

Statistical Methods in Diagnostic Medicine using SAS[®] Software

Jay N. Mandrekar, Ph.D., Sumithra J. Mandrekar, Ph.D.

Division of Biostatistics, Mayo Clinic, Rochester, MN

ABSTRACT

An important goal in diagnostic medicine research is to estimate and compare the accuracies of diagnostic tests, which serve two purposes 1) providing reliable information about a patient's condition and 2) influencing patient care. In developing screening tools, researchers often evaluate the discriminating power of the screening test by concentrating on the sensitivity and specificity of the test and the area under the ROC curve. We propose to give a gentle introduction to the statistical methods commonly used in diagnostic medicine covering some broad issues and scenarios. In particular, power calculations, estimation of the accuracy of a diagnostic test, comparison of accuracies of competing diagnostic tests, and regression analysis of diagnostic accuracy data will be discussed. Some existing SAS[®] procedures and SAS[®] macros for analyzing the data from diagnostic studies will be summarized. These concepts will be illustrated using datasets from clinical disciplines like radiology, neurology and infectious diseases.

INTRODUCTION

The purpose of a diagnostic test is to classify or predict the presence or absence of a condition or a disease. The clinical performance of a diagnostic test is based on its ability to correctly classify subjects into relevant subgroups. Essentially, these tests help answer a simple question: if a person tests positive, what is the probability that the person really has the disease / condition, and if a person tests negative, what is the probability that the person is really disease / condition free? As new diagnostic tests are introduced, it is important to evaluate the quality of the classification obtained from this new test in comparison to existing tests or the Gold Standard.

In this review paper, we discuss the different methods used to quantify the diagnostic ability of a test (sensitivity, specificity, the likelihood ratio (LR), area under the receiver operating curve (ROC)), the probability that a test will give the correct diagnosis (positive predictive value and negative predictive value), and regression methods to analyze diagnostic accuracy data. We will also discuss comparisons of areas under two or more correlated ROC curves and provide examples of power calculations for designing diagnostic studies. These concepts will be illustrated using SAS[®] macros and procedures.

SIMPLE MEASURES OF DIAGNOSTIC ACCURACY

The accuracy of any test is measured by comparing the results from a diagnostic test (positive or negative) to the true disease or condition (presence or absence) of the patient (Table 1).

Table 1: Cross Classification of Test Results by Diagnosis

Test Results	Disease / Condition	
	Present	Absent
Positive	True Positive (TP)	False Positive (FP)
Negative	False Negative (FN)	True Negative (TN)

The two basic measures of quantifying the diagnostic accuracy of a test are the sensitivity (SENS) and specificity (SPES) (Zhou et al., 2002). *Sensitivity* is defined as the ability of a test to detect the disease status or condition when it is truly present, i.e., it is the probability of a positive test result given that the patient has the disease or condition of interest. *Specificity* is the ability of a test to exclude the condition or disease in patients who do not have the condition or the disease i.e., it is the probability of a negative test result given that the patient does not have the disease or condition of interest. In describing a diagnostic test, both SENS and SPES are reported as they are inherently linked in that as the value of one increases, the value of the other decreases. SENS and SPES are also dependent on the patient characteristics and the disease spectrum. For example, advanced tumors are easier to detect than small benign lesions and detection of fetal maturity may be influenced by the gestational age of the patient (Hunink et al., 1990).

In clinical practice, it is also important to know how good the test is at predicting the true positives, i.e., the probability that the test will give the correct diagnosis. This is captured by the predictive values. The *positive predictive value* (PPV) is the probability that a patient has the disease or condition given that the test results are positive, and the *negative predictive value* (NPV) is the probability that a patient does not have the disease or condition given that the test results are indeed negative.

To illustrate these concepts, consider an example where results from a diagnostic test like x-ray or computer tomographic (CT) scan and the true disease or condition of the patient is known (Altman and Bland, 1994a). The different measures discussed above along with the 95% exact binomial confidence intervals for each estimate can be calculated (see Table 2A).

