UW Biostatistics Working Paper Series

1-7-2005

Limitations of the Odds Ratio in Gauging the Performance of a Diagnostic or Prognostic Marker

Margaret S. Pepe
University of Washington, mspepe@u.washington.edu

Holly Janes
University of Washington, hjanes@u.washington.edu

Gary M. Longton
Fred Hutchinson Cancer Research Center, glongton@fhcrc.org

Wendy Leisenring
Fred Hutchinson Cancer Research Center, wleisenr@fhcrc.org

Polly Newcomb
Fred Hutchinson Cancer Research Center, pnewcomb@fhcrc.org

BACKGROUND

The idea of using information about a subject to detect subclinical disease states and to predict future health events has great appeal. The notion is currently motivating much biotechnologic medical research. We hope to use biomarkers derived from new proteomic and genomic technologies to identify subjects that have or are very likely to develop cancer or other diseases [1]. In addition we hope to use these modern technologies and others to make precise diagnoses and more accurate prognoses of patients with disease, to help with decisions about treatment, and to monitor response to treatment. The use of biomarkers and risk factors in this way is not a new notion in medical practice. Prediction risk scores are commonly used. Examples include the Framingham risk score for cardiovascular events [2] and the Gail model risk score for breast cancer [3]. Even more common, epidemiologists have identified a myriad of disease specific risk factors that have been used alone, or in combination in public health practice. Clinical epidemiologists have analogously identified a multitude of factors that are associated with the clinical course of patients diagnosed with disease.

The statistical evaluation of factors, scores and biomarkers for assessing an individual's current status or future health outcome is the topic of this paper. We use the generic term 'marker' for the factor, score or biomarker and 'outcome' for that which is predicted or detected. We show in this paper that strong statistical *associations* between outcome and marker do not necessarily imply that the marker can usefully discriminate between those individuals likely to have the outcome from those who do not. Traditional statistical methods used by epidemiologists to assess etiologic associations are not adequate to determine the potential performance of a marker for classifying or predicting risk for individuals. This important point is not widely appreciated and may explain to some extent the disappointing performance of many identified "markers" when they are used to predict outcome for individuals. As we proceed to develop technologically sophisticated tools for individual level prediction and classification, for so-called 'personalized medicine,' we must be

careful to use appropriate statistical techniques in evaluating research studies aimed at assessing their performance. We describe some appropriate techniques in this paper.

The performance of a marker may change with the circumstance in which it is applied. Characteristics of the population, or the assay technique (if the marker is a biomarker) for example, may lead to better or worse performance. It is important to understand and quantify variations in the performance of a marker. We show how such questions can be addressed statistically and the pitfalls of using some common epidemiological methods for this purpose. In addition we discuss the evaluation of a marker in the presence of other information that is predictive of outcome. This may include risk factor data or other markers. The question is how to evaluate the incremental value of a marker for distinguishing cases from controls. Again we show that traditional epidemiologic methods can lead to false conclusions and we describe more appropriate statistical methods.

ASSOCIATION VERSUS CLASSIFICATION

Consider a binary risk factor as a marker. For example, unopposed estrogen replacement therapy is considered to be a strong risk factor for development of endometrial cancer [4]. It has a relative risk of about 3.0 associated with it, which is to say that in a case-control study the odds ratio for comparing cases versus controls in regards to 'ever having used estrogen replacement therapy' is about 3.0. Reporting the odds ratio or relative risk as a measure of association is typical in epidemiologic studies of etiologic risk factors and is now unfortunately commonly employed for studies of predictive markers as well. See for example Cui et al. (2003) in a recent issue of *Science* and Rhodes et al. (2003) in a recent issue of *Journal of the National Cancer Institute* [5, 6]. Examples from other popular journals are Ridker et al (2000), Zhang et al. (2001), Liou et al. (1999), and Hogue et al. (2001) [7, 8, 9, 10]. We note that the goals of etiologic risk factor studies are quite different from the sorts of studies we consider here, where markers are to be used for classifying individuals. The statistical considerations therefore also differ between such studies.

The accuracy or validity of a binary marker for classifying individuals is better summarized in a case-control study by reporting its true positive fraction (TPF, also known as sensitivity) and its false positive fraction (FPF, also known as 1-specificity):

$$\text{TPF} = \text{Prob}[\text{marker positive} \mid \text{outcome positive}]$$

and

$$\text{FPF} = \text{Prob}[\text{marker positive} \mid \text{outcome negative}].$$

A perfect marker will have TPF = 1 and FPF = 0. How close TPF should be to 1 and how close FPF should be to 0 in order for the marker to be useful for individual level classification depends on the context. Nevertheless the values cannot be too far away from the ideal values in order for an individual to believe the prediction.

The odds ratio (OR) can be written as a simple function of (FPF, TPF) [11]:

$$\text{OR} = \left(\frac{\text{TPF}}{1-\text{TPF}}\right) \times \left(\frac{1-\text{FPF}}{\text{FPF}}\right)$$

Figure 1 shows this relationship. Accuracy points (FPF, TPF) that yield the same value of the odds ratio are shown. Observe that an odds ratio of 3.0 is not consistent with an "accurate" marker. Suppose for example that a marker labels as many as 10 percent of controls (outcome negatives) as positive and has an odds ratio of 3.0 associated with it. We see from Figure 1 that it identifies only about 25 percent of the cases (outcome positives). That is, 75 percent of the cases are not picked up by the marker. As another example, suppose that a marker with OR=3 detects 80 percent of cases. The plot shows that it must mislabel almost 60 percent of the controls. Clearly this is not a marker that is useful for individual level classification or prediction. The figure shows that even weakly accurate markers are associated with odds ratios (or relative risks) that are far larger than those traditionally considered strong in epidemiologic studies of association.

When the marker, denoted now by X, is continuous, its association with outcome status, $D = 1$ for case and $D = 0$ for control, is also often summarized with an odds ratio. Consider the logistic regression model

$$\text{Prob}(D = 1|X) = \exp(\alpha + \beta X)/\{1 + \exp(\alpha + \beta X)\}$$

The odds ratio per unit increase in X is given by $\exp(\beta)$. The size of the odds ratio depends on the units in which X is measured. In Figure 2 we have scaled X so that a unit increase represents the difference between the 16^{th} and 84^{th} percentiles of X in the controls (i.e., 2 standard deviations = 1 unit). The distribution of X in controls is represented as normal with mean 0, which is general in the sense that data can always be transformed to this scale. Assuming that, for cases, X is also normally distributed with the same standard deviation as controls, Figure 2 shows their separation from controls for various values of the odds ratio. We see again that values of the OR considered large in traditional epidemiologic studies are derived from marker distributions that are largely overlapping (Figure 2).

