nal of the American Statistical Association, 93, 1282–1293.

Keiding, N., Begtrup, K., Scheike, T. H., and Hasibeder, G. (1996), "Estimation from Current-Status Data in Continuous Time," *Lifetime Data Analysis*, 2, 119–129.

Kodell, R. L., and Nelson, C. J. (1980), "An Illness-Death Model for the Study of the Carcinogenic Process Using Survival/Sacrifice Data," *Biometrics*, 36, 267–277. Corr: 37, 875.

Kooperberg, C., and Stone, C. J. (1992), "Logspline Density Estimation for Censored Data," *Journal of Computational and Graphical Statistics*, 1, 301–328.

Krewski, D., and Brown, C. (1981), "Carcinogenic Risk Assessment: A Guide to the Literature," *Biometrics*, 37, 353–366.

Kupper, L. L., Portier, C., Hogan, M. D., and Yamamoto, E. (1986), "The Impact of Litter Effects on Dose-Response Modeling in Teratology (C/R: V44 p 305–309)." *Biometrics*, 42, 85–98.

Lefkopoulou, M., Moore, D., and Ryan, L. (1989), "The Analysis of Multiple Correlated Binary Outcomes: Application to Rodent Teratology Experiments," *Journal of the American Statistical Association*, 84, 810–815.

Liang, K. Y., and Zeger, S. L. (1986), "Longitudinal Data Analysis Using Generalized Linear Models," *Biometrika*, 73, 13–22.

Lindsey, J. C., and Ryan, L. M. (1993), "A Three-State Multiplicative Model for Rodent Tumorigenicity Experiments," *Applied Statistics*, 42, 283–300.

Liu, C. and Rubin, D. R. (1998), "Ellipsoidally Symmetric Extensions of the General Location Model for Mixed Categorical and Continuous Data," *Biometrika*, 85, 673–688.

Lu, F. C. (1996), "Basic Toxicology," *Taylor and Francis.*

McKnight, B., and Crowley, J. (1984), "Tests for Differences in Tumor Incidence Based on Animal Carcinogenesis Experiments," *Journal of the American Statistical Association*, 79, 639–648.

Moolgavkar, S. H., Leubeck, E. G., de Gunst, M., Port, R. E., and Schwetz, M. (1990), "Quantitative Analysis of Enzyme-Altered Foci in Rat Hepatocarcinogenesis Experiments 1: Single Agent Regimen," *Carcinogenesis*, 11, 1271–1278.

Peto, R. (1974), "Guidelines on the Analysis of Tumor Rates and Death Rates in Experimental Animals," *British Journal of Cancer*, 29, 101–105.

Piegorsch, W. W., and Bailer, A. J. (1997), *Statistics for Environmental Biology and Toxicology*, Boca Raton, FL: Chapman & Hall.

Portier, C. J., and Bailer, A. J. (1989), "Testing for Increased Carcinogenicity Using a Survival-Adjusted Quantal Response Test," *Fundamental and Applied Toxicology*, 12, 731–737.

Regan, M., and Catalano, P. J. (1999), "Likelihood Models for Clustered Binary and Continuous Outcomes: Application to Developmental Toxicology," *Biometrics*, 55, 760–768.

Seilkop, S. K. (1995), "The Effect of Body Weight on Tumor Incidence and Carcinogenicity Testing in B6C3F1 Mice and F344 Rats," *Fundamental and Applied Toxicology*, 24, 247–259.

United States Environmental Protection Agency (1995), "Proposed Guidelines for Neurotoxicity Testing," *Federal Register*, 60, 52032–52036.

Waterman, M. (1998), *Introduction to Computational Biology: Maps, Sequences, and Genomes*, Chapman & Hall.

Westfall, P. H., and Young, S. S. (1993), "On Adjusting p-values for Multiplicity (Disc: p944–945)," *Biometrics*, 49, 941–944.

Williams, D. A. (1975), "The Analysis of Binary Responses from Toxicological Experiments Involving Reproduction and Teratogenicity," *Biometrics*, 31, 949–952.

