American Journal of Epidemiology

Vol. 159, No. 9
Printed in U.S.A.
DOI: 10.1093/aje/kwh101

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

Margaret Sullivan Pepe[1,2], Holly Janes[2], Gary Longton[1], Wendy Leisenring[1,2,3], and Polly Newcomb[1]

[1] Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA.
[2] Department of Biostatistics, University of Washington, Seattle, WA.
[3] Division of Clinical Research, Fred Hutchinson Cancer Research Center, Seattle, WA.

Received for publication June 24, 2003; accepted for publication October 28, 2003.

A marker strongly associated with outcome (or disease) is often assumed to be effective for classifying persons according to their current or future outcome. However, for this assumption to be true, the associated odds ratio must be of a magnitude rarely seen in epidemiologic studies. In this paper, an illustration of the relation between odds ratios and receiver operating characteristic curves shows, for example, that a marker with an odds ratio of as high as 3 is in fact a very poor classification tool. If a marker identifies 10% of controls as positive (false positives) and has an odds ratio of 3, then it will correctly identify only 25% of cases as positive (true positives). The authors illustrate that a single measure of association such as an odds ratio does not meaningfully describe a marker's ability to classify subjects. Appropriate statistical methods for assessing and reporting the classification power of a marker are described. In addition, the serious pitfalls of using more traditional methods based on parameters in logistic regression models are illustrated.

biological markers; diagnostic test; logistic regression; odds ratio; ROC curve; screening test

Abbreviations: FPF, false-positive fraction; ROC, receiver operating characteristic; TPF, true-positive fraction.

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 biotechnological medical research. We hope to use biomarkers derived from new proteomic and genomic technologies to identify subjects who 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. 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 commonly, 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 associated with the clinical course of patients diagnosed with disease.

Statistical evaluation of factors, scores, and biomarkers for assessing a person'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 that strong statistical *associations* between outcome and marker do not necessarily imply that the marker can discriminate between persons likely to have the outcome and 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 persons (4–7). 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 persons. As we proceed to develop technologically sophisticated tools for individual-level prediction and classification, for so-called

Correspondence to Dr. Margaret Sullivan Pepe, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue, North, MZ-B500, Seattle, WA 98109 (e-mail: mspepe@u.washington.edu).

personalized medicine, we must be careful to use appropriate statistical techniques when 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 epidemiologic methods for this purpose. In addition, we discuss evaluation of a marker in the presence of other information predictive of outcome, which 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 a strong risk factor for endometrial cancer (8). A relative risk of about 3.0 is associated with it; that is, in a case-control study, the odds ratio comparing cases with controls regarding "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 common in studies of predictive markers as well. Refer, for example, to studies by Cui et al. (9) in a recent issue of *Science* and by Rhodes et al. (10) in a recent issue of *Journal of the National Cancer Institute*. Examples from other popular journals are Ridker et al. (11), Zhang et al. (12), Liou et al. (13), and Hogue et al. (14). Note that the goals of etiologic risk factor studies are quite different from those in the sorts of studies we consider here, where markers are to be used for classifying persons. Therefore, the statistical considerations also differ between such studies.

The accuracy or validity of a binary marker for classifying persons 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). These are defined as follows: TPF = Prob[marker positive | outcome positive] and FPF = Prob[marker positive | outcome negative]. Because there are two types of errors (misclassifying positives and negatives), the study results should reflect both of these errors. A perfect marker will have TPF = 1 and FPF = 0. Obviously, to have confidence in the prediction that a marker makes, TPF and FPF should be close to these ideal values. The general public often expects that a marker offer reasonably accurate classification and confident prediction.

