Journal of Data Science **6**(2008), 189-205

On the Principles of Believe the Positive and Believe the Negative for Diagnosis Using Two Continuous Tests

Changyu Shen[1,2]

[1]*Indiana University and* [2]*Regenstrief Institute for Health Care*

Abstract: Believe the Positive (BP) and Believe the Negative (BN) rules for combining two continuous diagnostic tests are compared with procedures based on likelihood ratio and linear combination of the two tests. The sensitivity-specificity relationship for BP/BN is illustrated through a graphical presentation of a "ROC surface", which leads to a natural approach of choosing between BP and BN. With a bivariate normal model, it is shown that the discriminating power of this approach is higher when the correlation between the two tests has different signs for non-diseased and diseased population, given the location and variations of the two distributions are fixed. The idea is illustrated through an example.

Key words: Believe the Negative (BN), believe the Positive (BP), bivariate normal, (maximum) ROC curve, ROC Surface.

1. Introduction

The statistical properties of single diagnostic tests have been extensively studied. For a test with a continuous scale, the Receiver Operating Characteristic (ROC) curve is widely used as a method to characterize the accuracy of the diagnostic test. ROC displays the sensitivity versus 1-specificity achieved at different cut-off points applied to the test score. The area under the ROC curve (AUC) provides a measurement of the overall accuracy of the diagnostic test of interest by averaging the sensitivity at various specificity levels. Such an index also corresponds to the probability that a randomly selected diseased subject has higher test score than does a randomly selected non-diseased subject (assuming diseased population has higher mean test score than the non-diseased population).

When there is more than one diagnostic test, diagnostic accuracy can be improved by combining these tests. Simple methods of combining two diagnostic tests such as Believe the Positive (BP) and Believe the Negative (BN) have been widely used in many clinical settings. BP says that a subject is overall positive if either of the two tests is positive and BN says a subject is overall positive only if both tests are positive. When the measurements of the two individual tests are continuous, separate cut-off points are applied to each test. Then BP (BN) states that a subject is overall positive if either (both) test measurement(s) is (are) greater than the corresponding cut-off point(s). When the cut-points are fixed for the two tests, it is easy to see that (i) BP improves the sensitivity over either test at the cost of specificity and (ii) BN improves the specificity over either test at the cost of sensitivity. Another popular approach of combining two continuous tests is to construct a new measurement that is a linear combination of the two tests under consideration and a threshold is applied to the new measurement (Su and Liu, 1993). In Figure 1, we show the overall positive region in a two-dimensional plane for these three simple approaches, where the two dimensions correspond to the two test measurements.

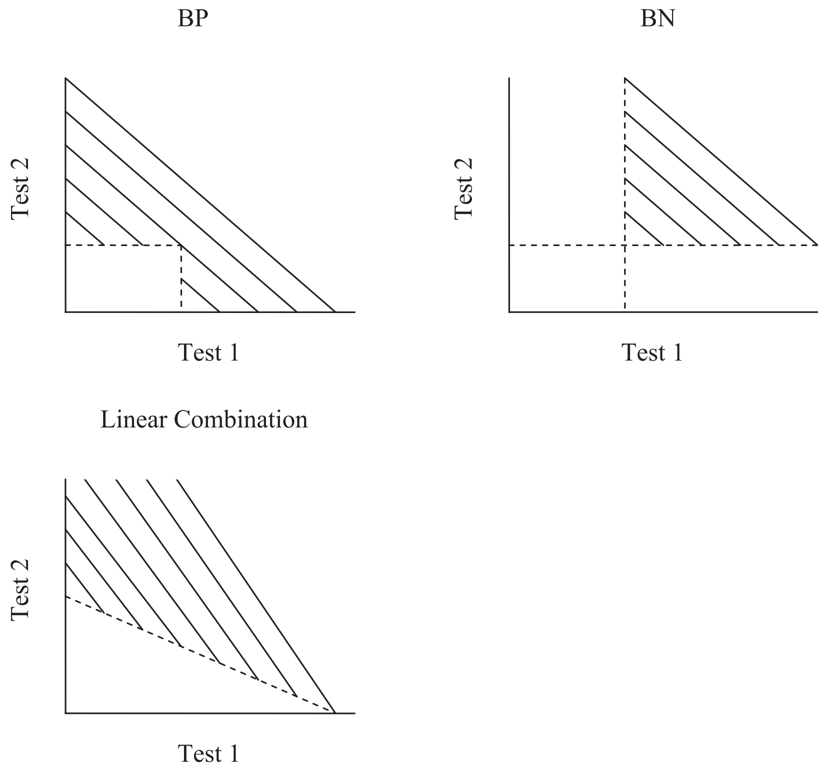

Figure 1: Overall positive region for BP, BN and Linear Combination rules. Shadowed area refers to the overall positive region.

