STATISTICS IN MEDICINE
Statist. Med. **17**, 2563–2578 (1998)

ANALYSIS OF REPEATED MARKERS USED TO PREDICT PROGRESSION OF CANCER

BIROL EMIR[1]*, SAM WIEAND[2], JOHN Q. SU[3] AND STEVE CHA[4]

[1] *Department of Statistics, Iowa State University, Ames, IA 50011, U.S.A.*
[2] *Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, U.S.A.*
[3] *Genentech INC, 460 Point San Bruno Blvd., South San Francisco, CA 94080, U.S.A.*
[4] *Cancer Center Statistics, Mayo Clinic, Rochester, MN 55905, U.S.A.*

SUMMARY

We consider methods for evaluating repeated markers to be used as a substitute for a clinical examination or to predict an outcome, in our case progression of breast cancer. We propose a definition of specificity and sensitivity for this setting and describe non-parametric estimators for these parameters. We then derive the theory required to obtain confidence intervals for the specificity and sensitivity of a marker and to define an asymptotically normal statistic for comparing the sensitivities of two markers at a fixed specificity. 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. The work allows for an approach that complements the frequently used time dependent Cox model, which we believe, will facilitate clinical interpretation of marker data. © 1998 John Wiley & Sons, Ltd.

1. INTRODUCTION

A problem of increasing interest to clinical trialists is the evaluation of repeated markers to be used as a supplement to or to replace a clinical examination or to predict an outcome. Perhaps the most common approach to this problem is to use repeated markers as an independent variable in a Cox[1,2] model with outcome as a dependent variable. The estimate of relative hazard associated with the marker becomes the basis for inference statements. This method is quite appropriate for determining if a marker is biologically associated with outcome. However, our experience has been that such analyses often do not meet the clinicians' need as relative hazard is not an optimal parameter for evaluating the utility of the marker.

This leads to two questions. The first is why we routinely use relative hazard in this setting. The second is whether a large relative hazard translates into acceptably high values of sensitivity, specificity and predictive values, the parameters normally used to evaluate the utility of a marker. We believe the answer to the first question follows from the fact that most clinical trialists are used to designing clinical trials to compare treatments, and, hence, become confortable using relative risk as a measure. This is particularly true when the endpoint is time to an event.

* Correspondence to: Birol Emir, Bayer Pharmaceutical Division, S&DS, Building 32, 400 Morgan Lane, West Haven, CT 06516, U.S.A.

CCC 0277–6715/98/222563–16$17.50

Received January 1997
Accepted February 1998

Table I. A hypothetical example for estimating relative risk

	Case	No case
Treatment A	30	30
Treatment B	10	50

Table II. Same hypothetical example for estimating parameters to evaluate the utility of a marker

	True positives	True negatives
Marker positive	30	30
Marker negative	10	50

This approach has carried over to the evaluation of repeated markers, in part because the time-dependent Cox model is an available tool which can easily be applied to estimate a relative hazard for a repeated marker.

The answer to the second question is that a high relative risk does not automatically translate into acceptable levels of sensitivity, specificity and predictive values, as is illustrated by the following hypothetical example.

Suppose 60 patients are randomized to each of two treatments and observed outcomes are as shown in Table I. If one defines the estimated relative risk associated with treatment A to be

$$\frac{\text{cases on treatment A}}{\text{patients on treatment A}} \times \frac{\text{patients on treatment B}}{\text{cases on treatment B}}$$

(which is analogous to relative hazard in a survival model), the relative risk is 3. This would almost certainly be sufficient information to conclude that treatment B is preferable to treatment A (in the absence of dangerous toxicities).

Suppose we consider the same numbers in the marker setting as illustrated in Table II. Of course, the relative risk associated with a positive marker is still 3. Notice, however, that the estimated specificity ((marker negatives $\cap$ true negatives)/true negatives) of the marker would be 0·625, sensitivity ((marker positives $\cap$ true positives)/true positives) would be 0·75 and positive predictive value ((marker positive $\cap$ true positives)/marker positives) would be 0·50. In many clinical settings, investigators might consider the latter three values to be too low for this marker to be useful. Thus, the large relative risk does not translate into sufficiently high measures of utility. We believe this example helps explain why clinicians are often disappointed when they use a marker that has been shown to have a highly significant elevated risk in previous trials as a diagnostic tool.

In Section 2, we propose definitions of specificity, sensitivity, and positive and negative predictive values which we have found to be applicable to the repeated marker problem. We will also provide methods for making inferences regarding these parameters. In Section 3, we will address the comparison of two continuous markers to determine if one is superior using the notions of sensitivity and specificity, a problem which was of a particular interest to our clinical

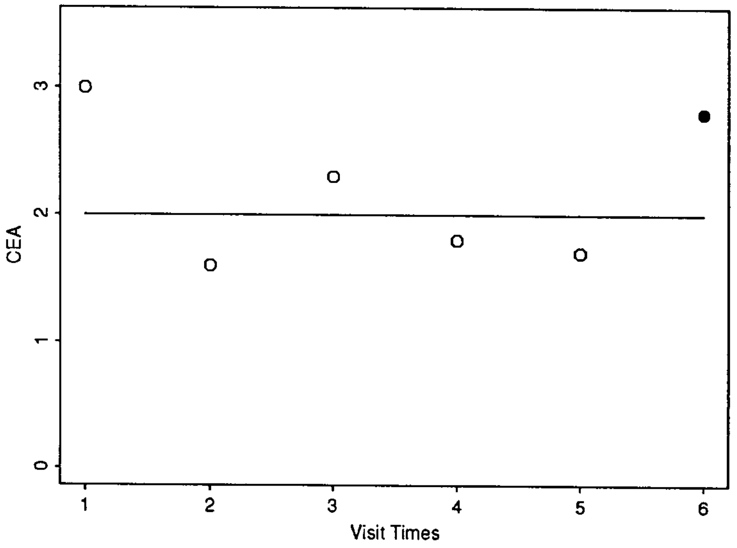

Figure 1. Marker values at 6 visits for one patient. Open circles denote the control visits, and solid circle denotes a progression visit

colleagues. In subsequent sections and appendices we will present the asymptotic and finite sample properties of the statistic designed to address this problem, apply our results to data from a breast cancer trial and discuss simulations and generalizations of our work.

2. PROPOSED DEFINITIONS

In this section we propose definitions of sensitivity, specificity and predictive values when markers are repeated. We will also show how one can obtain confidence intervals for the above parameters. For simplicity, we will first assume that we have only one dichotomous marker (which could be formed by assigning a fixed cut-off to a continuous marker).