However, a marker can be useful even if FPF and TPF are less than ideal. The criteria by which the marker is judged useful depend entirely on the context in which it is to be used. For example, a marker for screening a healthy population for cancer needs to have an extremely low FPF because workup procedures such as biopsy that follow a positive screening test are generally invasive and expensive. Given that cancer is a rare disease in the population tested, even a low FPF will result in huge numbers of people undergoing unnecessary, costly procedures. Using a utility function, Baker (15) argues that FPF needs to be below 2 percent when screening for prostate cancer and recommends that TPF exceed 50 percent. The considerations are different for a prognostic marker, that is, a marker measured in people with disease used to predict an aspect of their prognosis. For example, van de Vijver et al. (16) evaluated a gene-expression profile of tumor tissue in stage I or II breast cancer patients as a prognostic marker for distant metastases within 5 years. A prognostic marker of poor outcome should have high sensitivity, particularly if additional therapy will be instituted in only those patients who test positive. van de Vijver et al. estimated TPF to be equal to 92 percent for the gene-expression signature. Unfortunately, the FPF estimate was rather high at 42 percent. Whether it would be acceptable for 42 percent of good-prognosis patients to undergo unnecessary additional therapy would be a key factor in deciding on the usefulness of the marker.

The odds ratio (OR) can be written as a simple function of (FPF, TPF) (5, 17): $OR = \{TPF/(1 - TPF)\} \times \{(1 - FPF)/FPF\}$. Figure 1 shows this relation. 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 10 percent of controls (outcome negatives) as positive and that the associated odds ratio is 3.0. 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 detected by the marker. As another example, suppose that a marker with an odds ratio of 3.0 detects 80 percent of cases. The plot shows that it must mislabel almost 60 percent of the controls. Clearly, this marker is not useful for individual-level classification or prediction. The figure shows that even weakly accurate markers are associated with odds ratios (or relative risks) far larger than those traditionally considered strong in epidemiologic studies of association. For reasonable classification accuracy of, say, FPF = 0.10 and TPF = 0.80, the odds ratio is huge: 36.0. Note, however, that even if an odds ratio as large as 36.0 is reported, one cannot conclude that the marker has good accuracy since a variety of (FPF, TPF) values are consistent with it. For example, FPF may be unacceptably large; note that (FPF = 0.50, TPF = 0.973) also yields an odds ratio of 36.0.

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 following 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 16th and 84th percentiles of X in the controls (i.e., two standard deviations = one 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 normally

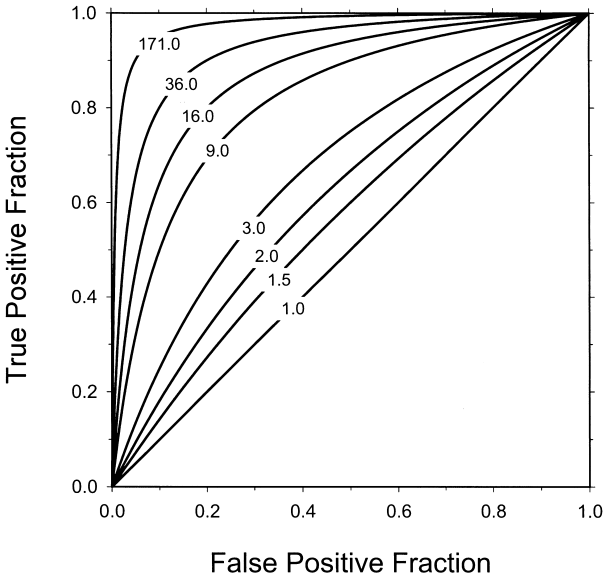

FIGURE 1. Correspondence between the true-positive fraction (TPF) and the false-positive fraction (FPF) of a binary marker and the odds ratio. Values of (TPF, FPF) that yield the same odds ratio are connected.

distributed with the same standard deviation, figure 2 shows their separation from controls for various values of the odds ratio. We see again that values of the odds ratio considered large in traditional epidemiologic studies are derived from marker distributions that largely overlap (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 are one point on the curve, and, by varying c from ∞ 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 many different markers represented on the same curve if their odds ratios are the same. The odds ratios shown here in figure 3 relate to a unit increase in a continuous marker, and the ROC curve concerns different decision criteria resulting from considering all possible thresholds for a single continuous marker.

