The receiver operating characteristic (ROC) curve

Shengping Yang PhD, Gilbert Berdine MD

Results from routine blood tests can be used potentially as biomarkers for identifying disease. An example would be using the hemoglobin concentration to identify patients with iron deficiency anemia. We want a binary (Yes/No) answer, but the values of these predictive tests are continuous; I am wondering how to use them to facilitate a diagnosis.

Originally developed for detecting enemy airplanes and warships during the World War II, the receiver operating characteristic (ROC) has been widely used in the biomedical field since the 1970s in, for example, patient risk group classification, outcome prediction and disease diagnosis. Today, it has become the gold standard for evaluating/comparing the performance of a classifier(s).

A ROC curve is a two-dimensional plot that illustrates how well a classifier system works as the discrimination cut-off value is changed over the range of the predictor variable. The x axis or independent variable is the false positive rate for the predictive test. The y axis or dependent variable is the true positive rate for the predictive test. Each point in ROC space is a true positive/false positive data pair for a discrimination cut-off value of the predictive test. If the probability distributions for the true positive and false positive are both known, a ROC curve can be plotted from the cumulative distribution function. In most real applications, a data sample will yield a single point in the ROC space for each choice of discrimination cut-off. A perfect result would be the point (0, 1) indicating 0% false positives and 100% true positives. The generation of the true positive and false positive rates requires that we have a gold standard method for identifying true positive and true negative cases. To better understand a ROC curve, we will need to review the contingency table or confusion matrix.

Corresponding author: **Shengping Yang**
Contact Information: Shengping.yang@ttuhsc.edu
DOI: 10.12746/swrccc.v5i19.391

THE CONFUSION MATRIX

A confusion matrix (also known as an error matrix) is a contingency table that is used for describing the performance of a classifier/classification system, when the truth is known.

In a confusion matrix, each column (or row) reports the numbers in a predicted class, e.g., the number of predicted disease or predicted normal, while each row (or column) reports the numbers in a true class, e.g., the number of true disease or true normal. In a typical 2×2 contingency table, four numbers are reported: 1) true positive (TP; also called sensitivity; a measurement of the proportion of positives, that are correctly predicted given it is truly positive), 2) false negative (FN; a measurement of the proportion of predicted negatives, given it is truly positive), 3) false positive (FP; a measure of the proportion of predicted positives, given it is truly negative), and 4) true negative (TN; also called specificity; a measure of the proportion of predicted negative, given it is truly negative). It is quite obvious that a better classifier is

Table 1. A confusion matrix

		Predicted condition	
		Disease	**Normal**
True condition	Disease	True positive (TP) (sensitivity)	False negative (FN)
	Normal	False positive (FP)	True negative (TN) (specificity)

expected to have both higher sensitivity and specificity. *Note that specificity is 1 – FP.*

The ROC curve and the area under curve (AUC)

If we choose a discriminating cut-off value for the predictive variable to be less than the lowest value observed, we generate the (0, 0) point in the ROC space. As we increase the discriminating cut-off value to include more and more data points, we generate a series of points within the ROC space that can be connected by a curve. A discriminating cut-off value greater than the highest value observed generates the (1, 1) point. The diagonal line connecting the (0, 0) point and the (1, 1) point indicates test predictions no better than random guesses. The further a point in the ROC space is above the diagonal line, the better the predictive value of the test.

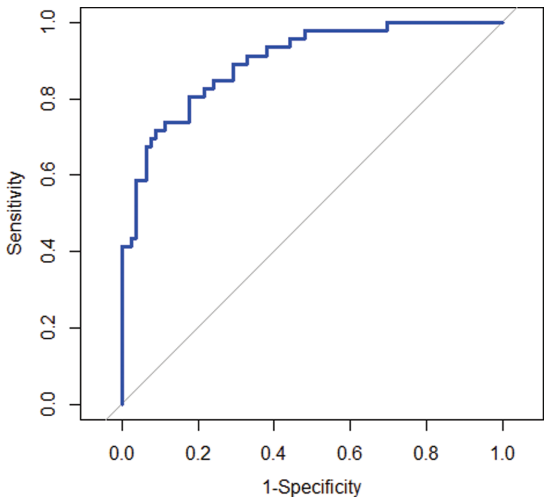

Figure 1 is a hypothetical ROC curve demonstrating the tradeoff between sensitivity and specificity. Particularly, sensitivity and specificity are inversely related, i.e., as the sensitivity increases, the specificity decreases, and vice versa. For example, if we use a lower hemoglobin cut-off value, more non-anemic patients will be considered as normal, and thus the true negative rate is higher (i.e., higher specificity); meanwhile, fewer anemia patients will be considered as having the disease, and thus the proportion of true positive is lower (i.e., lower sensitivity). Similarly, if we use a higher cut-off value, then we will have lower specificity and higher sensitivity.

