Special Proceedings of 20th Annual Conference of SSCA
held at Pondicherry University, Puducherry during January 29-31, 2018; pp 45-49

Some Aspects of Surrogate Biomarkers using ROC Curve analysis

K.V.S. Sarma
Sri Venkateswara Institute of Medical Sciences, Tirupati

Final Version Received on September 11, 2018

Abstract

Identification of biomarkers is an interesting topic in medical research. Confirmation of a disease is often made by the result of a lab test or some procedure. Some such tests are known as gold standards like histopathology to detect a malignant tumor. In practice, the researcher looks for alternative surrogate markers that can help early diagnosis at lower cost. The ability to distinguish between true and false positives is measured in terms of the sensitivity and specificity at cutoff value on the marker. The Receiver Operating Curve (ROC) is used to find the optimal cutoff. In this paper we revisit some interesting aspects of single and multiple biomarkers used in clinical studies and the way they are to be proposed. We also appraise the scientific potential of this area and the software support.

Kew words: Biomarkers, ROC Curve, Composite markers

1. Biomarkers

In medical diagnosis, a biomarker (or biological marker) is a broad subcategory of medical signs that can be measured accurately and reproducibly (Kyle and , 2010). They are quantifiable characteristics of a biological process and considered as reliable predictors of health status. Biomarkers differ from the notion of end points but in clinical trials they are considered as surrogates for clinically meaningful end points.

The health condition of an individual is normally assessed using two distinct states *viz*., Healthy (H) or having Disease (D). Each individual under study will be classified into one of these states uniquely by using a diagnostic procedure like a blood test or a scan. Such procedures serve as biomarkers to define the presence or absence of a disease. For instance, Waist Circumference is a marker for CAD. The question is whether these markers can classify the individual without error? If such a marker exists, it is called a *gold standard.* For instance, biopsy is a test that distinguishes between malignant and benign tumor. Similarly cardiac arrest distinguishes between alive and death. But we wish to predict such conditions well in advance and save life.

One area of clinical interest is to define alternative biomarkers that are affordable and perform closely to the gold standard. They are considered as surrogate markers. For instance, the Fine Needle Aspiration Cytology (FNAC) is a pre-operative test to evaluate breast lump. The US-guided CNB is currently recognized as a reliable alternative (surrogate) to surgical biopsy for the histological diagnosis of breast lesions (Rahman *et al.*, 2011. Moschetta *et al.*, 2014).

Corresponding Author: K.V.S. Sarma
E-mail: smskvs@gmail.com

The performance of a marker is expressed in terms of its ability to distinguish between D and H states. These conditions are also stated as *True* and *False* states. Given the biomarker and a cutoff, we can count the number of true and false positives made by the marker, in relation to the gold standard. There can be several markers for the same health condition but each of them shall be compared with the gold standard to know how best the marker can distinguish between the presence and absence of disease.

In the following section we discuss the statistical issues in the assessment of biomarkers. In section 3 we define the context of multiple biomarkers and the use of logistic regression to handle the analysis.

2. Performance Measures of Binary Classification

Let X denotes the test result (also called classifier) and c be a cutoff value so that an individual is test positive if $X > c$ and negative otherwise (this rule could be the other way also). In a cohort of N individuals, we can count the number of True Positives (TP), False Positives (FP), True Negatives (TN), False Negatives (FN) and arrange them as a table shown below.

Diagnosis (gold standard)	Test Result	
	Positive	Negative
Positive	TP (a)	FN (b)
Negative	FP(c)	TN (d)

We note that $(a+b+c+d) = N$ and the sum $(b+c)$ denotes the number of misclassifications by the marker at the given cutoff. Two other important measures of performance popularly used in medicinal diagnosis are

(i) Sensitivity (S_n) or True Positive fraction (TPF) = $a/(a+b)$ = P[$X > c$ | D] and
(ii) Specificity (S_p) or True Negative fraction (FPF) = $d/(c+d)$= P[$X <= c$ |H] and
(iii) False Positive Fraction (FPF) = 1- S_p

Both S_n and S_p takes values between 0 and 1. For a given marker at a given cutoff, suppose S_n = 0.74. It means the test can pick up 74% of those who really have the disease and 26% missing and if S_p = 0.48 then 52% of individuals having no disease will be show as positive (false). For the test results that are continuous (measurements), every value is a possible cutoff with corresponding S_n and S_p.

In practice *confirmatory tests* need a cutoff having greater specificity and lower sensitivity while *screening tests* need greater sensitivity and lower specificity.