We see from figure 3 that, when the odds ratio associated with a unit increase in 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 odds ratio per unit increase in X is at least 16.0, marker-based decision criteria seem very inaccurate. Even with an odds ratio of 16.0, a marker-based criterion that yields a 10 percent FPF at a threshold fails to detect over 40 percent of cases when that threshold is used. 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. For each of the marker distributions shown in figure 2, we also categorized the marker on the basis of the quartile cutpoints from the controls. The odds ratios for each quartile relative to the lowest quartile are shown in table 1. The solid-circle points on the ROC curves in figure 3 are those associated with using each of the three quartile cutpoints to classify subjects as positive or negative for disease. Similar to our observations for the binary and

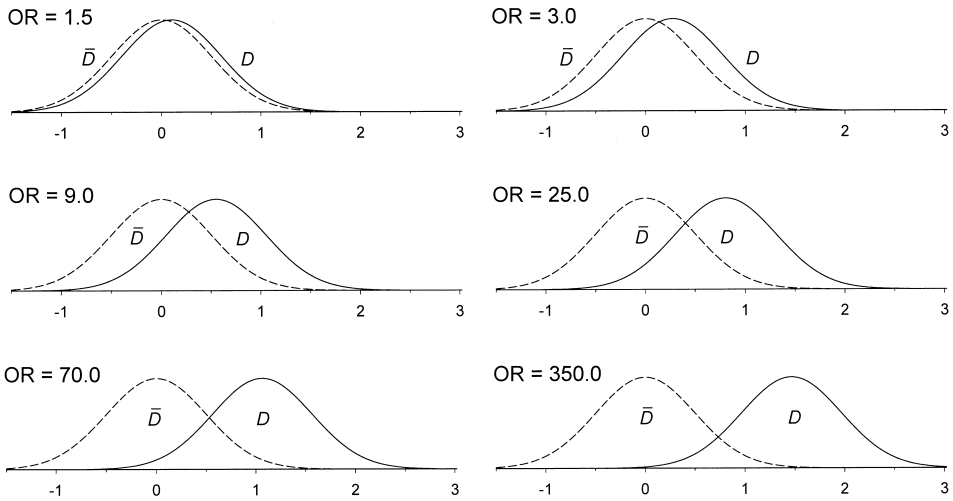

FIGURE 2. Probability distributions of a marker, X, in cases (solid curves) and controls (dashed curves) consistent with the logistic model $\text{logit}P(D = 1|X) = \alpha + \beta X$. It has been assumed that X has a mean of 0 and a standard deviation of 0.5 in controls so that a unit increase represents the difference between the 84th and 16th percentiles of X in controls. The marker is normally distributed, with the same variance in cases. The odds ratio (OR) per unit increase in X is shown.

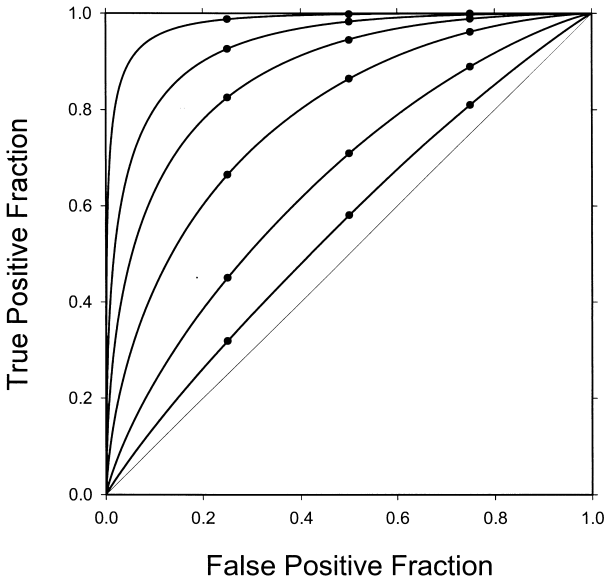

FIGURE 3. True- vs. false-positive fractions associated with dichotomous categorization of the continuous marker according to the decision criteria $X > c$ for the six scenarios shown in figure 2. Each curve corresponds to one scenario. Points on the curve correspond to different choices of threshold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

continuous marker settings, even if the magnitude of the odds ratio for an upper quartile versus the lowest quartile is large, the corresponding points on the ROC curves show a poor ability to classify cases and controls. For example, when the odds ratio for the upper quartile versus the lowest quartile is 4.1, and if we use the upper quartile to define a positive test result, then only 45 percent of cases are correctly classified whereas 25 percent of controls are incorrectly identified.

