The predictive receiver operating characteristic curve for the joint assessment of the positive and negative predictive values

BY SHANG-YING SHIU^{1,*} AND CONSTANTINE GATSONIS²

¹*Institute of Statistical Science, Academia Sinica, Taipei 115, Taiwan,
Republic of China*

²*Center for Statistical Sciences, Brown University, Providence, RI 02912, USA*

Binary test outcomes typically result from dichotomizing a continuous test variable, observable or latent. The effect of the threshold for test positivity on test sensitivity and specificity has been studied extensively in receiver operating characteristic (ROC) analysis. However, considerably less attention has been given to the study of the effect of the positivity threshold on the predictive value of a test. In this paper we present methods for the joint study of the positive (PPV) and negative predictive values (NPV) of diagnostic tests. We define the predictive receiver operating characteristic (PROC) curve that consists of all possible pairs of PPV and NPV as the threshold for test positivity varies. Unlike the simple trade-off between sensitivity and specificity exhibited in the ROC curve, the PROC curve displays what is often a complex interplay between PPV and NPV as the positivity threshold changes. We study the monotonicity and other geometric properties of the PROC curve and propose summary measures for the predictive performance of tests. We also formulate and discuss regression models for the estimation of the effects of covariates.

Keywords: predictive value; PROC curve; ROC curve; diagnostic test

1. Introduction

Diagnostic tests are evaluated on the basis of measures defined conditionally on the true disease status (sensitivity, specificity and receiver operating characteristic, ROC, curve) or conditionally on the test outcome (positive predictive value, PPV, and negative predictive value, NPV). Measures of test performance, such as sensitivity, specificity or ROC curves, provide the type of information that is typically needed for technology assessment and health policy purposes. Measures of predictive value provide the type of information that is typically needed for clinical decision-making, where clinicians and patients decide whether

* Author for correspondence (shiu@stat.sinica.edu.tw).

Electronic supplementary material is available at <http://dx.doi.org/10.1098/rsta.2008.0043> or via <http://journals.royalsociety.org>.

One contribution of 13 to a Theme Issue ‘Mathematical and statistical methods for diagnoses and therapies’.

to use a test or how to assess the implications of a test result. The clinical relevance of predictive value information notwithstanding, a large majority of diagnostic test evaluations continue to be designed with a primary focus on measures defined conditionally on the disease status. One of the reasons for this is the theoretical invariance of sensitivity and specificity to disease prevalence. Predictive values, on the contrary, vary across populations with different disease prevalence, making comparisons of diagnostic tests difficult.

The invariance to disease prevalence of quantities such as sensitivity and specificity is predicated on the assumption that the disease status is the only variable affecting test outcome. Although this simplification is useful as a building block of the theory of diagnostic test evaluation, in reality the test outcome may depend on a number of other characteristics beyond disease status (Moons & Harrel 2003). For example, in a clinical study where electrocardiographic stress test was used to diagnose coronary artery disease, the sensitivity and specificity of this stress test result varied substantially according to gender, relative workload as well as the number of diseased vessels (Moons *et al.* 1997). Moreover, variations in sensitivity, specificity and ROC curves are routinely observed in meta-analysis of diagnostic tests (Irwig *et al.* 1994; Rutter & Gatsonis 2001) and in studies of the performance of test interpreters (Ishwaran & Gatsonis 2000; Beam *et al.* 2003).

The available statistical methodology for the study of the predictive value of tests is less extensive than the corresponding methodology for measures defined conditionally on disease status (Bennett 1985; Pepe 2003). Copas (1999) proposed to use the logit rank plot as a summary of the effectiveness of risk scores. Summary measures of the resulting predictiveness curve were subsequently considered (Bura & Gastwirth 2001) and a thorough study of inference based on the full predictiveness curve was presented (Huang *et al.* 2007). Leisenring *et al.* (2000) discussed a model-based approach to the comparison of the predictive values of binary tests for paired designs. In this approach, a marginal regression modelling framework was used with disease status as the response variable and test indicator as an explanatory variable.

The effect of the threshold value used to declare a positive test result is a fundamental aspect of our understanding of the performance of diagnostic tests. This threshold is the conceptual basis for the well-known trade-off between sensitivity and specificity of a test that gives rise to the ROC curve. A significant body of statistical literature has discussed models with implicit or explicit thresholds for test positivity (Hanley & McNeil 1982; Hanley 1989; Hanley 1998; Pepe 2003). Although it is clear that the PPV and NPV of a test are also functions of the threshold for test positivity, the effect of this dependence has not been studied extensively. This dependence induces a close relation between the two quantities and implies that if the threshold is moved *both* will be affected. It follows that a complete characterization of the predictive power of a test requires the study of both quantities as a pair. Moskowitz & Pepe (2004) proposed a graphical method and a regression framework to estimate and compare predictive values of continuous prognostic factors as a function of the positivity threshold. In that work, the PPV and NPV are quantified and assessed separately. To the best of our knowledge, the *joint* evaluation of the two quantities has not been discussed in the literature. Although both quantities may not be of equal interest in a given practical setting, it is rarely the case that

interest lies exclusively in only one of them. However, the joint behaviour of the two quantities cannot necessarily be inferred from their marginal behaviour that would be assessed by a separate analysis of each.

In this paper, we undertake a systematic study of the effect of the positivity threshold on the pair of PPV and NPV of tests. Our emphasis is on the study of the interplay between the two types of predictive values of a test as the positivity criterion varies and on the development of summaries of a test's possible pairs of predictive values. In particular, we define the predictive receiver operating characteristic (PROC) curve that shows all possible pairs of the PPV and NPV of a test as the threshold varies. We study the geometric properties of the PROC curve, discuss methods for estimating the curve for continuous and ordinal valued tests and propose summary measures for the predictive performance of tests. We also formulate and discuss regression models for the estimation of the effects of covariates.

This paper is organized as follows. In §2 we define the PROC curve that can apply to tests with continuous or ordinal categorical outcomes. The geometric patterns and other properties of the PROC curve are discussed in §3. Details of mathematical derivations of PROC curve properties are presented in appendix B of the electronic supplementary material. Section 4 presents the estimation of the PROC curve. We discuss an indirect approach through ROC curve estimation for tests in general. We also propose a direct approach that jointly estimates the PPV and NPV for tests with ordinal outcomes. In §5 we describe methods for evaluating a test's predictive performance and comparing tests using the PROC curve. Illustrative examples with both continuous and ordinal test data are presented in §6, including assessing the ability of standardized uptake value based on lean body mass (SUV-lean; continuous test data) to predict axillary node involvement in women diagnosed with breast cancer and comparing predictive accuracy between digital and screen-film mammography (ordinal test data) for breast-cancer screening. Section 7 summarizes our conclusions.