Receiver operating characteristic (ROC) curves corresponding to each of the pairs of marker distributions in Figure 2 are shown in Figure 3. Each point on an ROC curve represents the decision criterion that is positive if X exceeds a threshold c. The FPF and TPF values associated with that criterion is one point on the curve, and by varying c from $+\infty$ to $-\infty$, the (FPF, TPF) points corresponding to all possible thresholds are shown. Although they appear similar, Figure 1 differs from Figure 3. Figure 1 concerns binary markers only, with odds ratios defined in the usual way for a binary marker. Many different markers are represented on the same curve in Figure 1 if their odds ratios are the same. The odds ratios here in Figure 3 relate to a unit increase in a continuous marker and the ROC curve concerns different decision criteria resulting from thresholding that single continuous marker.

We see from Figure 3 that when the OR associated with a unit increase in the continuous

marker X is 3.0, regardless of the threshold chosen, the (FPF, TPF) values associated with the corresponding decision criterion generally would not be adequate for individual level classification. In fact unless the OR per unit increase in X is at least 16, marker based decision criterion seem to be very inaccurate. Even with OR = 16, a marker based criterion that yields 10 percent FPF at a threshold fails to detect over 40 percent of cases using that threshold. As another example, it will falsely detect as many as 30 percent of controls if a threshold that yields 80 percent of the cases is used.

Frequently, continuous markers are grouped to form categorical covariates, allowing the analyst to avoid specific assumptions about the shape of the log (odds) function. For each of the marker distributions shown in Figure 2, we also categorized both cases and controls based on the quartile cut-points from the control population. The odds ratios for each quartile relative to the lowest quartile is calculated and displayed in Table 1. The solid circle points shown on the ROC curves in Figure 3 are those associated with using each of the three quartile cut-points to classify subjects as positive or negative for disease. Similar to our observations for the binary and continuous marker settings, even if the odds ratio for an upper quartile versus the lowest quartile is of a large magnitude, the corresponding points on the ROC curves for each cut-point show poor ability to classify cases and controls. For example, when the OR for the upper quartile vs. the lowest quartile is 4.1, if we use the upper quartile to define a positive test result, only 45 percent of cases are correctly classified while 25 percent of controls were incorrectly identified.

In summary, there are two key points that we wish to make based on Figures 1, 2 and 3. The first, as previously stated, is that markers with odds ratios of the order that are considered strong in traditional epidemiologic research, are not adequate for discriminating between those who do and do not have an outcome of interest. Extremely strong associations are needed. The second is that odds ratios in and of themselves do not characterize the discriminatory capacity of a marker.

The odds ratio is a simple scalar measure of *association* between marker and outcome. It does not characterize the discrimination between cases and controls that can be achieved by a marker since many different pairs of true and false positive fractions are consistent with a particular value of OR. In the next section we discuss alternatives to the odds ratio that can be used in case-control studies to evaluate markers.

HOW TO QUANTIFY DISCRIMINATION IN A CASE-CONTROL STUDY

Classification Error Rates

Characterization of the discriminatory capacity of a binary marker has already been addressed in Section 2. The true and false positive fractions provide a description. Although predictive values (PV) are also of great interest, where positive PV = P[$D = 1$| marker positive] and negative PV = P[$D = 0$| marker negative], these entities essentially require either a cohort study design where the sample prevalence reflects the population prevalence or some external estimates of prevalence that pertain to the population from which cases and controls are drawn. On the other hand, TPF and FPF are defined conditional on outcome status and so can be estimated from a case-control study where sampling is dependent on outcome. Diagnostic likelihood ratios have also been promoted as measures to characterize the accuracy of a binary marker [12] but are not very popular in practice.

It is interesting that the characterization of accuracy requires two parameters, TPF and FPF say, while association measures such as the odds ratio or correlation coefficient (for continuous markers) are generally one dimensional. The characterization with (FPF, TPF) acknowledges that false positive and false negative errors are not equivalent and must be reported separately [13].

The ROC curve is the natural generalization of (FPF, TPF) to accommodate settings where the marker is continuous. It describes the whole set of potential (FPF, TPF) combinations that are possible with positivity criteria based on the marker. The raw distributions and ROC curves for two pancreatic cancer biomarkers are shown in Figures 4 and 5 [14]. Observe that the ROC curve

does not depend on how the marker is coded. Changing the units in which the marker is measured has no impact on its ROC curve. This contrasts with logistic regression models where, as noted above, the odds ratio must be interpreted according to a unit increase in the value of X. Moreover, ROC curves provide a natural common scale for comparing different markers even when they are measured in completely different units. For example a marker that measures a concentration in serum can be compared with one that measures flight time per unit charge derived from protein mass spectrometry. In contrast, because odds ratios are interpreted per unit increase in the marker, odds ratios for two markers may not be comparable. This is a key advantage of ROC curves.

MODIFIERS OF PREDICTOR PERFORMANCE

A variety of factors (or covariates) may affect how well a marker performs. For example, higher breast density makes mammographic readings less accurate [15]. Factors beyond the subject are often important too. The assay technique or the expertise of the lab technician can affect how well a biomarker measure performs. In audiology, the location in which the hearing test is done can affect the capacity of the test to detect hearing loss. If one can establish which covariates influence the performance of a marker then one may use this information to optimize the marker measurement. On the other hand it can suggest settings or populations for which the marker is less useful and where alternative markers should be sought.

Not Traditional Effect Modification

We continue to denote the marker by X and denote covariates that may affect the performance of the marker by Z. In epidemiology one typically uses a logistic regression model with statistical interaction between covariates and the marker of interest to evaluate if there is 'effect modification.' Mathematically we write

$$\text{logit}P[D = 1|X, Z] = \alpha + \beta_1 X + \beta_2 Z + \beta_3 XZ.$$

Evaluating the size and significance of β_3, the interaction term, asks if the odds ratio associated

with X varies with Z. However, since we have already established that the odds ratio does not properly characterize marker performance, it follows that this approach does not address questions about Z affecting the performance of X as a marker.

As a simple example consider the data shown in Table 2 where X and Z are binary. One is considering to use the marker X as a screening device and, to make the discussion concrete, suppose that the covariate Z is gender ($Z = 1$ for females and $Z = 0$ for males). The odds ratio associated with X is exactly the same for males as for females. That is, Z is not an effect modifier in the sense of altering the *association* between X and disease D when association is parameterized by the odds ratio. It appears to these authors however, that at least for the purposes of disease screening, X performs better in females ($Z = 1$) than in males ($Z = 0$). Almost all cases are detected in both circumstances (TPF = 0.97 and TPF = 0.94) but twice as many controls screen positive in the male population where FPF=0.06 versus in the female population where FPF=0.03. For widespread screening of healthy populations it is critical to keep the number of false positive results extremely low in order that the program be affordable and acceptable to healthy people. Therefore the lower FPF observed in females makes it a better screening marker in that population.

