STATISTICS IN MEDICINE
Statist. Med. 2000; **19**:511–523

Comparison of diagnostic markers with repeated measurements: a non-parametric ROC curve approach

Birol Emir[1,*], Sam Wieand[2], Sin-Ho Jung[3] and Zhiliang Ying[4]

[1] *Bayer Pharmaceuticals, Statistics and Data Systems, West Haven, CT 06516, U.S.A.*
[2] *Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, U.S.A.*
[3] *Division of Biostatistics, Indiana University School of Medicine, Indianapolis, IN 46202, U.S.A.*
[4] *Department of Statistics, Rutgers University, New Brunswick, NJ 08903, U.S.A.*

SUMMARY

In this paper we study a class of non-parametric statistics for comparing diagnostic markers with repeated measurements. Using adapted definitions of specificity and sensitivity, we suggest methods to compare the average of sensitivities across all specificities or a range of specificities. The theory allows for correlations introduced by the fact that markers may be obtained from the same patient at multiple visits and that both markers being compared may be obtained from the same patient. Results of the Monte Carlo simulations and an example from a breast cancer setting are provided. Copyright © 2000 John Wiley & Sons, Ltd.

1. INTRODUCTION

Emir *et al.* [1] introduced a non-parametric approach to evaluate repeated markers which might be used to supplement or replace clinical examinations as a method for predicting an outcome. They brought concepts such as specificity and sensitivity into the analysis of repeated markers. Their methods allowed for estimating sensitivities, specificities and predictive values, obtaining confidence intervals (CIs) for them, and comparing sensitivities at fixed specificities in the presence of between- and within-marker correlations. They developed an asymptotically normal statistic for comparing the sensitivities of two markers at a fixed specificity. They also presented an example from a breast cancer study to show why it may be more appropriate to use parameters such as specificity and sensitivity than parameters that evaluate the relative risk associated with the high values of a marker (as is traditionally done by using a time-dependent Cox [2] model).

A generalization of interest for us is to define a statistic for comparing the sensitivities across a range of specificities, instead of being restricted to one specificity. This is equivalent to estimating the area under an ROC (relative operating characteristic) curve, which is the trace of

* Correspondence to: Birol Emir, Bayer Pharmaceuticals, 400 Morgan Lane, Statistics and Data Systems, C32, West Haven, CT 06516, U.S.A.

CCC 0277–6715/2000/040511–13$17.50

(1-specificity, sensitivity) in the unit square (Swets and Pickett [3], Swets [4], Metz [5, 6] and Hanley and McNeil [7, 8]). Our approach generalizes the work of Wieand *et al.* [9] by extending the results of Emir *et al.* [1], obtained for a fixed specificity p, over intervals $p \in [p_1, p_2]$, $0 \leqslant p_1 \leqslant p_2 \leqslant 1$.

2. BACKGROUND NOTATIONS AND DEFINITIONS

2.1. Background

Emir *et al.* [1] provided definitions of specificity and sensitivity in the repeated marker setting, with examples, that allow for varied applications. The simplest application was when the marker was used as a surrogate for an outcome. In that situation, a patient would be classified as a case or control according to the value of the marker and the gold standard would be the result of a simultaneous clinical evaluation. A more useful application might involve using the marker to predict a subsequent event or to use scores like the slope of two consecutive markers or the maximum of two different markers. Emir *et al.* [1] address ways to modify the definition of specificity and sensitivity in such situations. However, the methods of inference required to generalize their results are unaffected by these various definitions, so we will use the simplest application in this manuscript.

2.2. Preliminary notation

We will assume that we have a random sample of n patients who are being followed for the outcome of disease progression (which will occur at most once at which time the patient will go off the study). Let X_{jl} be the continuous random variable whose observations are the marker values obtained from the jth patient, $j = 1, \ldots, n$, at the lth non-progression evaluation, (determined by the gold standard – physical exam), $l = 1, \ldots, m_j$, where m_j is the number of non-progression evaluations for patient j. Similarly, let Y_j be the continuous random variable associated with values for the same marker from the jth patient at the progression visit. Also let $\delta_j = 1$ if patient j became a case (had a progression) and $= 0$ otherwise. If $\delta_j = 0$, we may assign any value to Y_j. Define $D = \sum_{i=1}^{n} \delta_i$ to be the total number of cases, which is of order $O(n)$. Let F and G be the distribution functions of X_{jl} and Y_j, respectively, that is, F is the distribution function of the marker of a patient who has not progressed and G is the distribution of a marker of a patient who has progressed. Assume that if the value of X_{jl} or Y_j exceeds a predetermined cut-off point c the marker will be considered positive (classifying the patient as a progression). Then the specificity and sensitivity of the marker are $\text{spe} = F(c)$ and $\text{sen} = H(c) \equiv 1 - G(c)$, respectively. We define $\hat{\text{spe}}_j(c) = (1/m_j)\sum_{l=1}^{m_j} I(X_{jl} \leqslant c)$. Similarly, let $\hat{\text{sen}}_j(c) = \delta_j I(Y_j > c)$ and a random vector for each patient j, as $\{\hat{\text{spe}}_j(c), m_j, \hat{\text{sen}}_j(c), \delta_j\}'$ for $j = 1, \ldots, n$. These n vectors are independently and identically distributed. We define the sample specificity $\hat{\text{spe}}(c)$, and sample sensitivity $\hat{\text{sen}}(c)$, estimates of $\text{spe}(c)$ and $\text{sen}(c)$ as