We start with an example. Suppose that a cancer patient has returned for six visits after entering a study, and the continuous marker (say, CEA) values obtained on these visits were 3·0, 1·6, 2·3, 1·8, 1·7 and 2·8. We define the marker to be positive, that is, predictive of cancer, when CEA is $\geqslant 2$ and assume that the patient recurred at the time the 2·8 was observed (Figure 1). The marker would be classified as positive at two of five non-progression visits and at the progression visit. For this patient, we would estimate the specificity to be $3/5 = 0{\cdot}6$, spe = pr(marker negative|physical exam negative), and the sensitivity to be $1/1 = 1{\cdot}0$, sen = pr(marker positive|physical exam positive). Also note that this patient would have one true positive out of three positive marker values and have all three negative marker values correctly classified. Hence, this patient's estimated positive predictive value would be $1/3 = 0{\cdot}33$, ppv = pr(physical exam positive|marker positive), and estimated negative predictive value would be $3/3 = 1{\cdot}0$, npv = pr(physical exam negative|marker negative). Restricting our attention to only sensitivity and specificity, we can summarize the data from this patient with an outcome vector as $(0{\cdot}6, 5, 1{\cdot}0, 1)'$ an observation of a random vector $\mathbf{S} = (\hat{\text{spe}}, N_1, \hat{\text{sen}}, N_2)'$ where $\hat{\text{spe}}(\hat{\text{sen}})$ represents the patient's estimated specificity (sensitivity) and N_1 and N_2 represent the number of times the patient is a *control* (non-progression) and a *case* (progression), respectively. (If we want to make inferences about predictive values rather than sensitivity and specificity, then we can summarize the data from this patient as $(1{\cdot}0, 3, 0{\cdot}33, 3)'$, an observation from a random vector $\mathbf{P} = (\text{n}\hat{\text{p}}\text{v}, M_1, \text{p}\hat{\text{p}}\text{v}, M_2)'$ where $M_1(M_2)$ represents the total number of times the marker is *negative* (*positive*) etc.)

We will be interested in making inferences for the above parameters from a random sample of n patients. For patient j we let $\mathbf{S_j} = (\text{s}\hat{\text{p}}\text{e}_j, N_{1j}, \text{s}\hat{\text{e}}\text{n}_j, N_{2j})'$. We assume that $(\mathbf{S_1}, \dots, \mathbf{S_n})$ or $(\mathbf{P_1}, \dots, \mathbf{P_n})$ are n identically and independently distributed (i.i.d.) random vectors and that the first two moments of N_{1j}, N_{2j}, M_{1j}, and $M_{2j}, j = 1, \dots, n$ are finite. We define the sample estimate of sensitivity, specificity, positive predictive value, and negative predictive value from n patients as

$$\text{s}\hat{\text{e}}\text{n} = \left(1/\sum V_j\right)\sum V_j\, \text{s}\hat{\text{e}}\text{n}_j,$$

$$\text{s}\hat{\text{p}}\text{e} = \left(1/\sum W_j\right)\sum W_j\, \text{s}\hat{\text{p}}\text{e}_j,$$

$$\text{p}\hat{\text{p}}\text{v} = \left(1/\sum V'_j\right)\sum V'_j\, \text{p}\hat{\text{p}}\text{v}_j,$$

$$\text{n}\hat{\text{p}}\text{v} = \left(1/\sum W'_j\right)\sum W'_j\, \text{n}\hat{\text{p}}\text{v}_j \tag{1}$$

where the weights (W_j, V_j, W'_j, V'_j) are bounded functions of $(N_{1j}, N_{2j}, M_{1j}, M_{2j})$, respectively. For example, we can define $W_j = N_{1j}$ or $W_j = I(N_{1j})$, where $I(x) = 1$ if $x > 0$ and 0 otherwise. We require that $W_j(V_j) = 0$ if $N_{1j}(N_{2j}) = 0$, as $\text{s}\hat{\text{p}}\text{e}_j(\text{s}\hat{\text{e}}\text{n}_j)$ are undefined if $N_{1j}(N_{2j}) = 0$. Then, using independence across patients we can show that as n goes to ∞, $\text{s}\hat{\text{p}}\text{e}$ will have an asymptotic normal distribution with mean spe and variance

$$\text{var}(\text{s}\hat{\text{p}}\text{e}) = (1/E(W_j))^2\,\text{var}((1/n)\textstyle\sum W_j^2 \text{spe}_j) + (\text{spe}/E(W_j))^2\,\text{var}((1/n)\sum W_j)$$
$$- 2(\text{spe}/(E(W_j))^2)\text{cov}((1/n)\textstyle\sum W_j \text{spe}_j, (1/n)\sum W_j).$$

A non-parametric estimate of the variance can be obtained by replacing spe with $\text{s}\hat{\text{p}}\text{e}$, $E(W_j)$ with $(1/n)\sum W_j$, $\text{var}(W_j)$ with sample variance of W_j's etc. Similar results will hold for sensitivity and predictive values.

3. COMPARISON OF MARKERS

Our next goal is to address the question of how one compares two continuous markers using the notion of sensitivity and specificity. An approach of interest to clinical colleagues of the authors was to compare the sensitivities of the two markers when they both had the same specificity.

Let X_{jk} be the continuous random variable whose observations are the marker values obtained from jth patient, $j = 1, \dots, n$, at the kth non-progression evaluation, (determined by the gold standard (physical exam)), $k = 1, \dots, N_{1j}$, where N_{1j} is the number of non-progressions for patient j. Similarly, let Y_{jl} be the continuous random variable associated with values for the same marker from the jth patient at the lth progression evaluation where $l = 1, \dots, N_{2j}$ and N_{2j} is the number of progressions for patient j. (In our examples we are only interested in first progression, so N_{2j} is 0 or 1, but this restriction is not required). Let F and G be the distribution functions of X_{jk} and Y_{jl}. This implies X_{jk}, $k = 1, \dots, N_{1j}$ are identically distributed but not necessarily independent. Similarly, Y_{jl}, $l = 1, \dots, N_{2j}$ are identically distributed but not necessarily independent. Assume that if the value of X_{jk} or Y_{jl} exceeds a predetermined cut-off point c the marker will be considered positive (classifying the patient as a progression at the k (or l)th evaluation). Then the specificity spe, and sensitivity sen of the marker are spe $= F(c)$ and sen $= 1 - G(c)$, respectively. We can obtain a non-parametric estimate $\hat{\text{spe}}_j(\cdot)$ of $F(\cdot)$ from the jth patient, namely $\hat{\text{spe}}_j(x) = (1/N_{1j})\sum I(X_{jk} \leqslant x)$. Similarly, let $\hat{\text{sen}}_j(x) = (1/N_{2j})\sum I(Y_{jl} > x)$ be a non-parametric estimate of $H(\cdot) \equiv 1 - G(\cdot)$ from the jth patient.

