ROC Curve Regression Analysis: The Use of Ordinal Regression Models for Diagnostic Test Assessment

Anna N. A. Tosteson,[1] Milton C. Weinstein,[2] Jack Wittenberg,[3] Colin B. Begg[4]

[1]Department of Medicine and Community and Family Medicine, Dartmouth Medical School, Hanover, New Hampshire; [2]Department of Health Policy and Management, Harvard School of Public Health, Cambridge, Massachusetts; [3]Department of Radiology, Massachusetts General Hospital, Harvard Medical School, Cambridge, Massachusetts; [4]Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, New York, New York

Diagnostic tests commonly are characterized by their true positive (sensitivity) and true negative (specificity) classification rates, which rely on a single decision threshold to classify a test result as positive. A more complete description of test accuracy is given by the receiver operating characteristic (ROC) curve, a graph of the false positive and true positive rates obtained as the decision threshold is varied. A generalized regression methodology, which uses a class of ordinal regression models to estimate smoothed ROC curves has been described. Data from a multi-institutional study comparing the accuracy of magnetic resonance (MR) imaging with computed tomography (CT) in detecting liver metastases, which are ideally suited for ROC regression analysis, are described. The general regression model is introduced and an estimate for the area under the ROC curve and its standard error using parameters of the ordinal regression model is given. An analysis of the liver data that highlights the utility of the methodology in parsimoniously adjusting comparisons for covariates is presented. — Environ Health Perspect 102(Suppl 8):73–78 (1994)

Key words: ROC curves, ordinal regression, sensitivity, specificity, diagnostic test assessment, rating experiment

Introduction

For a diagnostic modality, the receiver operating characteristic (ROC) curve presents a varied graphic display of all true positive and false positive rate pairs obtained as the decision threshold used to classify a group of diseased and nondiseased subjects as positive (i.e., having disease). Because the ROC curve depicts test performance across all decision thresholds, it is a more general measure of diagnostic performance than sensitivity (true positive rate) and specificity (true negative rate), which rely on a single threshold. Metz (*1*) reviewed the use of ROC curve analysis in the radiologic literature.

Smoothed ROC curve estimation is usually completed using a model from signal detection theory, which assumes the existence of an unobserved scale along which responses are distributed. The most widely used model, detailed by Green and Swets (*2*), assumes that the responses (or some monotonic transformation of them) are distributed according to a normal density function with the parameters of the density function differing according to true disease status (e.g., $N(\mu_0, \sigma_0^2)$ for nondiseased patients and distribution $N(\mu_1, \sigma_1^2)$ for diseased patients). Under this model, estimation of ROC curve parameters is accomplished using a maximum likelihood algorithm developed by Dorfman and Alf (*3*). Swets and Pickett (*4*) discuss in detail the uses of the Dorfman and Alf methodology for the analysis of diagnostic systems.

The primary limitation of the Dorfman and Alf approach is the difficulty that occurs when covariates must be considered in the analysis. Because the signal detection model provides no facility for incorporating covariates directly into the analysis, one must form subgroups and estimate ROC curves separately in each. This approach precludes the consideration of continuous covariates and often becomes impractical because of sparse data. The need to accommodate covariates is exemplified by data from the liver protocol of the magnetic resonance imaging (MR) collaborative working group*, which was designed to compare MR at several magnet field strengths with dynamic, enhanced computed tomography (CT) in the detection of liver metastases in patients with primary malignancies of the breast, colon, lung, and pancreas. Among the factors that may influence the accuracy of MR and CT and that must be considered in the analysis are patient-specific covariates such as primary tumor site, modality-specific covariates such as magnet field strength, and external covariates such as reviewer.

Recently, Tosteson and Begg (*5*) described an approach to ROC curve estimation which includes the Dorfman and Alf model as a special case of a much more general class of models. To estimate ROC curve parameters from rating experiment data, the general regression approach uses one of a class of ordinal regression models developed by McCullagh (*6*).

This paper was presented at the 4th Japan-US Biostatistics Conference on the Study of Human Cancer held 9–11 November 1992 in Tokyo, Japan.

This work was supported in part by National Cancer Institute Contract NO. I-CM-47564.

Address all correspondence to Anna N. A Tosteson, Clinical Research, HB 7505, Dartmouth-Hitchcock Medical Center, One Medical Center Drive, Lebanon, NH 03756. Telephone (603) 650-6390. Fax (603) 650-8208.