2. Definition: the PROC curve

Let D be a binary random variable representing the disease status (0 = non-diseased, 1 = diseased) with prevalence $p = Pr(D=1)$. Let T be a continuous random variable representing the underlying latent scale of test result. We denote by G and F the cumulative distribution functions of $T|D=1$ and 0, respectively. In this notation, the PPV and NPV for a positivity threshold c are given by

$$PPV(c) = Pr(D = 1 | T \geq c) = \frac{[1 - G(c)]p}{[1 - G(c)]p + [1 - F(c)](1 - p)}, \quad (2.1)$$

$$NPV(c) = Pr(D = 0 | T < c) = \frac{F(c)(1 - p)}{G(c)p + F(c)(1 - p)}. \quad (2.2)$$

Analogous to the definition of the ROC curve, the PROC curve is defined as follows:

$$\{(1 - NPV(c), PPV(c))\}_{c \in R},$$

where R is the set of all possible thresholds for test positivity. The PROC curve consists of points representing all possible combinations of the PPV and one minus the NPV generated by varying the threshold. This curve displays the interplay between the two quantities and provides a graphical display of the predictive performance of a diagnostic test over the range of possible thresholds. Strictly speaking, the above definition applies to the *theoretical* PROC curve of the test. The *empirical* PROC curve is simply the collection of the observed $(1 - \text{NPV}, \text{PPV})$ points connected by straight-line segments.

3. Properties of predictive values

The ‘ideal’ point in the PROC graph is the point $(0, 1)$ on the y -axis, at which both predictive values equal 1. Hence, in analogy to ROC curves, a PROC curve would indicate high predictive performance if small trade-offs in NPV would enable the test to reach high PPV. However, the geometric patterns of PROC curves can often be rather complex, at least in comparison with the patterns observed in ROC curves. It is therefore essential to study and attempt to characterize the geometric properties of PROC curves before undertaking an investigation of how the curves can be used to evaluate the performance of a diagnostic test. This section is devoted to an exploration of properties of the *theoretical* PROC curve, with a detailed investigation of the commonly used binormal ROC model, which assumes that some monotone increasing transformation of the test result follows a standard normal distribution, conditional on $D=0$, and follows a normal distribution with mean a and standard deviation b , conditional on $D=1$. Equivalently, parameter a represents the difference between the two means and parameter b represents the ratio of two standard deviations of the underlying normal distributions.

(a) PROC curve as a function of disease prevalence

In contrast to the ROC curve that is invariant to disease prevalence, the geometric characteristics of the PROC curve are partially determined by disease prevalence. As can be seen from (2.1) and (2.2), the positive predictive value increases when the prevalence p increases and the negative predictive value decreases. In other words, with an increasing prevalence, a point on a PROC curve moves towards its upper-right direction.

The two extreme points (when the positivity threshold approaches minus infinity and infinity) of a curve correspond to settings in which all cases are classified as either positive or negative. Unlike the ROC curve in which the extreme points are always $(0, 0)$ and $(1, 1)$, the PROC curve has extreme points that also depend on the prevalence. Figure 1 shows examples under the usual binormal ROC model for selected values of a , b and p . Assuming a and b fixed, the prevalence affects primarily the location of the curve and only secondarily the shape of the curve.

(b) Monotonicity of the PROC curve

Perhaps the most challenging complexity of PROC curves arises from the potential that two distinct values of PPV can correspond to the same value of NPV and conversely. To examine this issue, we begin by defining a *monotonic*

Figure 1. Predictive curves with $a=0.8$. (a) $b=0.7$, (b) $b=1$, (c) $b=1.5$. Solid line, high prevalence ($p=0.7$); dot-dashed line, low prevalence ($p=0.3$).

PROC curve as a curve that has a one-to-one correspondence between the PPV and NPV. Lack of monotonicity is extensively present even in the binormal ROC model. Monotonicity can hold globally, over all threshold values, or locally, over particular intervals of the threshold values. If the PROC is seen as the graph of PPV as a function of NPV, our definition of monotonicity corresponds to the usual notion of monotonicity of a function. Before discussing the definition of summary measures and optimal operation points, the issue of the monotonicity of the PROC curve needs to be addressed.

A mathematical characterization of monotonicity can be obtained by considering properties of the hazard rate. By taking the first derivatives of predictive values in (2.1) and (2.2) as given in §2, we obtain the following two inequalities:

$$\frac{d}{dc} \text{PPV}(c) \geq 0 \Leftrightarrow \frac{g(c)}{1-G(c)} \leq \frac{f(c)}{1-F(c)} \quad (3.1)$$

and

$$\frac{d}{dc} \text{NPV}(c) \geq 0 \Leftrightarrow \frac{g(c)}{G(c)} \leq \frac{f(c)}{F(c)}. \quad (3.2)$$

Pairs of random variables with cumulative distribution functions and densities satisfying inequality (3.1) for all c are said to be *hazard rate ordered*, while pairs satisfying (3.2) for all c are said to be *reversed hazard rate ordered* (Shaked & Shanthikumar 1994). A random variable X is smaller than Y in the hazard rate order if the slope of the logarithm of the survival function of X is uniformly smaller than the corresponding slope for Y . Further discussion and an example of when the condition does not hold are provided in appendix A of the electronic supplementary material. Returning now to the discussion of the monotonicity of the PROC curve, a necessary and sufficient condition for the PPV and NPV to be monotone is that $T|D=1$ and 0 are hazard rate ordered and reversed hazard rate ordered, respectively. Both hazard rate order and reversed hazard rate order are guaranteed by the *likelihood ratio order*, under which the ratio $f(c)/g(c)$ is a monotone function of c . The above implies that a sufficient condition for the monotonicity of the PROC curve is that the pair of, possibly latent, test result variables for truly positive and negative cases is likelihood ratio ordered. This property of monotone likelihood ratio is possessed by the well-known one-parameter exponential family with a monotone canonical link function.

Examples include exponential distributions, gamma distributions with the same shape parameters and logistic distributions with the same scale parameters.

We now investigate the monotonicity property for the commonly used binormal model. Technical details of the derivation are provided in appendix B of the electronic supplementary material. Proposition 3.1 identifies values of the parameters for which the binormal PROC curve is monotone.

