STATISTICS IN MEDICINE
*Statist. Med.* **18**, 455–471 (1999)

# PROBABILISTIC ANALYSIS OF GLOBAL PERFORMANCES OF DIAGNOSTIC TESTS: INTERPRETING THE LORENZ CURVE-BASED SUMMARY MEASURES

WEN-CHUNG LEE*†

*Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, Taipei, Taiwan, R.O.C.*

## SUMMARY

Several indices based on the receiver operating characteristic curve (ROC curve) have previously been found to possess probabilistic interpretations. However, these interpretations are based on some unrealistic diagnostic scenarios. In this paper, the author presents a new approach using the Lorenz curve. The author found that the summary indices of the Lorenz curve, that is, the Pietra index and the Gini index, can be interpreted in several ways ('average change in post-test probability', 'per cent maximum prognostic information', and 'probability of correct diagnosis'). These interpretations have a close tie with real-world medical diagnosis, suggesting that these indices are proper measures of test characteristics. Copyright © 1999 John Wiley & Sons, Ltd.

## INTRODUCTION

Recent decades have witnessed a rapid progress in statistical methodologies for medical diagnosis.[1,2] Undoubtedly, the receiver operating characteristic (ROC) curve analysis[3,4] stands at centre stage and receives most of the attention. One of the most important applications of the ROC curve analysis is to evaluate and compare the 'overall' performance of diagnostic or screening tests. The term 'overall' is emphasized, since we are interested in the global picture of a test but not in the diagnostic performance at a particular cut-off point. Such evaluation and comparison of the global performances are usually based upon the 'area under the curve' (AUC) index, a summary index of the ROC curve.[5] The greater the AUC, the better the overall performance of a test. The AUC *per se* also possesses a clear interpretation in that it is equal to the probability that the test result of a randomly selected diseased subject exceeds that of a randomly selected non-diseased subject.[5]

In a previous paper, we introduced two new summary indices of the ROC curve which, interestingly, also possess probabilistic interpretations.[6] The 'projected length of the curve' (PLC)

* Correspondence to: Dr. Wen-Chung Lee, Graduate Institute of Epidemiology, National Taiwan University, No. 1, Jen-Ai Rd, 1st Sec, Taipei, Taiwan, R.O.C. E-mail: wenchung@ha.mc.ntu.tw

† The author was also with the National Defense Medical Center, R.O.C., while the work was done.

CCC 0277–6715/99/040455–17$17.50

*Received July 1997*
*Accepted May 1998*

index corresponds to the probability of a correct diagnosis when one subject – with equal chance of being diseased or non-diseased – is presented and we make a diagnosis according to which posterior probability is higher. The 'area swept out by the curve' (ASC) index is related to the correct probability when a pair of subjects (one diseased and the other non-diseased) are presented and we make a diagnosis after obtaining the actual measurement of the one with lower test result. The conventional AUC index also fits into such a 'paired-subjects' scenario – it corresponds to making a diagnosis by *comparing* but not actually *measuring* their test results.

However, the above characterization of diagnostic performances in probabilistic terms, though interesting, is in fact unrealistic. In actual diagnostic or screening practices, the subjects seldom come one by one, each with 50 : 50 chance of being diseased *a priori*, let along arrive in pairs of exactly one diseased and one non-diseased. Also it is hard to imagine that anyone would really bother to randomly select a pair of subjects and compare their test results. In this paper, we adopt a different approach, turning our attention to the 'Lorenz curve'.[7–10] We first demonstrate how to construct a Lorenz curve and calculate its summary indices, that is, the Pietra and the Gini indices.[11] Next we show that these indices are arithmetically linked to the changes in the pre-test and the post-test disease probabilities. In the actual settings, we believe that these changes in probabilities are what the patients, the doctors and the epidemiologists really care about. Finally, we compare the Lorenz curve and the ROC curve and discuss briefly some statistical properties of the Pietra and the Gini indices.

## THE LORENZ CURVE AND ITS SUMMARY INDICES

The Lorenz curve has been widely used by economists to assess the distributional properties of family income and wealth[7] and by demographers to quantify the degree of population concentration.[8] Recently it has also been applied to analysing seasonal data in detecting and testing for temporal clustering of disease occurrences[9] and to characterizing exposure–disease association in human populations.[10] The technique, however, has yet to prove its usefulness in medical diagnosis.

We use the data of Hanley and McNeil[5] to illustrate the methodology. The data (see Table I) consist of the rating results of 109 computed tomographic images obtained from 51 diseased subjects and 58 non-diseased subjects. The rating is on a five-category scale of 'definitely normal', 'probably normal', 'questionable', 'probably abnormal' and 'definitely abnormal'. The likelihood ratios (LRs) at the various rating categories for this diagnostic test are also presented. The likelihood ratio[12,13] at category $t$ (denoted as $LR_t$) is defined as the ratio of the probability of having rating result $t$ for a diseased subject to the corresponding probability for a non-diseased subject. The LR provides the information of at-risk status. The greater the LR value at rating result $t$, the greater the risk of being diseased for those rated to be at $t$. To construct the Lorenz curve, the rating categories must first be re-arranged according to the values of the respective LRs (from the lowest to the highest). For these particular data, there is no need to do the re-arrangement since the LRs are already monotonically increasing. Next, the cumulative percentages for the diseased and the non-diseased are calculated (see Table I). The Lorenz curve (see Figure 1) is simply the plot of the cumulative percentage of the diseased against the cumulative percentage of the non-diseased (with straight lines connecting the points). Note that our usage of the Lorenz curve is actually different from that of economists or demographers (economists often plot the cumulative percentage of 'income' against the cumulative percentage of 'population',[7] while demographers plot the cumulative percentage of 'population number' against the

 

Table I. Rating of 109 computed tomographic images

| Rating categories | Diseased subjects | | | Non-diseased subjects | | | LR† |
|---|---|---|---|---|---|---|---|
| | Numbers | % | Cum.%* | Numbers | % | Cum.%* | |
| Definitely normal | 3 | 5·88 | 5·88 | 33 | 56·90 | 56·90 | 0·10 |
| Probably normal | 2 | 3·92 | 9·80 | 6 | 10·34 | 67·24 | 0·38 |
| Questionable | 2 | 3·92 | 13·73 | 6 | 10·34 | 77·59 | 0·38 |
| Probably abnormal | 11 | 21·57 | 35·29 | 11 | 18·97 | 96·55 | 1·14 |
| Definitely abnormal | 33 | 64·71 | 100·00 | 2 | 3·45 | 100·00 | 18·76 |
| Total | 51 | 100·00 | | 58 | 100·00 | | |

* Cumulative percentage
† Likelihood ratio.

