12. Salzman EW 1977 Platelets, prostaglandins, and cyclic nucleotides. In: Gaetano G, Garattini S (eds) Platelets: a multidisciplinary approach. Raven Press, New York, p 227
13. Simons TJB 1974 Resealed ghosts used to study the effect of intracellular calcium ions on the potassium permeability of human red cell membranes. J Physiol 246:52P
14. Skaer RJ, Peters PD, Emmines JP 1970 Platelet dense bodies: a quantitative microprobe analysis. J Cell Sci 20:441
15. Sneddon JM 1973 Blood platelets as a model for monoamine-containing neurons. In: Kerkut GA, Phillis JW (eds) Progress in Neurobiology, vol 1. Pergamon Press, Oxford, p 151
16. Steiner M, Tateishi T 1974 Distribution and transport of calcium in human platelets. Biochim Biophys Acta 367:232
17. Taylor PM, Heptinstall S 1980 The abilities of human blood platelets to bind extracellular calcium and to be aggregated by adenosine diphosphate are related. Br J Haematol 46:115
18. White JG 1970 Origin and function of platelet dense bodies. Ser Haematol 3:17

0031-3998/84/1809-0916$02.00/0
PEDIATRIC RESEARCH

Vol. 18, No. 9, 1984
Printed in U.S.A.

Application of Receiver-Operator Analysis to Diagnostic Tests of Iron Deficiency in Man

INSUN KIM, ERNESTO POLLITT, RUDOLPH L. LEIBEL, FERNANDO E. VITERI, AND EDMUNDO ALVAREZ

School of Public Health, The University of Texas Health Science Center at Houston, Houston, Texas 77225 [I.K., E.P.], Laboratory of Human Behavior and Metabolism, Rockefeller University, New York, New York, 10021 [R.L.L.], Pan American Health Organization, Washington, D. C., 20037 [F.E.V.], and Institute of Nutrition of Central America and Panama, Guatemala City, Guatemala [E.A.]

Summary

The objective of the present report is to demonstrate the use of receiver-operator characteristics (ROC) analysis in the selection of diagnostic tests for iron deficiency in a specific population. Conventional ROC curves were prepared with true positive fraction (TPF) and false positive fraction (FPF) determined by the application of different cut-off points for four indicators of iron status. ROC plots were then transformed into normal deviate scales. The advantages of Gaussian transformation of TPF and FPF when underlying decision functions are normally distributed are: (i) the ROC curve is a straight line; and (ii) the separation between the two distributions and shape of these distributions can be simply quantitated as intercepts and slopes. In the present study, pretreatment hemoglobin concentration was the most robust diagnostic indicator of iron deficiency as operationally defined by a response of hemoglobin to iron treatment. Free erythrocyte protoporphyrin was a more sensitive and specific predictor than either serum ferritin or transferin saturation when a stringent operational definition of iron deficiency was used. These findings illustrate the utility of ROC analysis in discriminating between diagnostic indicators having different degrees of accuracy.

Abbreviations

FN, false negative
FP, false positive
Hb, hemoglobin
***N*, normal individuals**
***D*, diseased cases**
ROC, receiver-operator characteristics
TPF, true positive fraction
FPF, false positive fraction
FEP, free erythrocyte protoporphyrin
SF, serum ferritin
TS, transferrin saturation

Received March 30, 1983; accepted February 8, 1984.
Address reprint requests to: Ernesto Pollitt, Ph.D., The University of Texas, School of Public Health, PO Box 20186, Houston, TX 77225.
This study was partially supported by National Institutes of Health Grant R01-HD 12843.

A wide range of laboratory tests is currently used to assess systemic iron status in man. However, normal biological variability, measurement error, and confounding factors such as intercurrent infection may adversely affect the diagnostic efficiency of these tests. Some of these problems are minimized when iron status is operationally defined by the degree of hematologic response to iron administration. In assessing an individual's iron status, a significant rise in circulating hemoglobin mass in response to iron treatment provides reliable evidence of antecedent iron deficiency. Hemoglobin response can also be used to monitor the diagnostic efficiency of other tests of systemic iron status.