If we have two markers, we define a random vector for each patient j, as

$$\{\hat{\text{spe}}_{1j}(c_1), \hat{\text{spe}}_{2j}(c_2), N_{1j}, \hat{\text{sen}}_{1j}(c_1), \hat{\text{sen}}_{2j}(c_2), N_{2j}\}'$$

for $j = 1, \dots, n$. These n vectors are independently and identically distributed. We define the sample specificity $\hat{\text{spe}}_i(c_i)$, and sample sensitivity $\hat{\text{sen}}_i(c_i)$, estimates of $F_i(c_i)$ and $H_i(c_i)$ for marker i, as

$$\hat{\text{spe}}_i(c_i) = \left(1/\sum W_j\right)\sum W_j \hat{\text{spe}}_{ij}(c_i) \tag{2}$$

$$\hat{\text{sen}}_i(c_i) = \left(1/\sum V_j\right)\sum V_j \hat{\text{sen}}_{ij}(c_i) \tag{3}$$

where the weights (W_j, V_j) are bounded functions of N_{1j}, N_{2j}, respectively. Note that $\hat{\text{spe}}_i(c_i)$, and $\hat{\text{sen}}_i(c_i)$ are pooled estimates of $F_i(c_i)$ and $H_i(c_i)$ based upon all patients.

We are ready to compare sensitivities of two markers at a fixed specificity p, $0 < p < 1$. Let c_i satisfy $F_i(c_i) = p$, $i = 1, 2$. We estimate $c_i = F_i^{-1}(p)$ $i = 1, 2$, by

$$\hat{c}_i = \inf(x : \hat{\text{spe}}_i(x) > p) \tag{4}$$

where $\hat{\text{spe}}_i$ is given by equation (2).

Theorem 3.1. Assume that F_i and G_i are twice differentiable at c_i with $F_i'(c_i) = f_i(c_i) > 0$ and $G_i'(c_i) = g_i(c_i) > 0$, $i = 1, 2$. Let n be the number of patients being followed and $\mathscr{K}$ be the maximum possible number of evaluations per patient. Also let $\hat{\text{spe}}_i(c_i)$ (2) and $\hat{\text{sen}}_i(c_i)$ (3) be the estimated specificity and sensitivity, respectively, for the ith marker and let $\hat{c}_i$ be the estimate of c_i given by (4), $i = 1, 2$. Furthermore assume that $(W_j, V_j) = \{\Psi_1(N_{1j}), \Psi_2(N_{2j})\}$ such that $E\{\Psi_i(N_{ij})\} > 0$ for $i = 1, 2$ where $\Psi_1(\cdot)$, and $\Psi_2(\cdot)$ are bounded non-negative functions satisfying $\Psi(0) = 0$. Then, letting $\hat{\Delta} = \hat{\text{sen}}_2(\hat{c}_2) - \hat{\text{sen}}_1(\hat{c}_2)$ and $\Delta = \text{sen}_2(c_2) - \text{sen}_1(c_1) = H_2\{F_2^{-1}(p)\} - H_1\{F_1^{-1}(p)\}$

$$\tau = \frac{n^{1/2}(\hat{\Delta} - \Delta)}{\sqrt{V(c_1, c_2)}} \sim N(0, 1) \tag{5}$$

as $n \to \infty$ where

$$\begin{aligned} V(c_1, c_2) = {} & \{h_1(c_1)/f_1(c_1)\}^2 \mathrm{var}\{\hat{\mathrm{spe}}_1(c_1)\} + \{h_2(c_2)/f_2(c_2)\}^2 \mathrm{var}\{\hat{\mathrm{spe}}_2(c_2)\} \\ & + \mathrm{var}\{\hat{\mathrm{sen}}_1(c_1)\} + \mathrm{var}\{\hat{\mathrm{sen}}_2(c_2)\} - 2\mathrm{cov}\{\hat{\mathrm{sen}}_1(c_1), \hat{\mathrm{sen}}_2(c_2)\} \\ & - 2\{h_1(c_1)h_2(c_2)\}/\{f_1(c_1)f_2(c_2)\}\mathrm{cov}\{\hat{\mathrm{spe}}_1(c_1), \hat{\mathrm{spe}}_2(c_2)\} \\ & + 2\{h_1(c_1)/f_1(c_1)\}[\mathrm{cov}\{\hat{\mathrm{spe}}_1(c_1), \hat{\mathrm{sen}}_2(c_2)\} - \mathrm{cov}\{\hat{\mathrm{spe}}_1(c_1), \hat{\mathrm{sen}}_1(c_1)\}] \\ & + 2\{h_2(c_2)/f_2(c_2)\}[\mathrm{cov}\{\hat{\mathrm{spe}}_2(c_2), \hat{\mathrm{sen}}_1(c_1)\} - \mathrm{cov}\{\hat{\mathrm{spe}}_2(c_2), \hat{\mathrm{sen}}_2(c_2)\}] \end{aligned}$$

and

$$H_i(\cdot) \equiv 1 - G_i(\cdot), \text{ and } h_i(x) = \partial H_i(x)/\partial x,\ i = 1, 2.$$

Variance and covariance terms and the suggested empirical estimates of those terms are presented in Appendix I. From those estimates one can obtain a consistent estimator, $\hat{V}(c_1, c_2)$, of $V(c_1, c_2)$. Then τ can be used to test the hypothesis that $\Delta = 0$ versus various alternatives such as $\Delta \neq 0$. The outline of the proof of the theorem is presented in Appendix II.

4. FINITE SAMPLE PROPERTIES OF τ

Assume in the absence of progression, patients will have two markers obtained every month for a total of at most six monthly visits per patient. For each patient, we generated 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 variance–covariance matrix Σ with $\mathrm{cov}(z_{ij}, z_{ik}) = \rho^{|j-k|}$, for $j, k = 1, \ldots, 6$, for $0 < \rho < 1$.

We define markers X_1 and X_2 as

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

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

Then λ is the between-marker correlation and Σ is the within-marker variance–covariance matrix.

In the study presented in Section 5, 30 per cent of the cancer patients were progression-free at six months. Hence, we generated failure times for the patients using an exponential distribution such that expected failure rate at six months was 70 per cent. (The exponential model was chosen because it has been our experience that the hazard rate for advanced breast cancer patients remains fairly constant during the first six months.)