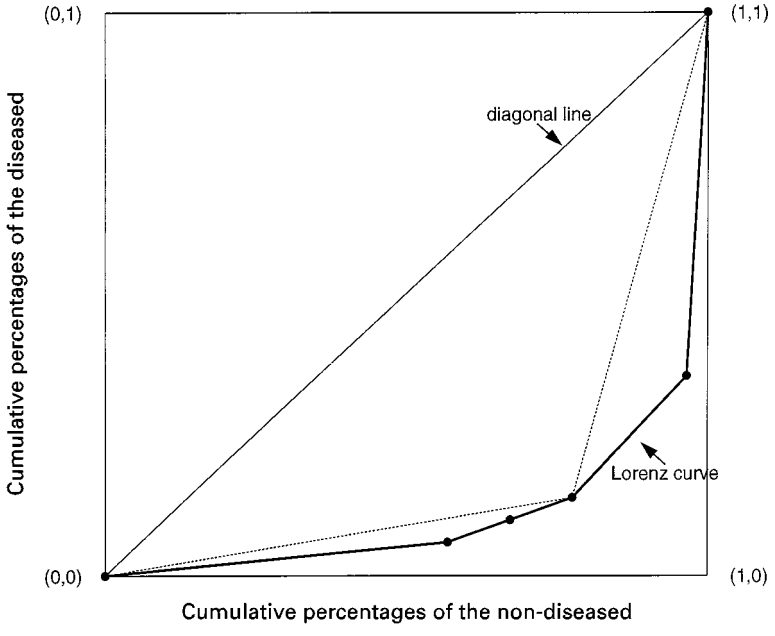


Figure 1. The Lorenz curve of the example in Table I. The Gini index is twice the area between the Lorenz curve and the diagonal line. The Pietra index is twice the area of the largest triangle (dotted line) inscribed in the Lorenz curve and the diagonal line

cumulative percentage of 'land area'.[8] However, we still call it the Lorenz curve, since they all involve the same procedures of categorizing, reordering, summing and plotting.

It is clear that for a useless diagnostic test which provides hardly any at-risk information, its LR will be 1 no matter what the test results would be, and its Lorenz curve will run along the diagonal line. On the other hand, if the diagnostic test perfectly separates the diseased from the non-diseased (that is, after the test result has been obtained, say $t$, the subject is known for sure either to be the diseased ($LR_t = \infty$) or to be the non-diseased ($LR_t = 0$)), the corresponding Lorenz curve would coincide with the $x$-axis throughout and then jump to the uppermost point (1, 1).

 

This observation suggests that between these two extreme cases, a more bowed Lorenz curve may indicate a better diagnostic test, while the flatter the curve, the less diagnostic a test is. It is desirable then to have summary indices able to quantify such 'bowedness', For the Lorenz curve, this is not difficult since the curve is constrained by the concavity and continuity properties.[11] We can define the Pietra index as twice the area of the largest triangle (see Figure 1) which can be inscribed in the area between the Lorenz curve and the diagonal line. We can also define the Gini index as twice the area between the Lorenz curve itself and the diagonal line. Equivalently, we may define the two indices as the ratios of the aforementioned two areas to the area of the triangle below the diagonal line. Clearly, these two indices are between 0 and 1.

Assume that the cumulative percentages of the non-diseased and the diseased subjects are denoted by $X_i$ and $Y_i$ ($i = 1, 2, \ldots, K$), respectively (note that $K$ represents the number of categories of the diagnostic test and that the data have been re-arranged according to the LRs). For the example in Table I, the indices of Pietra and Gini can be calculated as below (the straight lines indicate the determinant of the $2 \times 2$ matrix):

$$\begin{aligned}
\text{Pietra} &= \max_{1 \leqslant i \leqslant K-1} \left( \begin{vmatrix} X_i & Y_i \\ 1 & 1 \end{vmatrix} \right) \\
&= \max \left( \begin{vmatrix} 0{\cdot}5690 & 0{\cdot}0588 \\ 1 & 1 \end{vmatrix}, \begin{vmatrix} 0{\cdot}6724 & 0{\cdot}0980 \\ 1 & 1 \end{vmatrix}, \begin{vmatrix} 0{\cdot}7759 & 0{\cdot}1373 \\ 1 & 1 \end{vmatrix}, \begin{vmatrix} 0{\cdot}9655 & 0{\cdot}3529 \\ 1 & 1 \end{vmatrix} \right) \\
&= 0{\cdot}6386,
\end{aligned}$$

$$\begin{aligned}
\text{Gini} &= \sum_{i=1}^{K-1} \begin{vmatrix} X_i & Y_i \\ X_{i+1} & Y_{i+1} \end{vmatrix} \\
&= \begin{vmatrix} 0{\cdot}5690 & 0{\cdot}0588 \\ 0{\cdot}6724 & 0{\cdot}0980 \end{vmatrix} + \begin{vmatrix} 0{\cdot}6724 & 0{\cdot}0980 \\ 0{\cdot}7759 & 0{\cdot}1373 \end{vmatrix} + \begin{vmatrix} 0{\cdot}7759 & 0.1373 \\ 0{\cdot}9655 & 0{\cdot}3529 \end{vmatrix} + \begin{vmatrix} 0{\cdot}9655 & 0{\cdot}3529 \\ 1 & 1 \end{vmatrix} \\
&= 0{\cdot}7863
\end{aligned}$$

## INTERPRETATION OF THE PIETRA AND THE GINI INDICES

It is of interest to note that such geometrically defined indices can be interpreted in several ways. All suggest that they are ideal for evaluating and comparing the performances of diagnostic tests. These are explained in the following.