Proposition 3.1. *Assume that $a > 0$. For $b = 1$,*

- (i) *the positive predictive value increases as the threshold increases and*
- (ii) *the negative predictive value decreases as the threshold increases.*

Proposition 3.1 shows that the PROC curve is not only monotone, but also it shows a trade-off between PPV and NPV if and only if $b = 1$. When $b \neq 1$, we can consider only intervals of positivity threshold over which the PROC curve segments are monotone.

Proposition 3.2. *Assume that $a > 0$.*

- (a) *For $b > 1$,*
 - (i) *the positive predictive value is strictly decreasing on $c \in (-\infty, c_{\text{PPV}}^*)$ and is strictly increasing on $c \in (c_{\text{PPV}}^*, \infty)$ and*
 - (ii) *the negative predictive value is strictly increasing on $c \in (-\infty, c_{\text{NPV}}^*)$ and is strictly decreasing on $c \in (c_{\text{NPV}}^*, \infty)$.*
- (b) *For $b < 1$,*
 - (i) *the positive predictive value is strictly increasing on $c \in (-\infty, c_{\text{PPV}}^*)$ and is strictly decreasing on $c \in (c_{\text{PPV}}^*, \infty)$ and*
 - (ii) *the negative predictive value is strictly decreasing on $c \in (-\infty, c_{\text{NPV}}^*)$ and is strictly increasing on $c \in (c_{\text{NPV}}^*, \infty)$,*

where c_{PPV}^* and c_{NPV}^* are unique solutions of

$$\frac{\frac{\phi\left(\frac{c-a}{b}\right)}{b}}{1 - \Phi\left(\frac{c-a}{b}\right)} = \frac{\phi(c)}{1 - \Phi(c)}$$

and

$$\frac{\frac{\phi\left(\frac{c-a}{b}\right)}{b}}{\Phi\left(\frac{c-a}{b}\right)} = \frac{\phi(c)}{\Phi(c)},$$

respectively. Moreover, $c_{\text{PPV}}^* \leq c_{\text{NPV}}^*$.

Proposition 3.2 shows that both the PPV and NPV have only one local maximum (or minimum). We denote these positivity thresholds as c_{PPV}^* and c_{NPV}^* . The PROC curve is monotone and shows a trade-off between the PPV and NPV on $c \in (-\infty, c_{\text{PPV}}^*)$ and $c \in (c_{\text{NPV}}^*, \infty)$. On $c \in (c_{\text{PPV}}^*, c_{\text{NPV}}^*)$, the curve is also monotone but does not show such a trade-off. Indeed both the PPV and NPV either decrease (when $b < 1$) or increase (when $b > 1$) as the positivity threshold traverses this interval.

Figure 2. Each non-monotone PROC curve consists of three monotone curve segments defined on $(-\infty, c_{PPV}^*)$, $[c_{PPV}^*, c_{NPV}^*)$ and (c_{NPV}^*, ∞) , respectively. Circles denote points corresponding to operating thresholds at -1 and $a+b$, triangles denote points corresponding to operating thresholds at -2 and $a+2b$ and crosses denote c_{PPV}^* and c_{NPV}^* . (a) $a=1$, $b=0.5$, $p=0.5$; (b) $a=2$, $b=2$, $p=0.3$.

Figure 2 gives examples of PROC curves for selected values of a and $b \neq 1$. We note that some parts of the PROC curve, even though identified as strictly increasing or decreasing by proposition 3.2, are visually vertical or horizontal lines. The presence of these line segments is due to the fact that one of the PPV or NPV converges much faster than the other as the threshold approaches the infinity or the minus infinity. Moving the positivity threshold along these lines does not influence one of the PPV or NPV, while the other predictive value decreases (or increases). Figure 2 also locates the points at which the operating thresholds are ± 1 and ± 2 s.d. (of the distributions for the diseased and the non-diseased) away from the means. Note that a considerable part of the PROC curve corresponds to threshold values beyond ± 2 s.d.

To summarize, propositions 3.1 and 3.2 give monotonicity properties of the PROC curve. For tests with either continuous or ordinal outcome, the parameters of this curve can be estimated and the monotonicity of the estimated curve can be assessed. In practice, since a considerable part of the PROC curve may correspond to threshold values outside the usual range of interest, a partial PROC curve is often of more interest than a whole curve. This curve segment of interest may well be monotone.

(c) Further comments on the interpretation of the PROC curve

We note that the shape and location of the PROC curve are naturally affected by the degree of separation in the distribution of test results between positive and negative cases. Figure 3 shows the effect of this separation in the binormal model.

From propositions 3.1 and 3.2, we also note that the patterns of the shapes of the PROC curves can be summarized by three categories, which are patterns of $b > 1$, $b = 1$ and $b < 1$, as shown in figure 1. The apparent discontinuity of the binormal curve at the point $b = 1$, however, is the result of the behaviour of the curve at the extremes of the range of possible thresholds. Figure 4 shows PROC curve segments corresponding to threshold values on $[-3, a+3b]$ from $b < 1$ to $b > 1$ for selected values of a . In contrast to the whole curve, the curve segments in the figure show a continuous pattern of change.

Figure 3. The movement of the PROC curves with increasing a 's. (a) $b=0.7$, (b) $b=1$ and (c) $b=1.5$. Solid line, $a=5$; dotted line, $a=2$; dot-dashed line, $a=0.5$.

Figure 4. Each PROC curve segment corresponds to $c \in [-3, a + 3b]$, where the prevalence is 0.3. (a) $a=0.8$, (b) $a=2$. Solid line, $b=0.9$; dotted line, $b=0.95$; dot-dashed line, $b=1$; dashed line, $b=1.05$; triple dot-dashed line, $b=1.1$.

4. PROC curve estimation

A convenient way to derive estimates of the PROC curve is by first estimating the ROC curve and the prevalence for a set of data and then using a suitable transformation. The joint likelihood of the test outcome and the true disease status can be factored into two terms

$$L_{(T,D)} = L_{T|D} L_D,$$

where $L_{T|D}$ is constructed from the distribution of the test outcome conditional on the disease status and L_D is from the marginal distribution of the disease status. $L_{T|D}$ and L_D have mutually exclusive parameter sets; therefore, the maximization of L is equivalent to maximize $L_{T|D}$ and L_D separately. The conditional distribution of test results expressed by $L_{T|D}$ has been used for deriving the estimates of sensitivity, specificity and ROC curve (Tosteson & Begg 1988; Toledano & Gatsonis 1996; Metz *et al.* 1998; Pepe 1998, 2000b). The distribution of true disease status, L_D , is of interest in prospective studies and can also be modelled as a function of covariates.