*The magnetic resonance imaging collaborative working group was comprised of radiologists from Bowman-Gray School of Medicine, Cleveland Clinics, Duke University, Massachusetts General Hospital, and the University of California at San Francisco. This group was formed to assess the diagnostic accuracy of MR in the detection and characterization of disease at seven skeletal sites in comparison to the currently used standard modality (or modalities) at each disease site.

In this article, our goal is to illustrate how ROC regression analysis facilitates diagnostic test assessment. We first present a detailed description of issues in the analysis of data from the multi-institutional liver protocol. Next we introduce the ordinal regression model and use its parameters to give an estimate for the commonly used ROC curve summary statistic, the area under the curve, and its standard error. A detailed analysis of the liver protocol data is presented.

The Liver Protocol

In patients with cancer, it is important for patient management that the disease be followed closely and that the presence or absence of hepatic metastases be correctly diagnosed. To determine which modality is most accurate in the diagnosis of metastases the MR collaborative working group collected data according to the liver protocol described here.

Consecutive patients who were referred for investigation of possible liver metastases, or who had histologically documented liver metastases and had documentation (within 30 days of enrollment) of a primary malignant tumor of the breast, colon, lung, or pancreas, were prospectively enrolled in the study from 1985 through 1987 at Bowman-Gray School of Medicine, Cleveland Clinics, Duke University, Massachusetts General Hospital, and the University of California at San Francisco. Patients who did not have both MR and CT examinations completed within a 14-day period were excluded from the study. Each study subject was then followed for ultimate verification of true disease status, meaning in this case the presence or absence of liver metastases.

Each subject's disease status was determined by pathology or according to clinical criteria specified in the protocol, and was carefully documented on follow-up forms that were completed and submitted within six months of the original CT and MR examinations. The criteria for clinical proof of liver metastases were as follows: enlargement of focal metastasis on follow-up CT scan or MR image, abnormal liver function tests and presence of extrahepatic metastases, or abnormal liver function tests and subsequent treatment for liver metastases. Analogous criteria were used to document benign liver abnormalities, which were considered as normal for the purpose of assessing the accuracy of MR in the detection of liver metastases.

The Rating Experiment

Prior to the documentation of true disease status the CT, MR pairs for each subject were sent from each institution to a central review meeting. The central review process consisted of a rating experiment in which one radiologist from each institution participated.

The reviews were generally carried out over a 2-day period during which the five radiologists (reviewers) participated in four review sessions. For each radiologist a review session consisted of rating a batch of 20 to 40 CT scans or MR images. The review was designed so that each subject received two independent ratings of both their CT scan and MR image.

Reviewers were blinded to patient history and all patient characteristics except for primary tumor site. Primary tumor site was made available to the reviewing radiologists because it was judged that omission of this critical information would hamper the radiologists' review of both modalities and would therefore not fairly represent either modality as used in clinical practice.

In keeping with the blinded review process, reviewers were not allowed to rate examinations from their own institution, because their interpretation of MR and CT for these cases could be influenced by their knowledge of the patient's subsequent clinical course. In addition, with two exceptions (i.e., two sessions, one batch in each), no reviewer rated the CT, MR pair from an individual subject. These steps were taken in an effort to minimize bias in test interpretation (*7,8*). For example, if a reviewer was first to read the MR images for a group of subjects and later to read their CT scans, the ratings for the CT scans could be influenced by the reviewer's recall of the MR images. This could cause the CT ratings to appear more accurate than they would have been if rated independently, thus biasing the comparison of MR and CT.

The radiologists had been familiarized with the data form used in the rating experiment prior to the review, and were asked to give the likelihood on a scale from 0 to 100 that liver metastases were present.

Issues in the Data Analysis

The primary goal of the data analysis is to compare MR with CT in the detection of liver metastases. The liver data include images from five institutions, which operate MR units at four separate magnet field strengths, and afford a unique opportunity to assess the comparability of MR at each field strength with CT. In making these comparisons, we are concerned with factors that may influence either the accuracy of the modality under study or the manner in which it is interpreted. As outlined below, we consider three classes of factors: patient specific, modality specific, and external.