Statistical Assessment

How then should one assess if a covariate affects the performance of a marker? When the marker is binary one can simply determine to what extent the TPFs vary with Z and to what extent the FPFs vary with Z. We made this assessment already informally for the data in Table 2. Formal statistical techniques can be applied to test a hypothesis such as $H_0 : \text{FPF}(Z = 1) = \text{FPF}(Z = 0)$ or to quantify the difference between $\text{FPF}(Z = 1)$ and $\text{FPF}(Z = 0)$ on some scale. The FPFs and TPFs are binomial proportions and the usual techniques of Pearson chi-squared statistics and so forth can be applied. Inference about FPFs uses data only for controls ($D = 0$) and inference about TPFs uses data only for cases ($D = 1$). For example, using the data from Table 2 the comparison

of FPFs results in $p = 0.002$ from a chi-squared test. Writing

$$\text{FPF}(Z) = P(X = 1|D = 0, Z)$$

we see that logistic regression techniques can be applied to data for controls with the marker as the dependent variable and covariates Z as the independent variables to establish how the FPF varies with Z. Regression techniques may be preferable when there are multiple components to Z or if Z is continuous. Similarly logistic regression can be applied to data for the cases to establish how

$$\text{TPF}(Z) = P(X = 1|D = 1, Z)$$

varies with Z. We refer to Leisenring, Pepe and Longton (1997), Smith and Hadgu (1992), and Pepe (2003) (Section 3.5) [16, 17, 18], for illustrations and in-depth discussion.

For a continuous marker, one needs to determine if the ROC curves for X vary with Z. There are various ways to do this. If Z is dichotomous one can plot separate ROC curves for X using data for the two groups or circumstances defined by Z. Statistical techniques to compare ROC curves have been developed and are included in some software packages including Stata (2003) [19]. Data for prostate specific antigen (PSA) reported by Etzioni et al. (1999) [20] are shown in Figure 6 for men <65 years of age and men >65 years of age. Although the study measured PSA repeatedly over time we only use data for the last time point (prior to diagnosis for cases). The classic statistic for comparing two ROC curves is the difference in the areas under the empirical ROC curves. The difference is not statistically significant ($p = 0.44$). Thus, there is no evidence in this sample that age affects the capacity of PSA to distinguish cases with prostate cancer from age-matched controls.

Similar techniques can be used to compare the performances of two different markers in the same population. The ROC curves in Figure 4 for CA-125 and CA-19-9 [14] (Wieand et al., 1989) are statistically significantly different ($p < 0.01$). This p-value is based on the difference in empirical

areas under the ROC curves applied to paired data [21]. Methods based on comparing ROC curves over a relevant subinterval of false positive fractions are described in Pepe (2003, page 110) [18] and are probably more appropriate for comparing screening markers [13].

As mentioned earlier in the discussion about evaluating covariate effects on binary markers, regression techniques are appropriate when Z is multidimensional or continuous. The same is true for continuous markers but now regression models for ROC curves must be employed. Some regression modeling methods for ROC curves have been described. A variety of illustrations are provided in Pepe 2003 (Chapter 6)[18]. This is a relatively new area of statistical methodology and methods are evolving rapidly. We refer to Cai and Pepe (2002) and Dodd and Pepe (2003) [22, 23] for more recent work.

THE INCREMENTAL VALUE OF A MARKER

Now we consider another way in which covariates are often considered. Suppose that there are some established markers (or predictors) for the outcome that we denote by X_1. In considering a new candidate marker, X_2, we want to assess how much classification is improved by using X_2 in addition to X_1 [25]. Alternatively, we can ask if there is predictive information in X_2 that cannot be explained by associations with X_1. For the purposes of illustration suppose that CA-19-9 is an established biomarker for pancreatic cancer and we seek to determine the additional contribution of CA-125 to classification accuracy. Here X_1 is CA-19-9 and X_2 is CA-125. One common approach is to treat X_1 and X_2 as covariates in a logistic regression model.

$$\text{logit}P(D = 1|X_1, X_2) = \alpha + \beta_1 X_1 + \beta_2 X_2 \tag{1}$$

and interpret $\exp(\beta_2)$ as the odds ratio for the strength of association between X_2 and the outcome, D, after "accounting for" the associations with X_1. We do not dispute this. However, as mentioned earlier, a measure of association is not a characterization of the accuracy of prediction. Using the

data displayed in Figure 4 we estimate $\exp(\beta_2) = 2.54$ ($p = 0.002$), a statistically significant association between CA-125 and pancreatic cancer after controlling for CA-19-9.

In Figure 7 we show the ROC curves for classifying subjects as having pancreatic cancer or not using CA-19-9 alone and using the combination of CA-19-9 and CA-125 predictors. Assuming that equation (1) fits the data reasonably well, it is known that the linear combination $\beta_1 X_1 + \beta_2 X_2$ is the best combination of the markers for discriminating cases from controls [24]. We see that CA-125 adds little to the capacity of CA-19-9 to discriminate between pancreatic cancer cases and controls. For example, if we are content to accept a 5 percent false positive fraction, we can detect 68 percent of cases using CA-19-9 alone and 71 percent, using the combination. The tangible benefit of adding the new marker CA-125 to the existing CA19-9 marker appears to be minimal for the purposes of classification. That is, the independent contribution of CA-125 to classification is negligible. This is despite the fact that it has a strong association with disease status that is independent of its association with CA-19-9.

In our experience this is a rather common phenomenon, that a marker displaying an independent association considered strong by traditional epidemiologic standards, does not contribute meaningfully to improved classification. Another illustration is provided in Kattan (2003)[25]. This is quite consistent with the observations made earlier in Section 2. *Extremely* strong associations are required for meaningful classification accuracy. Again the important message is that the statistical evaluation of markers for classification should use techniques that directly address classification accuracy (e.g., ROC curves), rather than traditional logistic regression techniques for assessing associations.

DISCUSSION

The work in this paper was stimulated by the observation that many studies of predictive/diagnostic markers continue to use statistical methods based on the odds ratio or relative risk. This is despite

the fact that such methods are not suited to the task of evaluating classification accuracy. Others have mentioned that the odds ratio does not quantify the classification accuracy of a marker including Kattan (2003), Boyko and Alderman (1990), Emir et al. (2000), and Baker et al. (2002) [25, 26, 27, 28]. We have presented more detailed discussion, demonstrating the pitfalls of using the odds ratio for evaluating markers, for evaluating covariate effects on marker performance and for evaluating the incremental value of a marker over existing predictors. In addition, we have suggested more appropriate techniques that can be used to address these questions statistically. References to the literature hopefully will facilitate more widespread adoption of proper methods in practice.