$$\hat{\text{spe}}(c) = \sum w_j' \hat{\text{spe}}_j(c) = (1/M) \sum_{j=1}^{n} w_j \sum_{l=1}^{m_j} I(X_{jl} \leqslant c)$$

$$\hat{\text{sen}}(c) = (1/D) \sum \delta_j I(Y_j > c)$$

where $M = \sum_{j=1}^{n} m_j w_j$, the weights w_j are bounded functions of m_j, and $w_j' = (w_j m_j)/M$. We require that $w_j = 0$ if $m_j = 0$, as $\hat{\text{spe}}_j$ are undefined if $m_j = 0$.

Emir *et al.* [1] showed that $\sqrt{n}(\hat{\text{spe}} - \text{spe})$ (or $\sqrt{n}(\hat{\text{sen}} - \text{sen})$) is asymptotically normal under fairly general conditions (see Section 2.2.1) and obtained the asymptotic variance. We will present theorems that extend these results, but we first state conditions that allow us to prove the results.

2.2.1. *Conditions.*

1. F and G are twice differentiable at c with $F'(c) = f(c) > 0$ and $G'(c) = g(c) > 0$ where c is an arbitrary cut-off point.
2. The maximum possible number of evaluations per patient is bounded by a finite number $\mathscr{K}$.
3. $w_j = \Psi(m_j)$, such that $E(\Psi(m_j)) > 0$ where $\Psi(\cdot)$ is a bounded non-negative function satisfying $\Psi(0) = 0$.
4. Evaluation times are independent of the event times and the marker values.

For the rest of the manuscript we assume that these conditions hold for all markers.

2.3. *Area estimate*

In this section we present a statistic based on a weighted average of sensitivities. Let $\theta = \int_0^1 \text{sen}(\text{spe}^{-1}(t))\mathrm{d}P(t)$ be the average of sensitivities across specificities where $P(.)$ is the left continuous version of a cumulative density function on the unit interval, $[0, 1]$. If $P(t) = t$, $0 \leqslant t \leqslant 1$, θ would be the average of all sensitivities across all possible specificities and would correspond to the area under the ROC curve, that is, the curve defined by $(t, \text{sen}(\text{spe}^{-1}(t)))$, for all $t \in [0, 1]$. For $0 \leqslant t_1 \leqslant t_2 \leqslant 1$, let $P(t) = 0$ if $t \leqslant t_1$; $P(t) = (t - t_1)/(t_2 - t_1)$ if $t_1 < t < t_2$; $P(t) = 1$ if $t \geqslant t_2$. Then θ would correspond to averaging the sensitivities over the range of specificities between $[t_1, t_2]$. If P is a point mass at t_0, that is, $P(t) = I(t > t_0)$, then θ would correspond to the statistic discussed in Emir *et al.* [1]. For now, we assume that $P(t) = t$, $0 \leqslant t \leqslant 1$. A non-parametric estimate of θ is

$$\hat{\theta} = \int_0^1 \hat{\text{sen}}(\hat{\text{spe}}^{-1}(t))\mathrm{d}P(t) = \int_0^1 \hat{\text{sen}}(\hat{\text{spe}}^{-1}(t))\,\mathrm{d}t = \int_{-\infty}^{+\infty} \hat{\text{sen}}(u)\,\mathrm{d}\hat{\text{spe}}(u)$$

$$= (1/M) \sum_{l=1}^{n} w_j \sum_{l=1}^{m_j} \hat{\text{sen}}(X_{jl}) = (1/M) \sum_{j=1}^{n} w_j \sum_{l=1}^{m_j} \sum_{k=1}^{n} (\delta_k/D) I(Y_k > X_{jl})$$

$$= (1/(DM)) \sum_{j=1}^{n} \sum_{k=1}^{n} w_j \sum_{l=1}^{m_j} \delta_k I(Y_k > X_{jl}). \tag{1}$$

Suppose we have two markers being evaluated in the same set of patients. For marker i ($i = 1, 2$) and patient j ($j = 1, \ldots, n$), let $(X_{ij1}, X_{ij2}, \ldots, X_{ijm_j})$ denote the repeated marker values observed before progression and Y_{ij} denote those observed at the time of progression. We assume that $\{X_{ijl}, j = 1, \ldots, n, l = 1, \ldots, m_j\}$ are marginally identically distributed with CDF F_i, and X_{ijl} and $X_{ij'l'}$ are independent if $j \neq j'$. Also, among progressed patients ($\delta_j = 1$), we assume that Y_{ij} have CDF G_i and are independent from all other $X_{i'j'l}$ and $Y_{i'j'}$ if $j \neq j'$. Thus, the markers may be correlated within patients, but there will be no correlation between patients. The following two theorems provide the methodological framework for comparing the sensitivities of the two

markers across a range of specificities. A methodological approach for the less general problem of making inferences regarding a single marker will be obvious from the work as well.