——— (1982), "Extra-Binomial Variation in Logistic Linear Models," *Applied Statistics*, 31, 144–148.

World Health Organization (1962), *Sixth Report of the Joint FAD/WHO Expert Committee on Food Additives*, Geneva: WHO.

Receiver Operating Characteristic Methodology

Margaret Sullivan PEPE

1. INTRODUCTION

Diagnostic medicine has progressed tremendously in the last several decades, and the trend promises to continue well into the next millennium. Advances in technology provide new methods for detecting disease or physical impairment. Some examples include the use of biochemical serum markers such as prostate-specific antigen for prostate cancer and CA-125 for ovarian cancer, of radiographic imaging procedures such as mammography for breast cancer, and of electrophysical procedures such as brain stem response testing for hearing impairment. Research studies to assess the operating characteristics of diagnostic tests are clearly important to ensure that accurate and cost-effective procedures are selected for widespread use. Development of appropriate statistical methods for designing such studies and for analyzing data from them will be key to their success.

A statistical tool that is becoming popular for describing diagnostic accuracy is the receiver operating characteristic (ROC) curve. To define an ROC curve, first consider diagnostic tests with dichotomous outcomes, with positive outcomes suggesting presence of disease. For dichotomous tests, there are two potential types of error. A false-positive error occurs when a nondiseased individual has a positive test result, and conversely, a false-negative error occurs when a diseased individual has a negative test result. The rates with which these errors occur, termed the false-positive and false-negative rates, together constitute the operating characteristics of the dichotomous diagnostic test. Statisticians are already familiar with these concepts in the context of statistical hypothesis testing. ROC curves generalize these notions to nonbinary tests in the following fashion: Let D be a binary indicator of true disease status with $D = 1$ for diseased subjects. Let X denote the test result with the convention that larger values of X are more indicative of disease. For any chosen threshold value c, one can define a dichotomous test by the positivity criterion $X \geq c$, and calculate the associated error rates. A plot of 1 minus the false-negative rate (or true positive rate) versus the false-positive rate for all possible choices of c is the ROC curve for X. By definition, this is a monotone in-

Margaret Sullivan Pepe is Member, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, (E-mail: *mspepe@u.washington.edu*).

Journal of the American Statistical Association
March 2000, Vol. 95, No. 449, Vignettes

creasing function from [0, 1] to [0, 1], with higher curves associated with better tests.

The ROC curve is primarily a descriptive device displaying the range of trade-offs between true-positive and false-positive rates possible with the test. It transforms the test results to a scale that pertains to accuracy for detecting disease. This is particularly useful for comparing tests with numerical results that are on different measurement scales and for which no meaningful comparisons can be based on the raw data. Mathematically, the ROC curve can be written as $\mathrm{ROC}(t) = F_D(F_{\bar{D}}^{-1}(t))$ for $t \in (0, 1)$, where F_D and $F_{\bar{D}}$ denote the "survivor" functions for X in the diseased and nondiseased populations. It follows that the ROC curve is invariant to monotone-increasing transformations of the data.

2. ORDINAL RATING DATA

Although ROC analysis had its roots in the electronic signal detection theory developed in the 1950s (Green and Swets 1966), it was not until the early 1980s that it started to be used in biomedical applications. It became especially popular in radiology for characterizing the accuracy of diagnostic imaging modalities. Indeed, most of the statistical methodologic work on ROC analysis has been done in this context (see Hanley 1998 for a review). Why did it fit particularly well with the needs in radiology? Because image assessments are made subjectively by radiologists, and implicit criteria for assessments vary among radiologists. To see this, consider the classic setting where each reader assesses an image on an ordinal scale, $k = 1, \ldots, K$, with the lowest category labeled perhaps as "definitely no disease present" to the highest labeled as "disease definitely present." It is assumed (Metz 1986) that with each reading, Y, there is a continuous latent decision variable, X, and that the reader classifies the image in category k if the decision variable falls within the interval $(\tau_{k-1}, \tau_k), k = 1, \ldots, K$ with $\tau_0 = -\infty$ and $\tau_K = +\infty$. That is, $Y = k$ if $\tau_{k-1} < X < \tau_k$. Although one cannot know the value of X, K points on its ROC curve are identifiable, and the whole curve is identifiable under parametric modeling assumptions. Points for readers using different decision criteria (i.e., different threshold values $\{\tau_1, \ldots, \tau_{K-1}\}$) will simply fall at different locations on the ROC curve for X. In this way, ROC analysis accommodates variation among readers in their decision criteria and purports to disentangle this variation from the inherent discriminatory capacity of the diagnostic test.