Although the odds ratio does not characterize a marker's accuracy for classifying individuals, its relationship to the relative risk has long made it valuable for characterizing population variations in risk. A binary marker with relative risk = 3, say, can be used to identify a population with the risk factor that has 3 times the risk as does the population without the risk factor. This may be used for targeting prevention or screening strategies. Moreover, clinical trials can often be conducted more efficiently in such populations. However, as we have noted, such a marker will be a very inaccurate tool for classifying or predicting risk in individual subjects. Markers that are proposed for classifying or predicting risk in individual subjects must be held to a much higher standard than merely being associated with outcome. Their sensitivities and specificities must be shown to be adequate through appropriate statistical evaluations.

ACKNOWLEDGMENTS

This research was supported by NIH grants I-1 CA86368, R01 GM54438 and P01 CA18029.

References

[1] Srivastava S, Kramer BS. Early detection cancer research network. Laboratory Invest 2000;80(8):1147–1148.

[2] Wilson P, D'Agostino R, Levy D, Belanger A, Silbershatz H, Kannel W. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837–1847.

[3] Gail MH, Brinton LA, Byar DP, Corle DK, Green SB, Schairer C, Mulvihill JJ. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81(24):1879–1886.

[4] Newcomb P, Trentham-Dietz A. Patterns of postmenopausal progestin use with estrogen in relation to endometrial cancer (United States) Cancer Causes and Controls 2003;14:195–201.

[5] Cui H, Cruz-Correa M, Giardiello FM, Hutcheon DF, Kafonek DR, Brandenburg S, Wu Y, He X, Powe NR, Feinberg AP. Loss of IGF2 Imprinting: A Potential Marker of Colorectal Cancer Risk. Science 2003;299:1753–1755.

[6] Rhodes DR, Sanda MG, Otte AP, Chinnaiyan AM, Rubin MA. Multiplex biomarker approach for determining risk of prostate-specific antigen-defined recurrence of prostate cancer. J Natl Cancer Inst 2003;95:661–669.

[7] Ridker PM, Hennekens CH, Buring JE, Rifai N. C-reactive protein and other markers of inflammation in the prediction of cardiovascular disease in women. New England J Med 2000;342:836–843.

[8] Zhang R, Brennan ML, Fu X, et al. Association between myeloperoxidase levels and risk of coronary artery disease. J Am Med Assoc 2001;286(17):2136–2142.

[9] Liou SH, Lung JC, Chen YH, et al. Increased chromosome-type chromosome aberration frequencies as biomarkers of cancer risk in a blackfoot endemic area. Cancer Res 1999;59(7):1481–1484.

[10] Hogue A, Lippman SM, Boiko IV, et al. Quantitative nuclear morphometry by image analysis for prediction of recurrence of ductal carcinoma in situ of the breast. Cancer Epidemiol Biomarkers Prev 2001;10(3):249–259.

[11] Lachenbruch PA. The odds ratio. Controlled Clin Trials 1997; 8(4):381–382.

[12] Boyko EJ. Ruling out or ruling in disease with the most sensitive or specific diagnostic test: Short cut or wrong turn? Medical Decision Making 1994;14:175–179.

[13] 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;95:511–515.

[14] Wieand S, Gail MH, James BR, James KL. A family of nonparametric Statistics for Comparing Diagnostic Markers with Paired or Unpaired Data. Biometrika 1989;76:585–592.

[15] Carney PA, Miglioretti DL, Yankaskas BC, et al. Individual and combined effects of age, breast density, and hormone replacement therapy use on the accuracy of screening mammography. Ann Intern Med 2003;138:168–175.

[16] Leisenring W, Pepe MS, Longton G. Regression Modeling of Diagnostic Likelihood Ratios for the Evaluation of Medical Diagnostic Tests. Biometrics 1998;16:1263–1281.

[17] Smith PJ, Hadgu A. Sensitivity and Specificity for Correlated Observations. Stat in Med 1992;11:1503–1509.

[18] Pepe MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, United Kingdom, 2003.

[19] Stata 8.0. 2003. Stata Corporation, College Station, Texas.

[20] Etzioni R, Pepe, MS, Longton G, Hu C, Goodman G. Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. Med Dec Making 1999;19:242–251.

[21] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44: 837–845.

[22] Cai T, Pepe MS. Semiparametric ROC Analysis to Evaluate Biomarkers for Disease. J Am Statist Assoc 2002;97:1099–1107.

[23] Dodd LE, Pepe MS. Semiparametric regression for the area under the receiver operating characteristic curve. J Am Statist Assoc 2003; 00:0000–0000.

[24] McIntosh M, Pepe MS. Combining several screening tests: Optimality of the risk score. Biometrics 2002;58:657–664.

[25] Kattan MW. Judging new markers by their ability to improve predictive accuracy. J Natl Cancer Inst 2003;95:634–635.

[26] Boyko EJ, Alderman BW. The use of risk factors in medical diagnosis: opportunities and cautions. J Clin Epidemiol 1990;43:851–858.

[27] Emir B, Wieand S, Jung SH, Ying Z. Comparison of diagnostic markers with repeated measurements: a nonparametric ROC curve approach. Statistics in Medicine 2000;19(4):511–523.

[28] Baker SG, Kramer BS, Srivastava S. Markers for early detection of cancer: Statistical guidelines for nested case-control studies. BMC Medical Research Methodology 2002;2(1):4.

Table 1: Odds ratios for each quartile relative to the first quartile corresponding to the pairs of continuous marker distributions shown in Figure 2. Quartiles are based on the marker distributions in controls. Also shown are the odds ratios per 1 unit increase in X that are displayed in Figure 2.

	Quartile			
OR per unit of X	1	2	3	4
1.5	reference	1.2	1.4	1.7
2	reference	1.4	1.7	2.4
3	reference	1.6	2.3	4.1
9	reference	2.6	5.2	17.4
16	reference	3.2	7.9	38.7
25	reference	3.8	10.8	73.7

Table 2: Data showing that a covariate (Z) can affect the performance of X as a marker but is not necessarily an effect modifier in the usual sense of having odds ratios that vary with Z. For illustration we show 10,000 subjects with $Z = 0$ and 10,000 with $Z = 1$. The common odds ratio is (47×97)/9=506.6 when $Z = 0$ and $Z = 1$.

