Paper SP09-2009

ROC analysis for the evaluation of continuous biomarkers: Existing tools and new features in SAS® 9.2

Sanghyuk Shin, Beckman Coulter, Inc., San Diego, CA

ABSTRACT

Biomarkers have become essential tools for proper diagnosis and treatment of a wide range of illnesses, including cancer, diabetes, and infectious diseases. The growing need for rigorous evaluation of new biomarkers for medical practice has spurred the development and characterization of statistical methods for diagnostic accuracy. The receiver operating characteristic (ROC) curve is the standard analytical tool for evaluating diagnostic tests. However, recent studies suggest widespread use of inappropriate statistical methods for ROC analysis. In addition, while enhancements in SAS 9.2 have greatly simplified basic ROC analysis for SAS users, many common ROC techniques still require extensive additional programming for SAS users.

This paper provides an overview of statistical methods for evaluating continuous biomarkers using ROC analysis. For each statistical technique, existing tools available in SAS for performing the task are described, with particular emphasis on methods not addressed in previous SAS papers. The paper also introduces new features for ROC analysis that are now available as a standard component of the LOGISTIC procedure in SAS 9.2. Statistical techniques addressed in the paper include the comparison of the area under the ROC curve (AUC) of two or more biomarkers and the generation of confidence intervals for sensitivity given fixed specificity. Pancreatic cancer data (Wieand *et al*. 1989) that have been widely used as sample data in statistical literature are used in this paper to illustrate the application of statistical concepts.

KEYWORDS

ROC, biomarkers, sensitivity, specificity, diagnostic, Proc Logistic

INTRODUCTION

Statistical methods for evaluating the diagnostic accuracy of biomarkers have received increased scrutiny in recent years. In fact, systematic reviews of published studies on diagnostic tests have revealed numerous methodological shortcomings (Lijmer *et al*., 1999; Obuchowsky *et al*., 2004). Use of appropriate statistical techniques is essential in accurately evaluating the performance and clinical utility of a biomarker.

This paper discusses statistical techniques for evaluation of biomarkers and highlights methodological errors often found in medical literature. While the application of the methods presented in the paper is on biomarkers, much of the content of the paper is applicable to any diagnostic test with a continuous outcome. A common study design for biomarker evaluation involves comparison of the performance of a newly developed biomarker to a standard biomarker in the same sample population. Therefore, this paper focuses on “paired” or “correlated” analyses. Additionally, non-parametric techniques are emphasized, as they are more widely applied in biomarker evaluation than parametric ROC analysis involving binormal assumptions. Previous SAS papers have described other useful ROC techniques, including ROC analysis for ordinal outcome data and regression analysis to adjust for covariates (Mandrekar and Mandrekar, 2005; Gönen, 2006). The primary goal of this paper is to describe methods that have not been addressed by previous papers and also to highlight new tools in SAS 9.2 that simplify basic ROC analysis.

SENSITIVITY AND SPECIFICITY

The diagnostic accuracy of a biomarker is most commonly measured by calculating its sensitivity and specificity. Sensitivity is the proportion of patients who are correctly categorized as having disease among those who truly have the disease. Similarly, specificity is the proportion of patients who are correctly categorized as not having the disease among all patients who truly don’t have the disease. Since most diagnostic biomarkers provide results in the continuous scale (e.g. concentration of the CA19-9 tumor marker in serum measured in U/ml), the sensitivity and specificity of the test depends on the specific threshold selected. For example, for the CA19-9 biomarker in the sample data set (Wieand *et al.*, 1989), a value of 37 U/ml has been postulated as the threshold for positive result for pancreatic cancer. Sensitivity and specificity can be calculated for this threshold by constructing a 2 x 2 table with the FREQ procedure:

```
PROC FORMAT;
  VALUE tholdfmt  0 - 37 = "<37 U/ml"
                  37 - high = "37+ U/ml";

  VALUE pcafmt    0 = 'No Cancer'
                  1 = 'Cancer';
RUN;

PROC FREQ DATA=panca ORDER=formatted;
  FORMAT y1 tholdfmt. d pcafmt.;
  LABEL y1='CA19-9' d='Pancreatic Cancer';
  TABLES y1 * d / NOROW NOPERCENT;
run;
```

In the Pancreatic Cancer data set, d is a dichotomous variable for the patient's cancer status and y1 is a continuous variable of the biomarker CA19-9. For the 2 x 2 analysis, y1 is dichotomized by designating it with the THOLDFMT format for the appropriate threshold. The code produces the following output:

Table of y1 by d

y1(CA19-9) d(Pancreatic Cancer)

Frequency Col Pct	Cancer	No Cancer	Total
37+ U/ml	68 75.56	5 9.80	73
<37 U/ml	22 24.44	46 90.20	68
Total	90	51	141

The percent values on the Cancer column are used to determine sensitivity and the percent values on the No Cancer column are used to derive specificity using the CA19-9 threshold of 37 U/ml. Among the 90 cancer patients, 68 tested positive, giving us a sensitivity of 76%. Likewise, 46 of the 51 non-cancer patients tested negative for a specificity of 90%.