In this section we propose an alternative approach for PROC curve estimation when the diagnostic test is ordinal categorical. Instead of an indirect approach through the ROC curve estimation, we consider direct modelling of the two types of predictive values through a joint formulation. The model is henceforth referred to as *the predictive model*. Parameter estimation is based on a pseudo-likelihood approach. Estimates of the PROC curve can be obtained by parameter estimates from the predictive model. Models for regression analysis of PPV and NPV separately have been proposed in the literature (Leisenring *et al.* 2000; Moskowitz & Pepe 2004; Huang *et al.* 2007). The formulation here treats the two quantities jointly.

(a) *The predictive model: model specification*

Let T^* be a random variable representing the test result taking ordered categories $1, 2, \dots, K$ (1 =definitely normal, K =definitely abnormal). Then $\{T^* \geq k\} = \{T \geq \theta_k\}$, where θ_k 's denote the positivity thresholds on the underlying continuous scale T . Combining (2.1) and (2.2), under the usual binormal assumption, we obtain

$$Pr(D = 1|I_k) = \frac{1}{1 + rg(\theta_k, I_k, a, b)},$$

where $I_k = I_{\{T^* \geq k\}}$ is the binary test result based on the k th positivity threshold, $r = (1 - p)/p$ and $g(\theta_k, I_k, a, b) = (I_k - \Phi(\theta_k)) / (I_k - \Phi((\theta_k - a)/b))$. This predictive model jointly specifies the PPV and NPV. When $I_k = 1$ and 0 , the PPV and NPV are modelled, respectively.

The predictive model can be extended to include covariates such as those representing subject characteristics that may influence the predictive performance of a diagnostic test. In studies with more than one test, covariates can also be test indicators, thus making it possible to use the predictive model for test comparisons. Since the predictive performance of a diagnostic test depends on the location parameter a , the scale parameter b and the prevalence p , covariates may affect any one of these three parameters. The predictive model with covariates $\mathbf{x} = (1, x_1, \dots, x_J)$ is defined as

$$Pr(D = 1|I_k, \mathbf{x}) = \frac{1}{1 + (r\mathbf{x})g(\theta_k, I_k, \mathbf{a}\mathbf{x}, \mathbf{b}\mathbf{x})}. \quad (4.1)$$

In the following, we will present a pseudo-likelihood approach for estimating parameters in the predictive model assuming independent observations.

(b) *Formulation of the pseudo-likelihood function*

Let I_k be as defined in §4a, and let PPV_k and NPV_k be the PPV and NPV, respectively, using the k th positivity threshold. The disease status conditional on the test result follows a Bernoulli distribution

$$D|I_k \sim \text{Bernoulli}(q_k) \forall k,$$

with $q_k = PPV_k I_k + (1 - NPV_k)(1 - I_k)$. Assuming that each individual i is independent, the log-likelihood function for each positivity threshold k is

$$ll_k(\mathbf{a}, \mathbf{b}, \mathbf{r}, \theta_k) = \sum_{i=1}^N \ln f_k(D_i|I_{ik}) = \sum_{i=1}^N D_i \ln(q_{ik}) + (1 - D_i) \ln(1 - q_{ik}).$$

According to the predictive model (4.1), ll_k can be written as

$$ll_k(\mathbf{a}, \mathbf{b}, \mathbf{r}, \theta_k) = \sum_{i=1}^N -D_i \ln(1 + h_{ik}) + (1 - D_i) \ln\left(1 - \frac{1}{1 + h_{ik}}\right),$$

where $h_{ik} = \mathbf{r}\mathbf{x}_i g(\theta_k, I_{ik}, \mathbf{a}\mathbf{x}_i, \mathbf{b}\mathbf{x}_i)$.

For each ll_k , we can use the maximum-likelihood method to obtain estimates of parameters $\mathbf{a}$, $\mathbf{b}$, $\mathbf{r}$ and θ_k . However, since parameters $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{r}$ appear in all ll_k 's, maximum-likelihood solutions based on each ll_k are likely to differ. To address this problem, we formulate a single likelihood function consisting of all ll_k 's, and therefore will lead to a single set of parameter estimates. This likelihood function can be constructed by the following data replication.

Let each individual be replicated $K-1$ times as if there were another $K-1$ identical individuals in the data, except that the result of test positivity at each time of the replication is based on one of another $K-1$ positivity thresholds. Denote by $R_i = (R_{i1}, \dots, R_{iK})$ the disease status of individual i after replication, where $R_{ik} = D_i$ for all k 's. Denote by $I_i = (I_{i1}, \dots, I_{iK})$ the results of test positivity of individual i after replication. Then the log-likelihood function of the replicated data is

$$\sum_{i=1}^N \ln f(R_{i1}, \dots, R_{iK} | I_{i1}, \dots, I_{iK}), \quad (4.2)$$

where $f(R_{i1}, \dots, R_{iK} | I_{i1}, \dots, I_{iK})$ is the joint density of R_i given the results of test positivity I_i . Since by replication we introduce within-subject correlations, the joint density of $R_i | I_i$ is difficult to obtain. Hence, we consider the following approach to bypass the specification of the full likelihood.

Referring to the definition of the log of the pseudo-likelihood function (Aerts *et al.* 2002), we construct the following one for (4.2):

$$\sum_{i=1}^N \sum_{k=1}^K \delta_k \ln f_k(D_i | I_{ik}) = \sum_{k=1}^K \delta_k ll_k. \quad (4.3)$$

In this formulation, setting $\delta_k = 1 \ \forall k$ gives equal weight to each positivity threshold. Different weights may be desirable in some study settings.

(c) Maximum pseudo-likelihood estimators

The maximum pseudo-likelihood estimators are defined as the maximizer of (4.3). Under standard regularity conditions on $f_k(D | I_k)$, the estimators are consistent and asymptotically normal with empirically corrected variance $J^{-1} Q J^{-1}$ (Arnold & Strauss 1991; Aerts *et al.* 2002), where the matrix J is the negative matrix of expected second derivatives of the log pseudo-likelihood function, and the matrix Q is the expected product of the pseudo-likelihood scores. The standard regularity conditions are to ensure the existence of maximum pseudo-likelihood estimates and a positive definite matrix J , which are identified and proved in appendix C of the electronic supplementary material.