	Covariate			
	$Z = 0$		$Z = 1$	
Marker	$D = 0$	$D = 1$	$D = 0$	$D = 1$
$X = 0$	846	3	873	6
$X = 1$	54	97	27	94
(FPF, TPF)	0.06	0.97	0.03	0.94

Figure Legends

FIGURE 1. The correspondence between the (TPF, FPF) true and false positive fractions, of a binary marker and the odds ratio, OR. Values of (TPF, FPF) that yield the same OR are connected.

FIGURE 2. Probability distributions of a marker, X, in cases (solid curve) and controls (dashed curve) that are consistent with the logistic model logit$P(D = 1|X) = \alpha + \beta X$. We assume that X has mean 0 and standard deviation $\frac{1}{2}$ in controls so that a unit increase represents the difference between the 84^{th} versus 16^{th} percentiles of X in controls. The marker is normally distributed with the same variance in cases. The odds ratio per unit increase in X is shown.

FIGURE 3. True versus false positive fractions associated with thresholding the continuous marker with the criteria $X > c$ for the 6 scenarios shown in Figure 2. Each curve corresponds to one scenario. Points on the curve correspond to different choices of threshold $c \, \epsilon \, (-\infty, \infty)$. Solid circles represent points associated with using each quartile as the threshold criterion.

FIGURE 4. Frequency distributions of two markers for pancreatic cancer. Source Wieand et al. (1989).

FIGURE 5. ROC curves for the markers shown in Figure 4.

FIGURE 6. Total PSA for 71 prostate cancer cases and 70 age matched controls in the Beta-Carotene and Retin-A (CARET) study. ROC curves for subjects < 65 years and ≥ 65 years are shown.

FIGURE 7. ROC curves for CA-19-9 and its combination score with CA-125. The combination score is

$$\beta_1 X_1 + \beta_2 X_2 = 1.03\mathsf{log(CA\text{-}19\text{-}9)} + 0.93\mathsf{log(CA\text{-}125)}.$$