The classical approach for estimating an ROC curve in this context is to assume the binormal model; that is, that some monotone-increasing transformation of X has a standard normal distribution in the nondiseased population and a normal distribution with (mean, SD) = (a, b^{-1}) in the diseased population. The ROC curve for X then has the form $\mathrm{ROC}(t) = \Phi(a + b\Phi^{-1}(t))$ for $t \in (0, 1)$. Using readings of images from a set of diseased subjects, $\{Y_1^D, \ldots, Y_{n_D}^D\}$, and from a set of nondiseased subjects, $\{Y_1^{\bar{D}}, \ldots, Y_{n_{\bar{D}}}^{\bar{D}}\}$, the parameters (a, b) are estimated via maximum likelihood (Dorfman and Alf 1969). If different readers or different imaging modalities give rise to different ROC curves, then the data are stratified accordingly and stratum-specific parameters (a_s, b_s) are estimated. As described later, classically comparisons between curves are based on the area under the ROC curve, which for the binomial curve is given by $\Phi(a/(1 + b^2)^{1/2})$.

Tosteson and Begg (1988) proposed that ordinal regression modeling methods could be applied to radiology rating data to make inference about ROC curves. This opened up avenues for much more sophisticated and flexible ROC analysis than had been available previously. In the simplest setting, a location ordinal model is postulated,

$$P[X < \tau_k] = P[Y \leq k] = g(C_k - \alpha_1 D - \alpha_2 \mathbf{Z} - \alpha_3 \mathbf{Z} \cdot D),$$

where $\mathbf{Z}$ is a vector of covariates and g is a cumulative distribution function. The ROC curve corresponding to X conditional on $\mathbf{Z}$ is then $\mathrm{ROC}_{\mathbf{Z}}(t) = 1 - g(g^{-1}(1 - t) - \alpha_1 - \alpha_3 \mathbf{Z})$, which reduces to $\Phi(\Phi^{-1}(t) + \alpha_1 + \alpha_3 \mathbf{Z})$, for example, when g is the standard normal cdf. Thus covariates that have no interaction with disease status do not affect the ROC curve, but are only associated with shifting the operating points on the curve.

The ability to explore effects of covariates on test accuracy was a huge step forward in the analysis of radiology rating data. Important types of covariates include characteristics of subjects from which images are taken, characteristics of procedures used to process images, and characteristics of the readers such as experience or institutional affiliation. Other indicators of disease status such as clinical signs or symptoms, or indeed the results of another diagnostic test, could also be considered as covariates in the model. This strategy allows one to assess the incremental value of the imaging test over and above this information.

The general model proposed by Tosteson and Begg included the possibility for covariates to affect scale parameters as well and can be written as

$$P[Y \leq k] \\ = g(\{C_k - \alpha_1 D - \alpha_2 \mathbf{Z} - \alpha_3 D \cdot \mathbf{Z}\} / \exp\{\beta_1 D + \beta_2 \mathbf{Z} + \beta_3 D \cdot \mathbf{Z}\}).$$

Synthesis of covariate effects on the ROC curve is not as straightforward as in the location-only model. However, it does allow for estimation of covariate adjusted ROC curves. Note that the classic binormal model of Dorfman and Alf (1969) is a special case of the Tosteson and Begg model, when a probit link function is used and no covariates are included.