Patient-specific factors refer to particular patient and disease characteristics that may influence the accuracy of MR or CT or its interpretation. Some questions that we address are the following: Does primary tumor site (or age, gender, etc.) influence the criteria that reviewers used for assigning likelihoods of metastases to subjects? Alternatively, can any apparent differences between MR at varying magnet strengths and CT be explained by confounding due to patient and disease characteristics? The latter is important because, although the paired design of the protocol ensures balance of patient characteristics for the overall MR/CT comparison, this balance is not ensured for the comparisons between MR units using different magnet strengths and CT, because magnet field strengths differ across sites.

Modality-specific factors refer to characteristics of the modality that could influence its accuracy or interpretation, such as magnet field strength and the study year in which the examination was performed. Here we ask, does magnet field strength affect the accuracy of MR? Has there been a change in MR or CT over the three years in which the study was conducted due to modifications in the technology or due to improvements in the reviewers' interpretation of MR?

External factors encompass additional features of the study design that could confound the MR/CT comparisons. The effect that the reviewer has on the accuracy of MR and CT is of particular interest here. We ask: What is the nature of interobserver variability? Are some reviewers better than others or do all reviewers operate along the same ROC curve? The answers to these questions are of intrinsic interest and have strong implications for the comparison between MR and CT.

Although some of the issues discussed here could be accommodated by taking advantage of the paired design and limiting the analysis to a comparison of MR and CT within each site, doing so would result in a loss of power. Furthermore, because no reviewer read both the CT and MR of the same subject, the paired comparisons might be confounded by reviewer effects. Thus, the MR/CT comparison is ideally suited to a regression approach that allows us to control potential confounders. In the next section we introduce such a model and specify how it can be used to estimate ROC curve parameters.

Ordinal Regression Model

Prior to the analysis, the responses (likelihoods) assigned by the reviewers to each subject's MR image and CT scan are grouped into J ordered response categories, with the Jth category being most indicative of disease. Let Y be the ordinal variable representing these response categories (i.e., Y takes on the values $1,2,\ldots,J$). We wish to model $Pr[Y<j|X]$, the probability of a response in category j or lower for subjects with covariates X. To define an ROC curve it is necessary that one component of X be an indicator of true disease status. Assume momentarily that X is a scalar (i.e., there are no covariates other than disease status) and that $X=1$ if metastases are present and $X=0$ otherwise. Then the ROC curve is a plot of $1-Pr[Y<j|X=1]$, the proportion of true positives (TPR), against $1-Pr[Y<j|X=0]$, the proportion of false positives (FPR), as j is varied.

We model $Pr[Y<j|X]$ using one of a class of general ordinal regression models (*6*). The model is

$$Pr[Y<j|X]=\Phi[(\Theta_j-\alpha'X)\{\exp(\beta'X)\}^{-1}] \qquad [1]$$

where $\Phi(.)$ is the cumulative normal distribution function and the Θ_j are the $J-1$ estimated cutoff values for the boundaries between the J response categories. Note that this model accommodates both a location regression, $\alpha'X$, and a scale regression, $\exp(\beta'X)$, although the factors included in each regression need not be the same. A more general statement of Equation 1 would replace $\Phi(.)$ with any monotonically increasing link function. However, in the simple case where X represents a single covariate corresponding to true disease status, Equation 1 corresponds to the model given by Dorfman and Alf (*3*) as described by Tosteson and Begg (*5*).

Maximum likelihood estimates for the model parameters are obtained by an iteratively reweighted least squares algorithm (*6,9*). Additional computations using the estimated parameters α and β are necessary to generate the ROC curve coordinates (FPR, TPR) for curve plotting.

Model selection is guided by the estimated coefficients, α and β, their associated standard errors, and the deviance statistic, which is defined as minus twice the difference between the maximized log likelihood for the current model and the maximized log likelihood attainable for a perfect model (i.e., a full model, one using all the degrees of freedom available) (*9*). Hierarchical models are compared using the difference in the deviance statistic, which has a chi-square distribution with degrees of freedom equal to the difference between the number of parameters fit in each model.

The plot of the empirical ROC curve(s) that is obtained by varying the cutoff category for classifying subjects as having a positive rating and plotting the resulting true positive and false positive error rate pairs is also useful in checking the correspondence between the fitted curves and the raw data.