Note that although in §4b we presented the pseudo-likelihood approach as assuming independent observations, the same approach can easily apply to multiple correlated observations, such as data from studies of paired design. The empirically corrected variance can handle correlations introduced by data replication, as well as the within-subject correlations.

(d) *Estimating predictive values and the PROC curve*

Using the maximum pseudo-likelihood estimators with corrected standard errors as described in §4c, we can estimate the covariate-specific PPV and NPV by

$$\widehat{\text{PPV}}(\mathbf{x})_k = \frac{1}{1 + \hat{r}x \frac{1 - \Phi(\hat{\theta}_k)}{1 - \Phi\left(\frac{\hat{\theta}_k - \hat{\mathbf{a}}\mathbf{x}}{\hat{\mathbf{b}}\mathbf{x}}\right)}}, \quad \widehat{\text{NPV}}(\mathbf{x})_k = 1 - \frac{1}{1 + \hat{r}x \frac{\Phi(\hat{\theta}_k)}{\Phi\left(\frac{\hat{\theta}_k - \hat{\mathbf{a}}\mathbf{x}}{\hat{\mathbf{b}}\mathbf{x}}\right)}}.$$

Similarly, we can estimate the covariate-specific PROC curve by

$$\left\{ \left(\frac{1}{1 + \hat{r}x \frac{\Phi(c)}{\Phi\left(\frac{c - \hat{\mathbf{a}}\mathbf{x}}{\hat{\mathbf{b}}\mathbf{x}}\right)}}, \frac{1}{1 + \hat{r}x \frac{1 - \Phi(c)}{1 - \Phi\left(\frac{c - \hat{\mathbf{a}}\mathbf{x}}{\hat{\mathbf{b}}\mathbf{x}}\right)}} \right) \right\}_{c \in R}.$$

The standard errors of $\widehat{\text{PPV}}(\mathbf{x})_k$, $\widehat{\text{NPV}}(\mathbf{x})_k$ and the estimated PROC curve can be calculated using either the delta method, or bootstrap methods such as sampling with replacement from the original data and parametric bootstrap when the sample size is small.

5. The analysis of PROC curves

This section presents methods in evaluating and comparing the predictive performance of diagnostic tests by using PROC curves.

When a PROC curve or a segment of interest is monotone, summary measures such as the area under the full curve (full area) or a particular segment (partial area) can be used to summarize the predictive performance of a diagnostic test. The (partial) area can be interpreted as the average positive predictive value corresponding to a given range of negative predictive value. Comparisons of area under the curves can be made by the delta method.

In addition to summary measures that integrate over ranges of threshold values, we discuss approaches for defining ‘optimal’ points on the PROC curve and corresponding summary indices. A potential advantage of these approaches is that they may be interpretable in both monotone and non-monotone curve settings. As in the case of ROC analysis, it is possible to estimate test values that correspond to the optimal points on the curve using either the method for estimating the curve, for example as in Metz *et al.* (1998), Pepe (2000a) and Alonzo & Pepe (2002), or regression models as described in Zhou *et al.* (2002).

(a) *The r^* global measure of predictive performance*

We begin by proposing a statistic that measures how far away a diagnostic test is from perfect prediction, i.e. when both the PPV and NPV are equal to one. For each threshold c and the corresponding point on the PROC curve, this statistic calculates a distance from the point to the ideal point at the left upper corner

where $(1 - \text{NPV}, \text{PPV}) = (0, 1)$. Explicitly,

$$r(c) = [1 - \text{PPV}(c)] + [1 - \text{NPV}(c)]. \quad (5.1)$$

We also define a statistic r^* as the minimum distance between the PROC curve and the point of perfect prediction,

$$r^* = \min_c r(c).$$

The statistic r^* can be interpreted as the best performance that a diagnostic test can achieve and

$$c^* = \arg \min_c r(c)$$

is the optimal operating threshold. Note that both r^* and c^* depend on the prevalence.

An alternative derivation of r^* can be obtained as follows. Rewrite (5.1) as

$$\text{PPV}(c) = [1 - \text{NPV}(c)] + 1 - r(c). \quad (5.2)$$

Equation (5.2) represents a straight line on the PROC space with intercept equal to one minus $r(c)$ and slope equal to one. By minimizing $r(c)$, we maximize the intercept of this straight line. In other words, r^* can be obtained by moving this 45° line towards the upper left corner, until it reaches the highest point that intersects with the PROC curve.

The definition of $r(c)$ implicitly assumes that a gain in positive predictive value can make up for an equal amount of loss in negative predictive value and vice versa. However, there are situations when either the positive or negative predictive value is more important than the other. To accommodate such situations, the definition of $r(c)$ can be generalized to

$$r_\alpha(c) = [1 - \text{PPV}(c)] + \alpha[1 - \text{NPV}(c)], \quad (5.3)$$

where α indicates the ratio of the penalty due to the loss in PPV and NPV.

Figure 7 shows the distance to perfect prediction as a function of positivity threshold from the SUV-lean data that will be presented in §6a. Such a graph can provide valuable descriptive information about the behaviour of r in a particular range of threshold values and the overall variation of predictive performance. If in practice it is impossible to set a threshold explicitly, as is the case in reader interpretations of diagnostic imaging, a test with smaller variation in predictive performance might be preferable.

Figure 7 also presents information about r^* , the best predictive performance of a diagnostic test, as well as the optimal operating threshold on which the best performance is achieved. In practice, we need to evaluate whether the optimal threshold value is realistic. We also need to evaluate how far the test performance, under design settings that are currently the most common, is from the test's optimal performance. Moreover, it is useful to know how much moving the threshold influences the predictive performance. If predictive performance does not increase much by the change of the positivity threshold, then we should evaluate whether clinically it is worth the effort to choose and operate at the optimal threshold to gain only a certain increase in predictive accuracy.

(b) *Average of predictive performance*

A summary measure of predictive performance over a range of positivity threshold is provided by the integrated value of $r(c)$

$$R_\alpha = \int r_\alpha(c)f(c) \, dc, \quad (5.4)$$

where f is a weight function defined over values of the positivity threshold. If all positivity thresholds were equally important over a chosen interval, a uniform distribution over the interval would be appropriate. For a continuous diagnostic test, the weight function can be chosen as the empirical distribution of observed test results. For that choice, the average performance R is identical to two minus the sum of the areas under the PPV curve and the NPV curve proposed by Moskowitz & Pepe (2004). Note that $0 \leq r_\alpha(c) \leq 1$, and R_α is therefore bounded by $1 + \alpha$. R_α represents the average loss from perfect prediction, with smaller values indicating better average predictive performance.