The evaluation of a test's diagnostic efficiency requires assessment of its discriminative capacity in circumstances where the frequency and nature of its diagnostic errors can be unequivocally determined. A test's accuracy (ratio of correct decisions to total number of subjects tested) is of limited usefulness as a general index of diagnostic performance because it is strongly affected by disease prevalence (8).

If a test is to be used to discriminate iron-replete from iron-depleted subjects, some definitive diagnostic criterion is needed to allow evaluation of that test. In Figure 1, the performance of a hypothetical diagnostic test is examined. Diseased subjects, whose test result places them at the right of the cut-off point, *a*, will be FNs; normal individuals whose result is to the left of *a* will be FPs. The number of FPs can be reduced or eliminated by moving the cut-off *a* toward *b*, to the lower end of the distribution for *N*. However, as a result of eliminating FPs, the FN fraction will be increased. Likewise, the number of FNs can be eliminated by moving the cut-off *a* to *c*, the upper end of the distribution for *D*. The cut-off point can be positioned so as to maximize a test's diagnostic performance in a given clinical or epidemiologic context (8)

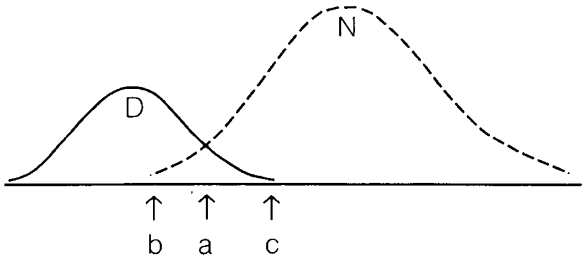

Fig. 1. Distribution of diseased (*D*) and nondiseased (*N*) subjects in a population. The probability of positive and negative tests is equal to the area under a distribution to either side of the cut-off point. When a cut-off point for a test is set at *a*, the TPF is equal to the area of *D* distribution and the FPF is equal to the area of *N* distribution to the left of a given cut-off point, *a*. *a*, *b*, and *c*, various cut-off points for a given discriminating test.

One important consideration in the performance of tests to screen populations for the presence of disease is the relationship between sensitivity, specificity, and the prevalence of the disease being screened. When the prevalence of a disease is very low (*e.g.*, 1 or 2%), even a highly sensitive and specific test will generate a large number of FPs. A small decrease in test specificity in this context will substantially increase the number of FPs. Unless offset by a large gain in sensitivity, the proportion of FPs will increase or, at best, remain unchanged. The effects on test performance of shifts in cut-off point can be usefully analyzed by examining the ROC of a particular indicator.

ROC analysis was developed as part of signal detection theory (7) and has been applied extensively to experimental studies of cognition (10). The ROC curve is a continuous plot of TPF *versus* FPF, both of which are independent of disease prevalence as the cut-off point for a diagnostic test is systematically varied.

The abscissa of the ROC curve represents the false positive fraction (FPF = 1 − specificity) and the ordinate represents the true positive fraction (TPF = sensitivity). Choice of an optimal cut-off point for a discriminating test will vary depending on the context of its application. If disease prevalence is low, then the FPF must be kept low; otherwise most positive tests will be false positives. A cut-off point at the lower part of the ROC curve will keep the FPF small, albeit at the cost of a low sensitivity. Conversely, if the same test is applied in circumstances in which disease prevalence is high (and, thus, FPs less prevalent), then a cut-off further up the ROC curve may be chosen to maximize the TPF while holding the false negative fraction relatively low (6). There may also be epidemiologic circumstances where location of all diseased individuals is more important than concerns about FPs (*e.g.*, tuberculin skin testing). In this instance, choosing a cut-off point to maximize sensitivity is more important than maximizing specificity. The ROC curve allows one to predict the extent to which the FPF will increase when the TPF is set at any desired level.

The objective of the present methodologic study was to demonstrate the use of ROC analysis in the selection of diagnostic tests and cut-off points for the diagnosis of iron deficiency in a discrete population. For this purpose, the response of hemoglobin concentration to oral iron administration was used as an operational definition of iron deficiency and as the referent to which various other indicators of systemic iron status were compared. Optimal cut-off points for each indicator of iron status were selected by intentionally varying the diagnostic cut-off point and observing the resulting changes in sensitivity and specificity of the test.