The Area Under the ROC Curve

The area under the ROC curve, which is related to the Wilcoxon or Mann-Whitney statistic (*10*), has become a commonly used summary measure in ROC curve analyses (*11,12*). Although this measure is global and not ideal in settings were ROC curves cross, we present an estimate for the area under the ROC curve using parameters derived from Equation 1 for completeness.

The area under any ROC curve is easily obtained from the estimated parameters α and β, using the notation and equations detailed below. Let $X=(X_1,X^*)$, where X_1 represents true disease status and X^* represents all other covariates. Then we define the covariate vectors for nondiseased and diseased subjects as $X_0=(0,X^*)$ and $X_1=(1, X^*)$, respectively. The vectors X_0 and X_1 differ depending on whether we are referring to the location regression, $\alpha'X$, or the scale regression, $\beta'X$. However, to simplify notation we assume that when associated with α, X is a vector of dimension $1\times N_1$ containing covariates included in the location regression and when associated with β, X is a vector of dimension $1\times N_2$ containing covariates included in the scale regression.

The area under the ROC curve, as described by Dorfman and Alf (*3*), and Swets and Pickett (*4*) is area $= A = \Phi[(\mu_1-\mu_0)/\{\sigma_1^2+\sigma_0^2\}^{1/2}]$, where μ_0, σ_0^2, μ_1, σ_1^2 are parameters for the normal density functions according to which responses from diseased ($N(\mu_0,\sigma_0^2)$) and nondiseased patients ($N(\mu_1,\sigma_1^2)$) are distributed. In terms of the ordinal regression model (Equation 1) parameters, the area estimator is

$$A = g(\alpha,\beta) = \Phi[(\alpha'X_1-\alpha'X_0)/\{\exp(2\beta'X_1)+\exp(2\beta'X_0)\}^{1/2}]\,. \qquad [2]$$

Let $N=N_1+N_2$ represent the total number of covariates in location and scale models, respectively, where a variable that appears in both locations is counted twice. Let $I_i=(0,\ldots,1,0,\ldots,0)$ represent a vector of appropriate dimension with zeroes everywhere except for the ith entry. Then the standard error for Equation 2 is derived using the delta method as $SE[g(\alpha,\beta)]=(G^*\Sigma^*G')^{1/2}$, where $G=[\partial g(\alpha,\beta)/\partial\alpha_i, \partial g(\alpha,\beta)/\partial\beta_i]$ is a $1\times N$ vector of derivatives of the form given below and where Σ is the variance-covariance matrix for the parameters α, β. The partial derivatives that comprise G are of the form:

$$\frac{\partial g(\alpha,\beta)}{\partial\alpha_i}=(I_iX_1-I_iX_0)\,t(\alpha,\beta)\,r(\beta)^{-1/2}$$

and

$$\begin{aligned}\frac{\partial g(\alpha,\beta)}{\partial\beta_i}=&(-1/2)(\alpha'X_1-\alpha'X_0)\\&\{2I_iX_1\exp(2\beta'X)\\&+2I_iX_0\exp(2\beta'X)\}\,t(\alpha,\beta)\,r(\beta)^{-3/2},\end{aligned}$$

where $t(\alpha,\beta)=(2\pi)^{-1/2}\exp(\{-{}^1/_2(\alpha'X_1-\alpha'X_0)\, r(\beta)^{-1/2}\}^2)$ and $r(\beta)=\exp(2\beta'X_1)+\exp(2\beta'X_0)$.

An estimate for the standard error of the difference between two area estimators is obtained analogously.

Description of Modeling Approach

In addition to true disease status, the patient-specific, modality-specific, and external factors described earlier must be included in our ROC regression. These factors may influence the ROC curves as either main effects or interactions with true disease status. Tosteson and Begg (*5*) describe in detail the impact that these terms have on the shape of ROC curves estimated using Equation 1. They show that, in the location regression, the covariate main effects do not affect the shape of the ROC curve, while covariate interaction terms define different curves for each level of the covariate. Covariate main effects in the location model are interpreted to imply movement along a single ROC curve. They indicate that a difference exists in the criteria for assigning likelihoods to subjects for differing levels of the covariate. That is, the interpretation of the modality is influenced by the covariate. In the scale regression, covariate main effects are shown to shift the curves to either the lower left or the upper right of the graph (away from symmetry), while interactions between covariates and true disease status define ROC curves that cross each other for varying levels of the covariate(s).