### Average Change in Post-Test Probability

Here we seek to quantify the diagnostic performance of a test by its ability to revise the disease probability. Before a test is applied, the best guess of whether a subject is the diseased or the non-diseased is to use the 'disease prevalence' (denoted by $P(D)$). However, this is a rough indicator, since the population to which the test is applied may consist of heterogeneous subjects, some of which are at a higher risk of being diseased, while others may be at a lower risk. After testing, depending on the test result $T_i$, the 'post-test disease probability' (denoted as $P(D|T_i)$) may be higher or lower than the 'pre-test probability' (or the prevalence) and therefore subjects

 

with varying degrees of risk are sorted out. It is conceivable that the larger the difference between the post-test and the pre-test probabilities, the better the test can revise the probability. Since the extent of this probability revision depends on the test results, we use the average value – the 'average absolute change in the disease probability provided by the testing' (denoted as $\Delta P$) – as an indicator of the global performance of a test. The formal definition of $\Delta$P is presented below:

$$\Delta P = \sum_{i=1}^{K} P(T_i)|P(D|T_i) - P(D)|$$

where the $P(T_i)$ denotes the probability of obtaining a test result $T_i$. Its role in the equation is to serve as a weighting factor for the averaging. Note that the $P(T_i)$ can be calculated by $P(T_i) = P(D)P(T_i|D) + [1 - P(D)]P(T_i|\mathrm{ND})$, where the ND denotes the 'non-diseased'. The $P(D|T_i)$ can be calculated using the conventional Bayes rule.[13]

For the example in Table I, we have calculated the $\Delta P$ assuming that the computed tomographic rating experiment was applied to different settings with $P(D) = 0{\cdot}2$, $0{\cdot}5$ and $0{\cdot}9$, respectively (see Table II). It turns out interestingly that it is related to the Pietra index by

$$\Delta P = 2P(D)[1 - P(D)]\,\mathrm{Pietra}.$$

A formal proof of the above relation is presented in the Appendix. For a binary test, the Pietra index is equal to SEN + SPE − 1, the Youden index[14] (we use 'SEN' and 'SPE' to denote 'sensitivity' and 'specificity'), and therefore $\Delta P = 2P(D)[1 - P(D)]$ (SEN + SPE − 1). This latter expression has appeared in two previous papers,[15,16] where the $\Delta P$ was dubbed the 'expected gain in certainty'[15] and the 'prognostic information',[16] respectively.

In the above, the global performance of a test is quantified by the average change between the post-test and the pre-test probabilities. However, another measure can also be considered, that is, the 'average absolute difference in post-test probabilities of two randomly selected subjects' (denoted as $\Delta P^*$). We present the formal definition below:

$$\Delta P^* = \sum_{i=1}^{K} \sum_{j=1}^{K} P(T_i)P(T_j)|P(D|T_i) - P(D|T_j)|.$$

For any two subjects, the likelihood of disease is the same before testing (they stand at the same starting point). However, as a consequence of testing, the two subjects become separated (the post-test probabilities become different). The larger the mean 'separation' (the $\Delta P^*$) a diagnostic test can attain, the more prognostic information it provides for the subjects being tested.

For the example in Table I, we also calculate the $\Delta P^*$ with $P(D) = 0{\cdot}2$, $0{\cdot}5$ and $0{\cdot}9$, respectively (see Table II). This time we found that it is related to the Gini index (the proof is given in the Appendix):

$$\Delta P^* = 2P(D)[1 - P(D)]\,\mathrm{Gini}.$$

**Per Cent Maximum Prognostic Information**

Let us consider a perfect test and examine the maximum values the $\Delta P$ and $\Delta P^*$ can attain (the maximum prognostic information). A perfect test when positive ($T_+$) indicates, without error, that the subject being tested is the diseased (that is, $P(D|T_+) = 1$) and when negative ($T_-$)

 

Table II. Performance indices of the computed tomographic rating experiments in diagnostic settings of different disease prevalences

| Rating categories | $P(D) = 0{\cdot}2$ | | $P(D) = 0{\cdot}5$ | | $P(D) = 0{\cdot}9$ | |
|---|---|---|---|---|---|---|
| | $P(T_i)$ | $P(D\|T_i)$ | $P(T_i)$ | $P(D\|T_i)$ | $P(T_i)$ | $P(D\|T_i)$ |
| Definitely normal | 0·4669 | 0·0252 | 0·3139 | 0·0937 | 0·1098 | 0·4820 |
| Probably normal | 0·0906 | 0·0866 | 0·0713 | 0·2749 | 0·0456 | 0·7733 |
| Questionable | 0·0906 | 0·0866 | 0·0713 | 0·2749 | 0·0456 | 0·7733 |
| Probably abnormal | 0·1949 | 0·2214 | 0·2027 | 0·5321 | 0·2131 | 0·9110 |
| Definitely abnormal | 0·1570 | 0·8243 | 0·3408 | 0·9494 | 0·5858 | 0·9941 |
| $\Delta P$ | 0·2044 | | 0·3193 | | 0·1149 | |
| $2P(D)[1 - P(D)]$ Pietra | 0·2044 | | 0·3193 | | 0·1149 | |
| $\Delta P^*$ | 0·2516 | | 0·3932 | | 0·1415 | |
| $2P(D)[1 - P(D)]$ Gini | 0.2516 | | 0.3932 | | 0.1415 | |

indicates the opposite ($P(D|T_-) = 0$). By definition we have

$$\begin{aligned}\max\Delta P &= P(T_-)|P(D|T_-) - P(D)| + P(T_+)|P(D|T_+) - P(D)| \\ &= [1 - P(D)]|0 - P(D)| + P(D)|1 - P(D)| \\ &= 2P(D)[1 - P(D)]\end{aligned}$$

and

$$\begin{aligned}\max\Delta P^* &= P(T_-)P(T_-)|P(D|T_-) - P(D|T_-)| \\ &\quad + P(T_-)P(T_+)|P(D|T_-) - P(D|T_+)| \\ &\quad + P(T_+)P(T_-)|P(D|T_+) - P(D|T_-)| \\ &\quad + P(T_+)P(T_+)|P(D|T_+) - P(D|T_+)| \\ &= [1 - P(D)]P(D)|0 - 1| + P(D)[1 - P(D)]|1 - 0| \\ &= 2P(D)[1 - P(D)].\end{aligned}$$

We obtain the following expressions for the Pietra and the Gini indices:

$$\text{Pietra} = \frac{\Delta P}{\max\Delta P}$$

and

$$\text{Gini} = \frac{\Delta P^*}{\max\Delta P^*}.$$

Thus we realize that both the Pietra and the Gini are 'ratio indices', which measure the 'per cent maximum prognostic information' provided by a less-than-perfect test (relative to the perfect test). Also, since the above expressions do not depend on $P(D)$, we see that these 'ratios' represent

 

the 'inherent' characteristics of a diagnostic test. 'Inherent' here means that they are the properties of the diagnostic tests *per se* and do not concern the diagnostic situations where the tests are administered. We note in passing that the traditional indices of SEN, SPE, LR and the summary indices of the ROC curve (AUC, PLC, ASC) do not depend on $P(D)$ either.

