Sensitivity, Specificity, Accuracy, Associated Confidence Interval and ROC Analysis with Practical SAS® Implementations

Wen Zhu[1], Nancy Zeng[2], Ning Wang[2]
[1]K&L consulting services, Inc, Fort Washington, PA
[2]Octagon Research Solutions, Wayne, PA

1. INTRUDUCTION

Diagnosis tests include different kinds of information, such as medical tests (e.g. blood tests, X-rays, MRA), medical signs (clubbing of the fingers, a sign of lung disease), or symptoms (e.g. pain in a particular pattern). Doctor's decisions of medical treatment rely on diagnosis tests, which makes the accuracy of a diagnosis is essential in medical care. Fortunately, the attributes of the diagnosis tests can be measured. For a given disease condition, the best possible test can be chosen based on these attributes. Sensitivity, specificity and accuracy are widely used statistics to describe a diagnostic test. In particular, they are used to quantify how good and reliable a test is. Sensitivity evaluates how good the test is at detecting a positive disease. Specificity estimates how likely patients without disease can be correctly ruled out. ROC curve is a graphic presentation of the relationship between both sensitivity and specificity and it helps to decide the optimal model through determining the best threshold for the diagnostic test. Accuracy measures how correct a diagnostic test identifies and excludes a given condition. Accuracy of a diagnostic test can be determined from sensitivity and specificity with the presence of prevalence. Given the importance of these statistics in disease diagnosis and the terms are easily confused, it is important to get familiar with how they work, it helps us better understand when to use, how to implement them, and how to interpret the results. The importance and popularity of these statistics urges for a thorough review along with practical SAS examples.

This paper will focus on the concepts of sensitivity, specificity and accuracy in the context of disease diagnosis: starting with a review of the definitions, how to calculate sensitivity, specificity and accuracy, associated 95% confidence interval and ROC analysis; followed by a practical example of disease diagnosis and related SAS macro code; then moving on to the common issues on interpreting the results of sensitivity, specificity and accuracy; ended by a final remark of the entire paper.

2. SENSITIVITY, SPECIFICITY AND ACCURACY, 95% CINFIDENCE INTERVAL AND ROC CURVE

2.1 SENSITIVITY, SPECIFICITY AND ACCURACY

Outcome of the diagnostic test	**Condition (e.g. Disease) As determined by the Standard of Truth**		
	Positive	**Negative**	**Row Total**
Positive	TP	FP	TP+FP (Total number of subjects with positive test)
Negative	FN	TN	FN + TN (Total number of subjects with negative test)
Column total	TP+FN (Total number of subjects with given condition)	FP+TN (Total number of subjects without given condition)	N = TP+TN+FP+FN (Total number of subjects in study)

Table 1. Terms used to define sensitivity, specificity and accuracy

The ideas described below are summarized in table 1.

There are several terms that are commonly used along with the description of sensitivity, specificity and accuracy. They are true positive (TP), true negative (TN), false negative (FN), and false positive (FP). If a disease is proven present in a patient, the given diagnostic test also indicates the presence of disease, the result of the diagnostic test is considered true positive. Similarly, if a disease is proven absent in a patient, the diagnostic test suggests the disease is absent as well, the test result is true negative (TN). Both true positive and true negative suggest a consistent result between the diagnostic test and the proven condition (also called standard of truth). However, no medical test is perfect. If the diagnostic test indicates the presence of disease in a patient who actually has no such disease, the test result is false positive (FP). Similarly, if the result of the diagnosis test suggests that the disease is absent for a patient with disease for sure, the test result is false negative (FN). Both false positive and false negative indicate that the test results are opposite to the actual condition.

Sensitivity, specificity and accuracy are described in terms of TP, TN, FN and FP.

Sensitivity = TP/(TP + FN) = (Number of true positive assessment)/(Number of all positive assessment)

Specificity = TN/(TN + FP) = (Number of true negative assessment)/(Number of all negative assessment)

Accuracy = (TN + TP)/(TN+TP+FN+FP) = (Number of correct assessments)/Number of all assessments)

As suggested by above equations, sensitivity is the proportion of true positives that are correctly identified by a diagnostic test. It shows how good the test is at detecting a disease. Specificity is the proportion of the true negatives correctly identified by a diagnostic test. It suggests how good the test is at identifying normal (negative) condition. Accuracy is the proportion of true results, either true positive or true negative, in a population. It measures the degree of veracity of a diagnostic test on a condition.

