UW Biostatistics Working Paper Series

1-22-2008

Estimation and Comparison of Receiver Operating Characteristic Curves

Margaret Pepe
University of Washington, Fred Hutch Cancer Research Center, mspepe@u.washington.edu

Gary M. Longton
Fred Hutchinson Cancer Research Center, glongton@fhcrc.org

Holly Janes
Fred Hutchinson Cancer Research Center, hjanes@scharp.org

Estimation and Comparison of Receiver Operating Characteristic Curves

Margaret Pepe, Gary Longton, and Holly Janes
Fred Hutchinson Cancer Research Center
Seattle, Washington, USA
mspepe@u.washington.edu

January 10, 2008

Abstract

The receiver operating characteristic (ROC) curve displays the capacity of a marker or diagnostic test to discriminate between two groups of subjects, cases versus controls. We present a comprehensive suite of Stata commands for performing ROC analysis. Non-parametric, semiparametric and parametric estimators are calculated. Comparisons between curves are based on the area or partial area under the ROC curve. Alternatively pointwise comparisons between ROC curves or inverse ROC curves can be made. Options to adjust these analyses for covariates, and to perform ROC regression are described in a companion article. We use a unified framework by representing the ROC curve as the distribution of the marker in cases after standardizing it to the control reference distribution.

1 Introduction

1.1 Definition of the ROC Curve

The receiver operating characteristic curve (ROC) displays the discriminatory capacity of a marker or test. Suppose $D = 0$ denotes controls and $D = 1$ denotes cases and assume without loss of generality that larger values of Y are more indicative of a subject being a case. The ROC curve for a marker, Y, is a plot of the true positive rate $\text{TPR}(c) = P[Y \geq c|D = 1]$ versus the false positive rate $\text{FPR}(c) = P[Y \geq c|D = 0]$ for the thresholding criterion '$Y > c$' where c varies from $-\infty$ to ∞. It is a monotone increasing function in the unit square tied down at the boundary points (0,0) and (1,1). A perfect classifier has an ROC curve that rises steeply along the left axis to the point (FPR=0, TPR=1), while an uninformative marker has an ROC curve that is the diagonal 45° line. Key attributes of the ROC curve are: (i) it does not depend on the raw measurement units for Y. It is invariant to monotone increasing transformations of Y; (ii) it provides a common scale for comparing performances of different markers; and (iii) it displays the range of possible performance levels that can be achieved by varying the threshold.

Figure 1 shows empirical ROC curves for 2 pancreatic cancer biomarkers (Wieand, Gail, James, et al. 1989). The data can be downloaded from the Diagnostic and Biomarker Statistical Center website (http://www.fhcrc.org/labs/pepe/dabs/), or loaded directly into a Stata session:

.use http://www.fhcrc.org/science/labs/pepe/book/data/wiedat2b

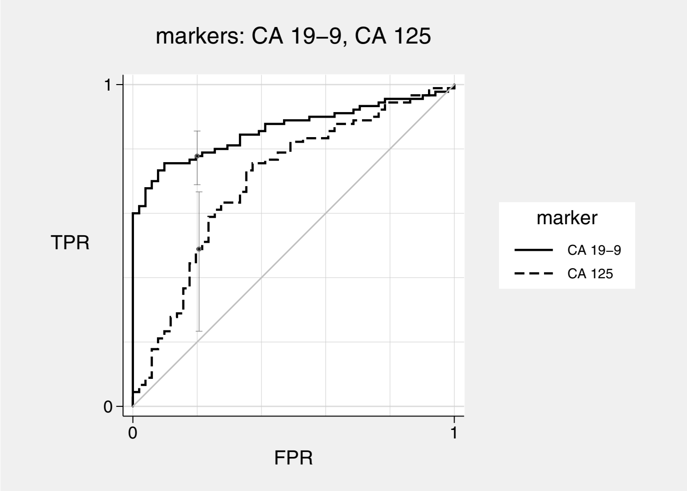

Figure 1: Non-parametric ROC curves for two markers of pancreatic cancer. 90% confidence intervals for ROC(0.2) are displayed.