Theorem 1. Assume the conditions in Section 2.2.1 hold for two markers, indexed by the subscript i. Suppose we want to compare the two markers using the total areas under their ROC curves. Let $P(t) = t$, $t \in (0, 1)$. Also let

$$\Delta = \theta_1 - \theta_2 = \int_0^1 [\text{sen}_1(\text{spe}_1^{-1}(t)) - \text{sen}_2(\text{spe}_2^{-1}(t))]\, \mathrm{d}P(t).$$

Let $\hat{\Delta} = \hat{\theta}_1 - \hat{\theta}_2$ be the corresponding non-parametric estimate of Δ obtained from equation (1). Then, as n tends to ∞, $(\hat{\Delta} - \Delta)/\hat{\sigma}_n$ tends to a normal distribution with mean zero and variance one where

$$\hat{\sigma}_n^2 = \sum_{j=1}^{n} (\hat{\varepsilon}_{1j} + \hat{\xi}_{1j} - \hat{\varepsilon}_{2j} - \hat{\xi}_{2j})^2 \tag{2}$$

$$\hat{\varepsilon}_{ij} = \delta_j (DM)^{-1} \sum_{k}^{n} w_k \sum_{l}^{m_k} \left\{ I(X_{ikl} \leqslant Y_{ij}) - D^{-1} \sum_{j'}^{n} \delta_{j'} I(X_{ikl} \leqslant Y_{ij'}) \right\} \tag{3}$$

and

$$\hat{\xi}_{ij} = w_j (DM)^{-1} \sum_{l}^{m_j} \sum_{k}^{n} \delta_k \left\{ I(X_{ijl} \leqslant Y_{ik}) - M^{-1} \sum_{j'}^{n} w_{j'} \sum_{l'}^{m_{j'}} I(X_{ij'l'} \leqslant Y_{ik}) \right\}. \tag{4}$$

A proof is presented in the Appendix.

Next we will define a statistic to compare the average sensitivities over a range of specificities. Let $P(t) = 0$ if $t \leqslant t_1$; $P(t) = (t - t_1)/(t_2 - t_1)$ if $t_1 < t < t_2$; $P(t) = 1$ if $t \geqslant t_2$. Then we suggest a non-parametric estimate of θ to be

$$\hat{\theta} \sim \int_0^1 \hat{\text{sen}}(\hat{\text{spe}}^{-1}(t))\, \mathrm{d}P(t)$$

$$= (1/M) \sum_{j}^{n} w_j \sum_{l}^{m_j} \hat{\text{sen}}(X_{jl}) I(\hat{\text{spe}}^{-1}(t_1) \leqslant X_{jl} \leqslant \hat{\text{spe}}^{-1}(t_2)). \tag{5}$$

Notice that $\theta(t_1 - t_2)$ in this case will be the area under the ROC curve between t_1 and t_2.

Theorem 2. Suppose we want to compare the partial areas under the two ROC curves when the specificity of interest is the interval $[t_1, t_2]$, where $0 \leqslant t_1 \leqslant t_2 \leqslant 1$. Let $P(t) = 0$ if $t \leqslant t_1$; $P(t) = (t - t_1)/(t_2 - t_1)$ if $t_1 < t < t_2$; $P(t) = 1$ if $t \geqslant t_2$. Also let

$$\Delta = \theta_1 - \theta_2 = \int_0^1 [\text{sen}_1(\text{spe}_1^{-1}(p)) - \text{sen}_2(\text{spe}_2^{-1}(t))]\, \mathrm{d}P(t).$$

Let $\hat{\Delta} = \hat{\theta}_1 - \hat{\theta}_2$ be the corresponding non-parametric estimate of Δ obtained from equation (5). Then, as n tends to ∞ $(\hat{\Delta} - \Delta)/\hat{\sigma}_B$ tends to a normal distribution with mean zero and variance one where we obtain $\hat{\sigma}_B$ by using the bootstrap technique.

An outline of the proof is provided in the Appendix. Note that in the non-repeated case, $\hat{\theta}$ in equations (1) and (5) will reduce to a class of statistics given by Wieand *et al.* [9].

3. SIMULATIONS AND BOOTSTRAP RESULTS

To examine finite sample properties of the hypothesis testing procedure in Section 2, we performed simulations that were similar to the ones done in Emir *et al.* [1]. In summary, in the absence of progression, patients will have two different markers, for example, CEA and CA15-3, obtained every month for a total of at most six monthly visits per patient. For each patient, we generate three independent multivariate normal random vectors $Z_i = (z_{i1}, \ldots, z_{i6})'$, $i = 1, 2, 3$, of size 6×1 with mean vector 0 and $\text{cov}(z_{ij}, z_{ik}) = \rho^{|j-k|}$, for $j, k = 1, \ldots, 6$. We define markers X_1 and X_2 as

$$X_1 = Z_1\sqrt{\lambda} + Z_2\sqrt{(1-\lambda)}$$

$$X_2 = Z_1\sqrt{\lambda} + Z_3\sqrt{(1-\lambda)}.$$

Note λ and ρ are the parameters that, respectively, introduce correlations between markers and between visits. We generate failure times for the patients using an exponential distribution such that the expected failure rate at six months is 90 per cent (a distribution which is consistent with the data used in an example in Section 4).