To minimize the complexity of model selection when both location and scale models are available, we developed a general approach in which we first fit true disease status in both location and scale models. True disease status was included in the scale model because of the finding that this term is usually of importance in modeling radiologic data (13). Next, expanded location models were considered and selection of hierarchical models was completed using the deviance statistic. Using this approach we either included or excluded all indicators for a particular covariate (e.g., primary tumor site would include indicators for three of the four groups). Once location modeling was completed, expanded scale models were considered. Scale modeling was done sparingly because the iterative fitting routine failed to converge for many complicated scale models.

Table 1. Distribution of patient charcteristics and percent with mestastases by characteristic.

	% with characteristic	Proportion of subjects with mestastases
Tumor site		
Breast	20	52
Colon	45	58
Lung	21	34
Pancreas	14	52
Total %	100	
Age		
<50	25	59
≥50	75	48
Total %	100	
Sex		
Male	47	51
Female	53	50
Total %	100	
Magnet field strength		
Low	31	51
Middle	42	50
High	27	51
Total %	100	

Table 2. Likelihood ratios for MR and CT based on rating table data (i.e., for unadjusted analysis).

	Likelihood ratio[a]	
Category	MR	CT
1	0.20	0.14
2	0.16	0.36
3	0.42	0.45
4	0.45	0.61
5	1.64	2.06
6	5.22	4.40
7	10.74	9.80

[a] The likelihood ratio is equal to the proportion of responses in a given category for diseased subjects divided by the proportion of responses in that category for nondiseased subjects.

Results of Data Analysis

A total of 502 patients were reviewed during the course of the study, with follow-up pathology available for 326 (65%). Although this raises the concern that verification bias (7,8) may pose a problem in the analysis, this is unlikely to be a problem because the majority of subjects without disease verification are those entered late in the study. Of those with verified disease status, 165 (51%) have documented liver metastases. The data are comprised of the following mix of primary malignancies: 20% breast, 45% colon, 21% lung, and 14% pancreas (Table 1).

Empirical and Unadjusted ROC Curves

For the analysis, the 100 point rating scale was collapsed into seven categories (i.e., 0, 1 to 19, 20 to 39, 40 to 59, 60 to 79, 80 to 99, 100). The rating data for MR and CT using this categorization are given in Table 2. The empirical data points and smoothed curves were fit to these data using Equation 1 (Figure 1). We note the apparent similarity between the unadjusted curves. Generally, crossing ROC curves, which are characteristic of models that include interaction terms in the scale model, indicate that no modality is clearly superior.

To answer questions regarding magnet field strength we divide the MR units into three groups: low (0.15 T [teslas] unit), mid (0.35 T, 0.5 T, and 0.6 T units), and high (1.5 T unit). These groups comprise 31, 42, and 27% of the sample, respectively. The empirical data points and curves estimated using Equation 1 for MR at each field strength in comparison to CT are given in Figure 2. The curve for high field strength MR is higher than the other curves, indicating that it may be superior to the other modalities in detecting liver metastases. The curves for low- and mid-field strength MR cross, with low MR having higher specificity than mid MR at sensitivities below approximately 70%. The areas under the ROC curves are quite similar at 0.87 ± 0.02 for CT, 0.85 ± 0.02, 0.87 ± 0.01, and 0.92 ± 0.02 for low-, mid-, and high-field strength MR, respectively.

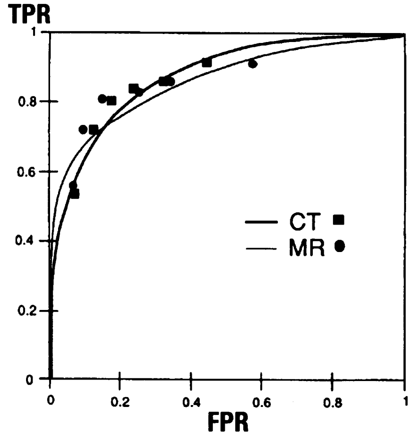

Figure 1. ROC curves for the diagnosis of liver metastases, CT versus combined MR. True positive rate (TPR) is shown on the ordinate and false positive rate (FPR) on the abscissa. Areas under the curves are not significantly different.