When a threshold is predefined as in the example above, confidence intervals for sensitivity and specificity can be calculated using standard statistical methods for binomial data. Again, PROC FREQ is a simple solution for this task. You will have to generate separate output for sensitivity and specificity by using the WHERE statement to select for cancer patients and non-cancer patients. The code below generates the 95% confidence intervals for sensitivity of the CA19-9 marker in the Pancreatic Cancer data set.

```
PROC FREQ DATA=panca ORDER=formatted;
  FORMAT y1 tholdfmt. d pcafmt.;
  LABEL y1='CA19-9' d='Pancreatic Cancer';
  TABLES y1 / BINOMIAL;
  EXACT BINOMIAL;
  WHERE d=1;
run;
```

SAS generates confidence intervals for the proportion in the first row of the output (in this instance, 75.56%). Therefore, you should make sure that the proportions are listed in the order that places sensitivity (or specificity) in the first row. The code above uses ORDER=formatted to instruct SAS to use formatted values for ordering. Otherwise, the output would have generated confidence intervals for 24.44%. The BINOMIAL option in the TABLES statement along with the EXACT BINOMIAL statement instructs SAS to produce confidence intervals using the normal approximation of the binomial distribution (asymptotic standard error or ASE) and the exact binomial distribution. SAS is able to efficiently calculate exact confidence intervals, which is the preferred method. See output below.

CA19-9

y1	Frequency	Percent	Cumulative Frequency	Cumulative Percent
37+ U/ml	68	75.56	68	75.56
<37 U/ml	22	24.44	90	100.00

Binomial Proportion for y1 = 37+ U/ml

Proportion (P)	0.7556
ASE	0.0453
95% Lower Conf Limit	0.6668
95% Upper Conf Limit	0.8443
Exact Conf Limits	
95% Lower Conf Limit	0.6536
95% Upper Conf Limit	0.8400

Test of H0: Proportion = 0.5

ASE under H0	0.0527
Z	4.8488
One-sided Pr > Z	<.0001
Two-sided Pr > |Z|	<.0001
Exact Test	
One-sided Pr >= P	6.246E-07
Two-sided = 2 * One-sided	1.249E-06

Sample Size = 90

One important note: the binomial method of calculating confidence intervals for sensitivity and specificity is only valid if the threshold is predefined. One common mistake in diagnostic medicine literature is the use of simple binomial methods when thresholds are not predefined (Obuchowsky *et al.*, 2004). Often, the relevant research question involves determining the sensitivity and its 95% confidence interval for a predefined specificity. In such instances, the corresponding threshold must be estimated and the sampling variability in estimating the threshold must be taken into account. The same principle holds for deriving specificity for a predefined sensitivity. Appropriate methods to address this problem are discussed below.

ROC ANALYSIS USING THE LOGISTIC PROCEDURE IN SAS 9.2

GENERATING THE ROC CURVE

The empirical ROC curve is the plot of sensitivity on the vertical axis and 1-specificity on the horizontal axis for all possible thresholds in the study data set. It is often used to explore thresholds for the application of a new biomarker in clinical practice or to visually assess the overall performance of the biomarker. With the release of SAS 9.2, ROC curves can be generated using standard ODS STATISTICAL GRAPHICS and simple LOGISTIC procedure statements. The code below generates an ROC curve for the Pancreatic Cancer data.

```
ODS GRAPHICS ON;
   PROC LOGISTIC DATA=panca PLOTS(ONLY)=ROC;
      MODEL d(EVENT='1') = y1;
   RUN;
ODS GRAPHICS OFF;
```

The PLOTS(ONLY)=ROC directs ODS STATISTICAL GRAPHICS to plot an ROC curve without plotting other standard graphs associated with PROC LOGISTIC. The MODEL statement is constructed using the standard PROC LOGISTIC syntax (dependent variable = covariates) with EVENT='1' specified as the outcome we want to predict. The code generates the ROC curve shown in Figure 1.

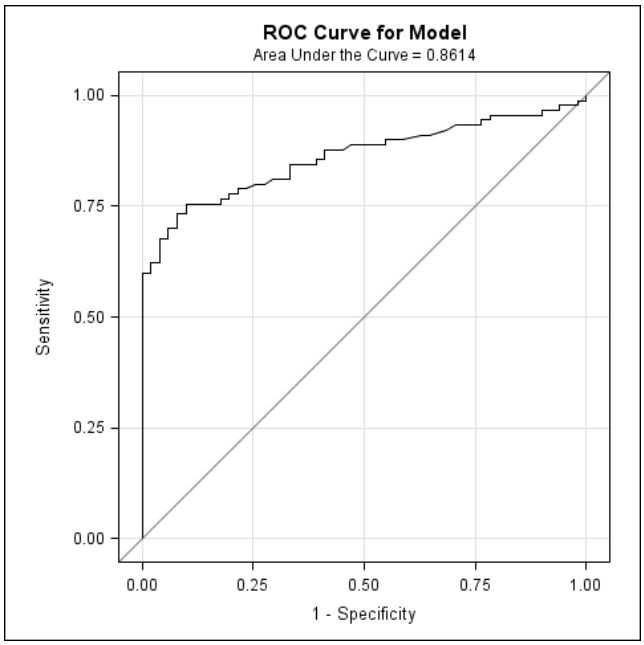

Figure 1. ROC curve for biomarker CA19-9.