(c) *Comparison between tests*

Differences in the predictive performance of tests may be due to differences in disease prevalence under which the tests were evaluated. This type of confounding is addressed automatically in studies using a paired design. In order to compare predictive performance estimates from other types of studies, an adjustment for differences in prevalence is needed. A PROC curve obtained in one population can be transformed to apply to a population with a different prevalence using the following formulae:

$$y_2(c) = \frac{y_1(c)(1 - p_1)p_2}{y_1(c)(1 - p_1)p_2 + (1 - y_1(c))p_1(1 - p_2)},$$

$$x_2(c) = \frac{x_1(c)(1 - p_1)p_2}{x_1(c)(1 - p_1)p_2 + (1 - x_1(c))p_1(1 - p_2)},$$

where $\Gamma_1(c) = \{(x_1(c), y_1(c))\}$ is the PROC curve of some test performed on a population with disease prevalence p_1 , and $\Gamma_2(c) = \{(x_2(c), y_2(c))\}$ is the PROC curve of the same test performed on a population with disease prevalence p_2 . This transformation is useful in comparisons of diagnostic tests applied to different populations.

We can compare tests on the basis of $r_\alpha(c)$, as well as on the basis of the average predictive performance R_α . Denote by $r_\alpha^A(c)$ and $r_\alpha^B(c)$ the $r_\alpha(c)$ statistic of test A and test B, respectively. We can observe whether the confidence bands for $r^A(c)$ and $r^B(c)$ overlap over a threshold range of particular interest. Similarly, we can observe whether the CIs for r_α^{A*} and r_α^{B*} , or R_α^A and R_α^B , overlap. If one test is uniformly better than the other over a specific range of threshold, or at a threshold value of particular interest, then we should further evaluate whether these thresholds are reasonable for conducting clinical studies.

6. Examples

To illustrate the ideas and methods developed in this paper, we present the following two examples. The first example used data from a prospective multi-centre study (Walh *et al.* 2004) to evaluate the predictive accuracy of SUV-lean in detecting axillary nodal metastases in women with primary breast cancer. The second example used data from the Digital Mammographic Imaging Screening Trial (DMIST; Pisano *et al.* 2005) to assess and compare the predictive performance of digital and film mammography for breast cancer detection. In both studies, participants were recruited prospectively and the estimates of the PPV and NPV generalize directly to the target populations.

(a) PROC analysis for SUV-lean data

In this study, 360 women with newly diagnosed invasive breast cancer were enrolled and imaged with FDG-PET. We used a subset of 308 cases with assessable axillae for this analysis. SUV-lean, which is a quantitative measurement of the single-pixel maximal standardized uptake value normalized to lean body mass, was determined only for patients with PET imaging score at least 2 (2=equivocal; 3=probably abnormal; 4=definitely abnormal). For patients with PET imaging score 0 (=normal) or 1 (=probably normal), the SUV-lean was imputed by 0.6 and 1. The choice of imputed values was made by the investigators based on the past experience with normal or probably normal cases. The reference standard information for the actual presence of axillary node involvement was derived from surgical exploration of the axilla and subsequent patient follow-up. Figure 5 plots the empirical predictive accuracy of the SUV-lean in terms of the PPV, NPV and the PROC curve.

To estimate the PROC curve, we first estimated the ROC curve of the continuously distributed SUV-lean using the maximum-likelihood method proposed by Metz *et al.* (1998). With parameter estimates $\hat{a} = 1.134$ (s.e. = 0.205) and $\hat{b} = 1.704$ (s.e. = 0.261), figure 6 presents the estimated PROC curve. The curve was derived from the likelihood method on the joint distribution of T and D , as described in the beginning of §4.

We are interested in the curve segment that corresponds to the range of observed SUV-lean. This segment has a range of positive predictive value [0.468, 1] and a range of negative predictive value [0.653, 0.791]. According to proposition 3.2 in §3b, this curve segment is monotone and displays trade-off between the PPV and NPV. Therefore, for this single dataset, we can use summary measures analogous to those developed for ROC analysis. The areas under this curve segment are $A1 = 0.113$ that corresponds to an average positive predictive value of 0.819 and $A2 = 0.116$ that corresponds to an average negative predictive value of 0.782. These results suggest that the positive and negative predictions by SUV simultaneously reach a reasonably satisfactory level of accuracy.

In addition to the average predictive performance, we also determine the optimal operating point at which the SUV-lean achieves its best predictive performance. Figure 7 shows the performance of SUV-lean by the cut-off point. The optimal performance occurred at 3.23 on the latent scale, which corresponds to an SUV-lean cut-off of 5.13. At the optimal point, the positive predictive value was 0.990 with 95% CI (0.975, 0.996) and the negative predictive value was 0.674

Figure 5. The empirical predictive performance of SUV-lean. (a) Positive predictive value; (b) negative predictive value; (c) PROC curve.

Figure 6. The estimated PROC curve of SUV-lean (dot-dashed line). The circles indicate the pairs of empirical PPV and NPV.

with 95% CI (0.614, 0.728). The small variation of $r(c)$ on the neighbourhood of c^* suggests that the performance of SUV-lean is insensitive to the estimation errors in this optimal threshold.

(b) Application to DMIST data

The DMIST conducted by the American College of Radiology Imaging Network was designed to measure the difference in diagnostic performance between digital and screen-film mammography for the purpose of breast-cancer screening (Pisano *et al.* 2005). In this example we used data from a retrospective reader study that randomly selected participants from DMIST study sample. Each participant underwent both digital and screen-film mammography, and mammograms from the same participant were interpreted by a radiologist using seven-point malignancy scale. Figure 8 shows empirical and estimated PROC curves from three readers. To make this example parsimonious and also to show

Figure 7. The estimated distance function of SUV-lean.

clear trade-off between the PPV and NPV, in the following we select data from reader 3 to present an analysis that compares the predictive accuracy of digital and film mammography.