Table 2A: Example: Test Results by Diagnosis (Prevalence = 75%)

Test Results	Disease Status / Condition		Total
	Present	Absent	
Positive	231 (TP)	32 (FP)	263
Negative	27 (FN)	54 (TN)	81
Total	258	86	344

$$\text{SENS} = 231/258 = 0.90 \text{ (95\% CI: } 0.85 - 0.93\text{)}$$

$$\text{SPES} = 54/86 = 0.63 \text{ (95\% CI: } 0.52 - 0.73\text{)}$$

$$\text{PPV} = 231/263 = 0.88 \text{ (95\% CI: } 0.83 - 0.92\text{)}$$

$$\text{NPV} = 54/81 = 0.67 \text{ (95\% CI: } 0.55 - 0.77\text{)}$$

The 95% exact binomial confidence intervals can be calculated using the `%bnmci` macro from the Mayo Clinic (see reference 1 under SAS[®] macros resource) using the following call statement:

```
%bnmci(width=, x=, n=);
```

where `x` = observed number of successes in `n` trials and `width` = width of the CI (default is 95). For example,

```
%bnmci(x=231, n=258); RUN;
```

would give the 95% CI for the SENS estimate in the above example.

Prevalence is defined as the prior probability of the disease before the test is carried out. For example, the estimate of the prevalence of the disease considered in Table 2A is 75% (258/344). The PPV and the NPV are dependent on the prevalence of the disease in the patient population being studied (Altman and Bland, 1994b). To put this in perspective, suppose that the prevalence of the disease considered in Table 2A is actually 25% (Table 2B). The PPV and the NPV are $77/173 = 0.45$ and $162/171 = 0.95$, but the SENS and the SPES remain unaltered.

Table 2B: Example: Test Results by Diagnosis (Prevalence = 25%)

Test Results	Disease Status / Condition		Total
	Present	Absent	
Positive	77 (TP)	96 (FP)	173
Negative	9 (FN)	162 (TN)	171
Total	86	258	344

Both SENS and SPES can be applied to other populations that have different prevalence rates, unlike the predictive values, which are dependent on the prevalence of the disease or condition being tested. It is therefore not appropriate to apply universally the PPV and the NPV obtained from one study without information on prevalence. For instance, the rarer the prevalence of the disease, the more sure one can be that a negative test result indeed means that there is no disease, and less sure that a positive test result indicates the presence of a disease. Also, the lower the prevalence, greater is the number of people who will be diagnosed as FP, even if the SENS and the SPES are high, as seen in example given by Table 2B.

The *Likelihood Ratio* (LR) is yet another simple measure of diagnostic accuracy, given by the ratio of the probability of the test result among patients who truly had the disease / condition to the probability of the same test among patients who do not have the disease/condition. In other words, the LR is really the ratio of SENS / (1-SPES). The LR for the example considered above is 2.4. Clearly it is also a measure that is independent of prevalence of the disease / condition. The magnitude of the LR informs about the certainty of a positive diagnosis. As a general guideline, a value of LR=1 indicates that the test result is equally likely in patients with and without the disease/condition, values of LR > 1 indicate that the test result is more likely in patients with the disease / condition and values of LR < 1 indicate that the test result is more likely in patients without the disease / condition (Zhou et al., 2002).

The LR can also be defined in terms of the pre-test and post-test probabilities of the disease / condition. In the example given by Table 2A, the pre-test probability of the disease / condition (or the pre-test odds of disease) = $0.75 / (1-0.75) = 3.0$ (since the prevalence of the disease is 0.75). The post-test probability of the disease / condition is given by $(231/263) / (1-(231/263)) = 0.878 / (1-0.878) = 7.2 = 3.0 \times 2.4 = \text{pre-test odds of disease} \times \text{LR}$. Thus, the LR can also be interpreted as the ratio of the post-test probability of disease/condition to the pre-test probability of the disease/condition.

AREA UNDER THE ROC CURVE

Both SENS and SPES require a cutpoint in order to classify the test results as positive or negative. The SENS and SPES for a diagnostic test are therefore tied to the diagnostic threshold or cutpoint selected for the test. Many times the results from a

diagnostic test may be on an ordinal or numerical scale rather than just a binary outcome of positive or negative. In such situations, the SENS and SPES are based on just one cutpoint when in reality multiple cutpoints or thresholds are possible. An ROC curve overcomes this limitation by including all the decision thresholds possible for the results from a diagnostic test.