1.2 Representation in terms of percentile values

Let F denote the right continuous cumulative distribution of Y in the control population, $F(y) = P(Y < y|D = 0)$. We define a standardization of Y:

$$\text{pv}_i = F(Y_i)$$

is the proportion of the control population with values below Y_i. In lay terms, $pv_i \times 100$ is the percentile of Y_i when the controls are considered the reference population against which to standardize the marker. We now show that the ROC curve can be written as the distribution of these standardized marker measurements in cases (Pepe and Cai, 2004; Pepe and Longton, 2005). This identity suggests simple algorithms for implementing standard ROC methods and also gives rise to some new methods (Huang and Pepe, 2007).

<u>Result</u>

The ROC curve is the cumulative distribution of $1 - \mathrm{pv}_D$,

$$\mathrm{ROC}(f) = P[1 - \mathrm{pv}_D \le f],$$

where pv_D denotes the standardized marker for a case.

Proof

Let y be the threshold that corresponds to a false positive rate f. By definition, a proportion f of the controls have marker values above y, $F(y) = 1 - f$. Since F is monotone increasing

$$\begin{aligned}\mathrm{ROC}(f) &\equiv P[Y_D \ge y] \\ &= P[F(Y_D) \ge F(y)] \\ &= P[pv_D \ge 1 - f] = P[1 - \mathrm{pv}_D \le f]\end{aligned}$$

2 Estimating the ROC Curve

The representation in Result 1 suggests that ROC curve estimation can be accomplished in two steps:

(i) Estimate the reference cumulative distribution function (CDF), F, using controls; and calculate corresponding standardized marker values for cases, and

(ii) Estimate the cumulative distribution of the standardized marker values for cases.

2.1 The Control Reference Distribution

The empirical estimator of the control reference distribution can be employed. Alternatively a parametric model can be assumed. The `roccurve` command allows one to use either the empirical method or a normal parametric distribution model.

Marker values for cases are standardized using the estimator $\widehat{F}$. Write the standardized values as

$$\widehat{\mathrm{pv}}_{Di} = \widehat{F}(Y_{Di}) \quad i = 1, \ldots n_D$$

where n_D is the number of case observations.

2.2 The CDF of Standardized Markers

The next step is to estimate the CDF of $1 - \mathrm{pv}_D$, denoted by H. The empirical CDF is a nonparametric option provided by `roccurve`. A parametric model can be used instead. This has the advantage of providing a smooth ROC curve instead of a step function. The parametric forms allowed by `roccurve` are:

$$H(f) = g(\alpha_0 + \alpha_1 g^{-1}(f))$$

where g is a CDF. Observe that this form acknowledges that the domain for H is restricted to (0,1). As a special case, when $g = \Phi$, the standard normal distribution, the corresponding ROC curve is binormal (Dorfman and Alf, 1969),

$$\mathrm{ROC}(f) = H(f) = \Phi(\alpha_0 + \alpha_1 \Phi^{-1}(f)).$$

The `roccurve` command also allows the logistic form, $g(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$, which gives rise to a biologistic ROC curve (Ogilvie and Creelman, 1968).

To fit these parametric models a set of discrete points on the FPR axis is chosen, $\{f_1, \ldots, f_{n_p}\}$. For each case i and for each f_k, a record is created that includes the binary variable, $U_{ki} = I[1 - \widehat{\mathrm{pv}}_{Di} \leq f_k]$, and covariate $g^{-1}(f_k)$. A binary regression model with link function g, outcome variable U and covariate $g^{-1}(f)$ yields estimates of (α_0, α_1) (Alonzo and Pepe 2002).

In some applications one may only want to model the ROC curve over a restricted FPR range, $(a, b) \subset (0, 1)$, in which case the FPR points $\{f_1, \ldots f_{n_p}\}$ should span the interval (a, b).

In figure 2 we display four different estimators applied to data on the pancreatic cancer biomarker CA-125. The first estimator is the standard empirical ROC curve that results from standardizing with the right continuous empirical control reference distribution and applying the empirical CDF for H. This is precisely the same as the empirical estimator that is provided by Stata's `roctab` command. The second estimator is the semiparametric binormal estimator that again calculates the standardized values with the empirical control distribution for Y but employs a probit link function for g. This rank invariant semiparametric estimator requires less computation than the binormal estimator provided by Stata's `rocfit` command and appears to have similar efficiency (Alonzo and Pepe 2002). The third estimator assumes that the marker is normally distributed in controls and is not rank invariant. It calculates standardized values as