We assume that patients are followed for six monthly visits and a marker is obtained at each visit. 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 $Y_2 = x_{24} + 1$.

In our example we set $n = 30$, 50, and 100 individuals with $(\rho, \lambda) = (0, 0)$ and $(0{\cdot}9, 0{\cdot}25)$. The last choice of (ρ, λ) most closely matches the correlation in the breast cancer data. Table III shows the

Table III. Proportion of times $|\tau| \geqslant 1{\cdot}96$ or $1{\cdot}645$ when $\Delta = 0$

Nominal type I error	Weights	$\lambda = \rho = 0$			$\lambda = 0{\cdot}25$ and $\rho = 0{\cdot}9$		
		$n = 30$	$n = 50$	$n = 100$	$n = 30$	$n = 50$	$n = 100$
$\alpha = 0{\cdot}05$	$I(N_{ij})$	0·072	0·049	0·045	0·047	0·041	0·053
$\alpha = 0{\cdot}05$	N_{ij}	0·068	0·067	0·053	0·061	0·048	0·052
$\alpha = 0{\cdot}10$	$I(N_{ij})$	0·130	0·094	0·089	0·085	0·090	0·095
$\alpha = 0{\cdot}10$	N_{ij}	0·105	0·126	0·101	0·101	0·094	0·102

Table IV. Empirical power levels

Nominal test size	$\rho = 0{\cdot}0$ $\lambda = 0{\cdot}0$		$\rho = 0{\cdot}0$ $\lambda = 0{\cdot}25$		$\rho = 0{\cdot}9$ $\lambda = 0{\cdot}0$		$\rho = 0{\cdot}9$ $\lambda = 0{\cdot}25$	
	$I(N_{ij})$	N_{ij}	$I(N_{ij})$	N_{ij}	$I(N_{ij})$	N_{ij}	$I(N_{ij})$	N_{ij}
5%	0·82	0·83	0·88	0·91	0·83	0·77	0·86	0·81
10%	0·89	0·90	0·93	0·94	0·89	0·84	0·92	0·89

proportion of times out of 1000 simulations that $|\tau| \geqslant 1{\cdot}96$ and $|\tau| \geqslant 1{\cdot}645$ when $\Delta = 0$ with both indicator weights $(W_j, V_j) = \{I(N_{1j}), I(N_{2j})\}$ and pooled weights $(W_j, V_j) = (N_{1j}, N_{2j})$.

Our conclusion from the above simulations is that the finite sample behaviour of τ is well approximated by asymptotic theory for sample sizes of 50 or more as nearly all the empirical rejection rates are within two standard deviations of the theoretical values (0·05 and 0·10) for the 1000 simulations. This remains true in the presence of correlations within (repeated) markers and between markers.

An issue of secondary interest was the relationship between weight functions and the power. One would anticipate that if there is no within (repeated) marker correlation, the weight function $(W_j, V_j) = (N_{1j}, N_{2j})$ would be optimal (since each observation is independent), while if the correlation is 1·0, the weight $(W_j, V_j) = \{I(N_{1j}), I(N_{2j})\}$ would be optimal, since there would effectively be only one marker per patient. To obtain the entries in Table IV, we simulated data using the same parameters as in Table III, except that we increased the expected value of the second marker to 1·77 at the time of failure. For example, if a patient failed at the fourth visit we let $Y_1 = x_{14} + 1$, but $Y_2 = x_{24} + 1{\cdot}77$. (The choice of 1·77 kept powers between 0·75 and 0·95 in Table IV). As anticipated the simulations result in higher power for the weight function $(W_j, V_j) = (N_{1j}, N_{2j})$ when the correlation is 0 and for $(W_j, V_j) = \{I(N_{1j}), I(N_{2j})\}$ when the correlation is near 1 (0·9).

The observed power is higher when there is a positive between-marker correlation than when there is no such correlation, as the theory would predict for paired data.

We did more simulations with other parameters to incorporate different schemes. For example, we used three visits instead of six visits per patient and we assumed a different failure proportion (50 per cent instead of 70 per cent). Some of these simulations were based on a colon cancer study which was not presented here. In all cases, results were similar to those already presented.

5. DATA AND APPLICATION

The North Central Cancer Treatment Group (NCCTG) and Mayo Clinic recently completed a study designed to permit an assessment of the role of monoclonal antibodies directed against soluble tumour antigens (CEA, CA15-3, TPS) as markers for progression of breast cancer. The marker study patients comprised a subset of the participants in a randomized trial which studied the efficacy of four chemotherapy treatment strategies. The design for the randomized trial and results of treatment strategies are discussed in Schaid *et al.*[3] and Ingle *et al.*[4]

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 the monoclonal antibodies. These scores were not provided to the treating physician, and 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.

Three soluble tumour antigens of primary interest to the investigators were CEA, which was commonly used as a cancer marker at the initiation of the study, CA15-3, and TPS. The latter two markers were new and had been promising in pilot studies. 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.

The first endpoint we address is the per visit specificity and sensitivity of a marker as a predictor for a detectable progression at that visit. For this endpoint, the result of a physical examination, progression or not, defines true positive or true negative, an applicable definition if the ultimate goal is to use the marker as a substitute for a physical examination. Initially, we define our marker value at each visit to be the ratio of the antibody at that visit divided by the baseline antibody value. Figure 2 presents the CA15-3 ratio values for all patients during the first ten visits. At each visit, the number of CA15-3 antibodies obtained with ratio to baseline greater than one are also provided.

There are several potential sources of bias associated with this choice of a data set, including subset selection and missing value. We will ignore these issues for now, as our immediate goal is to indicate how to apply our methods and to contrast our methods with the Cox model approach. However, we will address them in Section 6.