MATERIALS AND METHODS

Community. The clinical study was conducted in two lowland Guatemalan villages, Los Llanitos and Las Guacas, about 20 km from the port of San Jose on the Pacific coast and 85 km south of Guatemala City. The inhabitants of the community are exclusively "ladino," a mixture of Mayan Indian and Spanish.

The main causes of iron deficiency in the Guatemalan rural population appears to be a combination of low iron intake and low efficiency of absorption (12). Virtually all of the dietary iron is provided in the form of non-heme iron. Previous studies in Guatemala have shown that the administration of ferrous sulfate by mouth (6 mg/kg/day) for 4 months, and without any other intervention, markedly reduced the prevalence of anemia among preschool children (13). Hemoglobinopathies that would limit the effect of iron supplementation on hemoglobin mass are essentially nonexistent in this population (13).

Subjects and study design. The criteria used to select the children were a gestational age of 38 weeks or more, birthweight of 2,500 g or more, and no evidence of any chronic illness, severe malnutrition, or primary hematologic disorder that could adversely affect body hemoglobin mass. One hundred fifty-three children, ranging in age from 30 to 72 months, were assigned to one of three groups based on their initial venous hemoglobin values (<10.5 g/dl; 10.5–11.5 g/dl; >11.5 g/dl). Each of the two groups with the higher hemoglobin concentrations were then subdivided into two subgroups of equal size. Each member of these two subgroups was then randomly assigned to an iron or placebo treatment. All except two children with Hb < 10.5 g/dl received iron treatment.

Table 1 represents the distribution of the study subjects according to NCHS anthropometric standards (11). Waterlow's classification (14), using the parameters of height-for-age and weight-for-height, is used to describe the nutritional status of the subjects. Eighty-nine children (58.2%) are classified as normal. The remaining 64 children (41.8%) are classified as stunted, the condition that characterizes a population suffering from chronic undernutrition.

An oral ferrous sulfate solution (Fer-in-Sol) at a daily dose of 3 mg elemental iron per kg body weight was used as treatment. The placebo group received an indistinguishable solution prepared at the Institute of Nutrition of Central America and Panama and containing no iron or other nutrient. Both iron and placebo were administered daily by a field worker, 5 days a week, over a period of 11 weeks. Field personnel and parents had no knowledge of the child's initial hematologic status, nor of the content of the oral preparation. Table 2 presents the number of children, broken down by initial Hb level and treatment group assignments.

Laboratory procedure. A 6-ml venous blood sample was obtained in the morning at the beginning (T_1) and the end (T_2) of the study period from each child. An 1.2-ml aliquot of whole blood was separated for hematologic analysis; the remainder was immediately centrifuged and the serum was harvested. Serum samples were kept frozen at −20°C until analyzed. Serum ferritin (Clinical Assays, Boston, MA), serum iron (15), total iron-binding capacity (15), free erythrocyte protoporphyrin (1), and hemoglobin (5) concentration were determined on all samples.

Table 1. *Classification of physical growth according to reference anthropometric standards**

Height for age	Weight for height <80%	Weight for height ≥80%	Total
<90%	0 (0%) Stunted and wasted	64 (41.8%) Stunted	64 (41.8%)
≥90%	0 (0%) Wasted	89 (58.2%) Normal	89 (58.2%)
Total	0	153	153

* Values represent number of children.