THE AREA UNDER THE CURVE

The area under the ROC curve (AUC) is the average sensitivity of the biomarker over the range of specificities. It is often used as a summary statistic representing the overall performance of the biomarker. A biomarker with no predictive value would have an AUC of 0.5 (also represented by the diagonal "chance" line above), while a biomarker with perfect ability to predict disease would have an AUC of 1. The empirical AUC is calculated via the "trapezoidal" rule, where a trapezoid is constructed from the lines drawn for each two consecutive points on the curve. The sum of the areas of the trapezoids is the AUC. Mathematically, the AUC may be defined as

$$AUC = \int_0^1 Sn(t)dt$$

where t is the false positive rate (1-specificity) and the Sn(t) is the corresponding sensitivity (Pepe, 2003). In SAS 9.2, the empirical AUC is calculated and printed at the top of the ROC curve generated by PROC LOGISTIC. As shown in Figure 1, the CA19-9 biomarker has an AUC of 0.86 for the diagnosis of pancreatic cancer in the sample population.

The AUC of a biomarker is often compared to chance which has an AUC of 0.5. The statistical test involves estimating AUC_{test} - AUC_{chance} which is asymptotically normal. A p-value for the null hypothesis that the difference is 0 can be generated using standard Gaussian techniques. The code below performs this task using standard features in PROC LOGISTIC of SAS 9.2 for comparing ROC curves.

```
ODS GRAPHICS ON;
   PROC LOGISTIC DATA=panca PLOTS=ROC ROCOPTIONS(NODETAILS);
      MODEL d(EVENT='1') = y1 / NOFIT;
      ROC 'CA19-9' y1;
      ROC 'Chance';
      ROCCONTRAST REFERENCE('Chance') / ESTIMATE;
   run;
ODS GRAPHICS OFF;
```

The addition of ROCOPTIONS(NODETAILS) in the PROC LOGISTIC statement suppressed the printing of model fit statistics for the models specified in the ROC statements. As with the previous sample code, the MODEL statement is constructed using standard PROC LOGISTIC syntax. However, when the ROC statement is used, the actual model for each ROC curve to be compared is specified by the ROC statement. Therefore, the NOFIT option should be used to instruct SAS to ignore the model specified in the MODEL statement. The first ROC statement specifies the ROC curve labeled 'CA19-9' to model the biomarker variable y1. The second ROC statement is labeled 'Chance'

and no covariate is specified. The ROCCONTRAST statement provides details to SAS on comparing the ROC curves specified by the ROC statements. In the code presented above, the AUC for CA19-9 is to be compared to chance (AUC = 0.5). Finally, the ESTIMATE option instructs SAS to construct a table with details of each comparison. The following is the SAS output containing ROC-related statistics:

The LOGISTIC Procedure

ROC Association Statistics

	Mann-Whitney						
ROC	Area	Standard Error	95% Wald Confidence Limits		Somers' D (Gini)	Gamma	Tau-a
CA19-9	0.8614	0.0306	0.8015	0.9214	0.7229	0.7241	0.3362
Chance	0.5000	0	0.5000	0.5000	0	.	0

ROC Contrast Rows Estimation and Testing Results

Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
CA19-9 - Chance	0.3614	0.0306	0.3015	0.4214	139.6180	<.0001

The ROC Association Statistics table contains the AUC, standard error, and the 95% confidence interval for each biomarker specified in the ROC statement. In our example, the AUC of CA19-9 and chance are listed. The ROC Contrast Estimate table contains the comparison results for each ROC curve to the reference. In our example, the estimated AUC for CA19-9 is statistically greater than 0.5, providing evidence that the CA19-9 biomarker is useful for correctly classifying pancreatic cancer patients and patients without pancreatic cancer.

COMPARISON OF TWO AUC'S

Paired sample statistical techniques have been developed for the comparison of two biomarkers administered on the same sample population. The method exploits the mathematical equivalence of the AUC to the Mann-Whitney U-statistic (DeLong *et al.*, 1988). Under this framework, the ROCs of any two biomarkers can be compared by evaluating the difference of the AUCs which is asymptotically normal. The paired comparison is accomplished using similar code as presented above for the AUC comparison to chance. The following code compares the AUC of CA19-9 results to the AUC of biomarker CA125 results taken from the same sample population.

```
ODS GRAPHICS ON;
ODS SELECT ROCOVERLAY ROCASSOCIATION ROCCONTRASTESTIMATE;
   PROC LOGISTIC DATA=panca PLOTS=ROC;
      MODEL d(EVENT='1') = y1 y2 / NOFIT;
      ROC 'CA19-9' y1;
      ROC 'CA-125' y2;
      ROCCONTRAST REFERENCE('CA-125') / ESTIMATE;
   run;
ODS GRAPHICS OFF;
```