$$\mathrm{pv}_{Di} = \Phi((Y_{Di} - mean)/sd)$$

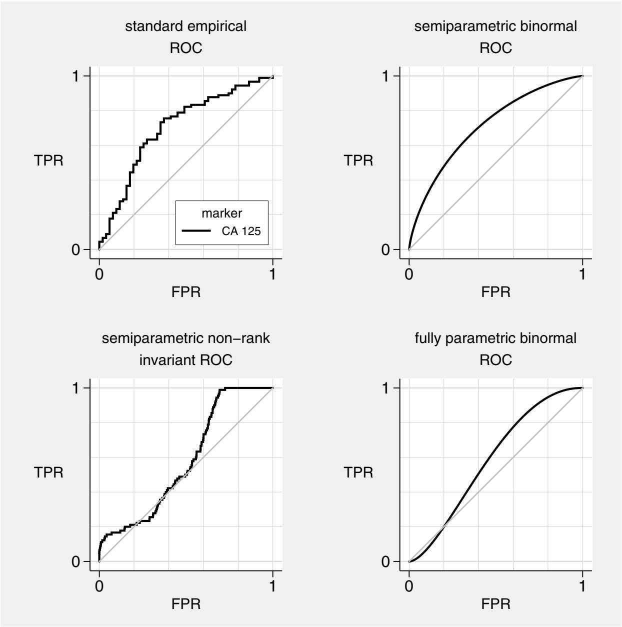

Figure 2: ROC curves for CA-125 as a marker of pancreatic cancer.

where $(mean, sd)$ are the sample mean and standard deviation of the control observations. The fourth estimator is fully parametric. In addition to modelling the control reference distribution as normal it assumes the ROC curve is binormal. The two assumptions taken together are equivalent to assuming markers for both cases and controls are normally distributed. In practice the rank invariant estimators are more popular. Parametric models for the reference distribution have a more prominent role in settings where covariates affect marker distributions and covariate-specific distributions are difficult to estimate empirically (Janes, Longton and Pepe, 2007).

3 Sampling Variability

We use bootstrap resampling to calculate pointwise confidence intervals for the ROC curve, $\mathrm{ROC}(f)$, and for its inverse, $\mathrm{ROC}^{-1}(t)$. In particular, if f is the false positive rate, the $(1-\alpha/2)$ and $\alpha/2$ quantiles of the bootstrap distribution of $\widehat{\mathrm{ROC}}(f)$ are delivered as the $(1-\alpha)$ confidence limits.

The resampling must reflect the study design. If selection to the study was outcome dependent, that is if a case-control design was employed as is common in early phase studies (Pepe, Etzioni, Feng, et al. 2001), then resampling is done separately within case and control strata. On the other hand, if subjects were enrolled without regard to their outcome status, resampling is done accordingly from the entire dataset. In addition, if observations are clustered, for example if subjects contribute several observations to ROC curve estimation, the `cluster()` option can be used to identify resampling clusters.

4 The `roccurve` Command

4.1 Syntax

The syntax for the `roccurve` command is

```
roccurve disease_var test_varlist [if] [in] [, options]
```

where `disease_var` gives the name of the binary outcome variable, $D = 1$ for a case and $D = 0$ for a control and `test_varlist` gives the names of markers or tests for which ROC curves are to be calculated

4.2 Options

4.2.1 Standardization Method

`pvcmeth`(*method*) specifies how $\widehat{F}$ is estimated. Options include *empirical* (the default), where $\widehat{F}$ is the empirical control marker distribution, and *normal*, that assumes a normal distribution and estimates the control mean and variance with the sample mean and variance.

`tiecorr` indicates that a correction for ties between case and control values is included in the empirical pv calculation. The correction is important in calculating summary indices such as the area under the ROC curve that is discussed later. The tie corrected pv for a case with marker Y_i is the proportion of control values $Y_{\bar{D}} < Y_i$ plus 1/2 the proportion of control values $Y_{\bar{D}} = Y_i$.

4.2.2 ROC calculation

`rocmeth`(*method*) specifies whether the *empirical*(default) or a *parametric* model for the ROC is used.

`link`(*link*) is relevant for a parametric ROC model. For a binormal model, link is specified as *probit* while the link is specified as *logit* for the bilogistic model.

`interval`(a b n_p) specifies the interval (a, b) and number of points (n_p) in the interval over which the parametric ROC model is to be fit. The program uses equally spaced points in the interval. Default values are $a = 0$, $b = 1$, and $n_p = 10$.

roc(f) specifies the false positive rate, f, for calculation of point estimates for ROC(f) and confidence intervals.

rocinv(t) specifies the true positive rate, t, for calculation of point estimates for $\mathrm{ROC}^{-1}(t)$ and confidence intervals.