Table 2. *Number of children in iron- and placebo-treated groups versus initial hemoglobin concentration levels*

Initial Hb level (g/dl)	Iron-treated	Placebo	Total
<10.5	40	2	42
10.5–11.5	12	10	22
>11.5	44	45	89
Total	96	57	153

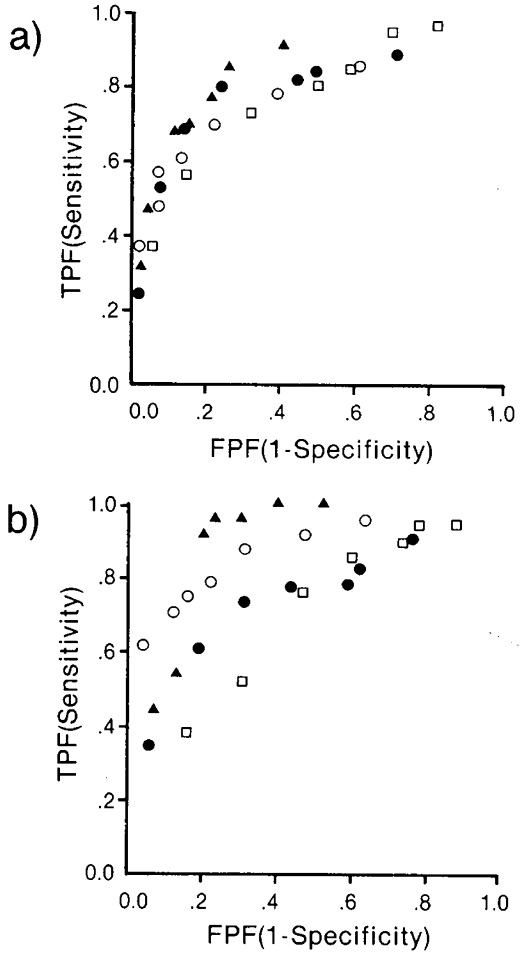

Fig. 2. ROC curves for various indicators of iron nutrition. Iron deficiency is operationally defined as (*a*) ΔHb ≥ 1.0 g/dl and (*b*) ΔHb ≥ 2.0 g/dl following iron administration. The *values* in *parentheses* represent the cut-off values of the respective indicators. HGB, hemoglobin.

RESULTS

Ninety-four of the 96 children who received iron treatment for an average of 11 weeks were chosen for the main analysis in the present study; two cases did not have complete data at both time periods. Figure 2 presents two groups of ROC curves based on a Hb response to iron treatment ≥1.0 g/dl (*a*) and ≥2.0 g/dl (*b*); each curve shows the performance data for all the iron indicators used. There are seven entries per indicator curve; each entry is determined by the use of a different cut-off point. Figure 2, *a* and *b*, illustrates the TPF and FPF obtained at each cut-off point for each indicator. A simple visual comparison of the ROC plots (*a* and *b*) shows that as the referent criterion is increased from ΔHb ≥ 1.0 g/dl to ΔHb ≥ 2.0 g/dl, there is improvement in the diagnostic efficiency of these iron indicators.

The differences in the shapes of ROC curves are better appreciated in a plot drawn on a double integrated normal chart or double probit paper which linearizes the area under the normal distributions. In Figure 3, the TPF is equal to the area of distribution *D* and the FPF is equal to the area of the distribution *N* to the left of a given cut-off point. Assuming that distributions *N* and *D* are both normal, the normal deviate scale of the TPF and FPF for a given cut-off point will be equivalent to the corresponding *Z* scores, relative to the nondiseased population, of the resulting distributions for the diseased and nondiseased populations, respectively. When a cut-off point for the diagnostic test is set at *a* in Figure 3, the normal deviate of the TPF is equal to +1.5, *Z* score for the area of *D* to the left of *a*, on the scale of the *Zd* and the FPF is equal to −1.0, *Z* score for the area of *N* to the left of *a*, on the scale of *Zn*. At a cut-off point that provides 50% FPF (*b* in Fig. 3) for example, the TPF is more than 99% or +3.0 on *Zd* scale which is represented by the intercept (read *Z* (TPF) at *Z* (FPF) = 0 or *Z* (TPF) at FPF = 50%) of the ROC plot on a double normal deviate scale. When a cut-off point is systematically varied by 1 unit of *Zn* (FPF), the increment of *Zd* (TPF) is 1.5 which corresponds to the slope of the ROC plot on a double normal deviate scale. Therefore, the intercept of the transformed ROC plot is equal to the difference in the mean values between the diseased and nondiseased populations expressed in units of standard deviation of the diseased. The slopes of these plots are proportional to the rate of the standard deviation of the nondiseased to the standard deviation of the diseased population. The area under the fitted ROC curve is theoretically equal to the probability of the correct decisions.