The AUC (also known as the c-statistic) can be used to evaluate the diagnostic ability of a test to discriminate the true disease status of a patient. In general, the rule of thumb for interpreting AUC value is:

AUC=0.5	No discrimination, e.g., randomly flip a coin
0.6≥AUC>0.5	Poor discrimination
0.7≥AUC>0.6	Acceptable discrimination
0.8≥AUC>0.7	Excellent discrimination
AUC>0.9	Outstanding discrimination

In addition, AUC can also be used to find the optimal cut-off value for a specific test, as well as compare the performance between two or more alternative tests.

Determining the optimal cut-off value for a test

Since a ROC curve presents sensitivity and specificity calculated with varying cut-off values, it is critical to determine the optimal cut-off value, so that the classifier has the best performance. Some cut-off value selection methods give equal weight to sensitivity and specificity in the calculation, thus they are easy to understand and simple to implement. However, most of time, they are built upon unrealistic assumptions, i.e., they do not take into account of the difference in disease prevalence or the ethical and financial costs associated with misclassification. To address this issue, methods incorporating costs for correct and false diagnosis have been developed to adjust for such differences. In general, if a disease has high prevalence and the associated costs for false positive are low, then a low cut-off value can be used; otherwise, a high cut-off value can be used. Unfortunately, determining the costs associated with ethnical and/or financial considerations is a complex problem and is beyond the scope of this article.

Comparing two diagnostic tests

Very often, more than one test can be used for diagnosing a certain disease; it is thus reasonable to compare these tests to see which one outperforms the others. In general, the bigger the AUC is, the better the test as a classifier. However, it can be shown that two tests with the same AUC value can have very different performance. For example, one test might have better performance in the high sensitivity range, and another test in the low sensitivity range, and therefore, it would be meaningful to choose the preferred test based on the sensitivity and specificity preference pertinent to specific diagnostic situations.

Note that the costs associated with different tests might be another factor to be considered in determining the preferred test, and the discussion on this is beyond the scope of this article.

Generating a ROC curve

Many statistical software packages can be used for generating ROC curves.

i. In SAS:

By adding the *plot=roc* option in the *proc logistic* statement, the ROC curve can be automatically generated as part of the procedure, and the AUC will be estimated and included in the ROC curve plot.

```
proc logistic descending plot=roc;
model Y = predictor <other covariates>;
run;
```

ii. In R:

The *roc* function in the R package *pROC* can be used. A ROC curve will be generated with the `plot=TRUE` option.

```
roc(outcome ~ predictor, data=data,
plot=TRUE)
```

A formal comparison of two ROC curves (two predictors) is also straightforward. By first calculating the ROC curves for each predictor, a comparison can be made using the `roc.test` function.

```
roc1=roc(outcome ~ predictor1,
data=data, plot=TRUE)
roc2=roc(outcome ~ predictor2,
data=data, plot=TRUE)
roc.test(roc1, roc2)
```

Pitfalls and issues associated with ROC curves

ROC graph is an ideal platform for visualizing and evaluating classifiers. However, there are some limitations and pitfalls we might want to be aware of:

1. To calculate AUC, sensitivity and specificity values are summarized over all possible cut-off values, and this can be misleading because only one cut-off value is used in making predictions.
2. Different study populations might have different patient characteristics; a ROC model developed using data generated from one population might not be directly transferred to another population. A training and a validation set approach can be used to evaluate the performance of a classifier.
3. Depending on disease prevalence and costs associated with misclassification, the optimal classifier might vary from one situation to another.
4. ROC curves are most useful when the predictors are continuous.

Author affiliation: Department of Pathology at Texas Tech University Health Sciences Center in Lubbock, TX (SY) and Department of Internal Medicine (GB).
Submitted: 4/11/2017
Conflicts of interest: none

References

1. Berrar D, Flach P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). Brief Bioinform 2012; 13: 83–97.
2. Hosmer DW, Lemeshow S. Applied Logistic Regression. 2nd ed. John Wiley & Sons, Inc. 2000. Pp. 156-164.
3. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8:283-298.