4.2.3 ROC plot

nograph suppresses the ROC plots; when only returned numerical results are desired.

bw specifies that black line types be used to distinguish between ROC curves rather than solid colors (default). The graphics scheme *s1mono* is used if bw is specified; the *s1color* scheme is used otherwise. Either scheme can be overridden by explicitly specifying any other graphics scheme as a separate twoway option.

twoway options various graph options overriding default axis options, titles, and overall graph appearance. Exceptions include specific ROC line and marker type options and the by() option.

offset(#) specifies the x-axis offset from f for placement of second and subsequent CIs for ROC(f) or $\mathrm{ROC}^{-1}(t)$ to avoid overlap of interval bars for different markers.

4.2.4 Sampling Variability

This is only relevant if either of the roc(f) or rocinv(t) options are specified.

nsamp(#) specifies the number of bootstrap replications to be performed for estimating confidence intervals. The default is 1000 replications.

noccsamp specifies that bootstrap samples be drawn from the combined sample rather than sampling separately from cases and controls; case-control sampling is the default.

cluster(varlist) specifies variables identifying bootstrap resampling clusters.

level(#) specifies the confidence level, as a percentage, for confidence intervals.

4.2.5 Additional Options

There are options to create new variables.

genrocvars generates new pairs of variables, fpr# and tpr# for each marker in the test_varlist, with ROC coordinates for corresponding marker values. Point resulting from the empirical rocmeth() are plotted as a right-continuous step function. New variable names are numbered (#) according to variable order in the `test_varlist`.

`genpcv` generates variables, pcv#, to hold percentile values for each marker in the `test_varlist`. The numbers (#) correspond to marker variable order in the `test_varlist`.

`replace` requests that existing variables fpr# , tpr# or pcv# be overwritten by `genpcv` or `genrocvar`.

There are also options to adjust the ROC curve estimates for covariates. These options are described in another article in this journal (Janes, Longton and Pepe, 2007).

4.2.6 Example

The following code produced the plot in Figure 1:

```
use http://www.fhcrc.org/science/labs/pepe/book/data/wiedat2b
roccurve d y1 y2, roc(.2) level(90)
```

The 4 panels in Figure 2 were produced using the following 4 commands:

```
roccurve d y2, pvcmeth(empirical) rocmeth(nonparametric)
roccurve d y2, pvcmeth(empirical) rocmeth(parametric) link(probit)
roccurve d y2, pvcmeth(normal) rocmeth(nonparametric)
roccurve d y2, pvcmeth(normal) rocmeth(parametric)
```

5 Summary Indices

5.1 Area and partial Area

Measures derived from the ROC curve are used to summarize discriminatory accuracy. More importantly, they serve as the basis for test statistics to compare ROC curves. The most popular index is the area under the ROC curve (AUC), also known as the c-index or probability of correct ordering, AUC= Prob($Y_D > Y_N$) where (Y_D, Y_N) are a random pair of case and control marker values. We and others (Pepe 2003, pg 78; Cook, 2007) have argued against using the AUC as a key summary measure because it is not clinically relevant. Subjects do not present clinically as pairs and typically the clinical problem is not to decide which member of such a pair is the case.

For clinical applications we prefer use of the ROC (or ROC^{-1}) at a specific point. Consider ROC(f). Given that one is willing to accept a false positive rate (f), what proportion of cases will be detected? This is relevant to clinical practice. However, fixing one FPR of interest can be difficult. A compromise is the partial AUC that averages the ROC curve over a range of false positive rates (McClish 1989, Thompson and Zucchini 1989). Since low FPR are typically of interest, one can calculate the partial area between 0 and the largest acceptable FPR, denoted by f_0:

$$\text{pAUC}(f_0) = \int_0^{f_0} \text{ROC}(f) df.$$

Interestingly, the classic nonparametric estimator of the AUC can be written as the sample mean of the *nonparametric* case percentile values (Delong et al 1988; Hanley and Hajian-Tilaki, 1997).

$$\widehat{\text{AUC}}_e = \sum_{i=1}^{n_D} \widehat{\text{pv}}_{Di}/n_D \tag{1}$$