In the present paper, a least squares estimation of a linear regression equation was obtained for each ROC curve plotted on a double normal deviate scale. The respective slopes and intercepts are shown in Figure 4. χ^2 assessment of goodness-of-fit for a straight line was tested on each transformed ROC curve. The assumption of linearity was therefore reasonable in all cases.

When ROC data are transformed to normal deviate scales, parallel lines are interpreted as indicating that one indicator has higher sensitivity than the other at all FPFs. However, the rate of improvement in TPF with respect to FPF (*i.e.*, slope) is the same for both indicators. Intersection of two lines indicates the comparative performance of the indicators changes as a function of the specific cut-off point used. In such cases, the relative quality of performance of each indicator must be judged in the context of its diagnostic application.

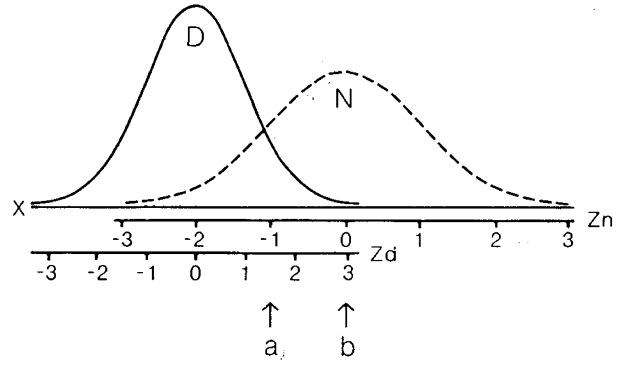

Fig. 3. Distribution of diseased (*D*) and nondiseased (*N*) subjects in a population. Assuming the normality of both distributions, normal deviates of the TPF and the FPF are equal to the respective *Z* scores of the distributions of diseased and nondiseased subjects for a given cut-off point. The difference in the mean values between these two distributions is represented by the intercept of the transformed ROC plot on a double normal deviate scale. The improvement of TPF, when a cut-off point is systematically varied by 1 unit of FPF, is represented by the slope of ROC plot. *X*, scale of test values for a given diagnostic test; *Zd*, normal deviate scale of the distribution for diseased subjects; *Zn*, normal deviate scale of the distribution for nondiseased subjects; *a* and *b*, cut-off points for a given diagnostic test.