If a simulated failure time is greater than six months, we define the patient to be a control at all six visits and use all six values of X_1 and X_2 for the marker values. If a simulated failure time occurs before six months, we assume that the failure is detected clinically at the next visit. For example, if a failure occurs between the third and fourth visit, the simulated markers for the first three visits are $(x_{11}, x_{12}, x_{13})'$ and $(x_{21}, x_{22}, x_{23})'$. We assume the expected value of the marker is increased by 1 at the time of failure, hence we define the markers at this fourth visit to be $Y_1 = x_{14} + 1$ and $Y_2 = x_{24} + 1$. With this set-up, both markers are equivalent. Using the above parameters, if we have 100 patients, on average we will have 90 cases and 192 controls.

In all simulations we used two different weight functions: $w_j = \min(0, 1/m_j)$, which gives equal weight to each patient, and $w_j = 1$, which gives equal weight to each marker.

3.1. Total area

We first assume that the correlations $\rho = \lambda = 0$, that is, each of the patients and the marker values are independent of each other. Under the normality assumption we calculate the area to be

$$\theta_i = \int_0^1 \{1 - \Phi(\Phi^{-1}(t) - 1)\}\, \mathrm{d}t = 0.760$$

for $i = 1, 2$, where Φ is the standard normal distribution. If we let $w_j = 1$ be the weight, it can be shown that the asymptotic variance of the area is the same as for the non-repeated marker case with $M = \sum_{j=1}^{n} m_j$ controls and $D = \sum_{j=1}^{n} \delta_j$ cases. For the independent case, we can use the work of DeLong *et al.* [10] and Wieand *et al.* (Wieand, Gail and Hanley, 1983, unpublished) to show that $\text{var}(\hat{\theta}_2 - \hat{\theta}_1) = 0.00183$.

The performed 500 Monte Carlo (MC) simulations to evaluate the performance of the test statistic suggested in theorem 1. Each MC simulation contained a random sample of 100 patients whose marker values and failure times were generated according to the above schema. Total areas were estimated by using equation (1) for both markers and for each MC simulation.

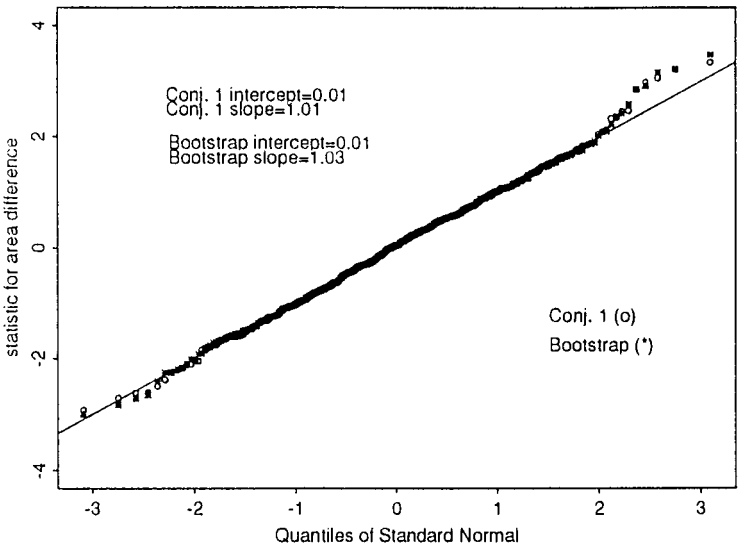

Figure 1. The Q-Q plot of the statistics from 500 Monte Carlo simulations to compare the areas under the ROC curves. The straight line has intercept 0 and slope 1, hence represents the theoretical Q-Q plot for a N(0, 1) distribution.

The variance of the area was estimated by two methods, namely (i) the empirical estimates in (2) to (4), and (ii) by using 250 bootstrap samples for each MC simulation.

Out of 500 MC simulations, the average area estimates were 0.761 and 0.760 for the two markers. The MC simulation average of the variance estimate from theorem 1 was 0.00180 and MC simulation average of the bootstrap variance for the difference was 0.00183. Normal probability (Q-Q) plots of $(\hat{\theta}_2 - \hat{\theta}_1)/\hat{\sigma}$ where $\hat{\sigma}$ is estimated using the empirical estimates and using the bootstrap, respectively, indicate that the standardized statistic is nearly N(0, 1) when the sample size is 100 (see Figure 1). We obtained similar results for different combinations of correlations (ρ, λ) and weights w_j (not presented).

We extend the MC simulations to establish the finite sample behaviour of the test statistic under the null hypothesis. We used random samples of size 30, 50 and 100 patients for each of the MC simulations with a different combination of correlations (ρ, λ) and weights. Each entry in Table I is the proportion of times the absolute value of the t-like statistic exceeds 1.645 and 1.96 out of 2000 MC simulations. For as few as 50 patients the statistic performs well.

Simulated power results are presented in Table II. In this case we used $Y_2 \sim \mathrm{N}(1.65, 1)$ and $X_2 \sim \mathrm{N}(0, 1)$. This alternative was chosen so that the power would be in a range of interest (0.70, 1.00). One would expect that if there were no within (repeated) marker correlation, the weight function $w_j = 1$ would be optimal (since each observation is independent), but that as the correlation approached 1, the weight $w_j = \min(0, 1/m_j)$ might be better. In fact if the correlation was 1.0, the weight $w_j = \min(0, 1/m_j)$ would be optimal, since there would effectively be only one marker per patient. As anticipated the simulations resulted in higher power for the weight $w_j = 1$ when $\rho = 0$ and for the weight $w_j = \min(0, 1/m_j)$ when the $\rho = 0.9$. One would also anticipate that

Table I. Proportion of times $|\sqrt{n}\hat{\Delta}/\hat{\sigma}_n| \geqslant 1.96$ or 1.645 when specificities range in [0, 1].