The numerical values of sensitivity represents the probability of a diagnostic test identifies patients who do in fact have the disease. The higher the numerical value of sensitivity, the less likely diagnostic test returns false-positive results. For example, if sensitivity = 99%, it means: when we conduct a diagnostic test on a patient with certain disease, there is 99% of chance, this patient will be identified as positive. A test with high sensitivity tents to capture all possible positive conditions without missing anyone. Thus a test with high sensitivity is often used to screen for disease.

The numerical value of specificity represents the probability of a test diagnoses a particular disease without giving false-positive results. For example, if the specificity of a test is 99%. It means: when we conduct a diagnostic test on a patient without certain disease, there is 99% chance; this patient will be identified as negative.

A test can be very specific without being sensitive, or it can be very sensitive without being specific. Both factors are equally important. A good test is a one has both high sensitivity and specificity. A good example of a test with high sensitive and specificity is pregnancy test. A positive result of pregnancy test almost for sure suggests the subject who took the test is pregnant. A negative result almost certainly rules out the possibility of being pregnant.

In addition to the equation show above, accuracy can be determined from sensitivity and specificity, where prevalence is known. Prevalence is the probability of disease in the population at a given time:

Accuracy = (sensitivity) (prevalence) + (specificity) (1 - prevalence).

The numerical value of accuracy represents the proportion of true positive results (both true positive and true negative) in the selected population. An accuracy of 99% of times the test result is accurate, regardless positive or negative. This stays correct for most of the cases. However, it worth mentioning, the equation of accuracy implies that even if both sensitivity and specificity are high, say 99%, it does not suggest that the accuracy of the test is equally high as well. In addition to sensitivity and specificity, the accuracy is also determined by how common the disease in the selected population. A diagnosis for rare conditions in the population of interest may result in high sensitivity and specificity, but low accuracy. Accuracy needs to be interpreted cautiously.

2.2 ASYMPTOTIC AND EXACT 95% CONFIDENCE INTERVAL

The sensitivity, specificity and accuracy are proportions, thus the according confidence intervals can be calculated by using standard methods for proportions[1]. Two types of 95% confidence intervals are generally constructed around proportions: asymptotic and exact 95% confidence interval. The exact confidence interval is constructed by using binomial distribution to reach an exact estimate. Asymptotic confidence interval is calculated by assuming a normal approximation of the sampling distribution. The choice of these two types of confidence interval depends on whether the sample proportion is a good approximation of normal distribution. If the number of event is very small or if the sample size is very small, the normal assumption cannot be met. Thus, exact confident interval is desired. SAS example for both types of 95% confidence intervals will be provided in section 3.

2.3 RECEIVER OPERATING CHARACTERISTICS (ROC) ANALYSIS

For a given diagnostic test, the true positive rate (TPR) against false positive rate (FPR) can be measured, where

TPR= TP/(TP+FN)

And

FPR = FP/(FP+TN)

As we can see from the above equations, TPR is equivalent to sensitivity and FPR is equivalent to (1 – specificity). All possible combinations of TPR and FPR compose a ROC space. One TPR and one FPR together determine a single point in the ROC space, and the position of a point in the ROC space shows the tradeoff between sensitivity and specificity, i.e. the increase in sensitivity is accompanied by a decrease in specificity. Thus the location of the point in the ROC space depicts whether the diagnostic classification is good or not. In an ideal situation, a point determined by both TPR and FPF yields a coordinates (0, 1), or we can say that this point falls on the upper left corner of the ROC space. This idea point indicates the diagnostic test has a sensitivity of 100% and specificity of 100%. It is also called perfect classification. Diagnostic test with 50% sensitivity and 50% specificity can be visualized on the diagonal determined by coordinate (0, 0) and coordinates (1, 0). Theoretically, a random guess would give a point along this diagonal. A point predicted by a diagnostic test fall into the area above the diagonal represents a good diagnostic classification, otherwise a bad prediction. A graphic presentation of what described above is shown in figure 1.