In the view of hypothesis test, diagnosis can be thought of as a simple hypothesis test such that the null hypothesis is that the score of a subject is drawn from the non-diseased population and the alternative hypothesis the test score is drawn from the diseased population (Pepe, 2003). Then sensitivity and 1-spectificy correspond to the power and the type I error rate. For the screening phase of medical study of a rare disease, often we want to maximize the yield of the screening given a tolerable specificity level. Then the problem becomes seeking the most powerful test at a fixed type I error rate, which can be achieved by the likelihood ratio test based on the Neyman-Pearson lemma. However, even for simple parametric models of the distribution of the two diagnosis test scores, such a procedure is rather cumbersome for practical application in the screening phase, in which a large number of subjects are examined. For example, in a large disease screening or clinical study, many of the tests are conducted in a biological lab, where the test results are first recorded on papers and then are put into an electronic database later on. For many diseases, some extra clinical procedures will be scheduled/conducted right away if the tests just completed imply disease status for ethical and screening/study efficiency reasons. Therefore, a quick and straight forward way to make such a decision based on these tests is in need, where numerical calculations should be minimized. Moreover, in most scenarios, a parametric model is difficult to conjecture. Thus, the practical problem is to seek a simple sub-optimal procedure to combine more than one test. Therefore, research on the properties of simple methods as shown in Figure 1 is of great practical significance.

The ROC generated by the likelihood ratio test obviously is the optimal one because its sensitivity is maximized at each specificity level. Since likelihood ratio test under many parametric univariate models (e.g. normal distribution) reduces to applying the cutoff point to the test score, sensitivity is often maximized for a given specificity level under such models. In the scenario of two continuous diagnostic tests, however, it is not clear how the simple rules in Figure 1 compare with the optimal one. Moreover, there are few guidelines on the usage of BP and BN rules, though there has been some literature on the optimal linear combination of two continuous tests that maximizes the AUC (Pepe and Thompson, 2000; Su and Liu, 1993) among all possible linear combinations. In this paper, we try to address some properties of BP and BN as to when and how to apply them to maximize the power of disease detection, using the model of bivariate normal distribution. Specifically, we will discuss the following issues:

(a) How is the ROC curve based on BP/BN compared with the optimal one? When does BP/BN perform better than the best linear combination?

(b) When should we choose BP or BN?

(c) Under what condition (if possible) is BP/BN powerful in terms of their capability of discriminating diseased from non-diseased?

We address question (a) in Section 2 and (b) and (c) in Section 3. We conclude this paper with a discussion section.

2. A Comparison of ROC Curves

In this section, we compare the ROC curves based on the likelihood ratio test (LR), BP, BN and the best linear combination of the two tests (LIN) (Su and Liu, 1993). The aims are to seek some guidelines as to how much power we lose using BP, BN or LIN as compared with LR and which of these three rules are better. We first introduce some notation to facilitate the discussion in this and later sections. Let $X = (X_1, X_2)$ be a vector of two continuous tests such that diseased subjects have larger mean values on both tests than non-diseased subjects. Let D and N denote diseased and non-diseased population, respectively; $F_{1,N}(\cdot), F_{2,N}(\cdot)$ and $F_N(\cdot,\cdot)$ denote the cumulative distribution function (CDF) of X_1, the CDF of X_2 and the joint CDF of X_1 and X_2 for non-diseased population, respectively; $F_{1,D}(\cdot), F_{2,D}(\cdot)$ and $F_D(\cdot,\cdot)$ are similarly defined for the diseased population; $S_N(\cdot,\cdot)$ and $S_D(\cdot,\cdot)$ are the joint survival function of X_1 and X_2 (e.g. $S_N(c_1, c_2) = Pr[X_1 \geq c_1, X_2 \geq c_2 \,|\, N]$) for the non-diseased and diseased population, respectively. For bivariate normal distribution, we will use $\mu_N = (\mu_{N1}, \mu_{N2})^T$ to denote the mean vector and Σ_N to denote the variance-covariance matrix composed of variances of X_1 and X_2 (σ^2_{N1} and σ^2_{N2}) and their covariance based on correlation ρ_N for the non-diseased population. Similarly, we use (μ_D, Σ_D) to denote the same set of parameters for the diseased population.