By embedding ROC methodology in the framework of ordinal regression, issues of correlated data that had previously been difficult to deal with now became much easier because of the growth in methods for handling clustered data that also occurred in the 1980s. Toledano and Gatsonis (1995) chose to model marginal probabilities and account for correlation with GEE, whereas Gatsonis (1995) suggested incorporating random reader effects into the ordinal regression framework. An advantage of the latter approach is that it allows for assessment of inter-rater variability in decision criteria and in accuracy, which is of considerable interest in diagnostic medicine (Beam, Layde, and Sullivan 1996).

Studies of diagnostic or screening tests often suffer from so-called verification bias. This occurs when subjects with certain test results, X, or other characteristics, $\mathbf{Z}$, indicative of disease are more likely to be assessed for disease, D, with the gold standard than are other subjects. This sort of selection makes practical clinical sense when the gold standard is invasive or costly, such as requiring surgery for cancer detection. However, it introduces bias into estimates of test accuracy unless adjustments are made in the analysis. Although some attempts at ROC adjustments had been made previously, embedding ROC analysis into the ordinal regression framework again provided access to a rich variety of missing-data techniques that were already available for generalized linear regression analysis. These include inverse probability weighting and EM-based maximum likelihood methods.

3. CONTINUOUS DATA

Concepts of ROC have more recently gained popularity in biomedical applications involving tests with results on non-ordinal scales. Many biochemical measurements—for example, including serum antigen or enzyme concentrations (Zweig and Campbell 1993)—are continuous in nature. At first glance, one might expect that ROC analysis for continuous data would be more straightforward than it is for ordinal data, because in this setting the decision variable itself is available rather than just a categorized version. However, new challenges present themselves with continuous data.

Consider, for example, estimation of the ROC curve from data $\{Y_i^D, i = 1, \ldots, n_D, : Y_j^{\bar{D}}, j = 1, \ldots, n_{\bar{D}}\}$, where Y_i^D denotes the test result for the ith diseased study unit and $Y_j^{\bar{D}}$ that for the jth nondiseased unit. The curve $\widehat{\mathrm{ROC}}(t) = \hat{F}_D(\hat{F}_{\bar{D}}^{-1}(t))$, where $\hat{F}_D$ and $\hat{F}_{\bar{D}}$ are empirical estimators of F_D and $F_{\bar{D}}$ based on $\{Y_i^D, i = 1, \ldots, n_D\}$ and $\{Y_j^{\bar{D}}, j = 1, \ldots, n_{\bar{D}}\}$, is a jagged curve. Curves that incorporate reasonable smoothness assumptions can be based on parametric or smoothed estimators of F_D and $F_{\bar{D}}$. However, these estimated ROC curves do not enjoy a fundamental property of ROC curves—namely, invariance to monotone-increasing transformations of the data. Metz, Herman, and Shen (1998) noted that classical smooth (parametric) ROC curve estimators for ordinal data, on the other hand, do have this property. Indeed, they proposed that smooth ROC curve estimators for continuous data be based on application of ordinal data methods to categorized versions of the continuous data.

The fact that decision criteria are explicit with continuous data rather than implicit also impacts on analysis. One has the capacity in practice to specify the decision criterion and thus to control the false-positive rate. Because higher ranges of false-positive rates may be of no interest for future application of the test, it may be appropriate to focus the analysis on a restricted range, $\{\mathrm{ROC}(t), t \in [0, t_0]\}$, for some $t_0 < 1$. Wieand, Gail, James, and James (1989) proposed test statistics for comparing ROC curves over restricted intervals.