Figure 1

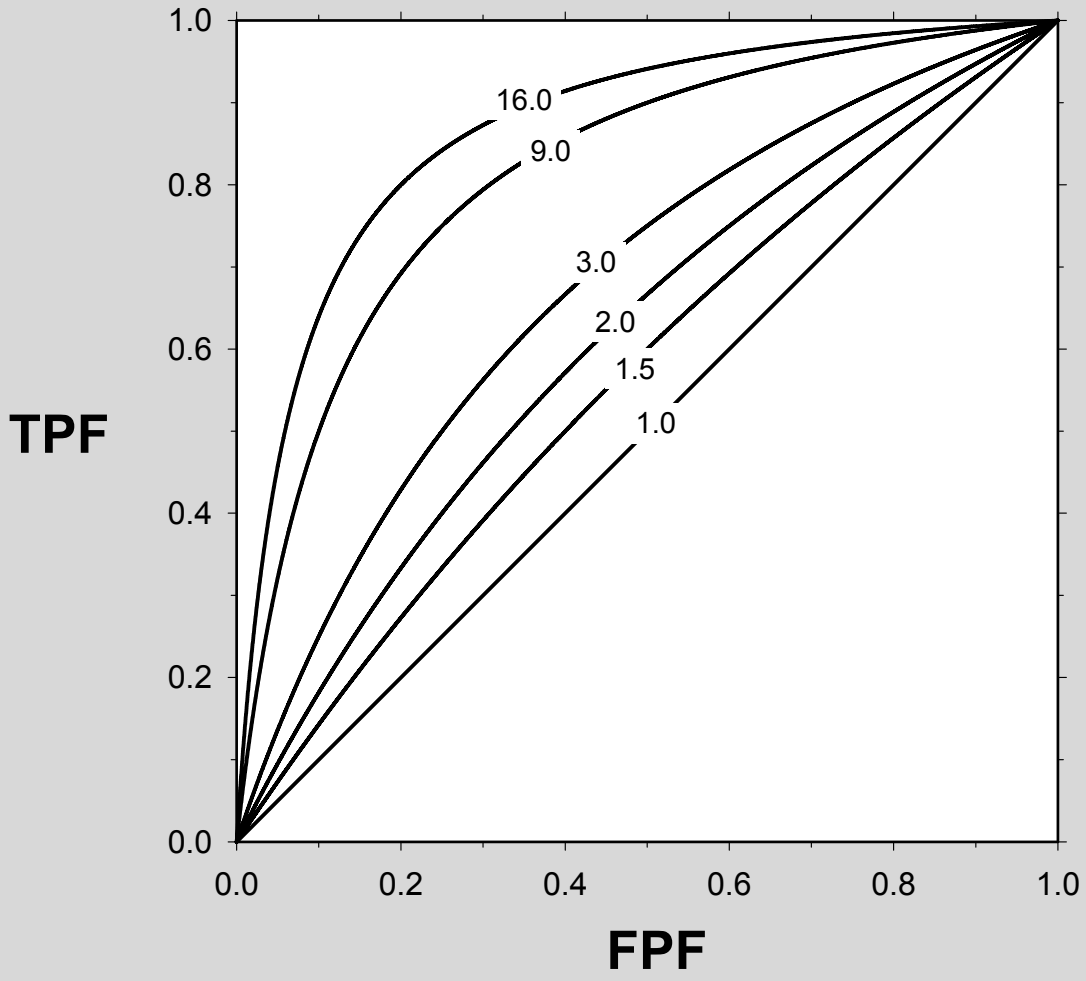


Figure 2

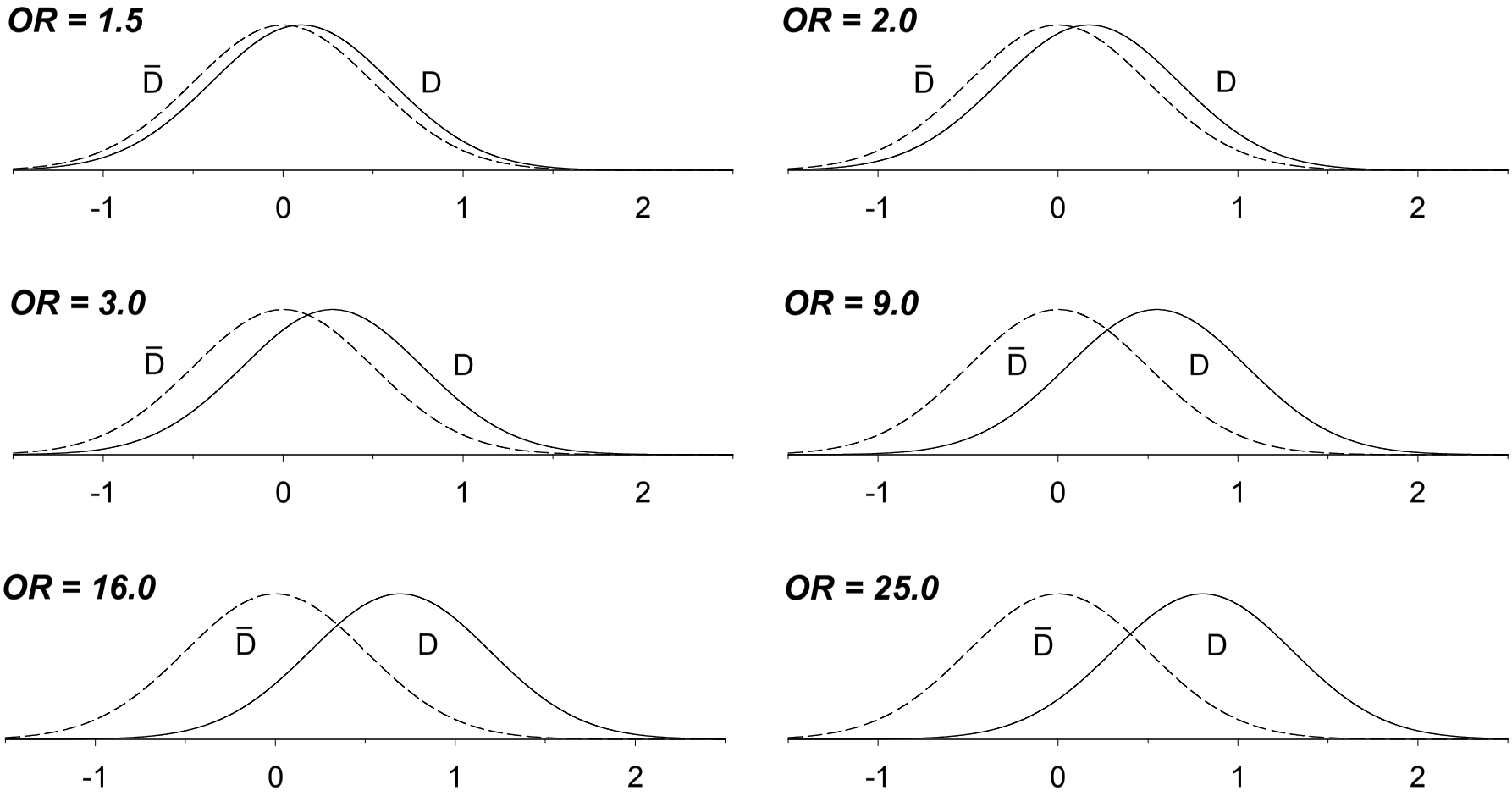


Figure 3

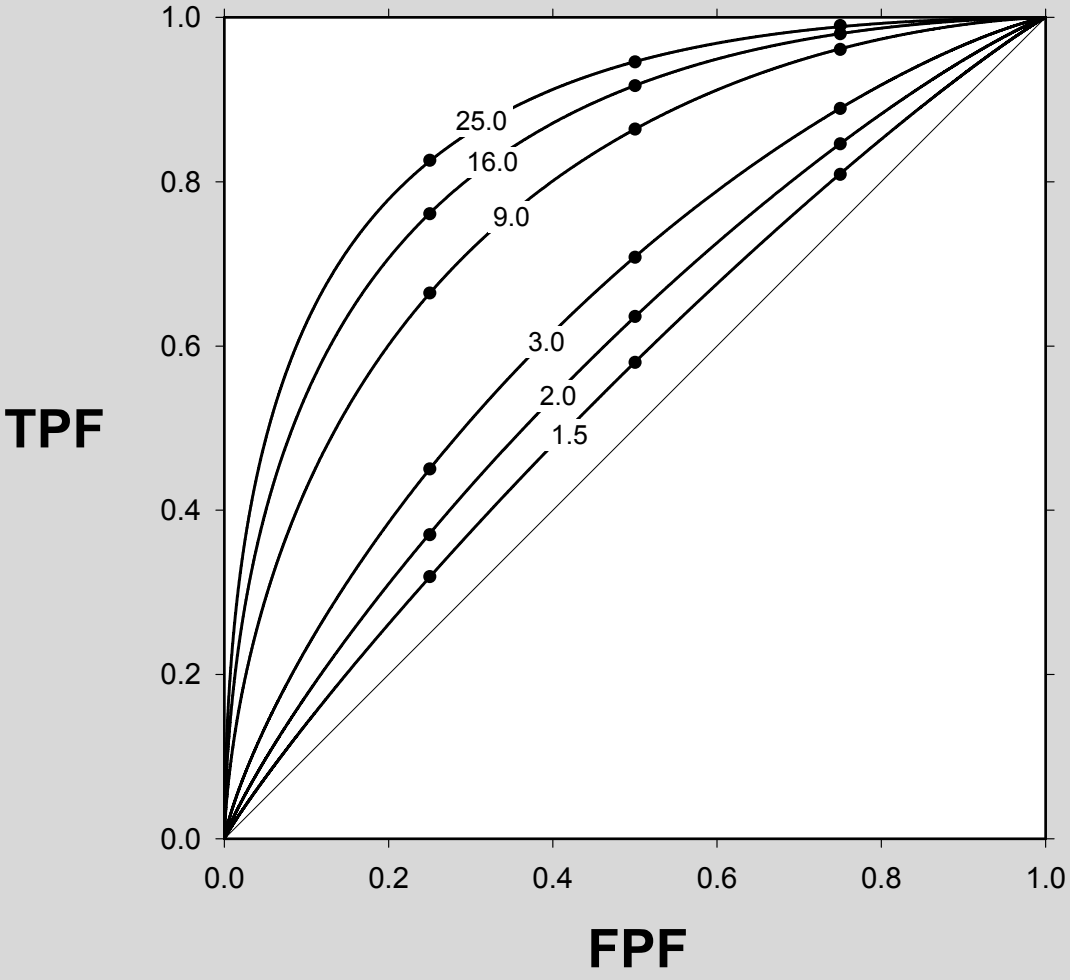