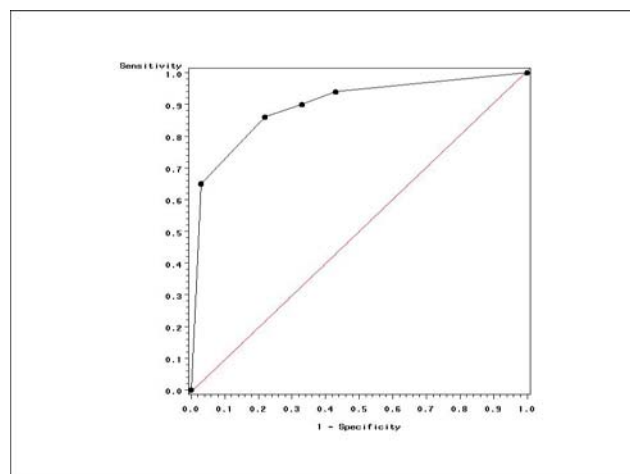
An ROC curve is a plot of the SENS versus (1-SPES) of a diagnostic test, where the different points on the curve correspond to different cutpoints used to determine if the test results are positive. As an illustration, consider ratings of CT images from 109 subjects by a radiologist, given by Table 3 (Hanley and McNeil, 1982). Clearly, in this example, multiple cutpoints are possible for classifying a patient as normal or abnormal based on the CT scan. The designation of a cutpoint to classify the test results as positive or negative is relatively arbitrary. Suppose that the ratings of 4 or above indicate, for instance, that the test is positive, then the SENS and SPES would be 0.86 and 0.78. In contrast, if the ratings of 3 or above are considered as positive, then the SENS and SPES are 0.90 and 0.67 respectively. This illustrates that both SENS and SPES are specific to the selected decision threshold.

Table 3: True Disease Status by CT Ratings

True Disease Status	CT Ratings					Total
	1: Definitely Normal	2: Probably Normal	3: Unsure	4: Probably Abnormal	5: Definitely Normal	
Normal	33	6	6	11	2	58
Abnormal	3	2	2	11	33	51
Total	36	8	8	22	35	109

A better way to represent this data is an ROC curve (Figure 1), which does not require the selection of a particular cutpoint. Figure 1 essentially has 2 components, the empirical ROC curve that is obtained by joining the points represented by the (SENS, 1-SPES) for the different cutpoints, and the chance diagonal, which is the 45 degree line drawn through the coordinates (0,0) and (1,1). If the test results diagnosed patients as positive or negative for the disease / condition by pure chance, then the ROC curve will fall on the diagonal line. An ROC curve can be considered as the average value of the sensitivity for a test over all possible values of specificity or vice versa. A more general interpretation is that given the test results, the probability that for a randomly selected pair of patients with and without the disease/condition, the patient with the disease/condition has a result indicating greater suspicion (Hanley and McNeil, 1982). Sometimes a fitted (smooth) ROC curve based on a statistical model can also be plotted in addition to the empirical ROC curve. A binormal distribution is commonly used to fit ROC curves, and comprises of one normal distribution to describe the test results of patients with the disease and another normal distribution to describe the test results of patients without the disease (for more details, see Zhou et al., 2002).

Figure 1: The ROC curve for the CT data



Following are the SAS® commands for generating Figure 1.

```
GOPTIONS reset=all cback=white border;
DATA new;
INPUT _SENSIT_ SPEC;
_1MSPEC_ =1-SPEC;
CARDS;
1 0
.94 .57
```

```

.9 .67
.86 .78
.65 .97
0 1
;
DATA anno;
FUNCTION='move'; xsys='1'; ysys='1'; x=0; y=0; OUTPUT;
FUNCTION='draw'; xsys='1'; ysys='1'; color='red'; x=100; y=100; OUTPUT;
RUN;
PROC GPLOT data=new;
PLOT _SENSIT*_lMSPEC_ / anno=anno HAXIS=axis1 VAXIS=axis2;
SYMBOL1 i=j v=dot c=black;
AXIS1 length=5 in;
AXIS2 length=5 in;
LABEL _SENSIT_='Sensitivity';
LABEL _lMSPEC_='1 - Specificity';
RUN;