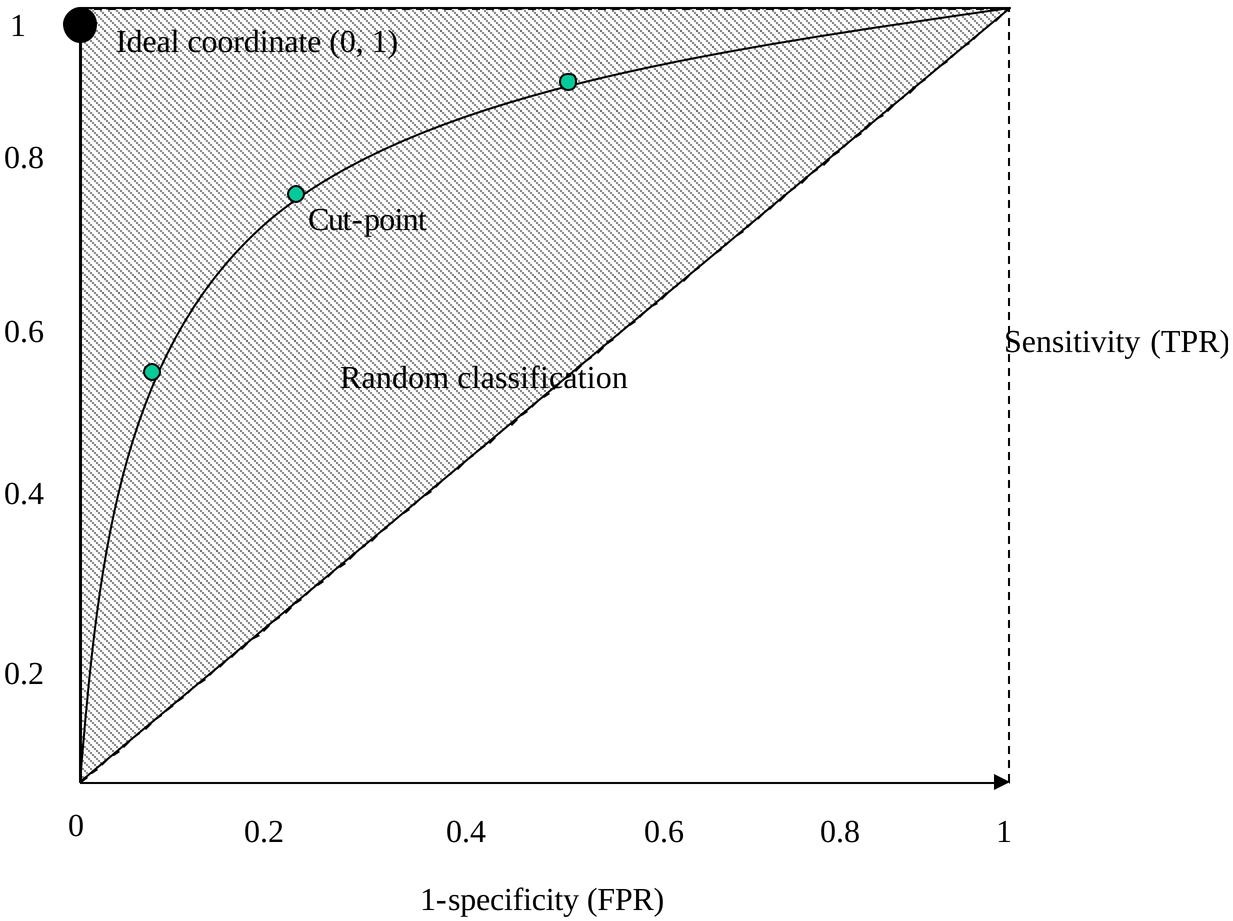

Figure 1: ROC Space: shadow area represents better diagnostic classification

A single cut-point of a diagnostic test defines one single point in the ROC space; however, different possible cut-points of a diagnostic test determine a curve in ROC space, which is also called ROC curve. Like a single point in the

ROC space, ROC curve is often plotted by using true positive rate (TPR) against false positive rate (FPR) for different cut-points of a diagnostic test, starting from coordinate (0, 0) and ending at coordinate (1, 1). FPR (1 – specificity) is represented by x-axis and TPR (sensitivity) is represented by y-axis. Thus, ROC curve is a plot of a test's sensitivity vs. (1-specificity) as well. The interpretation of ROC curve is similar to a single point in the ROC space, the closer the point on the ROC curve to the ideal coordinate, the more accurate the test is. The closer the points on the ROC curve to the diagonal, the less accurate the test is. In addition, (1) the faster the curve approach the ideal point, the more useful the test results are; (2) the slope of the tangent line to a cut-point tells us the ratio of the probability of identifying true positive over true negative, i.e. likelihood ratio (LR) for the test value: LR = sensitivity/(1-specificity), if the ratio is equal to 1, the selected cut-point doesn't add additional information to identify true positive result. If the ratio is greater than 1, the selected cut-point help identify true positive result. If the ratio is less than 1, it

decreases disease likelihood (3) the area under ROC curve (AUC) provides a way to measure the accuracy of a diagnostic test. The larger area, the more accurate the diagnostic test is. AUC of ROC curve can be measured by the following equation, Where t = (1 – specificity) and ROC (t) is sensitivity.