Figure 4

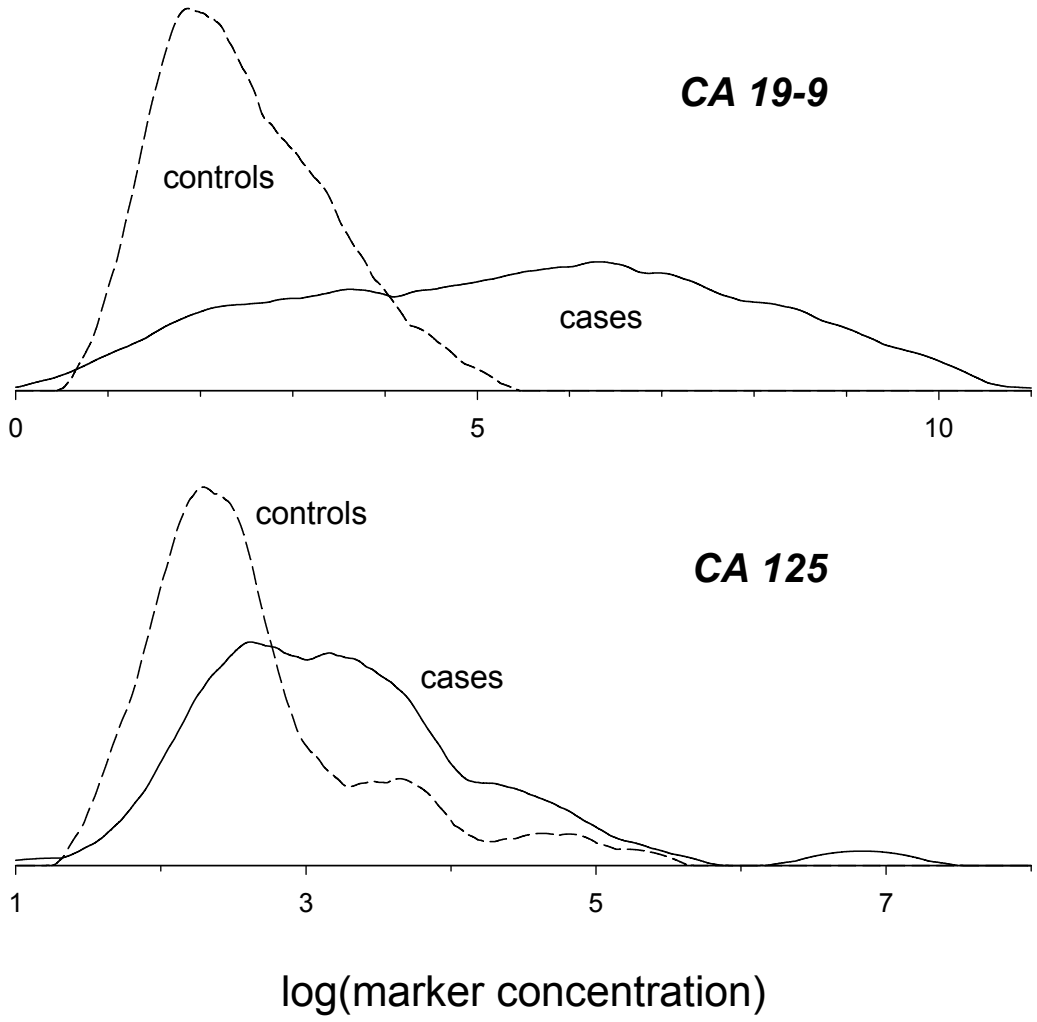


Figure 5

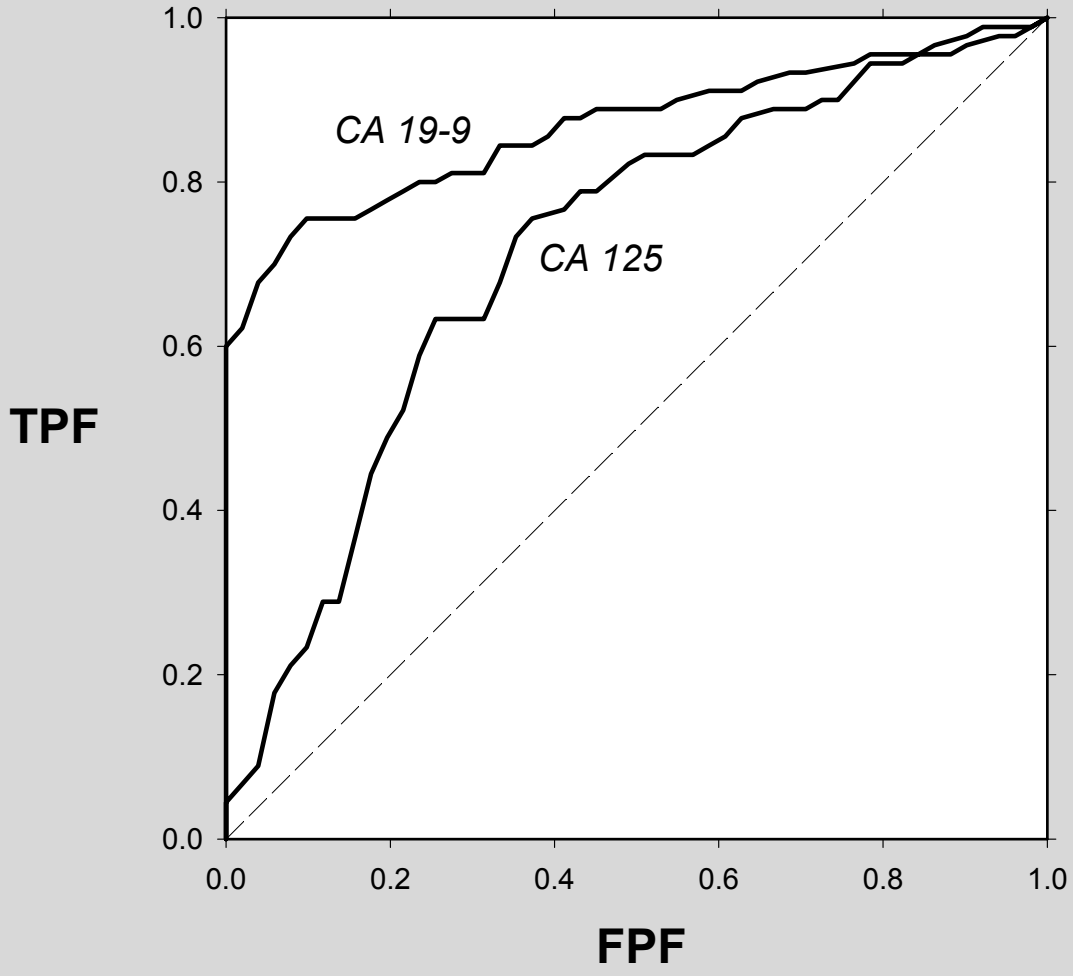


Figure 6

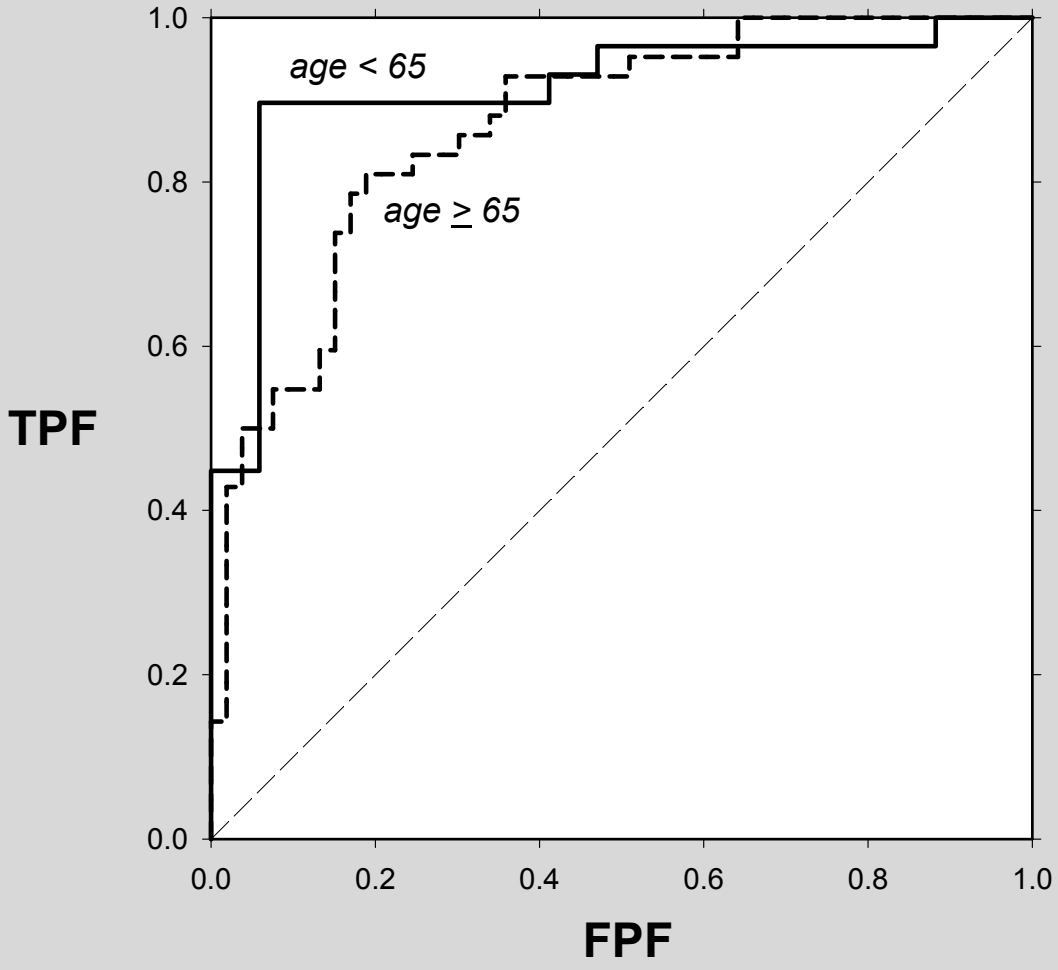
