE could be reversed, in which case one would use the PSA results to refine the assessment of DRE. This, plus the monotonicity assumption, would result in our Table 7. Like our Table 2, our Table 7 displays a pair of measures of each of sensitivity and specificity. Of course, the columns of Table 7 could be permuted if the monotonicity assumption were reversed, but we do not show this table.

Whether using the DRE to refine the assessment of the PSA or using the PSA to refine the assessment of the DRE, a natural question might arise concerning the possible usefulness of combining the two measures of sensitivity and combining the two measures of specificity. A weighted average of the sensitivity for high-risk cancer and the sensitivity for low-risk cancer would likely be better at identifying biomarkers for further study than either single sensitivity would alone. One can construct such a weighted average based on a suitable measure of how "close" the middle category is to one extreme category relative to how "close" it is to the other extreme category. The issue becomes one of the amenability to (or need for) treatment of a DRE– tumor relative to that of a DRE+ tumor. Specifically, define the amenability to treatment to be 1 ("fully amenable to treatment") for DRE+ tumors, z ("partially amenable to treatment") for DRE– tumors, and 0 ("not amenable to treatment at all") for no cancer, with $0 \leq z \leq 1$. The quantity z is not known, but with data from a screening study using both

TABLE 5. Prostate-specific antigen (PSA) assessment, assuming digital-rectal-examination-negative (DRE–) tumors are most in need of detection

	DRE– cancer	DRE+ cancer	No cancer	Total
PSA+	42	152	192	386
PSA–	12	23	322	357
Total	54	175	514	743

TABLE 6. Assessment of prostate-specific antigen (PSA) based on splitting off the digital-rectal-examination-negative (DRE–) column

	DRE– cancer	No cancer or DRE+ cancer	Total
PSA+	42	344	386
PSA–	12	345	357
Total	54	689	743

TABLE 7. Digital rectal examination (DRE) assessment, accounting for the prostate-specific antigen (PSA) status of cancers detected and missed

	PSA+ cancer	PSA– cancer	No cancer	Total
DRE+	152	23	236	411
DRE–	42	12	278	332
Total	194	35	514	743

PSA and DRE, such as the ongoing PLCO [Prostate, Lung, Colon, and Ovarian Cancer] Trial (13), one could estimate z as indicated in the next paragraph. The PLCO Trial (13) specifies that each subject with biopsy-confirmed cancer (DRE– or DRE+) is to be treated according to the discretion of his or her physician.

For some subjects, the decision will be to undergo watchful waiting, whereas for others (presumably, for those deemed in need of treatment) the decision will be to treat. Let P{T | DRE+} and P{T | DRE–} be the proportion of subjects, among those with DRE+ cancer and with DRE– cancer, respectively, to undergo treatment. These quantities estimate the proportion among each group considered to need treatment. Therefore, one reasonable estimate of z would be the quotient P{T | DRE–}/P{T | DRE+}. A value for z would suggest a weight to be used for a weighted average of the two measures of sensitivity. To see this, consider again our Table 1, with its general cell counts, that is, where symbols are used in place of numbers, to make the structure of the table applicable to future studies no matter what numbers they obtain. For simplicity, suppose that severe cancer is in greatest need of detection, moderate cancer is intermediate, and mild cancer does not need to be detected at all. Then the sensitivity could be defined as $s_1 = x_1/n_1$ if $z = 0$, or as $s_2 = (x_1 + x_2)/(n_1 + n_2)$ if $z = 1$. But in the more realistic case that $0 < z < 1$, neither measure of sensitivity will suffice. The overall sensitivity may be defined as $(x_1 + zx_2)/(n_1 + zn_2) = hs_1 + (1–h)s_2$, where the weight h can be found as $(1–z)n_1/[(1–z)n_1 + z(n_1 + n_2)]$. This measure reduces to the right quantities when $z = 0$ or $z = 1$, and assigns partial credit for DRE– tumors in the more realistic case that $0 < z < 1$. One could also start with a desired weight h on the sensitivities, and work backwards to determine the induced value of z that corresponds to h as $z = [n_1(1–h)]/[n_1 + n_2h]$.