5.1. Comparison of the Cox model approach to the methods of Section 2

For each of the markers CEA, TPS and CA15-3, a time-dependent covariate was defined by letting $Z(t) = 0$ if the last marker ratio (current value divided by its baseline) was less than 1·5 and $Z(t) = 1$ otherwise. A patient did not enter the risk set until a marker ratio was available. We then performed three analyses, one for each marker, by using $Z(t)$ as a covariate in a time-dependent Cox model. We obtained relative risks of 2·3 ($p = 0{\cdot}0002$), 1·5 ($p = 0{\cdot}06$) and 1·3 ($p = 0{\cdot}15$) for the increased ratio of CA15-3, CEA and TPS, respectively, and concluded that CA 15-3 was biologically associated with progression of disease while there was a suggestion that CEA and TPS had such an association as well. The results for CA15-3 were encouraging and we applied the

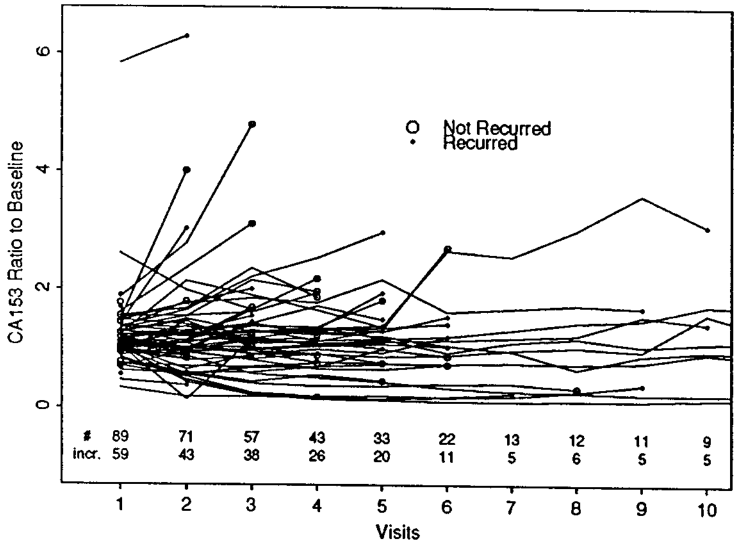

Figure 2. CA15-3 ratio values for all patients during the first 10 visits. At each visit, the number of CA15-3 antibodies obtained with ratio to baseline greater than one are also provided

procedures described in Section 2 to the CA15-3 data to see if the relative risk of 2·3 translated into high measures of clinical utility.

Our estimate of specificity was 0·82 with 95 per cent confidence interval (0·74, 0·89); estimated sensitivity was 0·30, with confidence interval (0·17, 0·43). The estimated positive predictive value of the marker ratio was a disappointing 0·27 with confidence interval (0·21, 0·33), indicating that the marker ratio (using a cut point of 1·5) would be an unsatisfactory substitute for a clinical examination. Thus, application of the methods of Section 2 gave a new and strikingly different perspective to the CA15-3 data than one might have gained from analysis using only the time dependent Cox model.

The largest CA15-3 ratio (6·30) occurred for a case, thus the empirical estimate of the positive predictive value for a cut-off that classified only one patient as a case was 1 (with negative predictive value 0·81). Such a high cut-off would be of minimal interest since the sensitivity (using this cut-off) would be 0·02. For all other cut-offs, the positive predictive value was 0·5 or less.

Note that in all of our calculations we use the weight $(W_j, V_j) = \{I(N_{1j}), I(N_{2j})\}$ as our markers had within-marker correlations near 0·9.

5.2. Two applications of the methods in Section 3

One might question why we used the ratio of the antibody value to its baseline value for our marker rather than the actual value of the antibody. Again, using the CA15-3 data we plotted the sensitivity estimates for both possible markers across all cut-offs (Figure 3). For nearly all specificities, the sensitivities are higher using the ratio than using the original value. Applying the methods of Section 3, we find that these differences in sensitivity are statistically significant (two-sided $p < 0{\cdot}05$) for all specificities between 0·65 and 0·85. For example, at specificity 0·80 the sensitivity of the ratio approach is 0·34 (using a cut-off of 1·46) while the sensitivity using the

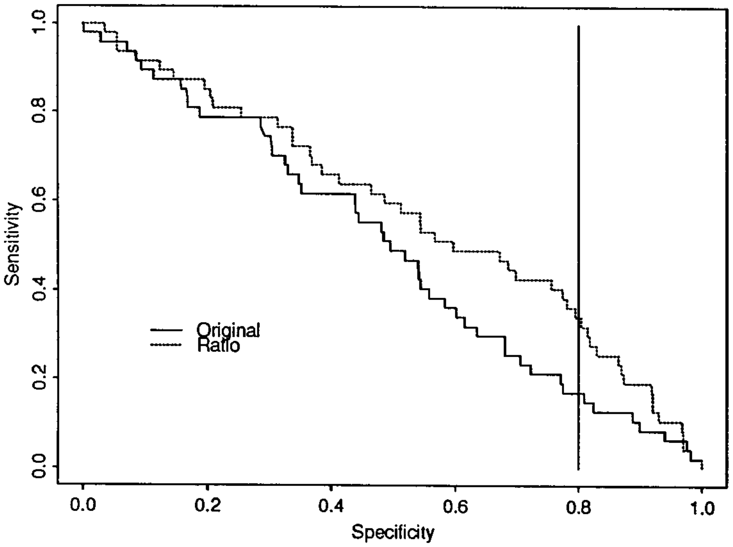

Figure 3. Sensitivity versus specificity curve for CA15-3 original values versus CA15-3 ratio over marker values using indicator weight

actual CA15-3 value (at a cut-off of 368) is 0·17 ($p = 0{\cdot}02$). Thus it does appear that the ratio of current value to baseline is a better marker than absolute value of the antibody, at least for CA15-3. It is noteworthy that out methods can be applied to compare these two highly correlated markers based on the same antibody since our theory allows for such correlation.

Our original reason for developing the theory in Section 3 was to have a tool for comparing CEA, TPS and CA15-3 as markers for progression. The disappointing utilities of these three markers lessened our interest in the question, but we discuss it briefly for illustrative purpose. Figure 4 plots the sensitivities of the three marker ratios across all sensitivities. The differences in the sensitivities are not significant between any of the markers for specificities of 0·80 or greater (which was the range of primary interest to the investigators), although the sensitivities for CA15-3 are significantly higher than for TPS at specificities 0·65 ($p = 0{\cdot}05$), 0·70 ($p = 0{\cdot}01$) and 0·75 ($p = 0{\cdot}01$). From these results, we would conclude that CA15-3 and CEA look more promising than TPS as markers, but that clear superiority for any marker has not been established.