Since we are interested in only the ROC-related output from a long list of results and graphs, it is often worthwhile to specify the ODS tables and graphs to be displayed. The ODS SELECT statement accomplishes this by allowing us to specify the three output elements of interest: the plot of the two ROC curves (ROCOVERLAY), the estimated AUCs and their 95% confidence intervals (ROCASSOCIATION), and the test of the difference of the AUCs (ROCCONTRASTESTIMATE). In the code above, both variables, Y1 and Y2, representing test results for the two biomarkers are specified in the MODEL statement. As before, the NOFIT option is used since we are not interested in a logistic regression model fitted for both covariates. ROC statements are constructed for both biomarkers and the ROCCONTRAST statement specifies the CA-125 marker as the reference. Below is the resulting output.

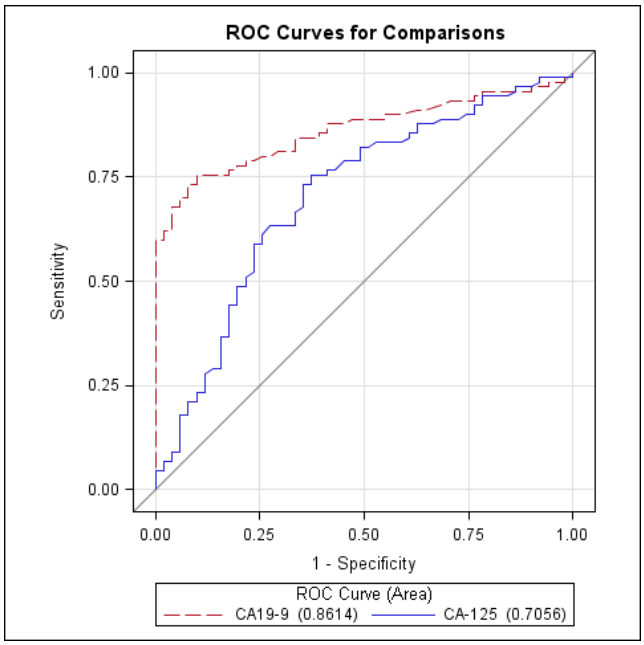

Figure 2. ROC curves for biomarker CA19-9 and CA-125.

The LOGISTIC Procedure

ROC Association Statistics

ROC	Area	Mann-Whitney Standard Error	95% Wald Confidence Limits		Somers' D (Gini)	Gamma	Tau-a
CA19-9	0.8614	0.0306	0.8015	0.9214	0.7229	0.7241	0.3362
CA-125	0.7056	0.0468	0.6138	0.7973	0.4111	0.4123	0.1912

ROC Contrast Rows Estimation and Testing Results

Contrast	Estimate	Standard Error	95% Wald Confidence Limits		Chi-Square	Pr > ChiSq
CA19-9 - CA-125	0.1559	0.0573	0.0436	0.2681	7.4096	0.0065

As displayed in Figure 2, CA19-9 appears to perform better than CA-125, particularly in the area of the curve representing high specificity. Overall, the AUC of 0.86 for CA19-9 is significantly greater than the AUC of 0.71 for CA-125 (p=0.0065).

SENSITIVITY AT FIXED SPECIFICITY

ESTIMATING SENSITIVITY AT FIXED SPECIFICITY

As previously mentioned, ROC analysis of a new continuous biomarker often involves determining the specificity at a fixed sensitivity or vice versa. This task involves: 1) estimating the threshold at the predefined specificity, and 2) estimating the sensitivity at the estimated threshold. The following notation represents this task (Pepe, 2003):

$$t \xrightarrow{n_{\bar{D}}} \text{estimated threshold} = \hat{S}_{\bar{D}}^{-1}(t) \xrightarrow{n_D} \text{estimated sensitivity} = \hat{S}n(t) = \hat{S}_D(\hat{S}_{\bar{D}}^{-1}(t))$$

At a fixed specificity of t, the estimated threshold, $\hat{S}_{\bar{D}}^{-1}(t)$, is the t^{th} quantile of the sample population with no disease. The estimated sensitivity is derived by calculating the survival function of the population with disease at the estimated threshold. Several methods have been proposed to estimate the variance of $\hat{S}n(t)$ that incorporates sampling variability for both steps. The following is a modified notation for the empirical estimation of the variance for large

samples, n_D and $n_{\overline{D}}$ (Pepe, 2003; Linnet, 1987):

$$\text{var}(\hat{S}n(t)) = \text{var}(\hat{S}_{\overline{D}}^{-1}(t)) + \text{var}(\hat{S}_D(\hat{S}_{\overline{D}}^{-1}(t))),$$

where

$$\text{var}(\hat{S}_{\overline{D}}^{-1}(t)) = \left(\frac{f_D(\hat{S}_{\overline{D}}^{-1}(t))}{f_{\overline{D}}(\hat{S}_{\overline{D}}^{-1}(t))} \right)^2 \frac{t(1-t)}{n_{\overline{D}}}$$

and

$$\text{var}(\hat{S}_D(\hat{S}_{\overline{D}}^{-1}(t))) = \frac{\hat{S}n(t)(1-\hat{S}n(t))}{n_D}.$$

The variance of the second component is simply the binomial variability of the estimated sensitivity at the fixed threshold, $\hat{S}_{\overline{D}}^{-1}(t)$. The variance of the estimated threshold involves the probability densities (f) of the biomarker at the threshold for the population with and without the disease. In fact, $\frac{f_D(\hat{S}_{\overline{D}}^{-1}(t))}{f_{\overline{D}}(\hat{S}_{\overline{D}}^{-1}(t))}$ is the slope of the ROC curve at the fixed specificity (Pepe, 2003). The estimation of the probability densities is often performed using kernel density estimates or normal distribution assumptions.