In preparation for when reasonable estimates of the quantities z(PSA) and z(DRE) become available, we consider, in our Table 8 and our Table 9, a range of values for each. For the Cooner data as presented by Baker (9), the sensitivities and specificities do not vary much with z(PSA) or z(DRE). This reflects the fact that the cell counts of the middle columns of our Table 2 and our Table 7 were small relative to the cell counts of the two other columns. This finding, in turn, is due to the fact that there were relatively few (only 77) subjects with a combination of a negative screening test and a positive biopsy. Part of the explanation for this phenomenon is that of 1 520 subjects with a negative screening test, only 503 had a biopsy, and thus an opportunity for a positive biopsy. If the other 1 017 subjects had also had biopsies, then we might have found more subjects with tumors, and with conflicting results from the two screening tests, that is, with one test detecting the tumor and the other test not detecting it. Note that of the 35 cancers missed by PSA, only 12 were also missed by DRE, and of the 54 cancers missed by DRE, only 12 were missed by PSA. So if all the subjects in the Cooner data set (9) had been biopsied, there likely would have been more disagreements between PSA and DRE. In that case, the ranges in our Table 8 and Table 9 would have expanded by virtue of the larger cell counts for the middle columns of our Table 2 and our Table 7. In such a case there would be more benefit in performing both screening tests and in presenting them in the more informative way that we have, that is, with the added third column.

TABLE 8. Sensitivity and specificity of prostate-specific antigen (PSA) for given values of V_{DRE}[a]

V_{DRE}	Sensitivity of PSA[b]	Specificity of PSA[c]
0.0	0.869	0.588
0.1	0.866	0.592
0.2	0.863	0.595
0.3	0.861	0.599
0.4	0.859	0.602
0.5	0.856	0.606
0.6	0.854	0.610
0.7	0.852	0.614
0.8	0.851	0.618
0.9	0.849	0.622
1.0	0.847	0.626

[a] V_{DRE} = value for digital rectal examination.
[b] Sensitivity of PSA = $[152+42V_{DRE}]/[175+54V_{DRE}]$.
[c] Specificity of PSA = $[334–12V_{DRE}]/[568–54V_{DRE}]$.

Mistry and Cable (1) recently found PSA and DRE sensitivities of 72.1% and 53.2%, respectively, and with specificities of 93.2% and 83.6%, respectively. This appears to contradict our findings, except that our findings were based on Cooner data that were published by Baker (9) after they had been conveyed to him (Baker) via personal communication. The threshold for classifying results as disease or not were not specified by Baker (9), and decreasing the threshold could certainly result in more positive findings, both true-positives and false-positives. This would serve to increase the sensitivity and decrease the specificity. Therefore, one possible explanation for our apparent disagreement with Mistry and Cable is that they used a different point on the ROC curve (12) for PSA (different cutpoint) and/or for DRE (different subjective decision). However, even if this turns out not to be the case, our findings are still not invalidated. That is because our contention is not that the PSA and DRE have certain sensitivities and specificities, but rather only that more informative data displays

TABLE 9. Sensitivity and specificity of digital rectal examination (DRE) for given values of V_{PSA}[a]

V_{PSA}	Sensitivity of DRE[b]	Specificity of DRE[c]
0.0	0.783	0.528
0.1	0.781	0.531
0.2	0.779	0.531
0.3	0.777	0.532
0.4	0.775	0.533
0.5	0.773	0.534
0.6	0.771	0.536
0.7	0.769	0.537
0.8	0.768	0.538
0.9	0.766	0.539
1.0	0.764	0.541

[a] V_{PSA} = value for prostate-specific antigen.
[b] Sensitivity of DRE = $[152+23V_{PSA}]/[194+35V_{PSA}]$.
[c] Specificity of DRE = $[290–12V_{PSA}]/[549–35V_{PSA}]$.

should be used. Note that our results cannot even be used to suggest that PSA is a better screening test than DRE (based on better sensitivities and specificities), because like quantities are not being compared. A better comparison of PSA and DRE would involve each being evaluated with the same categorization of tumors by amenability to treatment, such as the Gleason score, free PSA (14), or the ratio of total PSA to free PSA (14).