A final point of interest to the investigators was whether a marker might be useful in predicting progression shortly before the progression was to occur, rather than in having the marker act as a substitute for a clinical examination. To address this question, we modified our definitions of a case and control as follows. Each time a patient had a marker, it was determined whether or not the patient had a clinical evaluation 1 week to 5 weeks after the marker was observed (the time interval was suggested by the investigators). If the patient did not have such an evaluation, the marker was not used in the analysis. If a patient did have one or more clinical evaluations 1 to 5 weeks later, the patient was classified as a case for that marker if any of the clinical evaluations during the time period were positive and as a control (no progression) otherwise. The results were again disappointing. Using the ratio approach with a cut-off of 1·5 the positive predictive value

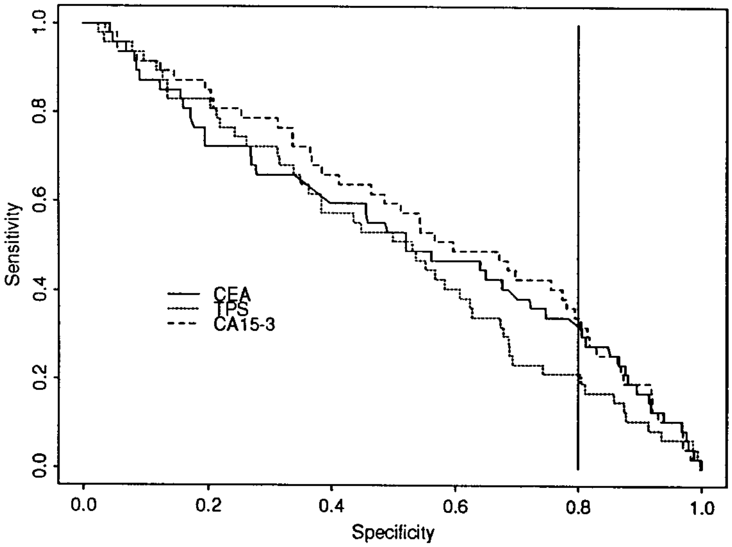

Figure 4. Sensitivity versus specificity curve for CEA, CA15-3 and TPS for ratio over marker values using indicator weight

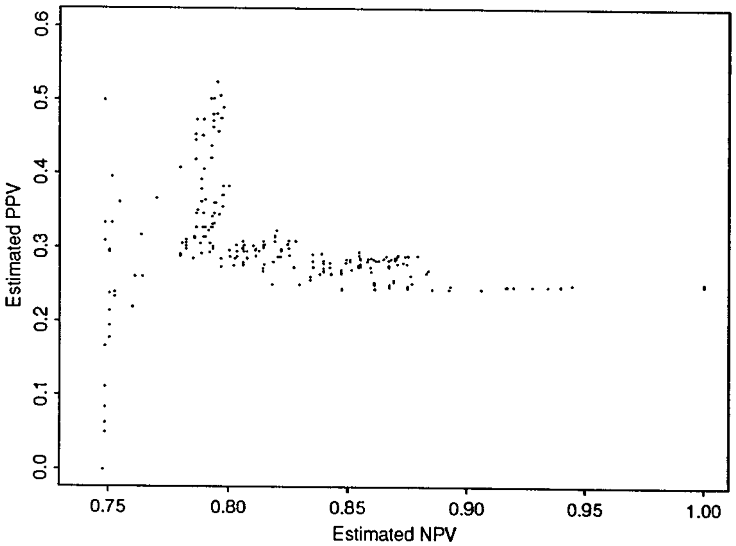

Figure 5. Estimated positive predictive values versus negative predictive values for the CA15-3 ratio one to five weeks prior to an event

was 0·47 with a 95 per cent confidence interval (0·45, 0·49). In fact the positive predictive value failed to exceed 0·53 for any cut-off value. We concluded that CA15-3, which was our most promising marker, would be of little value for predicting an imminent recurrence (Figure 5).

6. DISCUSSION

We have introduced a non-parametric approach to bringing concepts such as specificity and sensitivity into the analysis of repeated markers, since it is our experience that these concepts can be more useful than relative risk, particularly when trying to determine how to implement potential markers in clinical practice. We found the concept to be useful in analysing the breast cancer data described above, although it was the unfortunate case that the markers all had rather poor sensitivities for specificities which would be of interest. This knowledge in itself was useful, since time dependent Cox model analyses identified a reasonably large relative risk (2·3) associated with high (ratio $> 1{\cdot}5$) values of CA15-3.

Our methods allowed for estimating sensitivities, specificities and predictive values, obtaining confidence intervals for them, and comparing sensitivities at fixed specificities in the presence of between- and within-marker correlations. The method allows considerable flexibility. We used the ratio as a marker in two examples rather than the marker itself. This approach might reduce the between-patient variability, particularly if one suspected that each patient had a different mean value and that the standard deviation of the marker values were proportional to that mean. One could then use our methods to compare that approach to using the marker itself (as in Figure 3) or other variations such as the marker minus its baseline value (appropriate if one anticipated different mean values for each patient, but a constant within-patient variance).

In the initial evaluation of a marker, one might want to evaluate it in isolation. However, one might also wish to evaluate a panel of markers, or, ultimately, a decision rule which incorporates patient characters (possibly over time), for example, let the score from a time dependent Cox model be the marker. Our method can handle these cases. For example, we could have compared one marker to a combination of two markers, for example, CA15-3 to $\max[\{\text{CA153} - \min(\text{CA15-3})\}/\text{range}(\text{CA15-3}),\ \{\text{CEA} - \min(\text{CEA})\}/\text{range}(\text{CEA})]$ or defined a marker to be positive only if it increased on three consecutive visits. In fact, we tried all these approaches for the breast cancer data, but again obtained low utility measures for the markers. (A referee noted that these methods might be applicable in the monitoring of transplant patients, where there are a variety of markers for chronic rejection. We had not considered this application, but believe that the flexibility of our methods would be useful for evaluating such markers, or a panel of such markers.)

We are hoping to extend these results to allow comparisons across a range of specificities and to compare areas under the receiver operating characteristic (ROC) curves for repeated markers much as is done in Wieand *et al.*[5]

The statistic can easily be generalized to test for the equality of the sensitivity of more than two markers at a fixed specificity by obtaining a variance–covariance matrix for $\{\hat{H}_1(\hat{c}_1), \hat{H}_2(\hat{c}_2), \ldots, \hat{H}_k(\hat{c}_k)\}'$ and defining appropriate contrasts. For the breast cancer data described above, a global test of the hypothesis that all three sensitivities are equal at specificity 0·8 has p-value of 0·29.

As previously noted, we ignored some important bias issues in our example. The first was that only 89 of the 225 breast cancer patients who had participated in the randomized treatment trial had a baseline and repeated marker, and some of these patients had subsequent missing markers. However, the treatment trial was under way when the marker study was initiated, so some patients did not have the opportunity to participate in the marker trial. In other cases, serum samples were not submitted for logistic reasons. There was no evidence that a patient's condition was a factor in submission of markers, although we cannot completely rule out that possibility. It is unlikely that such a selection bias would have caused these disappointing results. Another point worth noting is that one might choose to be more selective in choosing risk sets in the time dependent Cox model. For example, one might exclude patients whose marker had been obtained more than a month prior to an event of interest. These issues were partially addressed at a conference (Ingle *et al.*[6]) and will be addressed in detail when a clinical paper is submitted.