Appendix A contains SAS code for macro %CALCSN, developed for estimating the sensitivity and 95% confidence intervals at fixed specificity. Macro %CALCSN is called below using the following parameters for the pancreatic cancer data set: **panca**, name of the data set; **95**, fixed specificity; **d,** name of the variable for disease classification; **1**, value of **d** that denotes a patient as having the disease; **y1**, the variable name for the marker; **higher**, higher or lower marker result that corresponds to greater probability of disease.

```
%calcsn(panca,95,d,1,y1,higher);
```

The resulting output is:

95% CIs for Sensitivity at Specificity=95, marker=y1

Marker	Fixed Specificity	Estimated Threshold	Estimated Sensitivity	Variance estimation method	Standard Error	Lower Limit	Upper Limit
y1	0.95	59.2	0.67778	Naive exact binomial	0.04926	0.57100	0.77248
y1	0.95	59.2	0.67778	Linnet adjusted	0.11700	0.42398	0.85737

The empirical estimation described by Linnet resulted in wider confidence intervals compared to the naïve estimation. This is consistent with the notation above which shows that the appropriate method must account for both sources of variability. Alternative methods for deriving confidence intervals include a parametric approach (Obuchowski, 2004) and various forms of the bootstrap technique (Platt *et al*., 2000; Zhou and Qin, 2005).

COMPARISON OF SENSITIVITIES AT FIXED SPECIFICITY

The comparison of two dichotomous biomarkers applied to the same sample of patients would normally involve the use of the McNemer's test to account for correlation of the two markers. For comparison of sensitivities at a fixed specificity, the sampling variability related to the threshold estimation must again be taken into account. For this task, the difference in estimated sensitivities of the two markers is evaluated as below:

$$\hat{\Delta} = \hat{S}n_1(t) - \hat{S}n_2(t).$$

The distribution of $\hat{\Delta}$ is asymptotically normal with

$$\mathrm{var}(\hat{\Delta}) = \mathrm{var}(\hat{S}n_1(t)) + \mathrm{var}(\hat{S}n_2(t)) - 2\,\mathrm{cov}(\hat{S}n_{12}(t)).$$

The variance estimate for $\hat{\Delta}$ must account for the variance of the estimated sensitivity at specificity t for both markers and the correlation between the two markers as represented by the covariance term. Wieand *et al.* described a non-parametric estimation of the covariance term which involves determining density functions and joint distributions of the two markers (Wieand *et al.*, 1989). This method is computationally complex and sensitive to the choice of smoothing parameters. A more robust approach involves the bootstrap technique (Qin *et al.* 2006) to evaluate the following parameter:

$$\hat{\Delta}_{adj} = \hat{S}n_{adj1}(t) - \hat{S}n_{adj2}(t)$$

where the sensitivity calculations are adjusted as below (Agresti and Caffo, 2000)

$$\hat{S}n_{adj,i}(t) = \frac{\sum_{k=1}^{n_D} I_{[Y_{ik} \geq \hat{S}_D^{-1}(t)]} + z_{1-\alpha/2}^2/2}{n + z_{1-\alpha/2}^2}, i = 1,2.$$

The bootstrap variance estimator is:

$$\mathrm{var}^*(\hat{\Delta}_{adj}) = \mathrm{var}^*(\hat{S}n_{adj1}(t)) + \mathrm{var}^*(\hat{S}n_{adj2}(t)) - 2\,\mathrm{cov}^*(\hat{S}n_{adj12}(t)).$$

$\mathrm{var}^*(\hat{S}n_{adj}(t))$ for each marker is derived by determining the variance of the mean adjusted sensitivity estimated for B bootstrapped samples. The covariance term is estimated by:

$$2\,\mathrm{cov}^*(\hat{S}n_{adj12}(t)) = \frac{1}{B-1}\sum_{b=1}^{B}(\hat{S}n^*_{adj1b}(t) - \bar{S}n^*_{adj1}(t))(\hat{S}n^*_{adj2b}(t) - \bar{S}n^*_{adj2}(t)).$$

The estimated variance of $\hat{\Delta}_{adj}$ can then be used to derive confidence intervals as below:

$$(\bar{\Delta}^*_{adj} - z_{1-\alpha/2}\sqrt{\mathrm{var}^*(\hat{\Delta}_{adj})},\ \bar{\Delta}^*_{adj} - z_{1-\alpha/2}\sqrt{\mathrm{var}^*(\hat{\Delta}_{adj})})$$