Among the 94 participants in the dataset, only one participant was rated as ‘7=definitely malignant’. Therefore, the seventh and sixth categories were combined for analysis purpose. The pseudo-likelihood function constructed in §4*b* was used to adjust for correlations induced by data replication and the paired study design. The delta method in the logit scale was used to calculate 95% CIs. [Figure 9](#) presents the results of pseudo-likelihood estimation. The film mammography showed uniformly higher estimates of predictive performance than the digital mammography over all positivity thresholds. However, the difference was not statistically significant as the CIs overlapped. [Figure 10](#) also presents the partial PROC curves without the vertical and horizontal segments that correspond to extreme positivity thresholds (beyond 15.03 and smaller than -23.71 on the latent scale). Both estimated curve segments have NPV ranging from 0.73, which is one minus the prevalence, to 1. Over this range, the average positive predictive value was 0.613 with 95% CI (0.411, 0.853) for film mammography and 0.598 with 95% CI (0.411, 0.853) for digital mammography. Lastly, the area under the loss function R over the range of ± 2 s.d. using $\alpha=1$ was 0.587 with 95% CI (0.471, 0.718) for film mammography and 0.632 with 95% CI (0.516, 0.762) for digital mammography. Overall, we conclude that there was no statistically significant difference in predictive accuracy between the digital and screen-film mammography.

7. Discussion

The unique aspect of the PROC curve is that the curve shows how PPV and NPV vary jointly as a function of the positivity threshold. Thus, the PROC curve shows all achievable pairs of PPV and NPV. In this sense, the PROC curve can be interpreted similarly to the ROC curve. The area under a monotone segment of the PROC curve can be interpreted as the average positive predictive value over a range of NPV. However, an analogue to the probabilistic interpretation of the area under the ROC curve does not seem to exist in the PROC context.

Figure 8. The empirical (solid line) and estimated (dot-dashed line) PROC curves of digital and screen-film mammography from three readers. The circles represent the pairs of empirical PPV and NPV. The number of participants, the prevalence and the estimates of parameters a and b are also provided. (a) Reader 1 film ($N=118$, $p=0.398$, $a=1.749$, $b=1.537$); (b) reader 2 film ($N=119$, $p=0.395$, $a=1.882$, $b=1.927$); (c) reader 3 film ($N=94$, $p=0.269$, $a=1.231$, $b=0.918$); (d) reader 3 digital ($N=94$, $p=0.269$, $a=1.031$, $b=1.044$).

In addition, in contrast to the simple trade-off between sensitivity and specificity, as captured by the ROC curve, the dependence of PPV and NPV individually and jointly on the positivity threshold does not follow simple patterns. A detailed characterization of the monotonicity of the PROC curve and its segments shows that the patterns are complex even for the binormal model. The complexity of the PROC curve reflects the complexity of the dependence between the PPV and NPV of a test. When both quantities are of interest, a separate analysis of each carries a notable element of risk for reaching misleading conclusions.

While in many situations, scientific inferences and clinical decision-making may not need to make use of the full PROC curve, such as the case when clinical interests lie in estimating the positive predictive value of a test given a certain range of NPV, or vice versa, the non-monotonicity of the full curve becomes less a concern and the monotonicity of a curve segment of interest is indeed more relevant. In situations when the full curve contains segments that correspond to positivity thresholds that are very unlikely to occur in practice, such as the visually vertical or horizontal segments in almost every PROC curve, the analyst can consider only segments of the curve instead of the full curve. Generally speaking, a PROC curve segment corresponding to positivity threshold of clinical

Figure 9. Pseudo-likelihood analysis of mammography data. (a) The estimated latent distributions of the diseased and non-diseased subjects under the binormal assumption with estimated operating thresholds $c_1, \dots, c_5$, (b) the estimated positive predictive value, (c) the estimated negative predictive value and (d) the estimated distance to the perfect prediction with 95% CIs. (a) Dotted line, non-diseased; solid line, diseased film; dot-dashed line, diseased digital. (b–d) Solid line, film; dot-dashed line, digital.

Figure 10. The estimated partial PROC curves of film (solid line) and digital mammography (dot-dashed line).

relevance should be considered as opposed to the use of the full curve, and the properties of segments of interest can be identified through proposition 3.2. These results, however, are valid only for the binormal model. Similar investigations can be undertaken under other distributional assumptions. Non-parametric models can also be considered and may lead to more robust conclusions, as long as it is possible to verify the conditions of hazard ordering.

This paper emphasizes the importance of including the positivity threshold in an analysis of predictive values. The summary measure $r(c)$, which retains threshold information, can be used to study the influence of the selection of a particular threshold on the predictive performance of a diagnostic test and to incorporate the threshold effect into test comparisons. The optimal operating point can also be obtained on the basis of this measure $r(c)$. In contrast to the ROC curve that typically shows a trade-off between sensitivity and specificity, we note that in situations when $b \neq 1$ the PROC curve has one monotone segment of $c \in [c_{\text{PPV}}^*, c_{\text{NPV}}^*]$, on which both the PPV and NPV can either increase (when $b > 1$) or decrease (when $b < 1$). This makes the selection of an optimal threshold on this interval a trivial task, as either the left or the right endpoint of this interval dominates the rest. However, clinical interest may not lie in this segment, as most of the observed points usually occur within the range of $c \in (-\infty, c_{\text{PPV}}^*)$ or of $c \in (c_{\text{NPV}}^*, \infty)$, depending on the value of parameter b .

We propose summary measures such as the average loss R and the area under the full curve or a curve segment to represent the overall predictive performance. The average positive predictive value over a specific range of negative predictive value, or vice versa, is also useful summary measures. Regardless of which measure is chosen for the analysis, however, an understanding of the shape of the PROC curve, based on which the inference would be drawn, is still essential.

The dependence of predictive values on disease prevalence is another concern associated with the use of predictive values in the evaluation of diagnostic tests. This issue is particularly problematic when tests are to be compared on the basis of data that did not arise from a study using a paired design, even if they were prospective studies. For example, this would be ordinarily the case if data for each of the tests come from separate studies. One approach to addressing this challenge is the transformation of curves described in §5c. Such a transformation makes it possible to compare PROC curves at any given prevalence of interest, and therefore may be very useful in situations such as disease screening and translation of test results from one population to another. It may also be possible to incorporate covariates related to the prevalence, for example, as in the model for ordinal data presented in §4. However, our experience with the use of such modelling in empirical investigations is so far limited to relatively simple settings with binary covariates.

In conclusion, we note that although the patterns of covariation of PPV and NPV are complex, the resulting difficulty does not imply that the dependence on threshold is not important in practice. Since the clinical use of a diagnostic test is primarily based on the predictive value of the test, a thorough assessment of predictive values and their properties can provide important information.