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

Commonly used classification using AUC for a diagnostic test is summarized in table 2:

Table 2: accuracy classification by AUC for a diagnostic test

AUC Range	Classification
0.9 < AUC < 1.0	Excellent
0.8 < AUC < 0.9	Good
0.7 < AUC < 0.8	Worthless
0.6 < AUC < 0.7	Not good

In short, ROC curve is a good tool to select possible optimal cut-point for a given diagnostic test.

3. AN PRACTICAL EXAMPLE WITH SAS CODE

A good example explains better. Suppose, if there is an existing test that can always identify the true positive and true negative in determining the presence or absence of haemodynamically relevant stenosis of the renal arteries. However it is very time-consuming and expensive. A more efficient and affordable test is discovered, the preliminary results shows high sensitivity and specificity. A trial is then carried out to confirm the efficacy of this diagnostic test. Assuming the existing test has 100% accuracy, it will be used as the standard of truth, in other words, we believe that the test results always reflect the true stenosis of the renal arteries. A population of patients is enrolled, and both existing test and trial test are performed on each patient in order to obtain comparable results. And the test results are record in the following dataset data1. And the SAS examples used to calculate the sensitivity, specificity, accuracy and associated asymptotic and exact confidence interval are provided below:

A dummy dataset is created below: data1, in which test1 is standard of truth, test2 is for trial test result. value 0 for test1 and test2 means disease absent, while 1 means disease present.

```
data data1;
       input id $ test1 test2;
       datalines;
sub01 1 0
sub02 1 0
sub03 0 0
sub04 0 1
sub05 1 1
sub06 1 1
sub07 1 1
sub08 0 0
sub09 1 0
sub10 1 1
sub11 1 1
sub12 1 1
sub13 0 0
sub14 0 0
```

```
sub15 1 1
sub16 1 1
sub17 1 1
sub18 1 0
sub19 0 0
sub20 0 1
sub21 0 1
sub22 1 1
sub23 0 0
sub24 0 0
sub25 1 1
sub26 0 0
sub27 0 0
sub28 0 0
sub29 0 0
sub30 1 1
sub31 1 0
sub32 0 0
sub33 0 0
sub34 0 0
sub35 1 1
sub36 0 0
sub37 1 1
sub38 1 0
sub39 0 0
sub40 0 0
;
run;
```

Compare the test result with the standard of truth and assigne TP, TN, FP, FN to each test result following the guidance of table 1.

```
data data2;
    set data1;
        if test1=1 then
            do;
                if test2=1 then result_c12="TP";
                else if test2=0 then result_c12="FN";
                    end;
        else if test1=0 then
            do;
                if test2=1 then result_c12="FP";
                else if test2=0 then result_c12="TN";
            end;
run;

proc sort data=data2;
       by test1 test2 ;
run;
```

Once the results are assigned to different categories, the sensitivity, specificity and accuracy can be easily calculated by using the formula provided in previous section of this table. The following SAS code can be used for the calculations.

This data step generates count for sensitivity and specificity.

```
data main1 (drop=id result_c12);
    set data2;
    by test1;
    retain tp tn fp fn;
    if (first.test1) then do;
        tp=0; tn=0; fp=0; fn=0;
    end;
    if (result_c12 in ("TP")) then tp=tp+1;
    if (result_c12 in ("TN")) then tn=tn+1;
    if (result_c12 in ("FN")) then fn=fn+1;
    if (result_c12 in ("FP")) then fp=fp+1;
    else ;
    if (last.test1) then output;
run;
```

This data step generates accuracy count.

```
data main2;
   set main1;
   tntp=tn+tp;
   fnfp=fn+fp;
run;
```

Total count of sensitivity, specificity and accuracy could be summed up by proc SQL.

```
proc sql;
   create table main3 as
   select sum(tp) as tp, sum(tn) as tn, sum(fp)as fp, sum(fn) as fn, sum(tntp) as
tntp, sum(fnfp) as fnfp
   from main2
       ;
quit;
```

3.1 SENSITIVITY, SPECIFICITY AND ACCURACY

```
proc sql;
       create table main4 as
       select tp/(tp+fn) as sensitivity, tn/(tn+fp) as specificity,
(tn+tp)/(tn+tp+fn+fp) as accuracy
       from main3
       ;
quit;
```

This MAIN4 dataset shows the sensitivity, specificity and accuracy of the diagnosis method.