```

The area under the ROC curve is an effective way to summarize the overall diagnostic accuracy of the test. It takes values from 0 to 1, where a value of 0 indicates a perfectly inaccurate test and a value of 1 reflects a perfectly accurate test. If the area under the ROC curve is 1, then it consists of two line segments joining the coordinates (0,0), (0,1) and (1,1). Clearly, this represents an ideal situation where both the SENS and the SPES of the test is 1. Likewise, if the area under the ROC curve is 0, then it represents a scenario where the test incorrectly classifies all patients with the disease/condition as negative and all patients without the disease/condition as positive. If the test results are reversed, then the perfectly inaccurate test can be transformed into a perfectly accurate test. The closer the ROC curve of a diagnostic test is to the (0, 1) coordinate, the better is the test. In general, a value of 0.5 for area under the ROC curve, i.e., the area below the 45 degree line, is considered as the lower bound. ROC curves above this diagonal line are considered to have reasonable discriminating ability to diagnose patients with and without the disease/condition. It is therefore natural to do a hypothesis test to evaluate if the area under the empirical ROC curve differs significantly from 0.5. Specifically, the null and alternate hypotheses are defined as H0: area under

ROC curve = 0.5, vs. H1: area under the ROC curve \neq 0.5. This test statistic given by $\left(\frac{\hat{AUC} - 0.5}{\text{Std. error}(\hat{AUC})} \right)$ is approximately normally distributed and has favorable statistical properties (Zhou et al., 2002).

For the example considered in this section (Table 3), the area under the ROC curve is computed to be 0.89 using the trapezoidal rule (Rosner, 2000). This means that the radiologist reading the CT scan has an 89% chance of correctly distinguishing a normal from an abnormal patient based on the ordering of the CT ratings. In the event of a tied rating, the assumption is that the radiologist will randomly assign one patient as normal and the other as abnormal. A formal hypothesis test of H0: area under ROC curve = 0.5, vs. H1: area under the ROC curve \neq 0.5 for this example yields a test statistic of 12.2, with a p-value <0.001, indicating that this test has excellent discriminating ability based on the guidelines specified by Hosmer and Lemeshow (2000).

An alternate SAS[®] code to generate Figure 1, as well as the area under the ROC curve is to use the PROC LOGISTIC procedure, where the c-statistic is the nonparametric estimate of the area under the ROC curve (Hosmer and Lemeshow, 2000). Assuming that the dataset “neuro” contains the information of the true disease status (“disease”) and the CT ratings (“ctrating”), following are the SAS[®] commands to obtain Figure 1 and the area under that ROC curve:

```

PROC LOGISTIC DESCENDING data=neuro;
MODEL disease = ctrating / outroc=roc1;
RUN;
SYMBOL1 i=join v=none c=black;
PROC GPLOT data=roc1;
PLOT _sensit*_lmspec_=1 / VAXIS=0 to 1 by .1 CFRAME=wh;
RUN;

```

An overall ROC curve is most useful in the early stages of evaluating a new diagnostic test. Once the diagnostic ability of a test is established, only a portion of the ROC curve may be of interest, for example, only regions with high SPES, and not the average SPES over all SENS values. Like the SENS and SPES, ROC curves are invariant to the prevalence of a disease, but dependent on the patient characteristics and the disease spectrum. An ROC curve does not depend on the scale of the test results, and can be used to provide a visual comparison of two or more test results on a common scale. The latter is not possible with SENS and SPES measures because a change in the cutpoint to classify the test results as positive or negative could affect the two tests differently (Turner, 1978).

ROC curves are useful for comparing the diagnostic ability of two or more screening tests for the same disease. While comparing two or more ROC curves based on tests performed on the same set of individuals (observational units), it is also important to account for the correlated nature of the data. A nonparametric approach to analyze such correlated data has been proposed based on generalized U-Statistics (DeLong et al., 1988). The test with the higher area under the ROC curve may be considered better, unless some specific value of SENS and SPES is clinically important for the comparison. In such instances, it is important to compare not the total area under the ROC curves, but area under the clinically important regions of the curve (partial area under the ROC curve). A comparison of areas under the 2 ROC curves can be done using the `%roc` SAS[®] macro available on the SAS[®] website (see reference 2 under SAS[®] macros resource). The macro call statement is as follows:

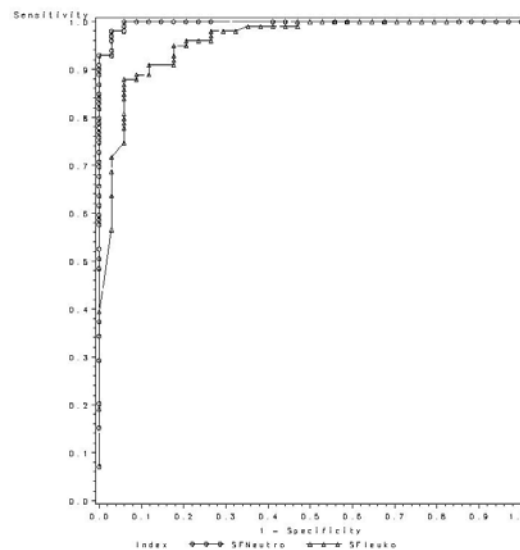
```
%roc(data=, var=, response=, contrast=%str(hypothesis matrix), details= );
```

where, `data` = name(s) of the dataset(s) to be analyzed separated by spaces, `var` = names of the predictors (diagnostic measures) to be compared (required), `response` = response variable with values of 0 or 1 only (required), `contrast=%str(hypothesis matrix)` specifies the desired hypothesis matrix, `L`, for testing the null hypothesis $L \cdot R = 0$, where `R` is the vector of ROC curve areas, and `details` = any additional output needed. Details on the definitions of each parameter used in the above macro call function can be found in the macro documentation online (see reference 2 under SAS[®] macros resource).

As an illustration of the comparison of areas under 2 correlated ROC curves, consider a study that was conducted at the Mayo Clinic between January 1998 and December 2003 on 133 patients with prosthetic joint infection without underlying inflammatory disease (Trampuz et al., 2004). The primary aim was to evaluate the accuracy of the synovial fluid leukocyte count and neutrophil percentage as diagnostic markers in these patients. Following are the SAS[®] commands to plot the ROC curves as well as to compute and compare the areas under the 2 ROC curves using the method of Hanley and McNeil (1983) and DeLong et al. (1988).

```
%roc(data=joint, details=yes, var=SFleuko SFNeutro, response=pji,contrast=%str(1 -1));
SYMBOL1 i=join v=circle c=black line=24;
SYMBOL2 i=join v=triangle c=black line=24;
PROC GPLOT data=joint;
LABEL index="Index";
PLOT _sensit_ * _lmspec_ = Index / VAXIS=0 to 1 by .1 HAXIS=0 to 1 by .1 CFRAME=wh;
RUN;
```

Figure 2: ROC curves for the synovial fluid leukocyte count and neutrophil percentage



The area under the ROC curve for leukocyte count is 0.958 and for neutrophil percentage is 0.997, suggesting that both methods have a good discriminating ability. In addition, a formal comparison of the areas under the 2 ROC curves gives a chi-square statistic of 5.13 with 1 degree of freedom with a p-value of 0.02. This suggests that there is a significant difference in the discriminating ability of the two diagnostic markers. Also, Figure 2 shows that an incremental increase in SENS is associated with a relatively smaller increase in 1-SPES for neutrophil percentage. This in addition to the higher area under the ROC curve for the neutrophil percentage suggests that it has a better diagnostic accuracy (Trampuz et al., 2004).

REGRESSION ANALYSIS

This is applicable in situations where we are interested in adjusting for multiple covariates when comparing areas under the correlated ROC curves. To illustrate this, we will consider data from a prostate cancer study (Hosmer and Lemeshow, 2000). The primary goal was to determine if the variables measured at baseline can be used to predict whether the tumor has penetrated the prostatic capsule. Among the 380 patients included in this modified version of the dataset, 153 patients had a cancer that penetrated the prostatic capsule. The response variable (capsule) is tumor penetration of prostatic capsule (Yes, No) and some of the predictor variables considered include results of the digital rectal exam (leftlobe, rightlobe, bilobe), prostatic specific antigen value (mg/ml) (PSA), and total Gleason score (gleason). For illustration purposes, we will compare the areas under the curve from three specific models containing; 1) Gleason score, PSA, and results of the digital rectal exam: Model 1, 2) Gleason score and PSA: Model 2, 3) Gleason score: Model 3. The areas under these correlated ROC curves can be computed and compared using %roc macro (see reference 1 under SAS[®] macros resource) that uses a nonparametric approach which is closely related to the jackknife technique (DeLong et al., 1988).