When ties between case and control marker values are present, a correction for ties is necessary in calculating the percentile values so that $\widehat{\text{AUC}}_e$ corresponds to the trapezoidal empirical AUC:

$$\widehat{\text{pv}}^c_{Di} = \widehat{\text{pv}}_{Di} + \frac{1}{2}\widehat{e}_i$$

where $\widehat{e}_i$ is the proportion of control marker values equal to Y_{Di}. The empirical estimator of the partial AUC (Dodd and Pepe 2003) can also be written as a sample mean

$$\text{p}\widehat{\text{AUC}}_e(f_0) = \sum_{i=1}^{n_D} max(\widehat{\text{pv}}_{Di} - (1 - f_0), 0)/n_D \tag{2}$$

again with the aforementioned tie correction for cases tied with controls.

By using a parametric model for the control reference distribution, the average of parametric case percentiles yields another estimator of the AUC. Analagously expression (2) with parametric case percentiles provides a semiparametric partial AUC estimate. Note that tie corrections are not necessary when the estimated reference distribution is continuous.

In general, calculation of areas and partial areas under parametric ROC curves requires numerical integration and are not output by our programs. The one exception is that the area under the binormal ROC curve has the closed form expression $\Phi(\alpha_0/\sqrt{1+\alpha_1^2})$. Stata's `rocfit` command provides this after fitting a binormal curve. Our programs do not. We only provide estimates that are non-parametric with regard to the shape of the ROC curve. This is also true for point estiamtes of ROC(f) and ROC$^{-1}(t)$ that are outbut by the `comproc` command.

5.2 Comparisons

To compare ROC curves we calculate a confidence interval for the difference between ROC summary indices. A Wald statistic, dividing the observed difference by its standard error is compared to the standard normal distribution in order to report p-value. Confidence intervals and standard errors are again derived from the bootstrap distribution of the estimators. The `comproc` command outputs results either for the AUC, ROC(f), ROC$^{-1}(t)$ or pAUC(f) where the fixed FPR=f or fixed TPR= t of interest are specified by the data analyst.

6 The `comproc` Command

6.1 Syntax

The syntax of the `comproc` command is

```
   comproc   disease_var   test_var1  [test_var2]   [if] [in] [, options]
```

where `disease_var` is the binary outcome status variable and `test_var1` and `test_var2` are the markers. If only one marker is specified, summary indices are output for that marker but no comparisons are made.

6.2 Options

Options for percentile value calculation and for dealing with sampling variability are the same as described above for the `roccurve` command. Options to include covariate adjustment in making comparisons are described in a companion paper (Janes, Longton and Pepe, 2007).

The options for summary indices to evaluate and to compare between markers are:

`auc`, the area under the ROC curve

`pauc`(f), the partial area under the ROC curve between 0 and f

`roc`(f), the ROC (f), the TPR value corresponding to FPR=f

`rocinv`(t), the ROC$^{-1}(t)$, the FPR value corresponding to TPR= t

6.2.1 Example

The `comproc` command applied to the pancreatic cancer marker data shown in Figure 1 yielded the following results:

```
. comproc d y1 y2, auc roc(0.2)

        Comparison of test measures
                                 test 1: CA 19-9
                                 test 2: CA 125

    percentile value calculation method: empirical
        percentile value tie correction: no

  bootstrap samples drawn
   separately from cases and controls

  # bootstrap samples: 1000

****************
  AUC estimates and difference,
    test 2 - test 1 (aucdelta)

Bootstrap results                               Number of obs      =       141
                                                Replications       =      1000

------------------------------------------------------------------------------
             |    Observed               Bootstrap
             |       Coef.       Bias    Std. Err.   [95% Conf. Interval]
-------------+----------------------------------------------------------------
        auc1 |   .86056644  -.0010577   .03067768    .8004393   .9206936    (N)
             |                                       .7964053   .9174292    (P)
             |                                       .7989107   .9185185   (BC)
        auc2 |   .70413947   .0007451    .0471203    .6117854   .7964936    (N)
             |                                       .6093682   .7955338    (P)
             |                                       .6069717   .7921569   (BC)
    aucdelta |  -.15642697   .0018028   .05788385   -.2698772  -.0429767    (N)
             |                                       -.266122  -.0415033    (P)
             |                                      -.2666667  -.0422658   (BC)
------------------------------------------------------------------------------
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval

test of Ho: auc1 = auc2
  z =    -2.7    p =   .0069

****************
 ROC estimates and difference,
    test 2 - test 1 (rocdelta)

  ROC(f) @ f = .2

Bootstrap results                               Number of obs      =       141
                                                Replications       =      1000

------------------------------------------------------------------------------
             |    Observed               Bootstrap
             |       Coef.       Bias    Std. Err.   [95% Conf. Interval]
-------------+----------------------------------------------------------------
        roc1 |   .77777779   .0011778   .04836552    .6829831   .8725725    (N)
             |                                       .6888889   .8777778    (P)
             |                                             .7   .8888889   (BC)
        roc2 |   .48888889  -.0091667   .13398627    .2262806   .7514971    (N)
             |                                       .2222222         .7    (P)
             |                                       .2333333   .7222222   (BC)
    rocdelta |   -.2888889  -.0103444   .14291224   -.5689918  -.0087861    (N)
             |                                      -.5777777  -.0444444    (P)
             |                                      -.5777777  -.0333334   (BC)
------------------------------------------------------------------------------
(N)    normal confidence interval
(P)    percentile confidence interval
(BC)   bias-corrected confidence interval

test of Ho: roc1 = roc2
  z =      -2    p =   .043
```