%BOOTSN, a SAS macro for the paired comparison of sensitivities of two biomarkers at fixed specificity is included in Appendix B. In the example below, macro %BOOTSN is called with the following parameters: **panca**, name of the data set; **80**, fixed specificity; **200**, number of bootstrap samples, **d,** name of the variable for disease classification; **1**, value of **d** that denotes a patient as having the disease; **y2**, the variable name for the first marker; **higher**, higher or lower marker result that corresponds to greater probability of disease for the first marker, **y1**, the variable name for the second marker; **higher**, higher or lower marker result that corresponds to greater probability of disease for the second marker.

```
%bootsn(panca,80,200,d,1,y2,higher,y1,higher);
```

The following output is generated:

Bootstrap 95% confidence intervals for difference in sensitivity at specificity=80

B = 200 samples

Sn1, y2	Sn2, y1	Sn2 - Sn1	Lower limit	Upper limit	Bootstrap p-value
0.48889	0.77778	0.28889	0.022756	0.54618	0.016568

The difference in the sensitivity of the two biomarkers is 0.29, with a 95% confidence interval of 0.02 to 0.55. The null hypothesis that there is no difference in sensitivities is rejected at p=0.017. The McNemar's test of the two biomarkers at their respective thresholds for 80% specificity results in p=0.0002, a much stronger rejection of the null. As mentioned above, inference based on a simple McNemer's test is not valid because it does not account for the sampling variability for determining the threshold.

CONCLUSION

Recent developments in ROC methodology provide a wide range of statistical tools that can be applied to evaluating the diagnostic accuracy of biomarkers. However, there continues to be a widespread use of inappropriate statistical techniques. A common mistake is to perform simple binomial analysis on continuous biomarkers when a pre-defined threshold does not exist. This paper presents some of the more common statistical problems in the evaluation of biomarkers and proposes solutions. Readers are encouraged to review the growing literature on appropriate ROC methodology prior to analysis. In a welcome development, SAS 9.2 contains useful standard features for ROC analysis. However, most ROC methods will still require additional programming. Fortunately, SAS has many features that can be adapted to perform ROC analysis. This paper presents SAS macros using standard SAS procedures and data steps. The SAS code contained in the macros should be easily modifiable for additional analytical needs.

REFERENCES

Agresti, Alan and B. A. Caffo. 2000. "Simple and Effective Confidence Intervals for Proportions and Difference of Proportions Result from Adding Two Successes and Two Failures." *The American Statistician* 54:280-288.

DeLong, Elizabeth R., D. M. DeLong, and D. L. Clarke-Pearson. 1988. "Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." *Biometrics* 44:837-845.

Gönen, Mithat. 2006. *Analyzing Receiver Operating Characteristic Curves with SAS®.* Cary, NC: SAS Institute Inc.

Lijmer, Jeroen G. *et al*. 1999. "Empirical Evidence of Design-Related Bias in Studies of Diagnostic Tests". *Journal of the American Medical Association.* 282:1061-1066.

Linnet, Kristian. 1987. "Comparison of Quantitative Diagnostic Tests: Type I Error, Power, and Sample Size". *Statistics in Medicine.* 6:147-158.

Mandrekar, Jay N. and Sumithra J. Mandrekar. "Statistical Methods in Diagnostic Medicine using SAS® Software". *Proceedings of the Thirtieth Annual SAS® Users Group International Conference.* April 2005. <http://www2.sas.com/proceedings/sugi30/211-30.pdf> (Jan. 14, 2008).

Obuchowsky, Nancy A., M. L. Lieber, and F. H. Wians, Jr. 2004. "ROC Curves in *Clinical Chemistry*: Uses, Misuses and Possible Solutions". *Clinical Chemistry* 50:1118-1125.

Pepe, Margaret S. 2003 *The Statistical Evaluation of Medical Tests for Classification and Prediction.* New York, NY: Oxford University Press Inc.

Platt, Robert W., J. A. Hanley, and H. Yong. 2000. "Bootstrap Confidence Intervals for the Sensitivity of a Quantitiatve Diagnostic Test". *Statistics in Medicine.* 19:313-322.

Qin, Gengsheng, Y. S. Hsu, and X. H. Zhou. 2006. "New confidence intervals for the difference between two sensitivities at a fixed level of specificity." *Statistics in Medicine.* 25:3487-3502.

Wieand, H. Samuel *et al.* 1989. "A Family of Nonparametric Statistics for Comparing Diagnostic Markers with Paired

or Unpaired Data". *Biometrika* 76:585-592.