Sensitivity	Specificity	Accuracy
0.7	0.85	0.775

3.2 ASYMPTOTIC AND EXACT 95% CONFIDENCE INTERVAL

Create a dataset for calculating 95% confidence interval; the dataset follows the format showing in table 3.2.

```
proc transpose data=main3 out=t_main;
       var tp tn fn fp tntp fnfp;
run;

data table32 (drop=_name_ col1);
       length group $20;
       set t_main;
       count=col1;
       if _name_="tp" then do;
              group="Sensitivity";
              response=0;
              output;
       end;
       else if _name_="fn" then do;
              group="Sensitivity";
              response=1;
              output;
       end;
       else if _name_="tn" then do;
              group="Specificity";
              response=0;
              output;
       end;
       else if _name_="fp" then do;
              group="Specificity";
              response=1;
              output;
       end;
       else if _name_="tntp" then do;
              group="Accuracy";
              response=0;
              output;
       end;
       else if _name_="fnfp" then do;
              group="Accuracy";
              response=1;
              output;
       end;
run;
```

Table 3.2 format of the dataset for calculating 95% confidence interval

Group	Response	Count
Sensitivity	0	TP
Sensitivity	1	FN
Specificity	0	TN
Specificity	1	FP
Accuracy	0	TN + TP
Accuracy	1	FN+FP

```
      proc freq data= table32;
                weight count;
                by group;
                tables response/alpha=0.05 binomial(p=0.5);
                exact binomial;
             run;
```

The 95% confidence interval outputted from the SAS procedure is listed below:

	Rate	Asymptotic 95% CI	Exact 95% CI
Sensitivity	0.7	(0.4992, 0.9008)	(0.4572, 0.8811)
Specificity	0.85	(0.6935, 1.0000)	(0.6211, 0.9679)
Accuracy	0.775	(0.6456, 0.9044)	(0.6155, 0.8916)

3.3 RECEIVER OPERATING CHARACTERISTICS (ROC) ANALYSIS

ROC analysis also allows analyzing sensitivity and specificity simultaneously at different cut-points; this approach better estimates the accuracy of a given trial test by using multiple pairs of sensitivity and specificity. In the previous example, a trial test is to measure the degree of occlusion (stenosis vs. non-stenosis). With ROC analysis, the trial test evaluates a continuous degree of stenosis, also called the cut-off points, these points are chosen from 20% upwards to 80% percent stenosis. Trial test at each cut-off point returns one sensitivity and one specificity. The ROC analysis plots all sensitivity vs. (1-specificity) at selected cut-offs points by placing each pair of sensitivity and (1-specificity) in ROC space. The area under the curve (AUC) is a parameter indicating the intrinsic accuracy of the diagnostic test in determining the haemodynamically relevant stenosis of renal arteries. The sensitivity and specificity for each cut-point is calculated in the same way described in previous section and recorded in a dataset named DATA_ROC (variable SENS is for sensitivity and spec1 is for 1-Specificity). The SAS example used to generate the ROC curve and AUC are provided below.

The dummy data that the example based on is as follow:

```
data data_roc;
   input order cut sens spec spec1;
    datalines;
1 0.80 0.11 0.95 0.05
2 0.70 0.30 0.90 0.10
```

```
3 0.50 0.60 0.88 0.12
4 0.30 0.76 0.80 0.20
5 0.25 0.85 0.66 0.34
6 0.20 0.97 0.15 0.85
       ;
run;
```

Cut-points	Sensitivity	(1-Specificity)
20%	0.93	0.85
25%	0.73	0.34
30%	0.67	0.20
50%	0.60	0.12
70%	0.30	0.10
80%	0.11	0.05

```
proc gplot data=data_roc;
       symbol1 v=square i=j;
       plot sens*spec1/ vaxis=0 to 1 by 0.1 haxis=0 to 1 by 0.1;
       label sens="Sensitivity" spec1="1-Specificy";
run;
```