This work is a part of the thesis of Shang-Ying Shiu at Center for Statistical Sciences, Brown University. The authors would like to thank the thesis committee members Joseph Hogan, Jeffrey Blume and Alicia Toledano for their invaluable comments that have helped to improve this work substantially. The authors would also like to thank the reviewers for their thoughtful comments and suggestions. This work was funded partly by grant CA 79778 from the National Cancer Institute.

References

- Aerts, M., Geys, H. & Molenberghs, G. 2002 *Topics in modelling of clustered data*, ch. 6. Monographs on Statistics and Applied Probability no. 96. Boca Raton, FL: Chapman & Hall.
- Alonzo, T. A. & Pepe, M. S. 2002 Distribution-free ROC analysis using binary regression techniques. *Biostatistics* **3**, 421–432. (doi:10.1093/biostatistics/3.3.421)
- Arnold, B. C. & Strauss, D. 1991 Pseudolikelihood estimation: some examples. *Sankhya Ser. B* **53**, 223–243.
- Beam, C. A., Conant, E. F. & Sickles, E. A. 2003 Association of volume and volume-independent factors with accuracy in screening mammogram interpretation. *J. Natl Cancer Inst.* **95**, 282–290.
- Bennett, B. M. 1985 On tests for equality of predictive values for t diagnostic procedures. *Stat. Med.* **4**, 535–539. (doi:10.1002/sim.4780040414)
- Bura, E. & Gastwirth, J. L. 2001 The binary regression quantile plot: assessing the importance of predictors in binary regression visually. *Biom. J.* **43**, 5–21. (doi:10.1002/1521-4036(200102)43:1<5::AID-BIMJ5>3.0.CO;2-6)
- Copas, J. 1999 The effectiveness of risk scores: the logit rank plot. *Appl. Stat.* **48**, 165–183.
- Hanley, J. A. 1989 Receiver operating characteristic (ROC) methodology: the state of the art. *Crit. Rev. Diagn. Imaging* **29**, 307–335.
- Hanley, J. A. 1998 Receiver operating characteristic (ROC) curves. In *Encyclopedia of biostatistics* (eds P. Armitage & T. Colton), pp. 3738–3745. New York, NY: Wiley.
- Hanley, J. A. & McNeil, B. J. 1982 The meaning and use of the area under a receiver characteristic (ROC) curve. *Radiology* **13**, 129–133.
- Huang, Y., Pepe, M. S. & Feng, Z. 2007 Evaluating the predictiveness of a continuous marker. *Biometrics* **63**, 1181–1188. (doi:10.1111/j.1541-0420.2007.00814.x)
- Irwig, I., Tosteson, A. N., Gatsonis, C. A., Lau, J., Colditz, G., Chalmers, T. C. & Mosteller, F. 1994 Guidelines for meta-analyses evaluating diagnostic tests. *Ann. Int. Med.* **120**, 667–676.
- Ishwaran, H. & Gatsonis, C. A. 2000 A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Can. J. Stat.* **28**, 731–750. (doi:10.2307/3315913)
- Leisenring, W., Alonzo, T. A. & Pepe, M. S. 2000 Comparisons of predictive values of binary medical diagnostic tests for paired designs. *Biometrics* **56**, 345–351. (doi:10.1111/j.0006-341X.2000.00345.x)
- Metz, C. E., Herman, B. A. & Shen, J.-H. 1998 Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Stat. Med.* **17**, 1033–1053. (doi:10.1002/(SICI)1097-0258(19980515)17:9<1033::AID-SIM784>3.0.CO;2-Z)
- Moons, K. G. & Harrel, F. E. 2003 Sensitivity and specificity should be de-emphasized in diagnostic accuracy studies. *Acad. Radiol.* **10**, 670–672. (doi:10.1016/S1076-6332(03)80087-9)
- Moons, K. G., van Es, G. A., Michel, B. C., Buller, H. R., Habbema, J. D. & Grobbee, D. E. 1997 Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology* **8**, 12–17. (doi:10.1097/00001648-199701000-00002)
- Moskowitz, C. S. & Pepe, M. S. 2004 Quantifying and comparing the predictive accuracy of continuous prognostic factors for binary outcomes. *Biostatistics* **5**, 113–127. (doi:10.1093/biostatistics/5.1.113)
- Pepe, M. S. 1998 Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* **54**, 124–135. (doi:10.2307/2534001)
- Pepe, M. 2000a An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* **56**, 352–359. (doi:10.1111/j.0006-341X.2000.00352.x)
- Pepe, M. 2000b Receiver operating characteristic methodology. *J. Am. Stat. Assoc.* **95**, 308–311. (doi:10.2307/2669554)
- Pepe, M. S. 2003 *The statistical evaluation of medical tests for classification and prediction*. Oxford Statistical Science Series, vol. 30. Oxford, UK: Oxford University Press.

- Pisano, E. D. *et al.* 2005 Diagnostic performance of digital versus film mammography for breast-cancer screening. *N. Engl. J. Med.* **353**, 1773–1783. (doi:10.1056/NEJMoA052911)
- Rutter, C. M. & Gatsonis, C. A. 2001 A hierarchical regression approach to meta-analysis of diagnostic test accuracy evaluations. *Stat. Med.* **20**, 2865–2884. (doi:10.1002/sim.942)
- Shaked, M. & Shanthikumar, J. G. 1994 *Stochastic orders and their applications*. London, UK: Academic Press.
- Toledano, A. & Gatsonis, C. 1996 Ordinal regression methodology for ROC curves derived from correlated data. *Stat. Med.* **15**, 1807–1826. (doi:10.1002/(SICI)1097-0258(19960830)15:16<1807::AID-SIM333>3.0.CO;2-U)
- Tosteson, N. A. & Begg, C. B. 1988 A general regression methodology for ROC curve estimation. *Med. Decis. Mak.* **8**, 204–215. (doi:10.1177/0272989X8800800309)
- Walh, R. L., Siegel, B. A., Coleman, R. E. & Gatsonis, C. G. 2004 Prospective multicenter study of axillary nodal staging by positron emission tomography in breast cancer: a report of the staging breast cancer with pet study group. *J. Clin. Oncol.* **22**, 277–285.
- Zhou, X., Obuchowski, N. A. & McClish, D. K. 2002 *Statistical methods in diagnostic medicine*. Wiley Series in Probability and Statistics, ch. 4.3.5, p. 152. New York, NY: Wiley.