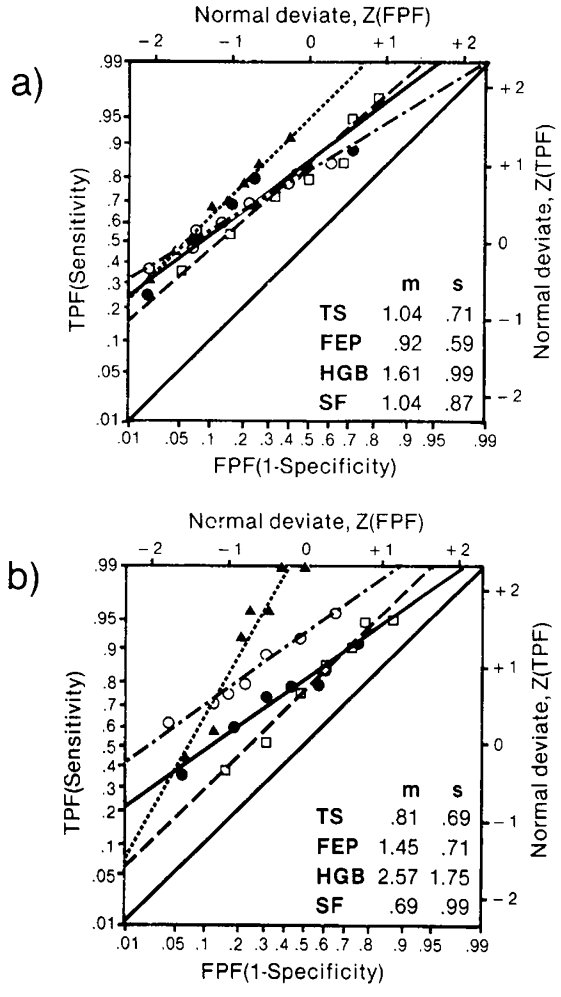

Fig. 4. ROC data replotted on a double normal deviate scale. Iron deficiency is operationally defined as (*a*) ΔHb ≥ 1.0 g/dl and (*b*) ΔHb ≥ 2.0 g/dl following iron administration. The *values* in the *parentheses* represent the cut-off values of the respective indicators. Δ*m*, intercept; *s*, slope; *HGB*, hemoglobin.

When iron deficiency is operationally defined as a Hb change of 1.0 g/dl or more following iron therapy, the intercept of the plot for initial Hb as an indicator is significantly higher than either FEP, SF, or TS. Moreover, the slope of the Hb plot is also significantly different from the slopes of the plots for either FEP or TS. In Figure 4, it can be seen, however, that the intersection of the Hb plot and that of the other two indicators occurs at a point 50% TPF. At the 50% FPF, Hb is already performing at a much higher degree of sensitivity than FEP and TS. Finally, there is parallelism between the SF and Hb plots indicating superior predictive performance of initial Hb at all cut-off points examined. Analysis of comparative performance of FEP, TS, and SF indicates that there is no statistically significant difference in their ability to predict reponse to iron treatment in this population.

When iron deficiency is defined as a change in Hb of 2.0 g/dl or more with iron therapy, except for the comparisons of intercepts between TS and SF and of slopes between TS and FEP, all other comparisons are statistically significant. These findings and Figure 4*a* indicate that, at 50% FPF, the TPF of initial Hb is significantly higher than that of all other indicators. Likewise, at 50% FPF level the performance of FEP is significantly better than that of SF and TS. However, there is no difference in the intercepts between SF and TS. At higher levels of FPF, the difference between Hb and the other indicators decreases, while the performance of FEP is maintained.

DISCUSSION

In the present study, more than one-fourth of the subjects were diagnosed as iron deficient when a stringent operational diagnostic criterion of iron deficiency was applied (*i.e.*, ΔHb ≥ 2.0 g/dl following iron therapy). When the criterion was set at ΔHb ≥ 1.0 g/dl, 47 (50%) were diagnosed as iron deficient. However, neither of these values reflects the actual prevalence of iron deficiency in the population studied. While all of the children with Hb less than 10.5 g/dl were treated with iron, only half of the children with Hb greater than 10.5 g/dl received iron. Therefore, the sample is skewed toward iron-deficient subjects and the true prevalence rate of iron deficiency in the population cannot be ascertained from this study.

The design of this study called for the treatment of only half of the children with possible mild iron deficiency (10.5 ≤ Hb ≤ 11.5 g/dl). Some children in the Hb > 11.5 g/dl group may also have been iron deficient, but none of these children were given iron. Thus, the group of children given iron in this study is not truly representative of the entire population of children in the community, and conclusions regarding the population's true prevalence rates for iron deficiency cannot be drawn from the present study. Where the sole purpose of such a study is to determine local prevalence rates of iron deficiency, the best strategy is to administer iron to a representative subsample of the entire community and to use response of Hb concentration as a *post hoc* means of classification. The present paper should be regarded as an exposition of a technique of data analysis, and not as a study of the prevalence of iron deficiency in a community. Likewise, because the distribution of starting Hb concentrations is experimentally created, the performance of the indicators of iron status used in this study cannot be extrapolated to the population of all children within the age range examined.