Nominal type I error	Weights	$\lambda = \rho = 0$			$\lambda = 0.25$, $\rho = 0.9$		
		$n = 30$	$n = 50$	$n = 100$	$n = 30$	$n = 50$	$n = 100$
$\alpha = 0.05$	$\min(0, 1/m_j)$	0.055	0.062	0.053	0.065	0.056	0.053
$\alpha = 0.05$	1	0.059	0.061	0.062	0.065	0.067	0.05
$\alpha = 0.10$	$\min(0, 1/m_j)$	0.119	0.119	0.103	0.120	0.110	0.110
$\alpha = 0.10$	1	0.113	0.117	0.113	0.126	0.122	0.107

Table II. Empirical power levels.

Nominal type I error	Weights	$\lambda = \rho = 0$		$\lambda = 0.25$, $\rho = 0.9$	
		$n = 50$	$n = 100$	$n = 50$	$n = 100$
$\alpha = 0.05$	$\min(0, 1/m_j)$	0.71	0.93	0.86	0.99
$\alpha = 0.05$	1	0.78	0.96	0.76	0.96
$\alpha = 0.10$	$\min(0, 1/m_j)$	0.80	0.96	0.92	1.00
$\alpha = 0.10$	1	0.85	0.97	0.86	0.98

the power would be higher when there was a between-marker correlation ($\lambda \neq 0$) than when $\lambda = 0$ as the variance would be reduced. This too was supported by the simulations.

3.2. *Partial area*

We did a second series of simulations to evaluate the statistic for comparing two markers over the range of specificities (0.6, 0.9) again letting $Y_1 = x_{14} + 1$ and $Y_2 = x_{24} + 1$. We used a random sample of 100 patients for each simulation with a different combination of correlations (ρ, λ) and weights. Each entry in Table III is the proportion of times the absolute value of the t-like statistic exceeds, 1.645 and 1.96 out of 500 MC simulations (using the bootstrap estimate, $\hat{\sigma}_B$).

4. APPLICATION TO BREAST CANCER DATA

Emir *et al.* [1] used data from a sub-study of a randomized clinical trial by the North Central Cancer Treatment Group and Mayo Clinic designed to assess the role of monoclonal antibodies directed against soluble tumour antigens (CEA, CA15-3, TPS) as markers for progression of breast cancer. The design for the randomized trial and results of treatment strategies are discussed in Schaid *et al.* [11] and Ingle *et al.* [12]. All the marker study patients had a lesion which was measurable or evaluable and could be followed for progression. They were to report for a physical examination on a regular schedule (every three to five weeks) at which time their disease status was assessed. During the physical examination, blood was drawn and sent to a central pathology laboratory for analysis. When the blood was analysed a numerical score was obtained for each of

Table III. Proportion of times $|\sqrt{n}\hat{\Delta}/\hat{\sigma}_B| \geqslant 1.96$ or 1.645 when specificities range in [0.8, 1].

Nominal type I error	Weights	$\lambda = \rho = 0$ $n = 100$	$\lambda = 0.25$, $\rho = 0.9$ $n = 100$
$\alpha = 0.05$	$\min(0, 1/m_j)$	0.056	0.048
$\alpha = 0.05$	1	0.056	0.058
$\alpha = 0.10$	$\min(0, 1/m_j)$	0.100	0.092
$\alpha = 0.10$	1	0.122	0.092

the monoclonal antibodies. These scores were not provided to the treating physician, hence played no role in patient management. Conversely, patient characteristics and outcomes were unknown to the laboratory personnel. Over 95 per cent of the patients in the study were followed to progression and the progression-free patients have at least three years of follow-up at this writing.

The analyses in this section utilize the data for monoclonal antibodies directed against these antigens for 89 patients, all of whom had at least a baseline value and one follow up value for each of the three antibodies. These patients had a total of 354 non-progression visits and 47 progression visits. Details regarding different ways to define markers using these antigens appear in Emir *et al.* [1]. For illustrative purposes, we first use a marker defined at each visit by the ratio of CA15-3 attained at that visit to the baseline CA15-3 (the value observed for the same patient at the initiation of the study).

Using equation (1), we calculated total area under the ROC curve for this marker to be 0.58 (95 per cent CI (0.504, 0.653)) using the weight $w_j = \min(0, 1/m_j)$. Note that the 95 per cent CI does not contain 0.5, indicating that CA15-3 has some correlation with outcome, although the area is still much less than one would expect for a very sensitive marker.

When the analysis of CA15-3 as a marker was first proposed, our fellow investigators were unsure whether it would be better to use the value of CA15-3 as observed or the ratio of that value to a baseline (as in the above illustration). To address this question we estimated the area under the ROC curve for the marker defined by the CA15-3 value at each visit; we obtained 0.49. A 95 per cent CI for the difference in the areas is (0.012, 0.159) favouring the ratio method, leading us to the conclusion that one obtained a better marker by using the ratio.