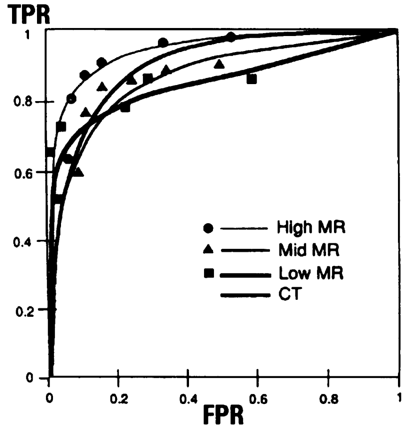

Figure 2. ROC curves for the diagnosis of liver metastases, CT versus categorized magnetic strength units. Areas under the curves are not significantly different among all of the imaging techniques.

Analyses of Single Covariates

Before undertaking formal comparisons, a series of analyses assessing the individual impact of factors on the ROC curves for MR and CT were completed to understand the broad associations present in the data. The results of two of these analyses are described.

Primary Tumor Site

An analysis considering primary tumor site indicated that both modalities detect liver metastases most accurately for colon primaries with decreasing accuracies evident for lung, breast, and pancreatic primaries, respectively. A single curve represents both combined MR and CT (Figure 3). These curves are representative of the family of curves produced for varying levels of a covariate (primary tumor site) when interaction effects are significant in the location model or when main effects are significant in the scale model.

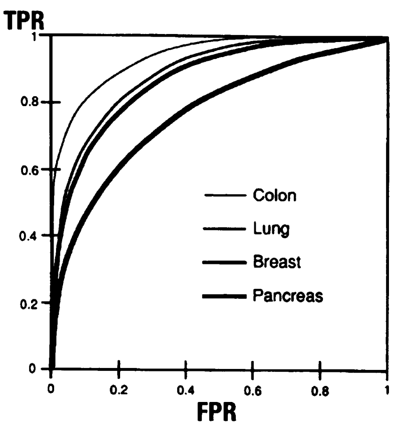

Figure 3. ROC curves for detection of liver metastases according to primary source of metastases. A single curve is used to represent both CT and combined MR. Area under curve for detecting colonic metastases is significantly greater than that for pancreatic metastases (p<0.001). No other significant differences were found.

Reviewer Effects

Over the course of the study, two of the five participating institutions changed reviewers. Thus, a total of seven reviewers participated in the rating experiment. When we examined the impact of interobserver variability on the ROC curve, using the ordinal regression model, we found reviewer effects that were significant when included in both the location and scale models as main effects and as interaction terms. The main effects for reviewers demonstrate that reviewers use different criteria or cutpoints for assigning likelihoods to subjects; more important, the significance of the interaction terms in the location model indicates that reviewers have different ROC curves along which they operate. That is, some reviewers are better diagnosticians than others.

Implications

Primary tumor site and reviewer are both important factors to control for in the MR/CT comparison. Because colon primaries are more easily identified than pancreatic primaries, if one MR field strength imaged a disproportionate share of patients with colon primaries, it could falsely appear superior to the other MR field strengths. Indeed a chi-square test of association between tumor site and magnet field strength was highly significant (p<0.001). Although this problem could be circumvented by restricting the comparisons to paired data within each institution, this would result in a loss of power. Furthermore, no reviewers read both the CT and MR for each case and paired comparisons could be confounded by the reviewer effects. That is, since some reviewers are "better" than others, if the "best" reviewer reads relatively more of one modality than another this could bias our comparisons.

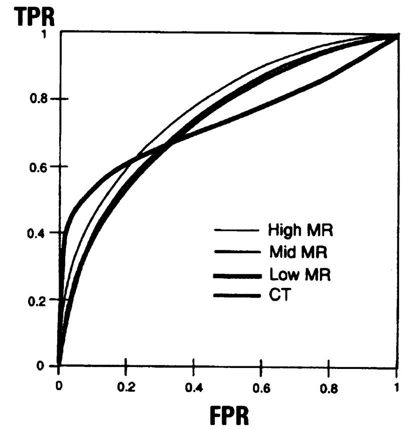

Figure 4. ROC curve for diagnosis of liver metastases adjusted for interobserver error and primary tumor site, CT versus categorized magnetic strength units. Area under curve comparisons are not significantly different.