In summary, there are two key points to make based on the information in figures 1, 2, and 3. The first, as stated previously, is that markers for which the odds ratios are considered strong in traditional epidemiologic research are not adequate for discriminating between those persons 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 TPFs and FPFs are consistent with a particular odds ratio value. Neither does it relate to the notion of utility (15). In the next section, we discuss alternatives to the odds ratio that can be used to evaluate markers in case-control studies.

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 this paper. The TPFs and FPFs provide a description. Although predictive values are also of interest, where positive predictive value = P[D = 1 | marker positive] and negative predictive value = P[D = 0 | marker negative], these entities essentially require either a cohort study design in which 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 in which sampling depends on outcome. Diagnostic likelihood ratios have also been promoted as measures to characterize the accuracy of a binary marker (18) but are not very popular in practice.

It is interesting that characterization of accuracy requires two parameters, for example, TPF and FPF, whereas association measures such as the odds ratio or correlation coefficient (for continuous markers) are generally one dimensional. Characterization with (FPF, TPF) acknowl-

TABLE 1. Odds ratios for each quartile* relative to the first quartile corresponding to the pairs of continuous marker distributions shown in figure 2

Odds ratio per unit of X†	Quartile			
	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

* Quartiles are based on the marker distributions in controls.
† Odds ratio per one-unit increase in X shown in figure 2.

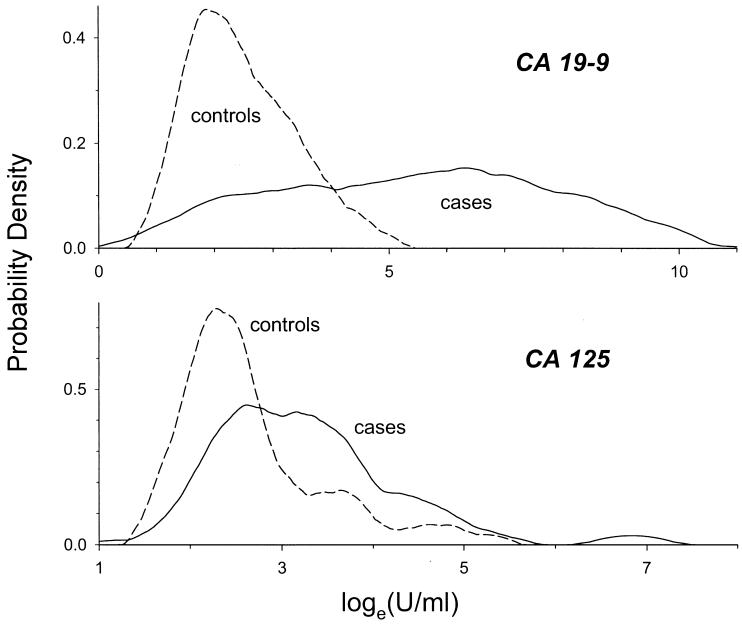

FIGURE 4. Frequency distributions of two markers for pancreatic cancer. Refer to Wieand et al. (20) for a description of source data.

edges that false-positive and false-negative errors are not equivalent and must be reported separately (19).

ROC curves

The ROC curve is the natural generalization of (FPF, TPF) to accommodate settings in which the marker is continuous. It describes the whole set of potential (FPF, TPF) combinations 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, respectively (20). 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 in contrast to logistic regression models in which, 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 serum concentration can be compared with one that measures spectral height at a given mass/charge ratio 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 (21). Factors other than those that are subject related are often important, too. The assay technique or the expertise of the lab technician can affect how well a biomarker performs. In audiology, the location in which the hearing test is conducted can affect the capacity of the test to detect hearing loss. If one can establish which covariates influence the performance of a marker, this