Finally, we want to note that DeLong *et al.*[7] introduced a method for estimating specificity (for a fixed cut-off) in the repeated marker case when each patient could have at most one progression, as in the breast cancer example. Their method required direct estimation of the sensitivity and the estimate of relative hazard from a time dependent Cox model. In fact, they had a slightly different setting, in which the time interval was divided into periods, such that each patient had a marker and a state in each time period. Ignoring that difference and using their formula for specificity, we obtained values quite close to those obtained with our non-parametric approach. Defining a marker to be positive when its ratio (current value divided by baseline) exceeded 1·5, their method would lead to specificity estimates of 0·85, 0·78 and 0·73 for CA15-3, CEA and TPS, respectively. The corresponding values obtained using our methods were 0·82, 0·80 and 0·65. Their estimates were all within the 95 per cent confidence intervals for our values. Although, DeLong and her colleagues did not have this setting in mind, it appears that their results also give a reasonable method for estimating specificity and sensitivity Their methods do not, however, extend to confidence intervals and hypothesis testing.

APPENDIX I: EXPLICIT VARIANCE AND COVARIANCE TERMS

For easy notation we let $\hat{F}$ be $\hat{\text{spe}}$ and $\hat{H}$ be $\hat{\text{sen}}$. Sample moments can be used to estimate all the variance and covariance terms given in Theorem 3.1 except the ratio $h_i(c_i)/f_i(c_i)$ and c_i itself. To illustrate the approach, let $\hat{F}(c) = \Sigma(W_j\hat{F}_j(c)/\Sigma W_j) = (\Sigma t_j/\Sigma W_j)$ where $t_j = W_j\hat{F}_j$. Defining $W^* = E(W)$ it is not difficult to show that

$$\operatorname{var}\{\hat{F}(c)\} \approx \{1/(W^*)^2\}\operatorname{var}(\bar{t}) - 2\{F(c)/(W^*)^2\}\operatorname{cov}(\bar{t}, \bar{W}) + \{F^2(c)/(W^*)^2\}\operatorname{var}(\bar{W}).$$

We use $\approx$ when term of order $o(1/n)$ are omitted. For fixed c, each of these terms can be estimated by the method of moments. For example, our estimates of $\hat{\operatorname{var}}(\bar{t})$ and $\operatorname{cov}(\bar{t}, \bar{V})$ are

$$\hat{\operatorname{var}}(\bar{t}) = (1/n)(\textstyle\sum t_j^2/n - \bar{t}^2)$$

and

$$\hat{\operatorname{cov}}(\bar{t}, \bar{W}) = (1/n)(\textstyle\sum t_jW_j/n - \bar{t}\bar{W}).$$

A similar approach can be used for each of the terms $\operatorname{var}\{\hat{H}_i(c)\}$, $\operatorname{cov}\{\hat{F}_i(c), \hat{H}_j(c)\}$.

At fixed specificity p, we define the estimate of $h_i(c)/f_i(c)$ to be

$$[\hat{H}_i\{F_i^{-1}(p+\delta)\} - \hat{H}_i\{F_i^{-1}(p-\delta)\}]/2\delta$$

where δ is $\min[\min\{p/3, (1-p)/3\}, 25/n]$ and n is the number of patients. In our simulations we used $\min\{p/3, (1-p)/3\}$ since our sample sizes were only 30, 50 and 100.

Finally, $F_i^{-1}(p-\delta)$, $F_i^{-1}(p+\delta)$ and $c = F_i^{-1}(p)$ must be estimated as well. We let $\hat{c} = \hat{F}_i^{-1}(p) = \inf(x : \hat{F}_i(x) > p)$, and define the estimates of $F_i^{-1}(p-\delta)$ and $F_i^{-1}(p+\delta)$ similarly. In all the above computations, we then replace c by $\hat{c}$. For two markers, examples of covariance terms are

$$\operatorname{cov}\{\hat{F}_1(c_1), \hat{F}_2(c_2)\} \approx \{1/W^*)^2\}\operatorname{cov}(\bar{t}_1, \bar{t}_2) - \{F_1(c_1)/(W^*)^2\}\operatorname{cov}(\bar{t}_1, \bar{W})$$
$$- \{F_2(c_2)/(W^*)^2\}\operatorname{cov}(\bar{t}_2, \bar{W}) + [\{F_1(c_1)F_2(c_2)\}/(W^*)^2]\operatorname{var}(\bar{W}).$$

Again, each component of these terms can be estimated using the method of moments. A program which computes these variance terms and the standardized statistic is available from the contact author.

APPENDIX II: OUTLINE OF THE PROOF

Our proof uses lemmas which are very similar to results which can be found in Serfling[8] (pp. 74–100). If we did not have repeated markers and if cases were independent from controls, we would be able to apply these lemmas directly. In that case $\hat{c}$ would be an order statistic, and $\hat{F}$ and $1 - \hat{H}$ would be (independent) empirical distribution functions and the proof would begin by noting that

$$\hat{H}(\hat{c}) - H(c) = [\{\hat{H}(\hat{c}) - \hat{H}(c)\} - \{H(\hat{c}) - H(c)\}] + \{H(\hat{c}) - H(c)\} + \{\hat{H}(c) - H(c)\}.$$

For $0 < p < 1$, suppose F is differentiable at c and $F'(c) = f(c) > 0$. Then with probability 1 (*wp*1)

$$|\hat{c} - c| \leqslant \{1/f(c)\}\{2n^{-1/2}(\ln(n))^{1/2}\}$$

for all sufficiently large n (see Serfling,[8] Lemma B, pp. 96). This fact and Bahadur's Lemma (see Serfling,[8] Lemma E, pp. 97–98), imply that *wp*1

$$[\{\hat{H}(\hat{c}) - \hat{H}(c)\} - \{H(\hat{c}) - H(c)\}] = O\{n^{-3/4}(\ln(n))^{3/4}\}, \text{ as } n \to \infty.$$

Hence

$$\sqrt{n}\{\hat{H}(\hat{c}) - H(c)\} \approx \sqrt{n}\{H(\hat{c}) - H(c)\} + \sqrt{n}\{\hat{H}(c) - H(c)\}.$$

Since

$$H(\hat{c}) - H(c) = (\hat{c} - c)h(c) + O\{(\hat{c} - c)^2\}.$$

Finally, since *wp*1, $\hat{F}(\hat{c}) = F(c) + O(1/n)$, as $n \to \infty$, we have *wp*1

$$(\hat{c} - c)f(c) = [F(c) - \hat{F}(c) + O\{n^{-3/4}(\ln(n))^{3/4}\}]$$

as $n \to \infty$, we have

$$\sqrt{n}\{\hat{H}(\hat{c}) - H(c)\} \approx \sqrt{n}[-\{h(c)/f(c)\}\{\hat{F}(c) - F(c)\} + \sqrt{n}\{\hat{H}(c) - H(c)\} \to \mathrm{N}(0, V(c))$$

where $V(c) = \{(h(c)/f(c))^2\}\operatorname{var}\{\hat{F}(c)\} + \operatorname{var}\{\hat{H}(c)\}$.