Analysis with Multiple Covariates

In an overall analysis using Equation 1 to compare MR at each field strength with CT, we considered adjusting for primary tumor site, reviewer, gender, age, and year of exam. The factors that influenced the accuracy of MR and CT were reviewer and primary tumor site (Table 3). In the location model, reviewer main and interaction effects and tumor site interaction effects were statistically significant. In practical terms, the importance of the main effects implies that the criteria for assigning a likelihood that liver metastases are present to each subject varied across reviewers. The importance of the interaction terms implies that even after controlling for all factors of importance, some reviewers are better diagnosticians than others and that it is easier to accurately assess the presence of liver metastases for colon primary tumors than for lung, breast, or pancreatic primaries. Larger coefficients for interaction terms are indicative of better diagnostic accuracy (*5*). After adjusting for reviewer and tumor, we found that no modality effects were of importance in the location model, indicating that no modality is clearly superior to any other.

In the scale model, reviewer and tumor main effects were the most important covariates. Main effects in the scale model have a similar impact to interactions in the location model; however, smaller coefficients are indicative of better diagnostic accuracy (*5*). An examination of the coefficients for tumor main effects indicates that after adjusting for all other factors, lung primaries were slightly, but not significantly, more accurately detected than colin primaries. As seen above, pancreatic primaries had the lowest (i.e., largest coefficient) ROC curve.

Table 3. Estimated coefficients and standard errors for final adjusted model.

Covariate	Location model	Scale model
Disease status	0.7383 (0.10)	0.1808 (0.11)
Reviewer main effects[a]		
Reviewer A	−1.3730 (0.34)	1.8300 (0.18)
Reviewer B	−0.5895 (0.14)	0.9987 (0.15)
Reviewer C	−0.5712 (0.18)	1.2490 (0.17)
Reviewer D	−0.1576 (0.09)	0.7562 (0.13)
Reviewer E	−0.5653 (0.12	0.9365 (0.14)
Reviewer F	−0.2380 (0.08)	0.1812 (0.13)
Reviewer interactions[a]		
Reviewer A	3.5440 (0.64)	
Reviewer B	1.2770 (0.24)	
Reviewer C	2.3640 (0.45)	
Reviewer D	0.8587 (0.16)	
Reviewer E	1.1980 (0.20)	
Reviewer F	0.2958 (0.13)	
Tumor main effects[b]		
Breast		0.1725 (0.09)
Lung		−0.0399 (0.09)
Pancreas		0.1801 (0.11)
Tumor interactions[b]		
Breast	−0.1121 (0.12)	
Lung	−0.1752 (0.11)	
Pancreas	−0.3654 (0.14)	
Modality main effects[c]		
Low MRI	−0.4231 (0.12)	
Mid MRI	−0.0003 (0.13)	
High MRI	−0.1842 (0.14)	
Modality interactions[c]		
Low MRI	0.6379 (0.18)	
Mid MRI	−0.0298 (0.17)	
High MRI	−0.0025 (0.19)	

[a] Comparisons are with Reviewer G. [b] Comparisons are with Colon Primaries. [c] Comparisons are with CT.

After adjustment for all other covariates, modality effects were of importance in the scale model. An examination of the modality coefficients indicates that only low field strength MR differs significantly from CT. However, the statistically significant coefficient for the interaction between low field strength MR and true disease status indicates the existence of crossing ROC curves. Thus, low-field strength MR is not uniformly superior or uniformity poorer than CT in detecting the presence of liver metastases.

To more fully understand the impact of the final model estimates (Table 3) on the ROC, curve a graphical presentation is useful. Given that there were seven reviewers and four primary tumor sites represented in the liver protocol data, the adjusted analysis produces 28 groups, in which there are four ROC curves (one for each modality). The relative position of the curves for each modality was the same within each reviewer and primary tumor site combination. Figure 4 shows composite adjusted curves, which were derived by including all parameter estimates from the final adjusted model in the regression model and weighting them equally to reflect what would be expected under a balanced design. The high field strength MR had higher sensitivity and specificity than both mid MR and CT for the entire range, but these differences were not statistically significant in terms of the regression parameter estimates or in terms of area under the ROC curve. As indicated by the importance of interaction terms in the scale model, we see that the ROC curve for low field strength MR crosses the other curves. We see that at lower sensitivities the low field strength MR was superior to all other modalities. It is noted that when using these modalities in screening patients for liver metastases, higher sensitivity is usually preferred to higher specificity.