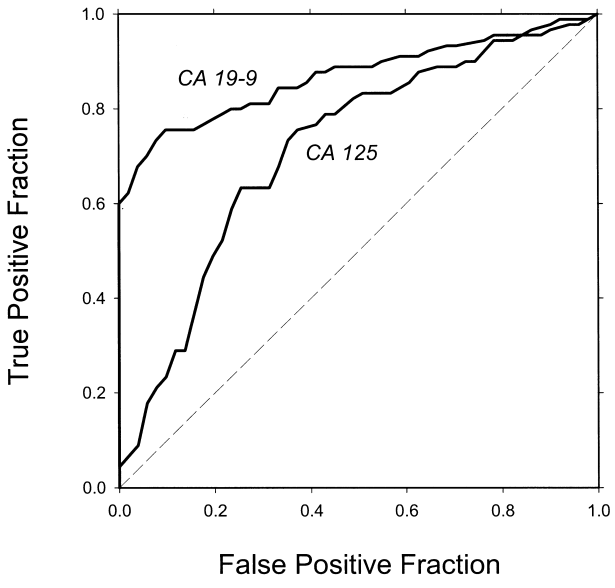

FIGURE 5. Receiver operating characteristic (ROC) curves for the pancreatic cancer markers shown in figure 4.

TABLE 2. Data* showing that a covariate (*Z*) can affect the performance of *X* as a marker (for disease *D*) but is not necessarily an effect modifier in the usual sense of having odds ratios that vary with *Z*

	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)	

* For illustration, data are shown for 1,000 subjects with $Z = 0$ and for 1,000 subjects with $Z = 1$. The common odds ratio is $(47 \times 97)/9 = 506.6$ when $Z = 0$ and $Z = 1$.

† FPF, false-positive fraction; TPF, true-positive fraction.

information may be used 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 the 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 whether "effect modification" occurs. 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 whether 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 an example, consider the data shown in table 2, where X and Z are binary. The marker X is being considered as a screening device, and, to make the discussion concrete, suppose that the covariate Z is gender ($Z = 1$ for females). The odds ratio associated with X is exactly the same for males and females. That is, Z is not an effect modifier in the sense that it alters the *association* between X and disease D when association is parameterized by the odds ratio. However, it seems that, at least for the purposes of disease screening, X performs better for females than for males. 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, and the lower FPF observed in females makes it a better screening marker in that population.

Statistical assessment

How then should one assess whether 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 already made this assessment informally for the data shown in table 2. Formal statistical techniques can be applied to test a hypothesis such as H_0 : FPF($Z = 1$) = FPF($Z = 0$) or to quantify the difference between FPF($Z = 1$) and FPF($Z = 0$). The FPFs and TPFs are binomial proportions, and the usual techniques of Pearson chi-square statistics and so forth can be applied. Inference about FPFs uses data for controls ($D = 0$) only; inference about TPFs uses data for cases ($D = 1$) only. For example, in table 2, comparison of FPFs yields $p = 0.002$ from a chi-square 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 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. Refer to Leisenring et al. (22), Smith and Hadgu (23), and Pepe (24, section 3.5) for illustrations.

For a continuous marker, one needs to determine whether the ROC curves for X vary with Z. If Z is dichotomous, one can plot separate ROC curves for X by 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 such as Stata (25). Data for prostate-specific antigen reported by Etzioni et al. (26) are shown in figure 6 for men less than 65 years of age and men 65 years of age or older. Although the study measured prostate-specific antigen repeatedly over time, we use only those 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 prostate-specific antigen 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 (20) 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 (27). Methods based on comparing ROC curves over a relevant subinterval of FPFs are described by Pepe (24, p. 110) and are probably more appropriate for comparing screening markers (19).

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 used. Some regression modeling methods for ROC curves have been described. A variety of illustrations are provided by Pepe (24, chapter 6). This area of statistical methodology is relatively new, and