Observe that the bootstrap results are output using Stata's `estat bootstrap` command

7 Remarks

Our programs rely on representing the ROC curve as the CDF of the case marker values after they are standardized to the control reference distribution. This representation gives rise to simple algorithms for calculating *standard* nonparametric estimators of the ROC, AUC, and pAUC(f). The representation also provides alternative estimators of the ROC and its summary indices that are semiparametric or fully parametric. In a companion article (Janes, Longton and Pepe, 2007) we describe methods for covariate adjustment and ROC regression. The percentile value representation is particularly useful in these settings.

Applications to continuous data are our focus. Though the methods can be applied to ordinal markers and diagnostic tests, some standard ROC methods for ordinal data are not included in our routines. In particular, our algorithm for fitting the binormal ROC model does not correspond to the Dorfmann and Alf algorithm (Dorfman and Alf, 1969) for ordinal data. In addition, the AUC corresponding to a fitted binormal model is not output. Rather non-parametric AUC estimates are provided. We recommend the `roctab` command in the main Stata package for fitting binormal models and calculating corresponding AUCs with ordinal data.

The DABS Center website is a repository of information for statistical evaluation of diagnostic tests and biomarkers. Included on the website are datasets. They can be used to gain familiarity with methods and software described here.

8 References

Alonzo, T.A., ans M. S. Pepe. 2002. Distribution-free ROC analysis using binary regression techniques. *Biostatistics*3: 421–32.

Janes, H., Longton, G. L., and Pepe, M. S. 2007. Accommodating covariates in ROC analysis. *The Stata Journal* (submitted).

Huang, Y. and Pepe, M. S. 2007. Biomarker evaluation using the controls as a reference population. *Biostatistics* (under revision)

Cook, N. R. 2007. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation* 115: 928–935.

DeLong, E. R., DeLong, D. M. and D. L. Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* 44: 837–45.

Dodd, L., and M. S. Pepe. 2003. Partial AUC estimation and regression. *Biometrics* 59(3): 614–23.

Dorfman, D. D., and E. Alf. 1969. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data *Journal of Mathematical Psychology* 6: 487–496.

Hanley, J. A., and K. O. Hajian-Tilaki. 1997. Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update *Academic Radiology* 4: 49–58.

Ogilvie, J. C. and C. D. Creelman. 1968. Maximum-likelihood estimation of receiver operating characteristic curve parameters. *Journal of Mathematical Psychology* 5: 377–391.

Pepe, M. S. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction* Oxford University Press, United Kingdom.

Pepe, M. S., and T. Cai. 2004. The analysis of placement values for evaluating discriminatory measures. *Biometrics* 60(2): 528–535.

Pepe, M. S., Etzioni, R., Feng, Z., Potter, J. D., Thompson, M. L., Thornquist, M., Winget, M., and Y.Yasui. 2001. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* 93(14): 1054–1061.

Pepe, M. S., and Longton, G. 2005. Standardizing diagnostic markers to evaluate and compare their performance. *Epidemiology* 16(5): 598–603.

Thompson, M. L. and W. Zucchini. 1989. On the statistical analysis of ROC curves. *Statistics in Medicine* 8: 1277–1290.

Wieand, S., Gail, M. H., James, B. R., and K. L. James. 1989. A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* 76: 585–592.