3. The Receiver Operating Characteristic Curve

We now need a measure of the performance of the test (as a classifier) at all possible cutoff values. The *Receiver Operating Characteristic Curve (ROC Curve)* is a graphical representation of the performance of a test or marker. It is a plot of TPF (Sensitivity) against FPF (1-Specificity) for different possible cutoffs. ROC curve is useful in several ways like

a) Finding the best threshold / cutoff value (in some sense) for the marker
b) Selection of the best marker among several available markers
c) Comparing two or more markers in terms of the Area Under the Curve (AUC)
d) Evaluating the positive and negative predictive values of the marker

All this is a data driven exercise with statistical inference.

Let the possible cutoff values be $c_1, c_2, \ldots, c_k$ and let the sample size be n_H and n_D for the healthy and diseased groups respectively. The ROC curve is constructed with the following steps and has interesting features as listed below.

1) At each c_i we can calculate S_n and S_p values by counting the TP and FP cases
2) Plot the S_n values against (1-S_p) values. The resulting ROC curve fits into the unit square and hence AUC is always <= 1
3) A diagonal line from left to right divides the area into two equal halves.
4) When the marker has only 50% chance of correct classification the curve matches with the diagonal line and the AUC will be 0.50
5) We need markers with AUC $\geq$ 0.5 and good if the value is closer to 1
6) ROC curves spanning below the diagonal line are of no interest
7) While comparing two or more markers, AUC is used as a measure. Higher the AUC better discrimination!

4. Sample ROC Curves

In clinical studies, we come across several biomarkers which are potential indicators of a health condition. Some markers will be categorical while some are measured. There will also be markers based on multiple conditions and serve as accepted references. Evaluation of a biomarker is however done with reference to only a gold standard. Here is an illustration.

Acute Physiology and Chronic Health Evaluation (APACHE) score and Sepsis Related Organ Failure Assessment (SOFA) score are two commonly used scoring systems to assess the status of a patient after admission into the Intensive Care Unit (ICU) of a hospital. These scores are calculated based on the patients parameters measured within 24/48 hours after admission into the ICU and higher score predicts death.

One can carry out the ROC curve analysis using standard software like MedCalc or SPSS. When the marker under consideration is continuous, every value is a possible cutoff and the ROC curve is built over these values. The AUC is a sample estimate of the true AUC and hence confidence interval will be provided for the true AUC. For instance in a sample study the results of ROC curve provide the following.

S.No.	Marker	AUC	95% Confidence Interval	Cutoff	S_n	S_p
1	APACHE	0.925	[0.815, 0.981]	> 17	100%	76.3%
2	SOFA	0.912	[0.798, 0.974]	> 10	83.3%	84.2%

For the marker APACHE the sensitivity of 100% indicates the score of >17 can predict the death almost sure but about 24% of individuals (100-76.3%) will be classified as false positive (wrongly predicted with the death as outcome)! But in the case of SOFA score 83% is the true positive rate at the cut off > 10 and only 16% were false positive.

Different markers will have different S_n and S_p values because the possible cutoffs are different. It is possible to find a cutoff value that balances these two indices and it is called the o*ptimal cutoff* which is found by one of the following methods

a) Value at which, the line joining the ROC curve with the top left corner of the unit square has shortest length
b) Value having the longest vertical distance from the curve to the horizontal axis (called Youden Index)

It is also possible to compare the statistical significance of the difference between two markers based on their AUC values. MedCac provides a module to compare ROC curves. A detailed discussion on ROC curve are well discussed (James and McNeil, 1982; Pepe, 2000; Krzanowski and David, 2009).

5. Design of New Markers

Designing of new biomarkers needs a systematic study by using statistically a valid protocol. Once a cutoff is proposed is made it is to be validated on known cases and tested on new cases before proposing the cutoff. The following steps are essential in this regard.

- Determine the sample size needed to achieve desired sensitivity and specificity taking into account the prevalence of disease.
- Collect data on known cases (already classified by gold standard)
- Divide the data into 'training group' and 'validation group'
- Perform ROC analysis and get AUC from training data
- Arrive at the 'new cutoff'
- Verify the results with this cutoff on the validation data
- Apply this cutoff on test data (new cases for whom status is unknown) and find the percentage of misclassifications.

A systematic study on deriving a new cutoff on an existing biomarker is carried out by Alladi Mohan *et al* (2016).