THE EXAMPLE OF NON-SMALL CELL LUNG CANCER

The Table 4 of an article by Pieterman et al. (2) displays the correct detection of mediastinal lymph node metastases with positron-emission tomography (PET) and with computed tomography (CT). Their data are reproduced in our Table 10, using the monotonicity assumption, that is, that CT+ metastases are in greater need of being detected than are CT– metastases. With this display in our Table 10, the sensitivity of PET, $(x_1 + zx_2)/(n_1 + zn_2)$, that is, $(22 + 7z)/(24 + 8z)$, is a linear combination of the sensitivities based on the 2x2 subtables that could be constructed, that is, $hs_1 + (1–h)s_2 = h(22/24) + (1–h)(29/32)$. If we instead used the futility assumption, that is, that CT– metastases are in greater need of being detected than are CT+ metastases, then the first two columns of Table 10 would be transposed. The resulting sensitivity would be $(7 + 22z)/(8 + 24z)$, which is a linear combination of the sensitivities based on the 2x2 subtables, that is, $hs_1 + (1–h)s_2 = h(7/8) + (1–h)(29/32)$. One could also refine the assessment of the CT based on consideration of the PET, as we did in our Table 11. Now the sensitivity is $(22 + 2z)/(29 + 3z)$, or $(2 + 22z)/(3 + 32z)$ if the first two columns are transposed.

DISCUSSION

Tautologically, any diagnostic test will detect each individual who falls into the group defined as those with a specific outcome from the test itself, but no individual who does not fall into that group. That is, one could *define* a disease by a positive test, and then this would ensure that the sensitivity and specificity would both always be 100%. But the question is whether or not it is useful to classify subjects according to their values from a diagnostic test. If we assume that certain groups of patients can be successfully treated, then it is worthwhile to identify members of this group. To the extent that a diagnostic test produces subject subgroups that

TABLE 10. Positron-emission tomography (PET) assessment, accounting for the computed tomography (CT) status of metastases detected in the case of non-small cell lung cancer staging

	CT+ metastases	CT– metastases	No metastases	Total
PET+	22	7	10	39
PET–	2	1	60	63
Total	24	8	70	102

TABLE 11. Computed tomography (CT) assessment, accounting for the positron-emission tomography (PET) status of metastases detected and missed in the case of non-small cell lung cancer staging

	PET+ metastases	PET– metastases	No metastases	Total
CT+	22	2	24	48
CT–	7	1	46	54
Total	29	3	70	102

approximate not necessarily a group that can be labeled with a given disease, but rather these treatable groups, the test is useful. The key is the extent to which a positive screening test may lead to treatment or to further diagnostics. With this in mind, we developed new data displays that exploit the ability to classify tumors by the extent to which they need to be treated. These displays are ideal for the situation in which one can identify those tumors most in need of treatment and distinguish them from those tumors less in need of treatment. A limitation of the data is that we did not have access to a data set that would allow for the substitution of the Gleason score or free PSA for the DRE. We are hopeful that this article has pointed out the benefit of such cross-tabulations, so that those in possession of such data will publish them, thereby addressing this lack of publicly available data.

Acknowledgements. The authors thank Stuart Baker, Phil Prorok, and the anonymous review team for offering helpful comments.

SINOPSIS

Modos de refinar los cálculos de sensibilidad y especificidad de las pruebas diagnósticas, aplicados al tamizaje del cáncer de próstata y a la estadificación del cáncer de pulmón no microcítico