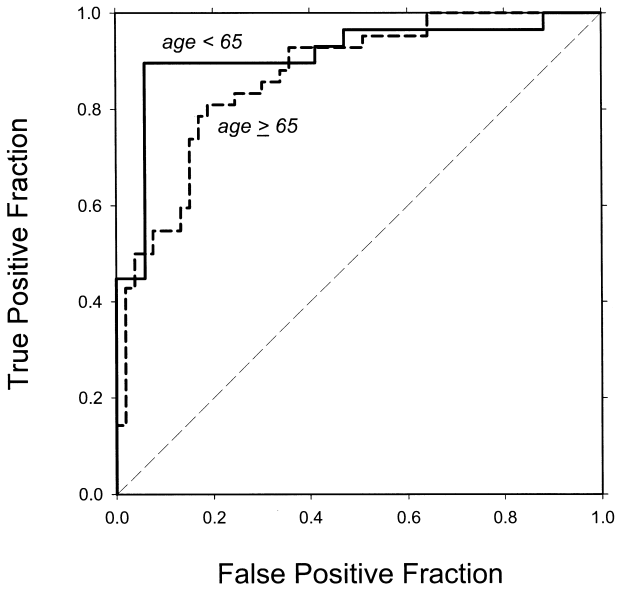

FIGURE 6. Total prostate-specific antigen for 71 prostate cancer cases and 70 age-matched controls in the Beta-Carotene and Retin-A (CARET) study, Fred Hutchinson Cancer Research Center, Seattle, Washington (coordinating center), 1985–1996. Receiver operating characteristic (ROC) curves for subjects aged <65 years and ≥65 years are shown.

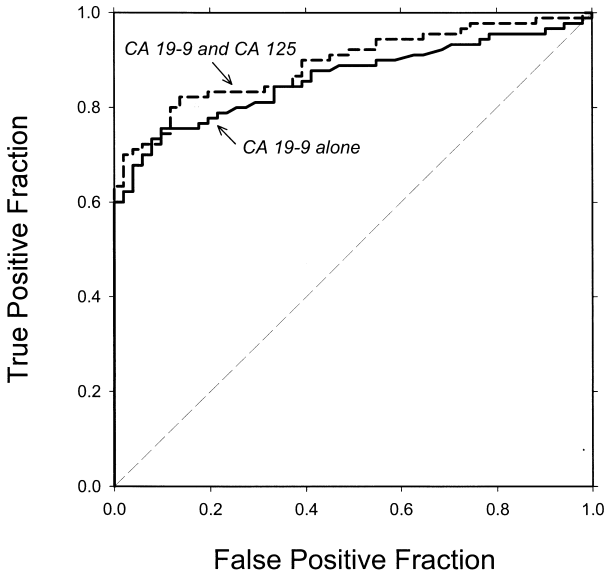

FIGURE 7. Receiver operating characteristic (ROC) curves for classifying subjects as having or not having pancreatic cancer by using the pancreatic cancer marker CA-19-9 alone and by using the combination of predictors CA-19-9 and CA-125. The combination score is $\beta_1 X_1 + \beta_2 X_2 = 1.03\log(\text{CA-19-9}) + 0.93\log(\text{CA-125})$.

methods are evolving rapidly. Refer to Cai and Pepe (28) and Dodd and Pepe (29) for more recent work.

THE INCREMENTAL VALUE OF A MARKER

Now we discuss 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 (4). Alternatively, we can ask whether 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 want 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,

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

and interpret $\exp(\beta_2)$ as the odds ratio for the strength of the association between X_2 and the outcome, D, after "accounting for" the associations with X_1. We do not dispute this approach. 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 that $\exp(\beta_2) = 2.54$ ($p = 0.002$), a statistically significant association between CA-125 and pancreatic cancer after controlling for CA-19-9.

Figure 7 shows the ROC curves for classifying subjects as having or not having pancreatic cancer by using CA-19-9 alone and by using the combination of predictors CA-19-9 and CA-125. If we assume 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 way to combine the markers for discriminating cases from controls (30). 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 FPF, we can detect 68 percent of cases by using CA-19-9 alone and 71 percent by using the combination. The tangible benefit of adding the new marker CA-125 to the existing marker CA-19-9 appears to be minimal for the purposes of classification. That is, the independent contribution of CA-125 to classification is negligible despite its strong association with disease status that is independent of its association with CA-19-9.

In our experience, this phenomenon—that a marker displaying an independent association considered strong by traditional epidemiologic standards does not contribute meaningfully to improved classification—is rather common. Another illustration is provided by Kattan (4). This finding is quite consistent with the observations made earlier. *Extremely* strong associations are required for meaningful classification accuracy. Again, the important message is that, for statistical evaluation of markers for classification, techniques should be used that directly address classification accuracy (e.g., ROC curves) rather than traditional logistic regression techniques for assessing associations.