As a first step, Models 1 through 3 as mentioned above are fit using the PROC LOGISTIC procedure, and the xbetas i.e., the logits ($\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$, where k = number of predictor variables in the model) are output to the datasets p1, p2, p3 respectively. The areas under the ROC curves given by c-statistics for these three models are respectively 0.82, 0.79 and 0.77. The appropriate SAS[®] commands for these are given below.

```
PROC LOGISTIC DESCENDING data=prostate;
MODEL capsule = leftlobe rightlobe bilobe psa gleason;
OUTPUT out=p1 xbeta=lpg;
RUN;
```

```
PROC LOGISTIC DESCENDING data=prostate;
MODEL capsule = psa gleason;
OUTPUT out=p2 xbeta=pg;
RUN;
```

```
PROC LOGISTIC DESCENDING data=prostate;
MODEL capsule = gleason;
OUTPUT out=p3 xbeta=g;
RUN;
```

As the second step, the %roc macro is used to compute the Mann-Whitney statistics and component-based estimates of their variance-covariance matrix. The test of equality of three ROC curve areas is done using a 2 degree of freedom test for the 2 contrasts.

```
%roc(data=p1 p2 p3, response=capsule, var=lpg pg g); RUN;
```

A chi-square statistic of 15.48 with 2 degrees of freedom suggests that at least 2 models differ significantly ($p=0.0004$). The %roc macro can now be used to do the 3 pair-wise comparisons using appropriate contrasts to assess which of the models differ, at an alpha level of 0.017, adjusting for multiple comparisons.

```
ods select l(persist) ctest(persist);
%roc(data=p1 p2 p3, response=capsule, var=lpg pg g, contrast=%str(1 -1 0));
%roc(data=p1 p2 p3, response=capsule, var=lpg pg g, contrast=%str(1 0 -1));
%roc(data=p1 p2 p3, response=capsule, var=lpg pg g, contrast=%str(0 1 -1));
ods select all;
RUN;
```

Significant p-values for the 3 pair-wise comparisons (0.0103, <.0001, 0.0183) indicate that the 3 models are different from each other. Based on these results and the area under the ROC curve values, Model 1 with Gleason score, PSA, and digital rectal exam results is considered to have the best ability to discriminate between the subjects (Hosmer and Lemeshow, 2000). However, the final decision should also be based on the clinical meaningfulness of such differences identified by statistical analysis. For example, if clinically 0.82, 0.79, and 0.77 are not different and if the Gleason score is both easier to measure and cost effective relative to other two factors, then one can go with a parsimonious model of using just the Gleason score as a predictor.

DESIGNING DIAGNOSTIC STUDIES

Up until this point, we have looked at ways to analyze data from diagnostic studies. Various methods to calculate the sample size and power for a diagnostic study are discussed in the literature. Chapter 6 from Zhou et al (2002) gives a comprehensive review of the different sample size calculation methods with examples. In this section, we discuss briefly the SAS[®] macros

available for power calculations.

The %ROCPOWER macro estimates the power of statistical tests involved in computing the area under ROC curves for various scenarios (Zepp, 1995). Specifically, this macro computes the power for comparing a single area to a value under null hypothesis as well as for comparing the areas under paired (correlated) and unpaired (uncorrelated) ROC curves (see reference 3 under SAS[®] macros resource). Asymptotic z-tests are used for the comparison of ROC areas. Standard errors are computed using the Hanley and McNeil method (1983) when the response is continuous and the Obuchowski method (1994) when the response is ordinal. The macro allows for both 1-tailed and 2-tailed tests. The complete macro call is given below:

```
%ROCPOWER(T1=, T2=, T0=, NA=, NN=, N=, PERCENT=, R=, ALPHA=, TAILS=, ORDINAL=, I=, J=);
```

T1, T2 and T0 are the hypothesized areas under the ROC curve, NA is number of abnormal patients to be tested, NN is number of normal patients, PERCENT is the percentage of abnormal patients in the total sample size, R is the correlation between T1 and T2 when the same patients are examined by both modalities, ALPHA is the type I error, TAILS is either 1 or 2 to specify one or two-tailed test, and ORDINAL is 0 for continuous and 1 for ordinal data. Further details on the macro parameters can be found in Zepp (1995). Below are three specific examples that illustrate the use of this macro.

Example 1: A study to assess ability of ultrasound to distinguish between benign and malignant breast tumors using biopsy findings as a gold standard (single modality)

```
%ROCPOWER(T1=.95, T2= 0, T0=.80, NN=50, NA=40, ALPHA=.01, TAILS=1,ORDINAL=0);
%ROCPOWER(T1=.95, T2= 0, T0=.80, NN=50, NA=40, ALPHA=.01, TAILS=1,ORDINAL=1);
```

The estimated power of the test is 0.937 and 0.844 based on whether the response is continuous or ordinal.