**Probability of Correct Diagnosis**

As mentioned earlier, the AUC, the PLC, and the ASC indices of the ROC curve can each be interpreted as the correct probability under a certain hypothetical diagnostic scenario.[6] It is of interest to find that the same is true for the indices of Pietra and Gini. Let us first consider the 'single-subject' scenario (with equal chance of being diseased and non-diseased). After testing, the post-test probability of this subject can be calculated. A reasonable strategy in this situation is to make a diagnosis that the subject is a diseased when his/her post-test probability is greater than 0·5, and is a non-diseased when below 0·5. The probability of correct diagnosis using this strategy (denoted as $P_c$) for the example in Table I can be calculated as below:

$$\begin{aligned} P_c &= \sum_{i=1}^{K} \max\left(\frac{x_i}{2}, \frac{y_i}{2}\right) \\ &= \max\left(\frac{0{\cdot}5690}{2}, \frac{0{\cdot}0588}{2}\right) + \max\left(\frac{0{\cdot}1034}{2}, \frac{0{\cdot}0392}{2}\right) + \max\left(\frac{0{\cdot}1034}{2}, \frac{0{\cdot}0392}{2}\right) \\ &\quad + \max\left(\frac{0{\cdot}1897}{2}, \frac{0{\cdot}2157}{2}\right) + \max\left(\frac{0{\cdot}0345}{2}, \frac{0{\cdot}6471}{2}\right) \\ &= 0{\cdot}8193 \end{aligned}$$

where the $x_i$ and $y_i$ denote the percentage of subjects with test result $T_i$ for the non-diseased and the diseased, respectively. It is found that such a correct probability is related to the Pietra index (see Appendix):

$$P_c = \frac{1}{2} + \frac{\text{Pietra}}{2}.$$

Readers can also check that the Pietra index of the Lorenz curve is actually the same index (up to a scaling factor) as the PLC of the ROC curve.[6]

Now, let us return to the paired-subjects scenario (one diseased and the other non-diseased). The AUC index amounts to making a diagnosis by simple *comparison* of the test outcomes of these two subjects, and this implies that we do not really have to send the two subjects for actual measurements; a laboratory test capable of *comparing* a paired sample suffices. The ASC index is a step further, corresponding to making a diagnosis not simply by comparison but also by actual *measurement* of one of the two subjects.[6] The Gini index calls for even more; diagnosis based on it requires the actual measurement of both subjects. After the measurement, we compare the post-test disease probabilities of the two subjects and identify the subject with greater post-test probability as the diseased and the one with lower probability as the non-diseased (in the case when the two subjects have the same post-test probability, we randomly select one subject as the diseased one). The correct probability of this new strategy (denoted as $P_c^*$) can also be calculated

 

using simple algebra:

$$\begin{aligned} P_c^* &= \sum_{i=1}^{K} x_i \left(1 - Y_i + \frac{y_i}{2}\right) \\ &= 0{\cdot}5690\left(1 - 0{\cdot}0588 + \frac{0{\cdot}0588}{2}\right) + 0{\cdot}1034\left(1 - 0{\cdot}0980 + \frac{0{\cdot}0392}{2}\right) \\ &\quad + 0{\cdot}1034\left(1 - 0{\cdot}1373 + \frac{0{\cdot}0392}{2}\right) + 0{\cdot}1897\left(1 - 0{\cdot}3529 + \frac{0{\cdot}2157}{2}\right) \\ &\quad + 0{\cdot}0345\left(1 - 1 + \frac{0{\cdot}6471}{2}\right) \\ &= 0{\cdot}8932 \end{aligned}$$

This time we found that it is related to the Gini index (see Appendix):

$$P_c^* = \frac{1}{2} + \frac{\text{Gini}}{2}.$$

As stated before, these last interpretations in terms of correct probability are in fact unrealistic. Nevertheless, we believe that such interpretations help clarify the relations between the indices of AUC, PLC, ASC, Pietra and Gini.

## COMPARISON BETWEEN LORENZ CURVE AND ROC CURVE

It is of interest to compare the Lorenz curve and the ROC curve. For the computed tomographic rating example, the Lorenz curve is just the 'upside-down' ROC curve (the ROC curve of this example can be found in figure 1 of reference 5). This is because the test has a monotone LR function such that it appears the same whether one reorders the data or not. In this case, there is one-to-one correspondence between the ROC-curve-based indices and the Lorenz-curve-based ones. That is, Pietra = MVD (maximum vertical distance between the ROC curve and the diagonal line) and Gini = 2AUC − 1. Thus one may resort to the conventional ROC curve for the estimation of the Lorenz curve parameters, when it is known *a priori* that the at-risk status increases (or decreases) with increasing values of test results.

However, the reordering procedure will make a difference when one deals with a non-monotone diagnostic test. In this case, the procedure in effect transforms a 'wiggly' ROC curve to a concave Lorenz curve (concavifying transformation).[17] Here we provide such an example. Shown in Table III is a 'binormal test' with $\Delta m = 1$ and $\alpha = 5$.[18] 'Binormal' implies that the test results of both the diseased and the non-diseased are normally distributed. The two parameters ($\Delta m$ and $\alpha$) describe their relative position. The $\Delta m$ measures the difference of the means of the two distributions in units of the standard deviation of the non-diseased, and the $\alpha$ is the ratio of the standard deviations of the diseased to the non-diseased. For ease of presentation we assume the distribution of the non-diseased to be the standard normal distribution and categorize the test results into 12 levels ($\leqslant -2{\cdot}5$, $-2{\cdot}5$– $-2{\cdot}0$, …, 2·0–2·5, $>2{\cdot}5$). Note that in Table III we have re-arranged the data according to the LRs. It can be seen that the ROC curve of this test (Figure 2) has portions that fall below the diagonal line. By contrast, the corresponding Lorenz curve (Figure 3) is concavified.