DISCUSSION

The work for 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, 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 (4–7). We have presented a more detailed discussion, demonstrating the pitfalls of using the odds ratio to evaluate markers, to evaluate covariate effects on marker performance, and to evaluate 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 persons, its association with the relative risk has long made it valuable for characterizing population variations in risk. A binary marker with a relative risk of 3, say, can be used to identify a population with the risk factor that has three times the risk as the population without the risk factor. This method may be used to target 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 for individual subjects. Markers 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 US National Institutes of Health grants U01 CA86368, R01 GM54438, and P01 CA18029.

REFERENCES

1. Srivastava S, Kramer BS. Early detection cancer research network. Lab Invest 2000;80:1147–8.
2. Wilson P, D'Agostino R, Levy D, et al. Prediction of coronary heart disease using risk factor categories. Circulation 1998; 97:1837–47.
3. Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst 1989;81:1879–86.
4. Kattan MW. Judging new markers by their ability to improve predictive accuracy. J Natl Cancer Inst 2003;95:634–5.
5. Boyko EJ, Alderman BW. The use of risk factors in medical diagnosis: opportunities and cautions. J Clin Epidemiol 1990; 43:851–8.
6. Emir B, Wieand S, Su JQ, et al. Analysis of repeated markers used to predict progression of cancer. Stat Med 1998;17: 2563–78.
7. Baker SG, Kramer BS, Srivastava S. Markers for early detection of cancer: statistical guidelines for nested case-control studies. BMC Med Res Methodol 2002;2:4.
8. Newcomb P, Trentham-Dietz A. Patterns of postmenopausal progestin use with estrogen in relation to endometrial cancer (United States). Cancer Causes Controls 2003;14:195–201.
9. Cui H, Cruz-Correa M, Giardiello FM, et al. Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. Science 2003;299:1753–5.
10. Rhodes DR, Sanda MG, Otte AP, et al. Multiplex biomarker approach for determining risk of prostate-specific antigen-defined recurrence of prostate cancer. J Natl Cancer Inst 2003; 95:661–9.
11. Ridker PM, Hennekens CH, Buring JE, et al. C-reactive protein and other markers of inflammation in the prediction of cardiovascular disease in women. N Engl J Med 2000;342: 836–43.
12. Zhang R, Brennan ML, Fu X, et al. Association between myeloperoxidase levels and risk of coronary artery disease. JAMA 2001;286:2136–42.
13. 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: 1481–4.
14. 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;10:249–59.
15. Baker SG. Identifying combinations of cancer markers for further study as triggers of early intervention. Biometrics 2000;56:1082–7.
16. van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999–2009.
17. Lachenbruch PA. The odds ratio. Control Clin Trials 1997;8: 381–2.
18. Boyko EJ. Ruling out or ruling in disease with the most sensitive or specific diagnostic test: short cut or wrong turn? Med Decis Making 1994;14:175–9.
19. 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;95:511–15.
20. Wieand S, Gail MH, James BR, et al. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika 1989;76:585–92.
21. 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;138:168–75.
22. Leisenring W, Pepe MS, Longton G. Regression modeling of diagnostic likelihood ratios for the evaluation of medical diagnostic tests. Biometrics 1998;16:1263–81.
23. Smith PJ, Hadgu A. Sensitivity and specificity for correlated observations. Stat Med 1992;11:1503–9.
24. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Oxford, United Kingdom: Oxford University Press, 2003.
25. Stata Corporation. Stata 8.0 software. College Station, TX: Stata Corporation, 2003.
26. Etzioni R, Pepe, MS, Longton G, et al. Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. Med Decis Making 1999;19:242–51.
27. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing

the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–45.
28. Cai T, Pepe MS. Semiparametric ROC analysis to evaluate biomarkers for disease. J Am Stat Assoc 2002;97:1099–107.
29. Dodd LE, Pepe MS. Semiparametric regression for the area under the receiver operating characteristic curve. J Am Stat Assoc 2003;98:409–17.
30. McIntosh M, Pepe MS. Combining several screening tests: optimality of the risk score. Biometrics 2002;58:657–64.