For regression modeling of covariate effects on ROC curves, the analog of Tosteson and Begg's approach is to model the test outcome with location and scale components that are functions of disease status and covariates. This approach again suffers from lack of invariance to monotone data transformations. Pepe (1997) proposed a fundamentally different approach to regression. She suggested that instead of modeling the test result and indirectly ascertaining induced covariate effects on the ROC curve, one could directly model the ROC curve itself. The general model takes the form

$$\mathrm{ROC}_{\mathbf{Z}}(t) = g(\mathbf{Z}\beta, h(\gamma, t))$$

for some specified functions g and h and unknown parameters (β, γ). A special case of this is the generalized linear model

$$\mathrm{ROC}_{\mathbf{Z}}(t) = g(\Sigma\gamma_l h_l(t) + \mathbf{Z}\beta),$$

where $h_1, \ldots, h_L$ are "basis" functions of t and g is a link function. Parameter estimation as proposed by Pepe (1997) is cumbersome, although newer approaches to inference using standard binary regression methods appear promising. Advantages of direct modeling of ROC curves over modeling the test results have been summarized (Pepe 1998) and include the ability to (a) restrict inference about ROCs to restricted ranges of false-positive rates; (b) incorporate interactions between covariate effects and t, so that covariates have different effects over different ranges of t; and, most important, (c) compare ROCs for tests with results of numerically different form and thus cannot be modeled sensibly in a single regression model for test results. Finally, by omitting the covariate component from the general model, it can be seen that this framework provides an avenue for developing smooth estimators of the ROC curve that are invariant to monotone data transformations.

4. THE AREA UNDER THE RECEIVER OPERATING CHARACTERISTIC CURVE

Traditionally, the area under the ROC curve (AUC) has been used as a summary index of test accuracy (Hanley 1989). Indeed, comparisons between two diagnostic tests are classically based on differences between estimated AUC's. This has long been the standard of practice in radiology where AUC is estimated with the binormal model (Swets and Pickett 1982) and for continuous data where nonparametric estimation of the AUC is possible with the Mann–Whitney U-statistic (Hanley and McNeil 1982). Regression analysis based on AUC statistics has been proposed. Here an AUC statistic is calculated as a derived variable from each subset of data that is homogeneous in regards to covariates, and then a regression model is fit to the AUC's (Obuchowski 1995). This regression framework, however, is more restrictive than others. It cannot, for example, accommodate continuous covariates.

The AUC statistic can be interpreted as the probability that the test result from a randomly chosen diseased individual is more indicative of disease than that from a randomly chosen nondiseased individual, $\mathrm{AUC} = P(X_i \geq X_j | D_i =$

$1, D_j = 0)$. Thus it can be thought of as simply a nonparametric measure of the distance between the distributions of test results for diseased and nondiseased individuals. Despite these interpretations, however, many investigators find the index unappealing, because it has no clinically relevant meaning. Another valid concern is that a large part of the area comes from the rightmost part of the curve that includes false-positive rates unlikely to be used in practice. If the curves for two tests cross, moreover, a meaningful difference between the tests over a range of interest might not be picked up by the AUCs. These considerations lead to consideration of a partial AUC, the area under the ROC curve in a restricted range of false-positive rates (Thompson and Zucchini 1989; Wieand et al. 1989).

5. CONCLUDING REMARKS

Methods for evaluating diagnostic tests have not received the same level of attention from biostatisticians as have, say, methods for evaluating new therapeutic strategies (Begg 1987). With impetus from regulatory agencies and from public health administrators, it appears that there will be an increasing demand for standards to be set for the design and analysis of studies evaluating new tests. ROC methodology is a popular statistical approach in this context. I consider ROC analysis most useful in the development phase of diagnostic testing. Its purpose is to determine whether a test has the capacity to effectively discriminate between diseased and nondiseased states. This is an important property, but it does not necessarily indicate its practical value for patient care (Zweig and Campbell 1993). Issues of cost, disease prevalence, and consequences of misdiagnosis will enter into the ultimate evaluation of test usefulness.