Zhou, Xiao-Hua and G. Qin. 2005. "Improved Confidence Intervals for the Sensitivity at a Fixed Level of Specificity of a Continuous-Scale Diagnostic Test." *Statistics in Medicine.* 24:467-477.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Sanghyuk Shin
Beckman Coulter, Inc.
10130 Sorrento Valley Road, Suite A
San Diego, CA 92121
Work Phone: (760) 438-6592
Fax: (858) 689-9288
E-mail: ssshin@beckman.com

Appendix A

```
*****************************************************************************
*  MACRO CALCSN – calculates sensitivity at fixed specificity;
*       uses method described by Linnet to incorporate sampling variability in estimating the
threshold;
*
*  Parameters:
*       dxdata - panca, name of the dataset;
*       sp - fixed specificity;
*       dis - name of the variable for disease classification;
*       pos - value of d that denotes a patient as having the disease (if DIS is a string
variable, POS must be specified in quotes);
*       marker - variable name for the marker result;
*       posdir - higher or lower marker result that corresponds to greater probability of
disease;

%macro calcsn (dxdata,sp,dis,pos,marker,posdir);

  %if "&posdir"="lower" %then
  %do;
    %let percut = %eval(100-&sp);
    %let format1=%nrstr(0 - %trim(&thresh) = "cancer");
        %let format2=%nrstr(>&thresh - high = "normal");
  %end;
  %else
  %do;
    %let percut = &sp;
    %let format1=%nrstr(0 - < &thresh = "normal");
        %let format2=%nrstr(&thresh - high = "cancer");
  %end;

        *Select threshold to give specificity=sp;
        proc univariate data=&dxdata;
          var &marker;
          output n=nd0 pctlpts=&percut pctlpre=p out=spthresh;
          where &dis ne &pos;
          title "Determine threshold at specificity=&sp, marker=&marker";
        run;

        data _null_;
          set spthresh;
       call symput("thresh",p&percut);
          call symput("nd0",nd0);
        run;

        proc format;
          value th&marker.fmt  %unquote(&format1)
                                                  %unquote(&format2);
        run;

    *Determine sensitivity and naive 95% CIs;
        proc freq data=&dxdata order=formatted;
          format &marker th&marker.fmt.;
          tables &marker / binomial;
          exact binomial;
          output out=outsp binomial;
          where &dis = &pos;
          title "Sensitivity at Specificity=&sp, marker &marker";
        run;

        *Dataset has 2 observations with naive and adjusted CIs;
        data sp&sp;
          format markernm $10. cimethod $25.;
          set outsp;
          markernm="&marker";
          thresh=%trim(&thresh);
          sp=&sp/100;
          cimethod='Naive exact binomial';
          if _bin_=1 then
          do; *corrects for 0% sensitivity, proc freq will erroneously output 100% sensitivity;
               sn=0;
               ll=1-xu_bin;
               ul=1-xl_bin;
```

```
         end;
         else
         do;
           sn=_bin_;
           ll=xl_bin;
           ul=xu_bin;
         end;
         se=e_bin;
         output;
         cimethod='Linnet adjusted';
         dd1=pdf('normal',probit(1-sn));    *density fxn;
      dd0=pdf('normal',probit(sp));
         slope = dd1/dd0;
         varsn=((sn*(1-sn))/n) + (slope**2) * ((sp*(1-sp))/&nd0);
         se=sqrt(varsn);
         logitll=log(sn/(1-sn))-1.96*sqrt(varsn)/(sn*(1-sn)); *logit transformation allows for
asymetric CIs;
         ll=exp(logitll)/(1+exp(logitll));
         logitul=log(sn/(1-sn))+1.96*sqrt(varsn)/(sn*(1-sn));
      ul=exp(logitul)/(1+exp(logitul));
         output;
         label sn='Estimated Sensitivity'
                     sp='Fixed Specificity'
                     se= 'Standand Error'
                     ll= 'Lower Limit'
                     ul= 'Upper Limit'
                     cimethod='Variance estimation method'
                     markernm='Marker'
                     thresh='Estimated Threshold';
         keep markernm thresh sn sp se ll ul cimethod;
       run;

       proc print data=sp&sp l noobs;
         var markernm sp thresh sn cimethod se ll ul;
         title "95% CIs for Sensitivity at Specificity=&sp, marker=&marker";
       run;

%mend calcsn;

%calcsn(panca,95,d,1,y2,higher);
%calcsn(panca,95,d,1,y1,higher);
```