Discussion

Our analysis demonstrates the role of ordinal regression models in diagnostic test assessment when tradeoffs between sensitivity and specificity must be captured through estimation of ROC curves. One clear advantage of the analysis that we present is the ease with which multiple covariates were assessed. Using traditional ROC curve methods, the subgroup analyses would have been tedious and in some cases impossible to complete or interpret due to sparse data.

The rating data used in our example represented ordinal ratings on a 0- to 100-point scale. In most rating experiments, a more limited set of categories is specified. When reviewers rated the likelihood that metastases were present, however, they tended to use a limited set of values (e.g., 10, 25, 50, etc.) for assigning ratings. To complete the analysis, the reviewers' ratings were grouped into seven response categories. The results of the analysis were unchanged when ratings were collapsed into twelve rather than seven response categories.

Swets and Pickett (*4*) describe two general approaches to the issue of multiple ratings (more than one review per subject) by different reviewers. The first involves pooling the ratings for each subject and treating them as independent observations. The second involves averaging parameters and summary measures from each reviewer's curve. In contrast, we approach the multiple rating problem by including reviewer effects in the ordinal regression model. The estimation of reviewer effects not only indicated that differences exist between reviewers, but also identified the nature of those differences. The estimated reviewer effects indicated that the decision criteria for the raters were quite different and that some reviewers were better diagnosticians than others, and therefore required us to control all modality comparisons for these important differences.

Our approach to these data, while offering several advantages, also has some limitations. First, our approach does not accommodate the pairing of the data. Second, the fixed-effects nature of the model used does not allow us to accommodate intrareader correlations. Further work is required before these limitations can be overcome.

Biases in radiologic test assessment are common and have been reviewed by Begg and McNeil (*8*). An attempt was made in our study to avoid test interpretation biases by the blinding procedures used in the rating experiment. Biases occurring as a result of disease spectrum, or differences in case mix were partially avoided by adjusting for patient-specific factors in the analysis. These adjustments are more easily made using the ordinal regression model than using standard ROC curve techniques, because of the facility provided by the ordinal regression model for subgroup analyses. As stated previously, we do not expect that verification bias is a serious problem in this analysis.

The potential over-representation of subjects with liver metastases in our data, relative to what would be expected in a population of patients with primary malignancies of the breast, colon, lung, and pancreas, is unlikely to bias our comparison of MR and CT. It does, however, make it likely that estimates of the sensitivity of both modalities in detecting liver metastases are overestimated using any form of ROC curve estimation.

In summary, the ordinal regression model provided a flexible and parsimonious method for analyzing the liver protocol data. The facility that it provides for appropriately adjusting ROC curves for potential confounders and other relevant factors enhances our capacity for carrying out complex comparisons when many factors must be controlled for in the analysis.

REFERENCES

1. Metz CE. ROC methodology in radiologic imaging. Inv Radiol 21:720–733 (1986).
2. Green DM, Swets DA. Signal Detection Theory and Psychophysics. New York:Krieger Huntington, 1974.
3. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals. Rating method data. J Math Psychol 6:487–496 (1969).
4. Swets JA, Pickett RM. Evaluation of Diagnostic Systems. Methods from Signal Detection Theory. New York:Academic Press, 1982.
5. Tosteson ANA, Begg CB. A general regression methodology for ROC curve estimation. Med Decis Making 8:204–215 (1988).
6. McCullagh P. Regression models for ordinal data (with discussion). J R Stat Soc B 42:109–142 (1980).
7. Begg CB. Biases in the assessment of diagnostic tests. Stat Med 6:411–423 (1987).
8. Begg CB, McNeil BJ. Assessment of radiologic tests: control of bias and other design considerations. Radiology 167:565–569 (1988).
9. McCullagh P, Nelder JA. Generalized Linear Models, 2nd ed. London:Chapman and Hall, 1989.
10. Bamber, D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol 12:387–415 (1975).
11. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 143:29–36 (1982).
12. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology 148:839–843 (1983).
13. Swets JA. Indices of discrimination or diagnostic accuracy: their ROCs and implied models. Psychol Bull 99:100–107 (1986).