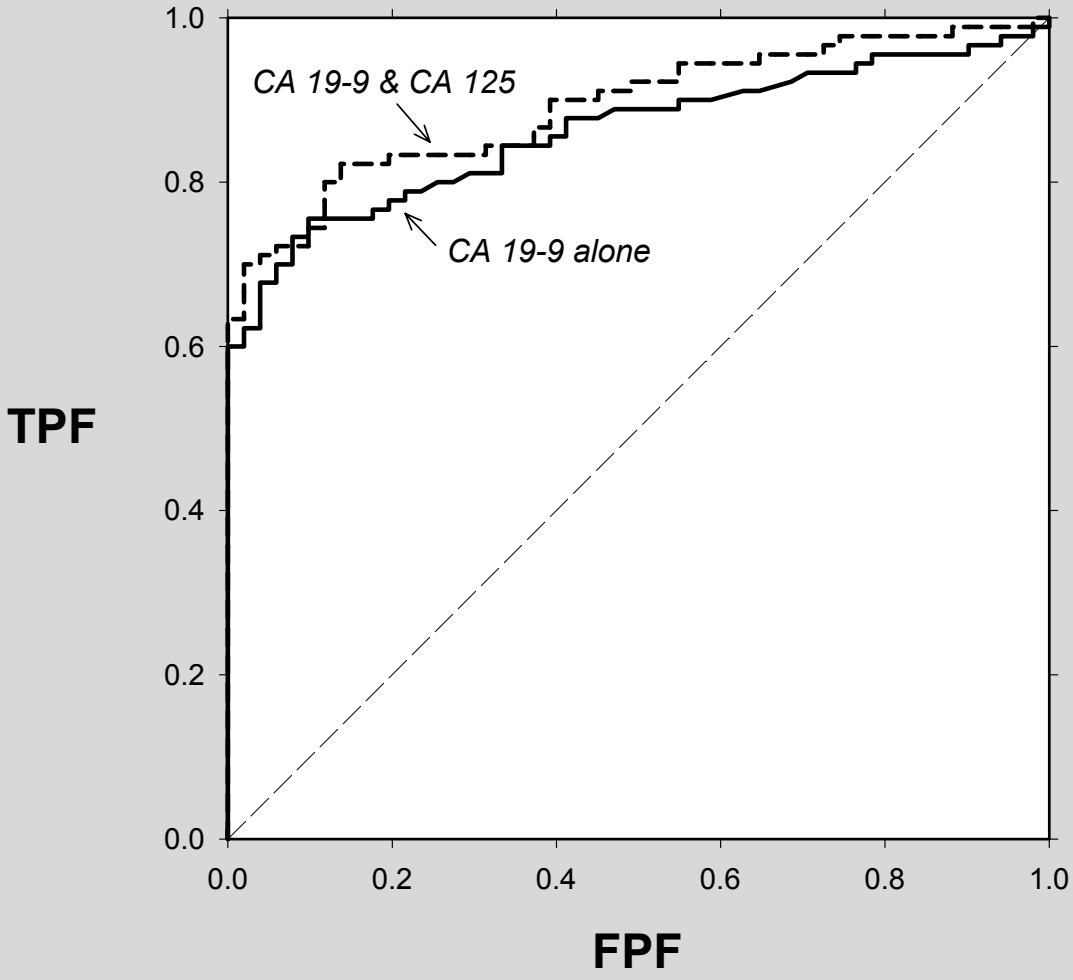


Figure 7


1.0
0.8
0.6
0.4
0.2
0.0
TPF
CA 19-9 & CA 125
CA 19-9 alone
0.0
0.2
0.4
0.6
0.8
1.0
FPF