Under the bivariate normal distributions, it is easy to show that the LR essentially claims test positive when

$$Y = (X - \mu_N)^T \Sigma_N^{-1} (X - \mu_N) - (X - \mu_D)^T \Sigma_D^{-1} (X - \mu_D) > c_1 \tag{2.1}$$

As for BP/BN, they claim test positive when

$$X_1 > c_2 \ \text{ or/and } \ X_2 > c_3 \tag{2.2}$$

Finally, the LIN claims test positive when

$$W = X_1 + aX_2 > c_4 \tag{2.3}$$

where $a = b_2/b_1$ and $(b_1, b_2) = (\Sigma_D + \Sigma_N)^{-1}(\mu_D - \mu_N)$. The sensitivity and specificity of the above methods will depend on the values of c_1 to c_4. The major difference of the BP and BN from the other two rules lies in that there is more than one sensitivity value corresponding to a fixed specificity, which we will illustrate with greater detail in a later section. Nevertheless, we can use the maximum of these sensitivity values and plot it against each 1-specificity value, which we will call the *maximum ROC (MROC)* curve (Thompson, 2003). Then

MROC reflects the capability of discriminating diseased from non-diseased for BP/BN. In summary, the (maximum) sensitivity (s) at a specificity value ($1-t$) of the four methods can be calculated as:

$$
\begin{aligned}
s^{LR} &= 1 - G_D(G_N^{-1}(1-t)) \\
s_M^{BP} &= \max_{F_N(c_2,c_3)=1-t} 1 - F_D(c_2,c_3) \\
s_M^{BN} &= \max_{S_N(c_2,c_3)=t} S_D(c_2,c_3) \\
s^{LIN} &= 1 - H_D(H_N^{-1}(1-t))
\end{aligned}
$$

where G and H are the CDFs of Y and W, respectively, and the subscript M for BP and BN stands for maximum. Under the bivariate normal distribution, the quantities listed above cannot be calculated analytically and numerical methods are required.

We created the ROC curves for LR and LIN and MROC curves for BP and BN under various parameter combinations. Under all calculations, we set $\mu_N = (0,0)^T, \mu_D = (0,0)^T$, and vary the variance-covariance structure for the diseased and non-diseased. First, we fix the correlation between the two diagnostic tests at 0 for both diseased and non-diseased. The result is shown in Figure 2. Under all four scenarios, at least one of BP and BN is quite close to the LR. Except when the two populations have the same variance-covariance matrix (a), at least one of BP and BN is at least as good as LIN at high specificity level (e.g. specificity > 0.9, or equivalently, $1-$ specificity < 0.1) uniformly. In particular, for quite a broad range of the specificity at high level end, BN beats LIN when the variances of the diseased are smaller than the non-diseased (b) and BP beats LIN when the variances of the diseased are greater than the non-diseased (c). LIN has the same ROC as the LR in (a). Simple algebra shows that this is because the LR reduces to LIN when the variance-covariance matrix is the same between the two populations. Second, we fix the standard deviations of the two variables for the two populations at 1 and vary the correlation. The result is shown in Figure 3. Except (h), either BP or BN is quite close to the LR. For (f) (positive correlation in non-diseased and negative correlation in diseased) and (g) (negative correlation in non-diseased and positive correlation in diseased), either BP or BN is at least as good as LIN for a broad range of specificity at high level end. Since the diseased and non-diseased populations have the same variance-covariance matrix in (e) and (h), LIN is the same as LR.