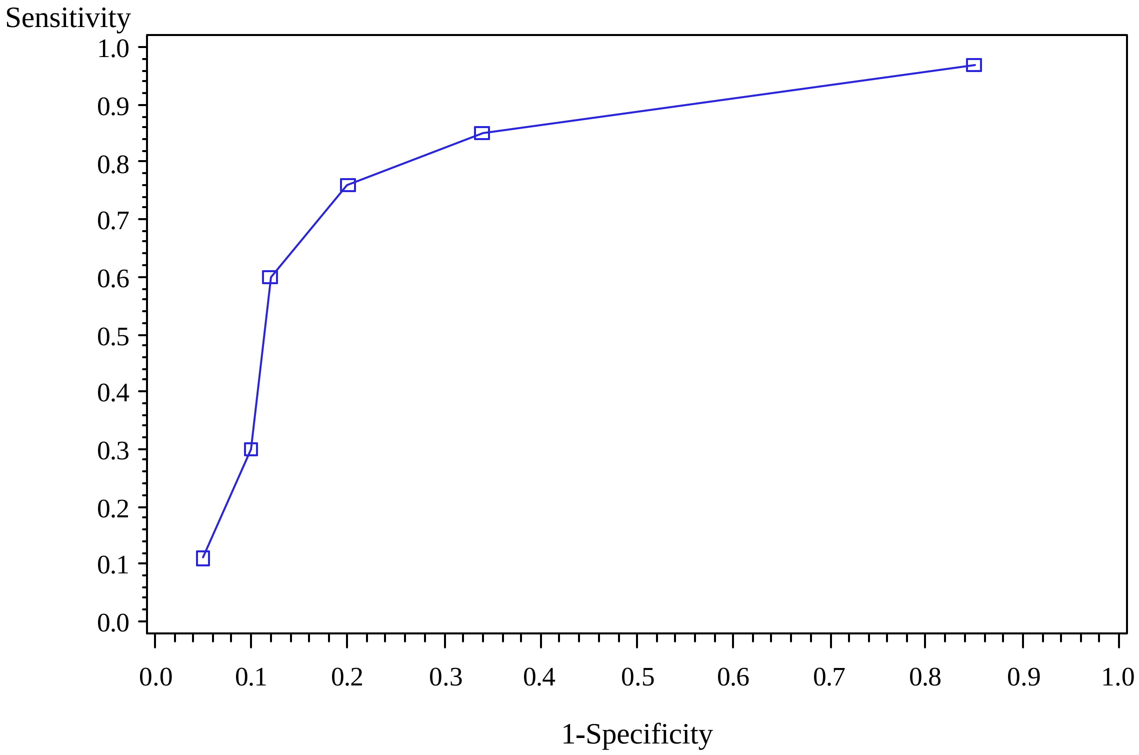

Figure 2. Sample ROC plot: x-axis = (1-specificity), y-axis = sensitivity. The area under the ROC curve represents accuracy of a trial test. ROC curve AUC is determined by multiple cut-points of the trial test, it gives better estimate of accuracy.

ROC curve AUC can be calculated by the following data step:

```
data auc;
       set data_roc end=eof;
       drop lagx lagy;
       lagx=lag(spec1);
       lagy=lag(sens);
       if order=1 then do;
              lagx=0;
              lagy=0;
       end;
       tpzd=(spec1-lagx)*(sens+lagy)/2;
       sumtpz+tpzd;
```

```
        if eof then do;
                roc_auc=sumtpz+(1-spec1)*(sens+1)/2;;
                output;
        end;
run;
```

The AUC obtained from the above code and data is 0.8010. According to table 2, the trial test has fairly good accuracy.

REFERENCE

1. Gardner MJ, Altman DG. Calculating confidence intervals for proportions and their differences. In: Gardner MJ, Altman DG, eds. Statistics with confidence. London: BMJ Publishing Group, 1989:28-33

ACKNOWLEDGEMENT AND COPYRIGHT INFORMATION

CONTACT INFORMATION

Questions and Feedbacks are welcome, please send them to:

Wen Zhu
K&L Consulting Services, Inc.
1300 Virginia Drive, Suite 103,
Fort Washington, PA 19034
201-949-4317
wen.zhu@klconsultingservices.com

Nancy Zeng
Octagon Research Solution, Inc.
585 East Swedeford Rd. Suite 200
Wayne, PA 19087
610-535-6500 ext 5803
nzeng@octagonresearch.com

Ning Wang
Octagon Research Solution, Inc.
585 East Swedeford Rd. Suite 200
Wayne, PA 19087
610-535-6500 ext 5633
nwang@octagonresearch.com