At the initiation of the breast cancer study, the investigators indicated that they were primarily interested in sensitivities of three markers (CA15-3, CEA, and TPS) when specificities were 0.80 or greater (Figure 2).

We calculated average sensitivities of the three markers (actually the ratios to baseline of the markers) for specificities in the range (0.8, 1.0) using equation (5). We performed 1000 bootstrap simulations to estimate the variance as well as constructing a 95 per cent bootstrap CI (BCI) for the average sensitivities, using the weights $w_j = \min(0, 1/m_j)$. Average sensitivities were 0.18 (95 per cent BCI (0.106, 0.277)) for CA15-3, 0.19 (95 per cent BCI (0.119, 0.259)) for CEA and 0.11 (95 per cent BCI (0.056, 0.172)) for TPS. A pairwise comparison yielded two-sided p-values of 0.12 between CEA versus CA15-3, 0.05 between CEA versus TPS, and 0.05 between CA15-3 versus TPS. After seeing these results the investigators concluded that none of the markers would be very useful for their purposes (a marker which had no association with outcome would have an

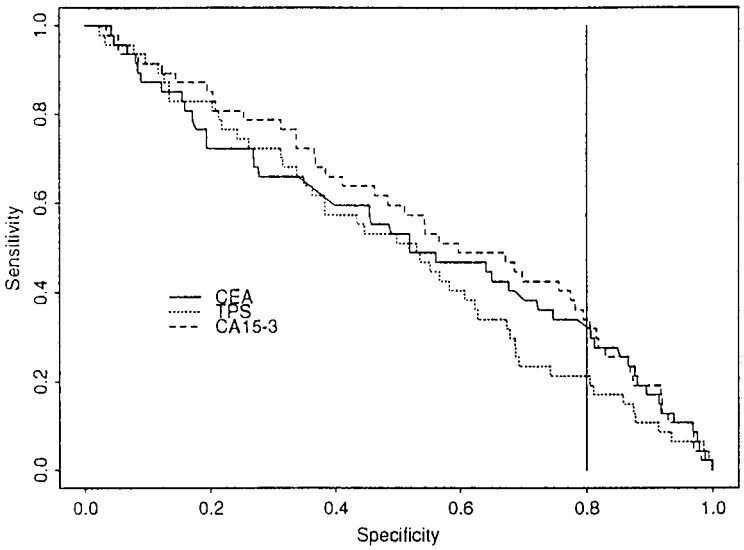

Figure 2. Sensitivity versus specificity curve for CEA, CA15-3 and TPS. Ratio of the markers to their baseline are used. The weight $\min(0, 1/m_j)$ results are presented.

average sensitivity of 0.10 over the specificity range (0.8, 1.0)) but that if they were to carry markers forward for further testing, they would use CA15-3 or CEA.

5. TIES

Even though our methods call for a continuous marker, it is very likely that in practice there will be some tie values. This changes the area computations if any marker values are identical for a non-progression and progression visit, that is, $Y_k = X_{jl}$ for any j, l and k. To adjust for ties we note that in the comparison of Y versus X we assign a score of 1 if $Y_k > X_{jl}$ and 0 otherwise. We extend this to the ties by giving a score of 1/2 when $Y_k = X_{jl}$. To accomplish this we create new control values as follows. We replaced each X_{jl} with X_{jl}^- and X_{jl}^+ where $X_{jl}^- = X_{jl} - \varepsilon$ and $X_{jl}^+ = X_{jl} + \varepsilon$, with ε being a small number say, $\varepsilon = 0.00001$. Then in equation (1) we use these $2M$ controls and let

$$\hat{\theta} = (1/(2DM)) \sum_{j=1}^{n} \sum_{k=1}^{n} w_j \sum_{l=1}^{m_j} \delta_k [I(Y_k > X_{jl}^-) + I(Y_k > X_{jl}^+)]. \tag{6}$$

In the presence of a tie, say $Y_1 = X_{11}$, $I(Y_1 > X_{11}^-) = 1$ and $I(Y_1 > X_{11}^+) = 0$. This returns a score of 1/2 to the comparison of Y_1 versus X_{11}. The bootstrap method can still be applied to estimate the variance.

Although the ratio of CA15-3 to its baseline did not contain a lot of ties (< 1 per cent of the time a control value was equal to a case value) we calculated the area using equation (6) with 1000 bootstraps to obtain an estimate of the variance. Results are similar when compared to the results of equation (1), The area = 0.58 with 95 per cent BCI (0.505, 0.657).

6. DISCUSSION

Wieand *et al.* [9] presented a family of non-parametric statistics for comparing diagnostic markers with paired or unpaired data. For repeated markers the parameters of utility have to be modified. Emir *et al.* [1] suggested new definitions for parameters such as sensitivity and specificity. They also extended the results of Wieand *et al.* [9] to allow comparison of sensitivities at a fixed specificity. The work in this manuscript extends the results of Emir *et al.* [1] to study the average sensitivity over a range of specificities. This applies to the more general definitions discussed in Emir *et al.* [1] to include: forming contrasts to compare more than two markers simultaneously; forming a marker by using a function of several antibodies, or using a marker obtained at time t to predict outcome in a later time interval, for example, 1 to 5 weeks after the marker is observed.