Los cálculos del rendimiento de las pruebas diagnósticas suelen presentarse en tablas de 2x2 con filas horizontales para los resultados positivos y negativos obtenidos con la prueba evaluada, y con columnas verticales para los resultados positivos y negativos obtenidos con el patrón de oro. Esta manera de presentar los datos visualmente, así como las sensibilidades y especificidades basadas en ella, le imprimen carácter binario a la prueba y al patrón de oro. Pero cuando los resultados de la prueba pertenecen a una de varias categorías ordenadas, a menudo se utilizan curvas de las características funcionales de la prueba (o curvas ROC, por receiver operator characteristic curve*) para indicar que esta no es binaria. Tratar el patrón de oro como si fuese binario también es problemático porque implica que toda enfermedad se comporta uniformemente, con el resultado de que a todos los casos se les trata como si fuesen intercambiables. No obstante, hay ciertos tumores, por ejemplo, que exigen más tratamiento que otros y que por lo tanto también exigen mayor detección. En el presente trabajo proponemos el uso de una tabla refinada que clasifica a los tumores en función de lo que se sabe de su susceptibilidad al tratamiento, con lo cual se pretende lograr una evaluación más informativa de las pruebas que la proporcionada por la tabla de 2x2. A manera de ejemplo presentamos una tabla de 2x3 en la cual se refina la medición del antígeno específico de la próstata (AEP) teniendo en cuenta el resultado de la palpación rectal. Dicho resultado se usa como indicador de la necesidad de tratar los cánceres prostáticos que se detectan o que no se detectan mediante la prueba del AEP. Un segundo ejemplo aplica los mismos conceptos a la tomografía por emisión de positrones y a la tomografía computadorizada cuando se usan para la estadificación del cáncer pulmonar no microcítico. Se usaría más información si se adoptara la estructura de 2x3 para configurar la tabla.*

Palabras clave: neoplasmas, técnicas y procedimientos diagnósticos, diagnóstico temprano, estadificación de neoplasmas, sensibilidad y especificidad, valor predictivo de los tests.

REFERENCES

1. Mistry K, Cable G. Meta-analysis of prostate-specific antigen and digital rectal examination as screening tests for prostate carcinoma. J Am Board Fam Pract. 2003;16(2):95–101.
2. Pieterman RM, van Putten JW, Meuzelaar JJ, Mooyaart EL, Vaalburg W, Koeter GH, et al. Preoperative staging of non-small-cell lung cancer with positron-emission tomography. N Engl J Med. 2000;343(4):254–61.
3. McNaughton M, Ransohoff DF, Barry MJ. Early detection of prostate cancer: serendipity strikes again. JAMA. 1997; 278:1516–9.
4. Baker SG. Identifying combinations of cancer biomarkers for further study as triggers of early intervention. Biometrics. 2000;56:1082–7.
5. Marcus PM, Bergstralh EJ, Fagerstrom RM, Williams DE, Fontana R, Taylor WF, et al. Lung cancer mortality in the Mayo Lung Project: impact of extended follow-up. J Natl Cancer Inst. 2000;92 (16):1308–16.
6. Baker SG, Kramer BS, Srivastava S. Markers for early detection of cancer: statistical guidelines for nested case-control studies. BMC Med Res Methodol. 2002;2(1):1–8.
7. Cooner WH, Mosley BR, Rutherford CL, Beard JH, Pond HS, Terry WJ, et al. Prostate cancer detection in a clinical urological practice by ultrasonography, digital rectal examination, and prostate specific antigen. J Urol. 1990;143(6):1146–54.
8. Berger VW. Improving the information content of categorical clinical trial endpoints. Control Clin Trials. 2002;23(5): 502–14.
9. Baker SG. Evaluating multiple diagnostic tests with partial verification. Biometrics. 1995;51:330–7.
10. DePuy V, Berger VW. Counterbalancing. In: Everitt B, Howell D, eds. Encyclopedia of statistics in behavioral science. New York: John Wiley and Sons; forthcoming 2005.
11. Permutt T, Berger VW. Rank tests in ordered 2xk contingency tables. Commun Stat Theory Methods. 2000;29(5):989–1003.
12. Baker SG. The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. J Natl Cancer Inst. 2003;95(7): 511–5.
13. Gohagan JK, Prorok PC, Kramer BS, Cornett JE. Prostate cancer screening in the prostate, lung, colorectal, and ovarian cancer screening trial of the National Cancer Institute. J Urol. 1994;152(5 Pt 2): 1905–9.
14. Catalona WJ, Southwick PC, Slawin KM, Partin AW, Brawer MK, Flanigan RC, et al. Comparison of percent free PSA, PSA density, and age-specific PSA cutoffs for prostate cancer detection and staging. Urology. 2000;56:255–60.