The discriminatory power of initial Hb concentration as a predictor of response to iron administration is not surprising on both statistical and biological grounds. First, pretreatment Hb values and response to treatment share common variance of the same measure. In this selected population, with a high prevalence of severe iron deficiency, Hb response to iron treatment obviously varied as a function of initial Hb values. Secondly, low Hb concentration represents an advanced state of the iron depletion process and large increments in Hb during iron treatment are most likely to occur in subjects with the lowest initial Hb mass. Thus, it is not surprising that initial Hb level is the most robust predictor of response to iron administration in the paradigm employed here. Both TS and FEP seem to perform best (relative to the best predictor, which is pretreatment Hb concentration) when mild iron deficiency is being sought. At more severe degrees of deficiency (as reflected by ΔHb with iron administration), these measures lose some of their power to predict the extent of Hb response to iron. This may be due to the relatively early plateauing of the response of these indicators to iron deficiency. Ferritin, though a poor predictor of response to iron therapy, does appear to show a more continuous (inverse) relationship to Hb response to the administration of iron. Such relationships may indicate something of the physiology of various iron compartments in man, indirectly confirming the observation that various compartments are differentially affected by progressively more severe degrees or iron deficiency.

Pretreatment Hb concentration was the most robust diagnostic indicator of iron deficiency as operationally defined by a response of Hb concentration to iron treatment. Moreover, FEP was a more sensitive and specific predictor than either SF or TS when a stringent operational definition (ΔHb ≥ 2.0 g/dl) of iron deficiency was used. These findings illustrate the utility of ROC analysis in discriminating between diangnostic indicators having different degrees of accuracy.

A problem in the statistical discrimination between each transformed ROC curves is the use of two parameters (intercept and slope) for comparative purposes. This may represent a particularly difficult obstacle when, for example, there are differences in the intercepts and the lines intersect above 50% TPF and below 50% FPF. A frequently used method of testing the differences between areas under the fitted ROC curves (4, 9) is also subject to qualitative judgments that may have to be made based on the particular context in which the test is to be applied. Conclusions drawn from the study are applicable only to the sample population on which the study is conducted.

REFERENCES

1. Blumberg WE, Eisinger J, Lamola AA, Zuckerman DM 1977 The hematofluorometer. Clin Chem 23:270
2. Collis LD 1981 Socioeconomic level, iron status and growth of children in rural Guatemala. MPH thesis, University of Texas
3. Dallman PR, Refino C, Yland MC 1982 Sequence of development of iron deficiency in the rat. Am J Clin Nutr 35:671
4. Dorfman DD, Alf E 1968 Maximum likelihood estimation of parameters of signal detection theory—a direct solution. Psychometrika 33:117
5. Drabkin DL 1949 The standardization of hemoglobin measurement. Am J Med Sci 217:710
6. Galen RS, Gambino SR 1975 Beyond Normality: the Predictive Value and Efficiency of Medical Diagnoses. John Wiley & Sons, New York
7. Green DM, Swets JA 1966 Signal Detection Theory and Psychophysics. John Wiley & Sons, New York
8. Leibel RL, Pollitt E, Kim I, Viteri F 1982 Studies regarding the impact of micronutrient status on behavior in man: iron deficiency as a model. Am J Clin Nutr 35:1211
9. Metz CE 1978 Basic principles of ROC analysis. Semin Nucl Med 8:283
10. Swets JA, Pickett RM, Whitehead SF, Getty DJ, Schnur JA, Swets JB, Freeman BA 1979 Assessment of diagnostic technologies. Science 205:753
11. United States Department of Health, Education, and Welfare 1977 NCHS growth curves for children, birth–18 years, United States. Vital and Health Statistics Ser 11, No 165, Publ No (PHS)76–1650
12. Viteri FE 1973 Hematologic status of the Central American population: iron and folate deficiencies. In: Pan American Health Organization 12th Meeting of the Advisory Committee on Medical Research. Pan American Health Organization, Washington, DC
13. Viteri FE, Garcia-Ibanez R, Torun B 1978 NaFeEDTA as an iron fortification compound in Central America: absorption studies. Am J Clin Nutr 31:961
14. Waterlow JC, Buzina R, Keller W, Lane JM, Nichaman MZ, Tanner JM 1977 The presentation and use of height and weight data for comparing the nutritional status of groups of children under the age of 10 years. WHO Bull 55:489
15. Yeh Y-Y, Zee P 1974 Micromethod for determining total iron binding capacity by flameless atomic absorption spectrophotometry. Clin Chem 20:360