I mention in closing some of the major statistical challenges for evaluating diagnostic tests in general and for applying ROC methodology in particular. First, in many settings a definitive gold standard assessment of disease status, D, is not available. Infection with *Chlamydia trachomis*, for example, can be assessed only imprecisely with standard bacterial culture techniques. How can inference for an ROC curve be accomplished in this setting? Second, the statistical literature on diagnostic testing typically assumes that the test result is a simple numeric value. However, test results may be much more complicated, involving several components. Do ROC curves have a role to play in determining how to combine different sources of information to optimize diagnostic accuracy? Third, disease status is often not a fixed entity, but rather can evolve over time. How can the time aspect be incorporated sensibly into ROC analysis? Finally, are there alternatives to the ROC curve for describing test accuracy? For binary outcomes, two ways of describing test accuracy are to report (a) true- and false-positive rates, and (b) positive and negative predictive values. ROC curves can be thought of as generalizing the former to continuous tests; that is, ROC curves generalize the binary test notions of true-positive and false-positive rates to continuous tests. Are there analogs of ROC curves that similarly generalize the notions of predictive values to continuous tests?

REFERENCES

Beam, C. A., Layde, P. M., and Sullivan, D. C. (1996), "Variability in the Interpretation of Screening Mammograms by U.S. Radiologists," *Archives of Internal Medicine*, 156, 209–213.

Begg, C. E. (1987), "Biases in the Assessment of Diagnostic Tests," *Statistics in Medicine*, 6, 411–423.

Dorfman, D. D., and Alf, E. (1969), "Maximum Likelihood Estimation of Parameters of Signal Detection Theory and Determination of Confidence Intervals—Rating Method Data," *Journal of Mathematical Psychology*, 6, 487–496.

Gatsonis, C. A. (1995), "Random Effects Models for Diagnostic Accuracy Data," *Academic Radiology*, 2, 514–521.

Green, D. M., and Swets, J. A. (1996), *Signal Detection Theory and Psychophysics*, New York: Wiley.

Hanley, J. A. (1989), "Receiver Operating Characteristic (ROC) Methodology: The State of the Art," *Clinical Reviews in Diagnostic Imaging*, 29, 307–335.

——— (1998), "Receiver Operating Characteristic (ROC) Curves," *Encyclopedia of Biostatistics*, 5, 3738–3745.

Hanley, J. A., and McNeil, B. J. (1982), "The Meaning and Use of the Area Under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, 143, 29–36.

Metz, C. E. (1986), "ROC Methodology in Radiologic Imaging," *Investigative Radiology*, 21, 720–733.

Metz, C. E., Herman, B. A., and Shen, J-H. (1998), "Maximum Likelihood Estimation of Receiver Operating Characteristic (ROC) Curves From Continuously Distributed Data," *Statistics in Medicine*, 17, 1033–1053.

Obuchowski, N. A. (1995), "Multireader, Multimodality Receiver Operating Characteristic Studies: Hypothesis Testing and Sample Size Estimation Using an Analysis of Variance Approach With Dependent Observations," *Academic Radiology*, 2, 522–529.

Pepe, M. S. (1997), "A Regression Modelling Framework for Receiver Operating Characteristic Curves in Medical Diagnostic Testing," *Biometrika*, 84(3), 595–608.

——— (1998), "Three Approaches to Regression Analysis of Receiver Operating Characteristic Curves for Continuous Test Results," *Biometrics*, 54, 124–135.

Swets, J. A., and Pickett, R. M. (1982), *Evaluation of Diagnostic Systems. Methods From Signal Detection Theory*, New York: Academic Press.

Thompson, M. L., and Zucchini, W. (1989), "On the Statistical Analysis of ROC Curves," *Statistics in Medicine*, 8, 1277–1290.

Toledano, A., and Gatsonis, C. A. (1995), "Regression Analysis of Correlated Receiver Operating Characteristic Data," *Academic Radiology*, 2, S30–S36.

Tosteson, A., and Begg, C. B. (1988), "A General Regression Methodology for ROC Curve Estimation," *Medical Decision Making*, 8, 204–215.

Wieand, S., Gail, M. H., James, B. R., and James, K. L. (1989), "A Family of Nonparametric Statistics for Comparing Diagnostic Markers With Paired or Unpaired Data," *Biometrika*, 76, 585–592.

Zweig, M. H., and Campbell, G. (1993), "Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine," *Clinical Chemistry*, 39, 561–577.