Sometimes a single marker may not give a satisfactory classification. Then it is possible to combine several markers into a model using statistical arguments. Logistic regression or Linear Discriminant Analysis is one approach to combine several markers and arrive at a *composite score* and compare its performance with the individual markers.

6. Composite Biomarkers and Biomarker Panels

Researchers now concentrate on *panel of markers* instead of single markers to distinguish between one health condition and the other. This concept is increasingly found applicable in the assessment of chronic diseases. The idea is to improve the chance of early detection of the disease with high sensitivity. In some cases a combination (weighted sum) of two or more markers using a statistical model and such markers are called composite markers.

A systematic development of composite markers is done as follows.

- Let $X_1, X_2, X_3,...X_k$ be markers whose AUC is already known.
- A new composite marker is defined as $Z = f(X_1, X_2, X_3,...X_k)$. One simplest form is a linear combination
  $Z = \beta_0 + \beta_1 X_{1+} \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \varepsilon$
- Use Binary Logistic Regression or Linear Discriminant Analysis to estimate the coefficients.
- The cutoff is of the form $Z > Z^*$ for classification.
- The power of the composite marker can be compared in terms of AUC

A simple example of a composite marker is the D-score obtained from the logistic regression model with SOFA and APACHE scores given by D-Score = -4.075+ 0.194*SOFA +0.139*APACHE. This marker has AUC = 0.962 with 95% CI : [0.865 to 0.996] and has higher AUC than that of the two individual scores.

Archana *et al.* (2016) have studied on the Validation of a biomarker panel and longitudinal biomarkers for early detection of ovarian cancer.

A new area of interest is the concept of longitudinal markers which are used to assess the disease progression leading to a state of severity or death over a period of time. Such markers are found to have specific interest in treatment of chronic diseases.

7. Conclusion

Design and implementation of biomarkers in health care is a specific area in statistical research as well as epidemiological studies. A strong theoretical basis is available for drawing inference on the design of biomarkers as well as measurement of their performance. The area of ROC curves has great potential in statistical decision making in the area of classification problems.

References

Alladi, M., Reddy, S.A., Sachan, A., Sarma, K.V.S., Kumar, D.P., Panchagnula, M.V., Srinivasa Rao, P.V.L.N., Siddhartha Kumar, B. and Krishnaprasanthi, P. (2016). Derivation and validation of glycosylated haemoglobin (HbA1c) cut-off value as a diagnostic test for type 2 diabetes in south Indian population. *Indian Journal of Medical Research*, **144**, 220-228.

Archana, R. Simmons, Charlotte, H. Clarke, Donna, B. Badgwell, Zhen Lu, Lori Sokoll, Karen, H. Lu, Zhen Zhang, Robert C. Bast and Steven J. Skates (2016), Validation of a biomarker panel and longitudinal biomarker performance for early detection of ovarian cancer. *International journal of gynecological cancer : official journal of the International Gynecological Cancer Society*, **26(6)**, 1070-1077. DOI:10.1097/IGC.0000000000000737.

James, H.A. and Mc Neil, B.J. (1982). A meaning and use of the area under a receiver operating characteristic (ROC) curves. *Radiology*, **143**, 29-36.

Krzanowski, J.W. and David J.H. (2009). ROC Curves for Continuous Data, Chapman & Hall/CRC.

Kyle, S. and Jorge A.T. (2010). What are biomarkers*? Current Opinion HIV AIDS*, **5**(6): 463–466.

Moschetta, M., Telegrafo, M., Carluccio, D.A., Jablonska, J.P., Rella, L., Serio, J., Carrozzo, M., Stabile Ianora, A.A. and Angelelli, G. (2014). Comparison between fine needle aspiration cytology (FNAC) and core needle biopsy (CNB) in the diagnosis of breast lesions. IL Giornale di Chrurgia, July-August, **35** (**7-8**), 171-176.

Pepe, M.S. (2000). Receiver operating characteristic methodology. *Journal of American Statistical Association*, **95**, 308-311.

Rahman, M.Z., Sikder, A.M. and Nabi, S.R. (2011). Diagnosis of breast lump by fine needle aspiration cytology and mammography. *Mymensingh Medical Journal*, **20**(**4**), 658-64.

Simmons, A.R., Clarke, C.H., Badgwell, D.B., Lu, Z., Sokoll, L.J., Lu, K.H., Zhang, Z., Bast Jr., R.C., and Skates, S.J. (2016). Validation of a biomarker panel and longitudinal biomarker performance for early detection of ovarian cancer. *International Journal of Gynecological Cancer*, **26**(**6**), 1070-1077.