 

Table III. A binormal diagnostic tests with $\Delta m = 1$ and $\alpha = 5$

| Test results* | | | Diseased subjects | | Non-diseased subjects | | LR‡ |
|---|---|---|---|---|---|---|---|
| | | | % | Cum.%† | % | Cum.%† | |
| − 0·5 | – | 0·0 | 3·87 | 3·87 | 19·15 | 19·15 | 0·20 |
| 0·0 | – | 0·5 | 3·94 | 7·81 | 19·15 | 38·29 | 0·21 |
| − 1·0 | – | − 0·5 | 3·75 | 11·56 | 14·99 | 53·28 | 0·25 |
| 0·5 | – | 1·0 | 3·98 | 15·54 | 14·99 | 68·27 | 0·27 |
| − 1·5 | – | − 1·0 | 3·60 | 19·15 | 9·18 | 77·45 | 0·39 |
| 1·0 | – | 1·5 | 3·98 | 23·13 | 9·18 | 86·64 | 0·43 |
| − 2·0 | – | − 1·5 | 3·43 | 26·56 | 4·41 | 91·04 | 0·78 |
| 1·5 | – | 2·0 | 3·94 | 30·50 | 4·41 | 95·45 | 0·90 |
| − 2·5 | – | − 2·0 | 3·23 | 33·73 | 1·65 | 97·10 | 1·95 |
| 2·0 | – | 2·5 | 3·87 | 37·59 | 1·65 | 98·76 | 2·34 |
| | < = | − 2·5 | 24·20 | 61·79 | 0·62 | 99·38 | 38·97 |
| | > | 2·5 | 38·21 | 100·00 | 0·62 | 100·00 | 61·53 |

* Test results are re-arranged according to LRs
† Cumulative percentage
‡ Likelihood ratio

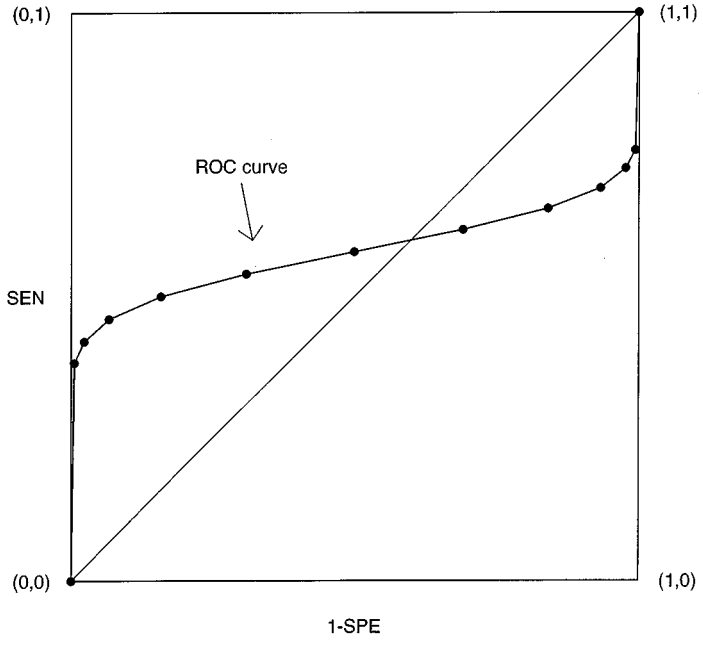


Figure 2. The ROC curve of the example in Table III

As another example with disparate ROC curve and Lorenz curve, Table IV presents the post-dexamethasone plasma cortisol concentrations observed in 215 melancholia patients and 152 patients of other diagnoses (the dexamethasone suppression test), taken from Figure 3 of Reference 19. It has been noted that the distribution of the melancholia patients follows a bimodal

 

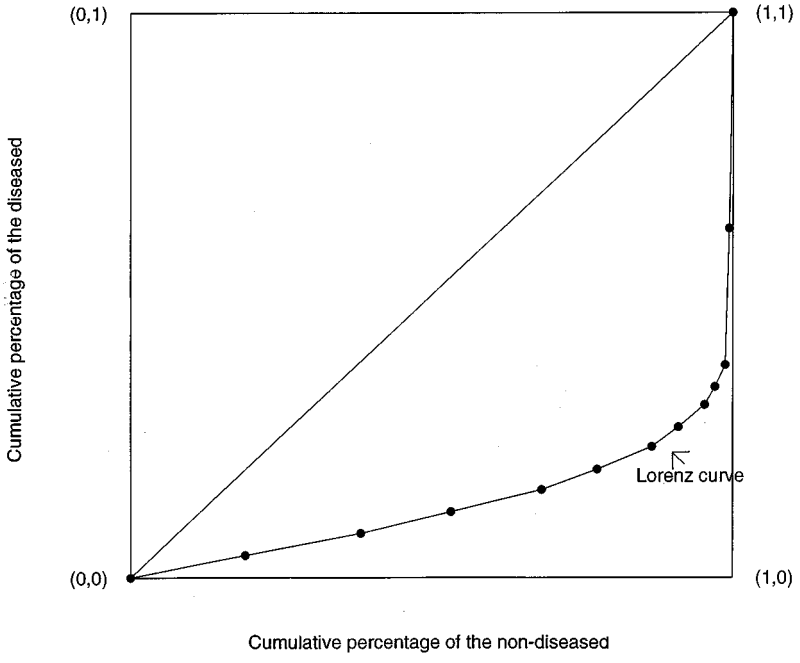


Figure 3. The Lorenz curve of the example in Table III

Table IV. Results of dexamethasone suppression test of 215 melancholia patients and 152 patients of other diagnoses

| Plasma cortisol concentrations* | Melancholia patients | | | Other diagnoses | | | LR‡ |
|---|---|---|---|---|---|---|---|
| | Numbers | % | Cum.%† | Numbers | % | Cum.%† | |
| $< = 0{\cdot}8$ | 28 | 13·02 | 13·02 | 38 | 25·00 | 25·00 | 0·52 |
| 0·8–1·5 | 46 | 21·40 | 34·42 | 45 | 29·61 | 54·61 | 0·72 |
| 1·5–3·0 | 38 | 17·67 | 52·09 | 56 | 36·84 | 91·45 | 0·48 |
| 3·0–6·0 | 29 | 13·49 | 65·58 | 10 | 6·58 | 98·03 | 2·05 |
| $> 6{\cdot}0$ | 74 | 34·42 | 100·00 | 3 | 1·97 | 100·00 | 17·44 |
| Total | 215 | 100·00 | | 152 | 100·00 | | |

* Post-dexamethasone plasma cortisol concentrations
† Cumulative percentage
‡ Likelihood ratio

pattern while the non-melancholia subjects are unimodal.[19] This results in LRs that are not monotonically increasing or decreasing (see Table IV) and thus the ROC curve and the Lorenz curve for this example will differ in shape (we omit the figures). As we have previously shown, the Lorenz curve maintains the desirable interpretation properties as compared to the conventional ROC curve analysis. For the above examples of the binormal test and the dexamethasone

 

suppression test, one should turn to the Lorenz curve-based Pietra and Gini indices for characterizing the test performance, rather than rely on the ROC curve-based MVD or AUC indices.