Appendix B

```
*******************************************************************************
*  MACRO %BOOTSN - boostrap comparison of sensitivities at fixed specificity;
*       uses method proposed by Zhou et al. to incorporate sampling variability in estimating the
difference in sensitivities;
*       requires the MACRO CALCSN to be defined prior to call;
*       could take a few minutes to run depending on the number of samples to be generated;
*
*  Parameters:
*       dxdata - panca, name of the data set;
*       sp - fixed specificity;
*       bootrep - number of bootstrap samples to be generated;
*       dis - name of the variable for disease classification;
*       pos - value of d that denotes a patient as having the disease (if DIS is a string
variable, POS must be specified in quotes);
*       marker1 - variable name for the result for continuous first continous marker;
*       posdir1 - higher or lower marker result that corresponds to greater probability of
disease for the first continous marker;
*       marker2 - variable name for the result for continuous second continous marker;
*       posdir2 - higher or lower marker result that corresponds to greater probability of
disease for the second continous marker;


%macro bootsn(dxdata,sp,bootrep,dis,pos,marker1,posdir1,marker2,posdir2);

  %calcsn(&dxdata,&sp,&dis,&pos,&marker1,&posdir1);

  proc datasets library=work;
    delete sp1&sp;
    change sp&sp=sp1&sp;
  run;

  %calcsn(&dxdata,&sp,&dis,&pos,&marker2,&posdir2);

  data delta;
    merge sp1&sp (keep=sp sn rename=(sn=sn1))
                 sp&sp (keep=sp sn rename=(sn=sn2));
        by sp;
        delta=sn2-sn1;
  run;

  *Perform McNemers test (not printed in the final output, but contained in the final data set
for comparison purposes);
  proc freq data=&dxdata noprint;
        format &marker1 th&marker1.fmt. &marker2 th&marker2.fmt. ;
        tables &marker1 * &marker2;
        exact mcnem;
        output mcnem out=mcnemout;
        where &dis=&pos;
  run;

  proc sql noprint;
    select count(*) into :Ndis
      from work.&dxdata
          where &dis=&pos;

    select count(*) into :Nnorm
      from work.&dxdata
          where &dis ne &pos;
  quit;

  proc surveyselect data=&dxdata (where=(&dis=&pos)) method=urs out=bootdis outhits rep=&bootrep
n=&Ndis noprint;
  run;

  proc surveyselect data=&dxdata (where=(&dis ne &pos)) method=urs out=bootnorm outhits
rep=&bootrep n=&Nnorm noprint;
  run;

  proc printto log=mylog print=myprint;
  run;

  proc datasets library=work;
```

```
  delete dxboot;
run;

%do i=1 %to &bootrep;

   data datarep;
        set bootdis (where=(replicate=&i))
                bootnorm (where=(replicate=&i));
      run;

      %calcsn(datarep,&sp,&dis,&pos,&marker1,&posdir1);

      proc datasets library=work;
     delete sp1&sp;
     change sp&sp=sp1&sp;
   run;

   %calcsn(datarep,&sp,&dis,&pos,&marker2,&posdir2);

      *calculate adjusted sensitivites;
      data snrepi;
         merge sp1&sp (keep=sp sn rename=(sn=sn1))
                      sp&sp (keep=sp sn rename=(sn=sn2));
            by sp;
            i1=sn1*&Ndis;
         btsn1adj=(i1+2)/(&Ndis+4);
            i2=sn2*&Ndis;
         btsn2adj=(i2+2)/(&Ndis+4);
            bootdadj=btsn2adj-btsn1adj;
            if _n_ = 1;
            drop i1 i2 sn1 sn2;
      run;

      proc append base=dxboot data=snrepi;
      run;

%end;


proc printto;
run;

proc univariate data=dxboot noprint;
  var btsn1adj btsn2adj bootdadj;
     output mean=meanbtsn1 meanbtsn2 meanbtd
               var=varbtsn1 varbtsn2 varbtd
        out=snboot;
run;

data bootci;
  retain cumv12 0 rvarbtsn1 rvarbtsn2 rmeanbtsn1 rmeanbtsn2 rmeanbtd dsn sn1 sn2 pmcnem;
     merge snboot dxboot delta (rename=(sn1=tempsn1 sn2=tempsn2)) mcnemout;
     if _n_=1 then
     do;     *Variables to be retained for final calculations at the end of the data step;
       dsn=delta;
       sn1=tempsn1;
       sn2=tempsn2;
       rvarbtsn1=varbtsn1;
       rvarbtsn2=varbtsn2;
       rmeanbtsn1=meanbtsn1;
       rmeanbtsn2=meanbtsn2;
       rmeanbtd=meanbtd;
       pmcnem=xp_mcnem;
     end;
     v12b=(btsn1adj-rmeanbtsn1)*(btsn2adj-rmeanbtsn2);
     cumv12=cumv12+v12b;
     if _n_=&bootrep then
     do;
       v12=cumv12/(&bootrep-1);
       v=rvarbtsn1+rvarbtsn2-(2*v12);
       dbootll=rmeanbtd-(1.96*sqrt(v));
       dbootul=rmeanbtd+(1.96*sqrt(v));
       pvalue=1-probnorm(abs(rmeanbtd)/sqrt(v));
       output;
```

```
        end;
        label sn1="Sn1, &marker1"
                  sn2="Sn2, &marker2"
                  dsn='Sn2 - Sn1'
                  dbootll='Lower limit'
                  dbootul='Upper limit'
                  pvalue='Bootstrap p-value'
                  pmcnem='McNemar p-value';
        drop meanbtsn1 meanbtsn2 meanbtd varbtsn1 varbtsn2 varbtd btsn1adj btsn2adj bootdadj v12b
tempsn1 tempsn2 delta _MCNEM_ P_MCNEM XP_MCNEM;
  run;

  proc print data=bootci l noobs;
    var sn1 sn2 dsn dbootll dbootul pvalue;
    title "Bootstrap 95% confidence intervals for difference in sensitivity at specificity=&sp";
    title3 "B = &bootrep samples";
  run;

%mend bootsn;
```