When we have two markers

$$n^{1/2}(\hat{\Delta} - \Delta)/\sqrt{V(c_1, c_2)} \to \mathrm{N}(0, 1)$$

as $n \to \infty$ where $\hat{\Delta} = \hat{H}_2(\hat{c}_2) - \hat{H}_1(\hat{c}_1)$, $\Delta = H_2(c_2) - H_1(c_1)$, and

$$V(c_1, c_2) = \{h_1(c_1)/f_1(c_1)\}^2\operatorname{var}\{\hat{F}_1(c_1)\} + \{h_2(c_2)/f_2(c_2)\}^2\operatorname{var}\{\hat{F}_2(c_2)\}$$
$$+ \operatorname{var}\{\hat{H}_1(c_1)\} + \operatorname{var}\{\hat{H}_2(c_2)\} - 2\operatorname{cov}\{\hat{H}_1(c_1), \hat{H}_2(c_2)\}$$
$$- 2\{h_1(c_1)h_2(c_2)\}/\{f_1(c_1)f_2(c_2)\}\operatorname{cov}\{\hat{F}_1(c_1), \hat{F}_2(c_2)\}$$

where $H_i(.) \equiv 1 - G_i(.)$, and $h_i(x) = \partial H_i(x)/\partial x$, $i = 1, 2$.

In fact this is the variance term for the fixed specificity case given on p. 587 of Wieand *et al.*[5] The change in the proof in the repeated marker case involves the following ideas:

(i) Because the same patient can be a case and a control, $\hat{F}_k$ and $\hat{H}_l$, $k = 1, 2$, $l = 1, 2$ can be correlated. Thus in the repeated marker case, $V(c_1, c_2)$ in Theorem 3.1 includes terms such as $-2\{h_k(c)/f_k(c)\}\text{cov}\{\hat{F}_k(c), \hat{H}_l(c)\}$. This has a minimal effect on the complexity of estimating the variance term as the covariance terms are obtained using methods similar to obtaining the variance term.

(ii) In the single marker case the contribution to the specificity from each individual is an indicator function. Although $\hat{F}_{ij} = (1/N_{1J})\Sigma \text{I}(X_{ij} > c)$ is no longer an indicator function, the $\hat{F}_{1j}$'s are still independent and identically distributed random variables bounded by 0 and 1 as are the $\hat{F}_{2j}$'s. In fact, if the within-patient correlation for a marker is less than 1 (if correlation is 1 then the problem simplifies to a single markers case), $\hat{F}_{ij}$ will have smaller variance than an indicator function and will still be unbiased.

(iii) Although $\hat{F}_i = (\Sigma W_j \hat{F}_{ij})/\Sigma W_j)$ is the ratio of two random variables, one can use the fact that $\bar{W} = (1/n)\Sigma W_j$ converges to W^* and replace $\hat{F}_i$ by $\hat{F}_i^* = \{1/(nW^*)\}\Sigma W_j \hat{F}_{ij}$ plus a term involving $\Sigma W_j - nW^*$ which approaches zero quickly enough that it does not affect the asymptotic distribution.

(iv) Although $\hat{F}_i^*$ is not an empirical distribution function, it is unbiased and many of the properties of an empirical distribution hold. In particular, the relevant bounds on convergence presented in Serfling[8] still are attained.

These ideas are used in a very intuitive way to emulate the results of Serfling[8] and obtain the variance term presented in Theorem 3.1. Unfortunately, the algebra associated with verifying the results in this setting is tedious and computations are too lengthy to be included here. A detailed proof of Theorem 3.1 is available from the first author.

Note that the proofs of the asymptotic normality of the terms in Section 2 are contained within the proofs of Theorem 3.1.

ACKNOWLEDGEMENTS

We want to thank two referees who made several insightful comments which resulted in improvements in the manuscript.

REFERENCES

1. Cox, D. R. 'Regression models and life tables (with discussion)', *Journal of the Royal Statistical Society, Series B*, **34**, 187–220 (1972).
2. Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*, Wiley, New York, 1980.
3. Schaid, D. J., Ingle, J. N., Wieand, S. and Ahmann, D. L. 'A design for phase II testing of anticancer agents within a phase III clinical trial', *Controlled Clinical Trials*, **9**, 107–118 (1988).
4. Ingle, J. N., Foley, J. F., Mailliard, J. A., Krook, J. E., Hartmann, L. C. M., Jung, S. H., Veeder, M. H., Gesme, D. H., Hatfield, A. K. and Goldberg, R. M., 'Randomized trial of Cyclophosphamide, Methotrexate, and 5-Fluorouracil with or without estrogenic recruitment in women with metastatic breast cancer', *Cancer*, **73**, 2237–2343 (1994).
5. Wieand, S., Gail, M. H., James, B. R. and James, K. L. 'A family of non-parametric statistics for comparing diagnostic markers with paired or unpaired data', *Biometrika*, **76**, 585–592 (1989).

6. Ingle, J. N., Ritts, R. E., Wieand, H. S. and Foley, J. F. 'Evaluation of a panel of potential serum tumor markers in women with metastatic breast cancer entered on a prospective chemotherapy clinical trial', A North Central Cancer Treatment Group Study, Proc XXIIIrd Meeting of the International Society for Oncodevelopmental Biology and Medicine, Montreal, Canada, 1995, p. 15.
7. DeLong, E. R., Vernon, W. B. and Bollinger, R. R. 'Sensitivity and specificity of a monitoring test', *Biometrics*, **41**, 947–958 (1985).
8. Serfling, R. J. *Approximation Theorems of Mathematical Statistics*, Wiley, New York, 1980.