## ESTIMATION OF THE PARAMETERS OF THE LORENZ CURVE

In actual practice, however, interpretability is not the only issue to be taken into account. Both the ROC and the Lorenz curves are estimated by the data and are subject to random errors. Furthermore, a direct use of the sample Lorenz curve will lead to biased estimation of the Pietra and Gini indices. This is because the same set of data is used twice – first to create a maximally bowed Lorenz curve and then to calculate this degree of bowedness. The situation is akin to the scenario when we use the same set of data to construct a model and to evaluate its prediction error. To demonstrate the bias incurred by summarizing a sample Lorenz curve, we perform a small simulation study on the above-mentioned binormal diagnostic test in Table III (the test has theoretical values of the Pietra and Gini indices of 0·6495 and 0·7442, respectively, when categorized into 12 levels). A series of 1000 computer simulations was performed, each with 100 diseased and 100 non-diseased subjects. The test results of these subjects are computer generated according to a 12-category multinomial distribution (probabilities taken from Table III). In each cycle of simulation, a Lorenz curve is constructed and the Pietra and Gini indices are calculated (using the same set of simulated data). The results for the Pietra and Gini indices are (mean $\pm$ standard error): 0·6693 $\pm$ 0·0014, 0·7828 $\pm$ 0·0014. Clearly, we see that this is an over-estimation of the true values.

To correct the bias from a sample Lorenz curve, we can use the bootstrap technique.[20] The idea is simple – avoid using the same of data in creating the Lorenz curve and calculating its summary indices. Here, we let the 'bootstrap sample' do the trick. To be precise, we re-select equal number of subjects from the original data set to be our bootstrap sample – randomly and with replacement. Then we let the bootstrap sample create the Lorenz curve (determine the reordering sequence and find the vertex of the largest inscribed triangle) but let the original data evaluate its bowedness (calculate the Pietra and Gini indices). The re-sampling can be done hundreds or thousands of times and the arithmetic means of the re-sampling results produce the 'bias-corrected' Pietra and Gini indices (denoted as $\text{Pietra}^c$ and $\text{Gini}^c$). For an illustration, we turn back to the example of the dexamethasone suppression test in Table IV. The uncorrected Pietra and Gini for this example are 0·3935 and 0·4588, respectively. To obtain the bias-corrected estimates, we perform 10,000 bootstrap re-sampling. In each cycle of simulation, a new set of 'case' subjects which also contains 215 melancholia patients is drawn by random sampling (with replacement) from the original melancholia patient population. Similarly, a new set of 'non-case' subjects containing 152 patients of other diagnoses is drawn from the original other-diagnosis population. These new set of case and non-case subjects are used to create a Lorenz curve, but its bowedness (Pietra and Gini indices) is evaluated using the original data (the data in Table IV). The sample means from these 10,000 bootstrap simulations produce the bias-corrected estimates: $\text{Pietra}^c = 0{\cdot}3895$ and $\text{Gini}^c = 0{\cdot}4467$. The need to perform a bootstrapping for bias correction may seem a disadvantage for the Lorenz curve analysis. However, with easy access to fast computation nowadays, we believe this is trivial even for practical concerns.

Besides the issue of biasedness, a sample Lorenz curve may be considerably less stable than a sample ROC curve. This is because the reordering procedure adds some extra-randomness into the estimation – the procedure is based on imprecise LRs (estimated from the data) and may

 

Table V. The results of bootstrap simulation for the example in Table IV

| Indices† | Bootstrapped means* | Bootstrapped standard errors* |
|---|---|---|
| Pietra$^c$ | 0·3893 | 0·0455 |
| Gini$^c$ | 0·4463 | 0·0547 |
| MVD$^c$ | 0·3927 | 0·0417 |
| 2AUC − 1 | 0·3984 | 0·0523 |

* based on 100,000 bootstrap simulations
† Pietra$^c$: bias-corrected Pietra index of the Lorenz curve
Gini$^c$: bias-corrected Gini index of the Lorenz curve
MVD$^c$: bias-corrected maximum vertical distance between the ROC curve and the diagonal line
2AUC − 1: two times the area under the ROC curve minus one

inadvertently rearrange adjacent categories of a graded test where it should not. The Pietra and the Gini indices (after correcting their biases) may thus be less precise compared to the ROC-curve-based MVD and AUC. To gauge the effect, one can resort to the bootstrap technique again. We also use the example of dexamethasone suppression test this time. A total of 100,000 bootstrapping simulations are performed, but for each simulation, two sets of bootstrap samples (instead of one) are drawn this time, one for creating the Lorenz curve and the other for evaluating its bowedness. The sample means and the sample standard deviations of the results of these 100,000 simulations are calculated (the bootstrapped means and the bootstrapped standard errors of the Pietra and Gini indices) and are presented in Table V. Note that the purposes of the above bootstrapping are twofold: the correcting of biases and the gauging of the stability of the estimates. For comparison, we also present the bootstrapped standard errors of the MVD$^c$ (bias-corrected MVD) and the 2AUC − 1 of the ROC curve. Note that unless corrected, the MVD of the ROC curve is also biased (the same set of data is used first to find the point in the ROC curve with maximum vertical distance from the diagonal line and then to measure the length of this maximum distance). As expected, we find that the standard errors of the Lorenz-curve-based indices are larger than those of the ROC-curve-based indices. To improve stability, one may consider smoothing the sample LRs using either parametric or non-parametric techniques.[21] The reordering procedure can then be based on these smoothed (and hence more stable) LRs.

## DISCUSSION

In this paper, we see that a Lorenz curve has very much in common with an ROC curve. The distinction lies solely in whether we choose to reorder test categories or choose not to. This may become a difficult decision, however, since non-monotonicity in the observed LRs may be real and thus reordering is advised, or else it may be due purely to random error and thus reordering is against. This problem is not entirely a statistical one. Rather, we should turn to our *a priori* knowledge about the test. We believe that the assumption of monotonicity is a reasonable in the majority of practical diagnostic test situations (so, do not bother to do the reordering!). However, in some special situations, the reordering procedure should be taken into serious consideration.

 

Here are three examples:

(i) Diagnostic tests with similar means but very different variances in the diseased and non-diseased groups (such as the example in Table III). Although such tests may seem 'eccentric', tests of this kind are not uncommon. Somoza[18] reported that of the 28 diagnostic tests found in recent literatures, three were found to be of this characteristic.
(ii) Diagnostic tests with bimodal distribution in either the diseased or the non-diseased group. The dexamethasone suppression test in Table IV is an example.
(iii) Diagnostic tests with the non-diseased subjects distributed symmetrically but with the diseased subjects distributed askew. An example can be found in Figure 2 of reference 6.

In all situations, the monotonicity assumption fails to capture the inherent characteristic of the test, and it is for tests of this special type that the Lorenz curve analysis will demonstrate added values over the traditional ROC curve analysis.