Example 2: Randomized comparison of magnetic resonance imaging and CT for detection of cerebral aneurysms (unpaired, where patients are randomized to receive only one modality)

```
%ROCPOWER(T1=.90, T2= .75, T0=.75, NN=50, NA=50, ORDINAL=0);
%ROCPOWER(T1=.90, T2= .75, T0=.75, NN=50, NA=50, ORDINAL=1);
```

The estimated power of the test is 0.605 and 0.538 based on whether the response is continuous or ordinal.

Example 3: Comparison of the CT colonography and colonoscopy for the detection of polyps and cancers of the colon (paired, where all patients receive both modalities)

```
%ROCPOWER(T1=.90, T2=.70, T0=.70, r=.5, N=150, PERCENT=.50, ORDINAL=0);
%ROCPOWER(T1=.90, T2=.70, T0=.70, r=.5, N=150, PERCENT=.50, ORDINAL=1);
```

The estimated power of the test is 0.999 and 0.998 based on whether the response is continuous or ordinal.

From the above examples, we can see that for the same sample size and hypothesized values, the power of the test is higher for a continuous response.

SUMMARY

Studies designed to measure the performance of diagnostic tests are important for patient care and health care costs. The literature in this area is growing and we have summarized a few of the scenarios and SAS[®] macros with intuitive examples. Attention must be paid to include proper representation of patients with the disease or condition of interest along with healthy participants to make the study results generalizable to the population of interest. We refer readers to the textbook by Zhou et al. (2002) for material on advanced topics such as adjustment for verification bias, imperfect gold standard and meta-analysis.

REFERENCES

1. Altman DG, Bland JM (1994a). Statistics Notes: Diagnostic tests 1: sensitivity and specificity. *British Medical Journal*, 308, 1552.
2. Altman DG, Bland JM (1994b). Statistics Notes: Diagnostic tests 2: predictive values. *British Medical Journal*, 309, 102.
3. DeLong ER, DeLong DM, Clarke-Pearson DL (1988). Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, 44, 837-845.

4. Hanley JA, McNeil BJ (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29-36.
5. Hanley JA, McNeil BJ (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148, 839-843.
6. Hosmer DW, Lemeshow S (2000). *Applied Logistic Regression*, 2nd Edition. New York: John Wiley and Sons.
7. Hunink MG, Richarson DK, Doubilet PM, Begg CB (1990). Testing for fetal pulmonary maturity: ROC analysis involving covariate, verification bias, and combination testing. *Medical Decision Making*, 10, 201-211.
8. Obuchowski NA (1994). Computing Sample Size for Receiver Operating Characteristic Studies. *Investigative Radiology*, 29, 238-243.
9. Rosner B (2000). *Fundamentals of Biostatistics*, 5th Edition. California: Pacific Grove.
10. Trampuz A, Hanssen AD, Osmon DR, Mandrekar J, Steckelberg JM, Patel R (2004). Synovial fluid leukocyte count and differential for the diagnosis of prosthetic knee infection. *The American Journal of Medicine*, 117, 556-562.
11. Turner DA (1978). An intuitive approach to receiver operating characteristic curve analysis. *Journal of Nuclear Medicine*, 19, 213-220.
12. Zepp RC (1995). A SAS[®] Macro for Estimating Power for ROC Curves One-Sample and Two-Sample Cases. *Proceedings of the 20th SAS Users Group International Conference (SUGI)*, Paper 223.
13. Zhou XH, Obuchowski NA, Obuchowski DM (2002). *Statistical Methods in Diagnostic Medicine*. New York: John Wiley and Sons.

SAS[®] MACROS RESOURCES

1. Mayo Clinic, Division of Biostatistics <http://www.mayo.edu/hsr/sasmac.html>
2. SAS Institute <http://ftp.sas.com/techsup/download/stat/roc.html>
3. Cleveland Clinic Foundation <http://www.bio.ri.ccf.org/Research/ROC/>

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Address all correspondences to:

Jay Mandrekar, Ph.D.

Mayo Clinic, Division of Biostatistics

200 First Street SW

Harwick 7

Rochester MN 55905

Phone: (507) 266 0573

Fax: (507) 284 9542

Email: mandrekar.jay@mayo.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.