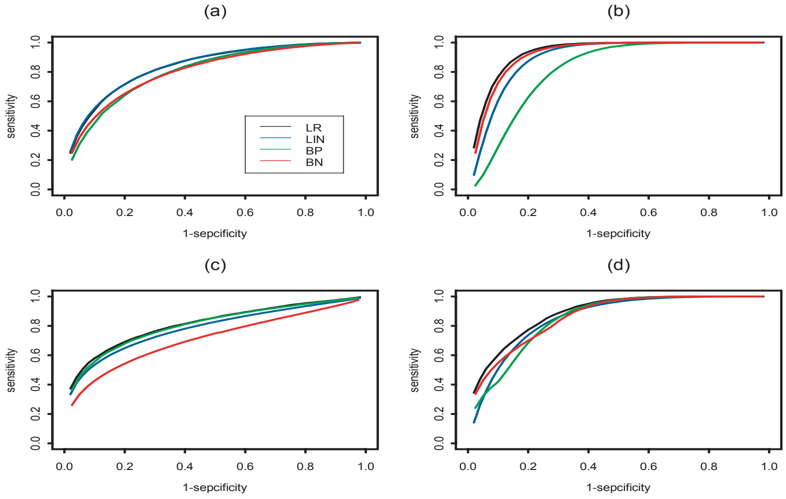

Figure 2: (M)ROC curves of LR, LIN, BP and BN for $\mu_N = (0.0)^T, \mu_D = (1,1)^T$ and $\rho_N = \rho_D = 0$, (a)$\sigma_{N1} = \sigma_{N2} = 1, \sigma_{D1} = \sigma_{D2} = 1$, (b) $\sigma_{N1} = \sigma_{N2} = 1, \sigma_{D1} = \sigma_{D2} = 0.5$, (c) $\sigma_{N1} = \sigma_{N2} = 1, \sigma_{D1} = \sigma_{D2} = 1.5$ and (d) $\sigma_{N1} = \sigma_{N2} = 1, \sigma_{D1} = 1.5, \sigma_{D2} = 0.5$.

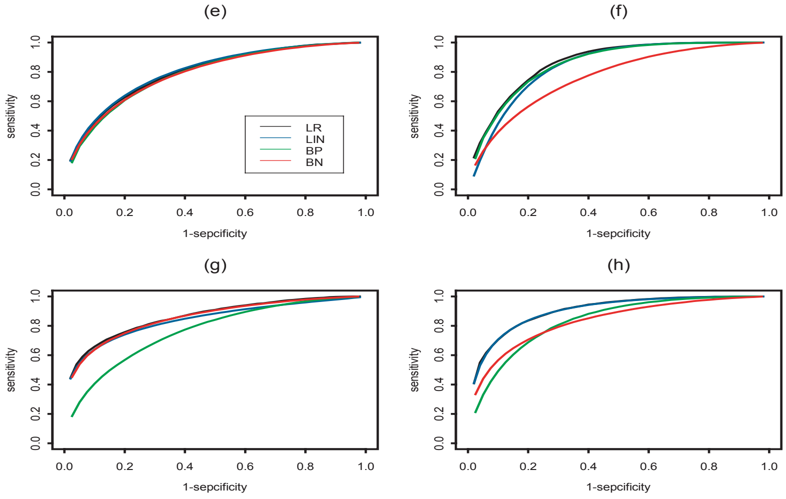

Figure 3: (M)ROC curves of LR, LIN, BP and BN for $\mu_N = (0,0)^T, \mu_D = (1,1)^T$ and $\sigma_{N1} = \sigma_{N2} = \sigma_{D1} = \sigma_{D2} = 1$, (e) $\rho_N = 0.4, \rho_D = 0.4$, (f) $\rho_N = 0.4, \rho_D = -0.4$, (g) $\rho_N = -0.4, \rho_D = 0.4$ and (h) $\rho_N = -0.4, \rho_D = -0.4$.

Since usually it is rather rare to observe the same variance-covariance matrix between the diseased and non-diseased population, Figure 1 and 2 suggest that either BP or BN is preferable to the LIN, at least at the high specificity level, which is usually what we desire. Then a natural thought would be to use the rule (BP or BN) that has a greater sensitivity value (at a tolerable specificity level). In next section, we discuss in further detail on this issue.

3. A Closer Look at BP and BN

3.1 ROC Surface

An ROC curve can be thought of as a graph of sensitivity versus the threshold if we replace the 1-specificity value with the threshold applied to the diagnostic test. It is natural to think that there should be a surface when there are two variables. We define an ROC surface in this subsection, which essentially is a stacking of infinitely number of ROC curves and allows one to visualize how these ROC curves change and where the sensitivity reaches its maximum for a fixed specificity level.

We first consider BP. Let $p = F_{1,N}(c_2)$ be the specificity when using X_1 as the only test at threshold c_2. Fix the specificity level at $1-t$ and we can obtain c_3 by solving the equation $F_N(F_{1,N}^{-1}(p), c_3) = 1-t$. We will denote the corresponding solution by $c_3 = Q(t, F_{1,N}^{-1}(p))$. Then we can rewrite s^{BP} as:

$$S^{BP} = 1 - F_D(F_{1,N}^{-1}(p), Q(t, F_{1,N}^{-1}(p))). \tag{3.1}$$

Hence, s^{BP} is a function of p and t. A plot of s^{BP} against the two arguments will be called an *ROC surface*, in which p is the X axis, t is the Y axis and s^{BP} is the Z axis. Each fixed p cuts through the surface and yields a ROC curve. On the other hand, each fixed t cuts through the surface and yields a curve from which one can identify the s_M^{BP}. Note that