In the presentation above, we are looking at the impact of the diagnostic test on the posterior probability of being diseased. However, exactly the same principle applies when we are interested in the probability of not being diseased – we simply re-arrange the data by *decreasing* LR and plot the cumulative percentage of the *non-diseased* against that of the *diseased*. The result is a mirror image of the original Lorenz curve and the values of the Pietra and Gini remain the same. This is reasonable since the probabilities of being diseased and of not being diseased are complementary and their expected (absolute) changes shall naturally be the same. Thus one need not worry about whether performance rating of a test will change when we are looking at not being diseased instead.

Geometrically, it is clear that Gini $\geqslant$ Pietra. This relation can also be inferred from the single-subject and paired-subjects scenarios. In the former setting, we can perform at most one measurement, while in the latter, we are entitled to perform twice (one measurement for each subject). Thus it is not too unexpected that the correct probability of the latter is no less than that of the former. However, this does not imply that the Gini is superior to the Pietra as a performance index. It is possible that the global performance of a particular diagnostic test may be rated superior to another test by the Pietra index, while at the same time rated inferior by the Gini. Actually, the indices of Pietra and Gini characterize the global performance of a diagnostic test from different perspectives – the Pietra quantifies the expected change in the disease probabilities as the consequences of testing, while the Gini reflects the degree of subject-to-subject variability in the post-test disease probabilities. Thus, for a subject waiting to be tested, his/her attending physician may wish to select a test with higher Pietra index so that the expectation of having a disease differs most before and after the testing. On the other hand, if separating as much as possible a group of subjects with equal pre-test disease probability is the desired end, a diagnostic test with higher Gini index may be preferable.

Lorenz-curve-based indices of Pietra and Gini have a close tie with real-world medical diagnosis – the connection with the expected gain in certainty and the prognostic information etc – and it is the purpose of this paper to demonstrate these simple and elegant relationships. In doing so, we have kept our presentation as simple and concise as possible. However, for practical applications, the statistical properties of the new indices should be investigated further. There are other issues worthy of study as well. First, the Pietra and Gini as shown are global measures of diagnostic performances. However, there are situations where interests centre on selecting an operating point to maximize utility[22,23] or on studying just a portion of the curve.[24,25] It is possible that the Lorenz curve analysis would provide added insights to these problems as well.

 

Second, the ROC curve analysis is limited to evaluating one diagnostic test a time. In actual practice however, a battery of tests are often administered simultaneously.[26] The question then is how to characterize the overall (not one-by-one) performance of these tests. The Lorenz curve analysis may hold the answer, since it reorders the data before plotting and after reordering (if can be properly done), it really does not matter whether the data come from a single test (uni-dimensional data) or multiple tests (multi-dimensional data). Finally, the performances of diagnostic tests in this paper are characterized using the language of probability—we are looking at the ability of a test in moving prior probability. However, a test can also be measured by its ability in reducing 'uncertainty' (entropy).[27,28] It seems worthwhile to explore the role of Lorenz curve analysis under this alternative criterion.

## APPENDIX

Assume that the diagnostic test has $K$ levels which have been rearranged according to their LRs (from the lowest to the highest) and are indexed by $i (i = 1, 2, \ldots, K)$. We let $x_i$ denote the percentage of the non-diseased subjects with testing results in the $i$th category, and $y_i$, the percentage of the diseased in the $i$th category. Their cumulative percentage are denoted by the upper-case letters: $X_i = \sum_{j \leqslant i} x_j$, $Y_i = \sum_{j \leqslant i} y_j$. We define $X_0 = 0$ and $Y_0 = 0$. The Lorenz curve is the plot of $Y_i$ against $X_i$. By definition

$$\text{Pietra} = \max_{1 \leqslant i \leqslant K-1} \left( \begin{vmatrix} X_i & Y_i \\ 1 & 1 \end{vmatrix} \right) \text{ and } \text{Gini} = \sum_{i=1}^{K-1} \begin{vmatrix} X_i & Y_i \\ X_{i+1} & Y_{i+1} \end{vmatrix} = 1 - \sum_{i=1}^{K} x_i (Y_i + Y_{i-1}).$$

Now, consider that the test has been applied to where disease prevalence (prior probability) is $\pi$. The probability of obtaining a test result $i(T_i)$ is $P(T_i) = \pi y_i + (1 - \pi) x_i$, and the post-test disease probability conditioned on observing $T_i$ is $P(D|T_i) = \pi y_i / [\pi y_i + (1 - \pi) x_i]$. The 'average absolute change in the disease probability provided by testing' ($\Delta P$) is

$$\begin{aligned}
\Delta P &= \sum_{i=1}^{K} P(T_i) |P(D|T_i) - \pi| \\
&= \sum_{i=1}^{K} |\pi y_i - \pi [\pi y_i + (1 - \pi) x_i]| \\
&= \pi (1 - \pi) \sum_{i=1}^{K} |y_i - x_i| \\
&= 2\pi (1 - \pi) \sum_{\text{LR}_i < 1} (x_i - y_i) \\
&= 2\pi (1 - \pi) \max_i (X_i - Y_i) \\
&= 2\pi (1 - \pi) \max_{1 \leqslant i \leqslant K-1} \left( \begin{vmatrix} X_i & Y_i \\ 1 & 1 \end{vmatrix} \right) \\
&= 2\pi (1 - \pi) \text{ Pietra.}
\end{aligned}$$

 

The 'average absolute difference in post-test probabilities of two randomly selected subjects' ($\Delta P^*$) is

$$\begin{aligned}
\Delta P^* &= \sum_{i=1}^{K}\sum_{j=1}^{K} P(T_i)P(T_j)|P(D|T_i) - P(D|T_j)| \\
&= \sum_{i=1}^{K}\sum_{j=1}^{K} |\pi y_i[\pi y_j + (1-\pi)x_j] - \pi y_j[\pi y_i + (1-\pi)x_i]| \\
&= \pi(1-\pi)\sum_{i=1}^{K}\sum_{j=1}^{K} |x_j y_i - x_i y_j| \\
&= \pi(1-\pi)\sum_{i=1}^{K} \{(X_{i-1}y_i - x_i Y_{i-1}) + [x_i(1-Y_{i-1}) - (1-X_{i-1})y_i]\} \\
&= \pi(1-\pi)\sum_{i=1}^{K} [2(X_{i-1}y_i - x_i Y_{i-1}) + (x_i - y_i)] \\
&= 2\pi(1-\pi)\sum_{i=1}^{K-1} \begin{vmatrix} X_i & Y_i \\ X_{i+1} & Y_{i+1} \end{vmatrix} \\
&= 2\pi(1-\pi)\,\text{Gini}.
\end{aligned}$$