Leisenring *et al.* [13] have published a comprehensive work addressing methods of modelling sensitivity, which includes, as one of several very useful applications, the problem of using marker results at time t to predict outcome at subsequent time points. Their model allows for a diminution or increase in effect over time, so that one could estimate the sensitivity for any time point s following observation of a marker at time t. Their goal was somewhat different from that of Emir *et al.* [1], as Emir *et al.* [1] were defining methods of evaluating a particular clinical strategy, while Leisenring *et al.* [13] provide methods for modelling the specificity and sensitivity as a function of $s - t$.

Although our goal in this paper was to describe how to generalize the sensitivity at fixed specificity to the average range of sensitivities, one of our examples provided insight into the difference between the two weight functions. Generally, the choice of a weight function has a minimal effect on the ROC curves. However, when we computed the total area under the ROC curve for the marker defined by the ratio of the CA15-3 value divided by the baseline value, the two methods differed more than one might expect (0.58 using the weight $w_j = \min(0, 1/m_j)$ and 0.62 using the weight $w_j = 1$). The within-patient correlation was high ($\rho = 0.9$ for adjacent visits) so we would be likely to have more confidence in weight $w_j = \min(0, 1/m_j)$ which gives equal weight to each patient. However, we did not expect a difference of 0.04 between the area estimates for the two weights. Data exploration showed that of 401 visits among the 89 patients, 111 came from seven patients. If these seven patients are excluded, the total areas are 0.57 95 per cent CI (0.495, 0.655)) for the weight $w_j = 1$ and 0.56 (95 per cent CI (0.486, 0.642)) for the weight $w_j = \min(0, 1/m_j)$, while analyses using only data from the seven patients (with 111 visits) results in areas of 0.73 for both weights. These seven patients have much more influence on the analysis using the weight $w_j = 1$ than that using the weight $w_j = \min(0, 1/m_j)$, which is why the weight $w_j = 1$ estimate is larger.

APPENDIX: PROOF

Let $\hat{G}_i(t) = D^{-1}\sum_{j=1}^{n} \delta_j I(Y_{ij} \leqslant t)$, and, for weights $(w_j, j = 1, \ldots, n)$, let

$$\hat{F}_i(t) = M^{-1} \sum_{j=1}^{n} w_j \sum_{l=1}^{m_j} I(X_{ijl} \leqslant t)$$

with $M = \sum_{j=1}^{n} m_j w_j$. Note that, for the pooled weight, $w_j = 1$; for the indicator weight, $w_j = \min(m_j, 1/m_j)$. Also let

$$\hat{\theta}_i = \frac{1}{DM} \sum_{j=1}^{n} \sum_{k=1}^{n} w_k \sum_{l=1}^{m_k} \delta_j I(X_{ikl} \leqslant Y_{ij})$$

$$= \int_0^{\infty} \hat{F}_i(t)\, \mathrm{d}\hat{G}_i(t)$$

be the area under ROC curve and $\theta_i = \int_0^{\infty} F_i \mathrm{d}G_i$. Now

$$\hat{\theta}_i - \theta_i = \int_0^{\infty} \hat{F}_i(t)\mathrm{d}\hat{G}_i(t) - \theta_i$$

$$= \int_0^{\infty} (\hat{F}_i - F_i)\mathrm{d}(\hat{G}_i - G_i) + \int_0^{\infty} F_i \mathrm{d}(\hat{G}_i - G_i) + \int_0^{\infty} (\hat{F}_i - F_i)\mathrm{d}G_i$$

$$\equiv W_1 + W_2 + W_3.$$

At first, we want to show that W_1 is the order $o_p(1/\sqrt{n})$. We know that $\hat{F}_i$ converges uniformly to F_i and that $\sqrt{n}(\hat{G}_i - G_i)$ converges weakly to a Gaussian process U_i with continuous sample path. So the strong embedding theorem (Shorack *et al.* [14], p. 47, Theorem 4) implies that, in a new probability space, $\{\hat{F}_i, \sqrt{n}(\hat{G}_i - G_i)\}$ converges almost surely to (F_i, U_i) in sup norm. In the new probability space, we apply integration by parts to get

$$W_1 = \{\hat{G}_i(t) - G_i(t)\}\{\hat{F}_i(t) - F_i(t)\}|_0^{\tau} - \int_0^{\tau} \{\hat{G}_i(t) - G_i(t)\}\, \mathrm{d}\{\hat{F}_i(t) - F_i(t)\}$$

$$= n^{-1/2} \int_0^{\tau} U(t)\mathrm{d}\{\hat{F}_i(t) - F_i(t)\} + o(n^{-1/2})$$

$$= o(n^{-1/2})$$

where the last equality follows from Helly's theorem (Serfling [15], p. 532). Hence, in the original space, $W_1 = o_p(n^{-1/2})$. Furthermore, $W_2 = \sum_j^n \varepsilon_{ij}$ and $W_3 = \sum_j^n \xi_{ij}$, where

$$\varepsilon_{ij} = \frac{\delta_j}{D} \int_0^{\infty} F_i(t)\mathrm{d}\{I(Y_{ij} \leqslant t) - G_i(t)\}$$

$$\xi_{ij} = \frac{w_j}{M} \sum_{l}^{m_j} \int_0^{\infty} \{I(X_{ijl} \leqslant t) - F_i(t)\}\mathrm{d}G_i(t).$$