Finally, we examine the probability of correct diagnosis under 'single-subject scenario' ($P_c$) and 'paired-subjects scenario' ($P_c^*$):

$$\begin{aligned}
P_c &= \sum_{i=1}^{K} \max\left(\frac{x_i}{2}, \frac{y_i}{2}\right) \\
&= \frac{1}{4}\sum_{i=1}^{K} [(x_i + y_i) + |x_i - y_i|] \\
&= \frac{1}{2} + \frac{1}{4}\sum_{i=1}^{K} |x_i - y_i| \\
&= \frac{1}{2} + \frac{1}{2}\sum_{\text{LR}_i < 1} (x_i - y_i) \\
&= \frac{1}{2} + \frac{1}{2}\max_i (X_i - Y_i) \\
&= \frac{1}{2} + \frac{1}{2}\max_{1 \leqslant i \leqslant K-1} \left(\begin{vmatrix} X_i & Y_i \\ 1 & 1 \end{vmatrix}\right) \\
&= \frac{1}{2} + \frac{\text{Pietra}}{2}
\end{aligned}$$

 

and

$$P_c^* = \sum_{i=1}^{K} x_i \left(1 - Y_i + \frac{y_i}{2}\right)$$
$$= \sum_{i=1}^{K} x_i - \frac{1}{2} \sum_{i=1}^{K} x_i (Y_i + Y_{i-1})$$
$$= \frac{1}{2} + \frac{\text{Gini}}{2}.$$

ACKNOWLEDGEMENT

The author wishes to thank Dr. Chuhsing Kate Hsiao for helpful comments.

## REFERENCES

1. Begg, C. B. 'Advances in statistical methodology for diagnostic medicine in the 1980's', *Statistics in Medicine*, **10**, 1887–1895 (1991).
2. Campbell, G. 'General methodology I: advances in statistical methodology for the evaluation of diagnostic and laboratory tests', *Statistics in Medicine*, **13**, 499–508 (1994).
3. Swets, J. A. 'Measuring the accuracy of diagnostic systems', *Science*, **240**, 1285–1293 (1988).
4. Erdreich, L. S. and Lee, E. T. 'Use of relative operating characteristic analysis in epidemiology: a method for dealing with subjective judgment', *American Journal of Epidemiology*, **114**, 649–662 (1981).
5. Hanley, J. A. and McNeil, B. J. 'The meaning and use of the area under a receiver operating characteristic (ROC) curve', *Radiology*, **143**, 29–36 (1982).
6. Lee, W. C. and Hsiao, C. K. 'Alternative summary indices for the receiver operating characteristic (ROC) curve', *Epidemiology*, **7**, 605–611 (1996).
7. Ekelund, R. B. and Tollison, R. D. *Economics*, Little, Brown and Company, Boston, 1986.
8. Shryock, H. S. and Siegel, J. S. *The Methods and Materials of Demography*, U.S. Government Printing Office, Washington, 1975.
9. Lee, W. C. 'Analysis of seasonal data using the Lorenz curve and the associated Gini index', *International Journal of Epidemiology*, **25**, 420–434 (1996).
10. Lee, W. C. 'Characterizing exposure-disease association in human populations using the Lorenz curve and Gini index', *Statistics in Medicine*, **16**, 729–739 (1997).
11. Butler, R. J. and McDonald, J. B. 'Using incomplete moments to measure inequality', *Journal of Econometrics*, **42**, 109–119 (1989).
12. Simel, D. L., Samsa, G. P. and Matchar, D. B. 'Likelihood ratios with confidence: sample size estimation for diagnostic test studies', *Journal of Clinical Epidemiology*, **44**, 763–770 (1991).
13. Simel, D. L., Samsa, G. P. and Matchar, D. B. 'Likelihood ratios for continuous test results: making the clinicians job easier or harder?', *Journal of Clinical Epidemiology*, **46**, 85–93 (1993).
14. Youden, W. J. 'An index for rating diagnostic tests', *Cancer*, **3**, 32–35 (1950).
15. Connell, F. A. and Koepsell, T. D. 'Measures of gain in certainty from a diagnostic test', *American Journal of Epidemiology*, **121**, 744–753 (1985).
16. Asch, D. A., Patton, J. P. and Hershey, J. C. 'Knowing for the sake of knowing: the value of prognostic information', *Medical Decision Making*, **10**, 47–57 (1990).
17. Hilden, J. 'The area under the ROC curve and its competitors', *Medical Decision Making*, **11**, 95–101 (1991).
18. Somoza, E. 'Classifying binormal diagnostic tests using separation-asymmetry diagrams with constant-performance curves', *Medical Decision Making*, **14**, 157–168 (1994).
19. Caroll, B. J., Feinberg, M., Greden, J. F., *et al.* 'A specific laboratory test for the diagnosis of melancholia: standardization, validation, and clinical utility', *Archives of General Psychiatry*, **38**, 15–22 (1981).
20. Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*, Chapman and Hall, New York, 1993.
21. Härdle, W. *Smoothing Techniques*: *with Implementation in S*, Springer-Verlag, New York, 1991.

 

22. Metz, C. E. 'Basic principles of ROC analysis', *Seminars in Nuclear Medicine*, **8**, 283–298 (1978).
23. DeNeef, P. and Kent, D. L. 'Using treatment-tradeoff preferences to select diagnostic strategies: linking the ROC curve to threshold analysis', *Medical Decision Making*, **13**, 126–132 (1993).
24. Wieand, S., Gail, M. H., James, B. R. and James, K. L. 'A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data', *Biometrika*, **76**, 585–592 (1989).
25. Thompson, M. L. and Zucchini, W. 'On the statistical analysis of ROC curve', *Statistics in Medicine*, **8**, 1277–1290 (1989).
26. Knottnerus, J. S. 'Application of logistic regression to the analysis of diagnostic data: exact modeling of a probability tree of multiple binary variables', *Medical Decision Making*, **12**, 93–108 (1992).
27. Diamond, G. A., Hirsch, M., Forrester, J. S. *et al.* 'Application of information theory to clinical diagnostic testing: the electrocardiographic stress test', *Circulation*, **63**, 915–921 (1981).
28. Somoza, E., Soutullo-Esperon, L. and Mossman, D. 'Evaluation and optimization of diagnostic tests using receiver operating characteristic analysis and information theory', *International Journal of Biomedical Computing*, **24**, 153–189 (1989).

 