Hence

$$\hat{\theta}_i - \theta_i \approx \sum_{j=1}^{n} (\varepsilon_{ij} + \xi_{ij}).$$

Under H_0:$\theta_1 = \theta_2$

$$\hat{\theta}_1 - \hat{\theta}_2 = (\hat{\theta}_1 - \theta_1) - (\hat{\theta}_2 - \theta_2) \approx \sum_{j=1}^{n} (\varepsilon_{1j} + \xi_{1j} - \varepsilon_{2j} - \xi_{2j}).$$

Since $\{(\varepsilon_{1j}, \xi_{1j}, \varepsilon_{2j}, \xi_{2j}), \quad j = 1, \ldots, n\}$ are zero-mean independent random vectors, $(\varepsilon_{1j} + \xi_{1j} - \varepsilon_{2j} - \xi_{2j}, \mathrm{j} = 1, \ldots, n)$ are zero-mean independent random variables. By the central limit theorem, under H_0, $\hat{\theta}_1 - \hat{\theta}_2$ is approximately normal with mean 0. Its variance can be estimated by

$$\hat{\sigma}_n^2 = \sum_{j=1}^{n} (\hat{\varepsilon}_{1j} + \hat{\xi}_{1j} - \hat{\varepsilon}_{2j} - \hat{\xi}_{2j})^2$$

where $\hat{\varepsilon}_{ij}$ and $\hat{\xi}_{ij}$ are obtained from ε_{ij} and ξ_{ij} by replacing F_i and G_i with $\hat{F}_i$ and $\hat{G}_i$.

The proof in the partial area case is similar but more tedious. In this case, it can be shown that

$$\theta = \Pr(Y > X, F^{-1}(t_1) < X < F^{-1}(t_2))$$

can be written as

$$\theta = \int_{F^{-1}(t_1)}^{F^{-1}(t_2)} F(t)\mathrm{d}G(t) + (t_2 - t_1) + t_1 G(F^{-1}(t_1)) - t_2 G(F^{-1}(t_2))$$

and that $\hat{\theta}$ from equation (2) is equivalent to

$$\int_{\hat{F}^{-1}(t_1)}^{\hat{F}^{-1}(t_2)} \hat{F}(t)\mathrm{d}\hat{G}(t) + (t_2 - t_1) + t_1 \hat{G}(\hat{F}^{-1}(t_1)) - t_2 \hat{G}(\hat{F}^{-1}(t_2))$$

where $\hat{F}^{-1}(.)$ is empirical estimate defined in Emir *et al.* [1]. If we again express the $\int_{F^{-1}(t_1)}^{F^{-1}(t_2)} F(t)\mathrm{d}G(t)$ in the form $W_1 + W_2 + W_3$ we can show that $W_1 \to 0$ faster than W_2 and W_3 and use a proof of asymptotic normality similar to that shown above. Unfortunately the resulting variance estimate involves numerous terms and is not robust for small samples, so we use the bootstrap estimator of the variance term in the applications.

REFERENCES

1. Emir B, Wieand S, Su J, Cha S. Analysis of repeated markers used to predict progression of cancer. *Statistics in Medicine* 1998; **17**:2563–2578.
2. Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**:187–220.
3. Swets JA, Pickett RM. *Evaluation of Diagnostic Systems*. Academic Press: New York, 1982.
4. Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. *Investigative Radiology* 1979; **14**:109–121.
5. Metz CE. Basic principles of ROC analysis. *Seminars in Nuclear Medicine* 1978; VII (4):283–298.
6. Metz CE. ROC methodology in radiologic imaging. *Investigative Radiology* 1986; **21**(9):720–733.
7. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; **143**:29–36.
8. Hanley JA, McNeil BJ. A method of comparing the areas under a receiver operating characteristics curves derived from the same cases. *Radiology* 1983; **148**:839–843.
9. Wieand S, Gail MH, James BR, James KL. A family of non-parametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 1989; **76**:585–592.

10. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristics curves: a nonparametric approach. *Biometrics* 1988; **44**:837–845.
11. Schaid DJ, Ingle JN, Wieand S, Ahmann DL. A design for phase II testing of anticancer agents within a phase III clinical trial. *Controlled Clinical Trials* 1988; **9**:107–118.
12. Ingle JN, Ritts RE, Wieand HS, Foley JF. Evaluation of a panel of potential serum tumor markers in women with meta-static breast cancer entered on a prospective chemotherapy clinical trial, a North Central Cancer Treatment Group Study. Proceedings of the XXIIIrd Meeting of the International Society for Oncodevelopmential Biology and Medicine, Montreal, Canada, 1995, 15.
13. Leisenring W, Pepe MS, Longton G. A marginal regression modeling framework for evaluating medical diagnostic tests. *Statistics in Medicine* 1997; **16**:1263–1281.
14. Shorack GR, Wellner GJA. *Empirical Processes with Applications to Statistics*. Wiley: New York, 1986.
15. Serfling RJ. *Approximation Theorems of Mathematical Statistics*. Wiley: New York, 1980.