$$1 - t \leq F_N(c_2, \infty) = F_{1,N}(c_2) = p.$$

In other words, the specificity of the BP is always less than the specificity by applying only one test as mentioned earlier. Then the domain of the ROC surface of the BP is the upper triangle of the unit square $[0,1] \times [0,1]$. Similar reasoning can be applied to the BN so that

$$S^{BN} = S_D(F_{1,N}^{-1}(p), R(t, F_{1,N}^{-1}(p))), \tag{3.2}$$

where $R(t, F_{1,N}^{-1}(p))$ is the solution of c_3 for equation $S_N(F_{1,N}^{-1}(p), c_3) = t$. Since $F_{2,N}(c_3) - F_N(c_2, c_3) \geq 0$,

$$t = 1 - F_{1,N}(c_2) - (F_{2,N}(c_3) - F_N(c_2, c_3)) \leq 1 - F_{1,N}(c_2) = 1 - p.$$

Thus, the domain of the ROC surface of BN is the lower triangle of the unit square $[0,1] \times [0,1]$. In Figure 4, we show the ROC surface of BP and BN with

$\mu_N = (0,0)^T$, $\sigma_{N1} = \sigma_{N2=1}$ and $\rho_N = 0$=0, and $\mu_D = (1,1)^T$, $\sigma_{D1} = \sigma_{D2} = 1$ and $\rho_D = 0$. In Figure 4, blue color represents the ROC surface of BP and orange color represents the ROC surface of BN. The orange curve corresponding to $p = 0$ and the blue curve corresponding to $p = 1$ are identical, both of which are the ROC curve when applying X_2 as the only test. The projection of the curve where the blue surface meets the orange surface onto the $t - s$ plane is the ROC curve when applying X_1 as the only test. We show in Figure 5 three ROC curves cut off from the ROC surface at different values of p.

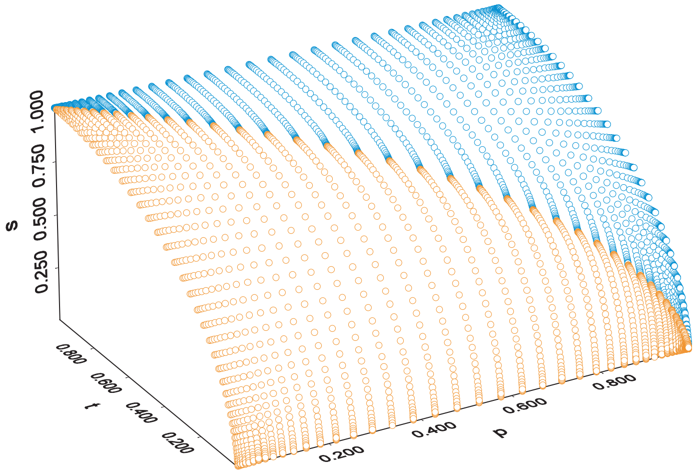

Figure 4: ROC surface BP and BN for $\mu_N = (0,0)^T, \sigma_{N1} = \sigma_{N2} = 1$ and $\rho_N = 0$, and $\mu_D = (1,1)^T, \sigma_{D1} = \sigma_{D2} = 1$ and $\rho_D = 0$ (blue: BP; orange: BN).

For BP/BN, the MROC curve is the projection of a path on their ROC surface to the t-s plane such that each point on the path corresponds to the maximum sensitivity value achieved at corresponding specificity value $(1 - t)$. It is well known that the AUC of the ROC of a single test is equal to the probability that the test score of a randomly selected diseased subject is higher than that of a randomly selected non-diseased subject. Then a natural question is whether or not there is a probabilistic interpretation of the AUC under the MROC as in the single test scenario. For BP, intuition suggests that this quantity might be equal to the probability (P) that a randomly selected diseased subject has at least one test score higher than that of a randomly selected non-diseased subject. Unfortunately, this is not true in general. A simple counter-example is as follows. Suppose X_1 and X_2 are independent for both diseased and non-diseased populations such that $F_{1,N}(x) = F_{2,N}(x) = 1 - e^{-x}$ and $F_{1,D}(x) = F_{2,D}(x) = 1 - 0.5e^{-0.5x}$. Thus X_1 and X_2 follow exponential distribution with mean 1 and 2 for non-diseased and diseased populations, respectively. Straight forward calculation shows that the AUC under the MROC is 11/15 and $P = 8/9$. Similarly, it is not true for BN, either.

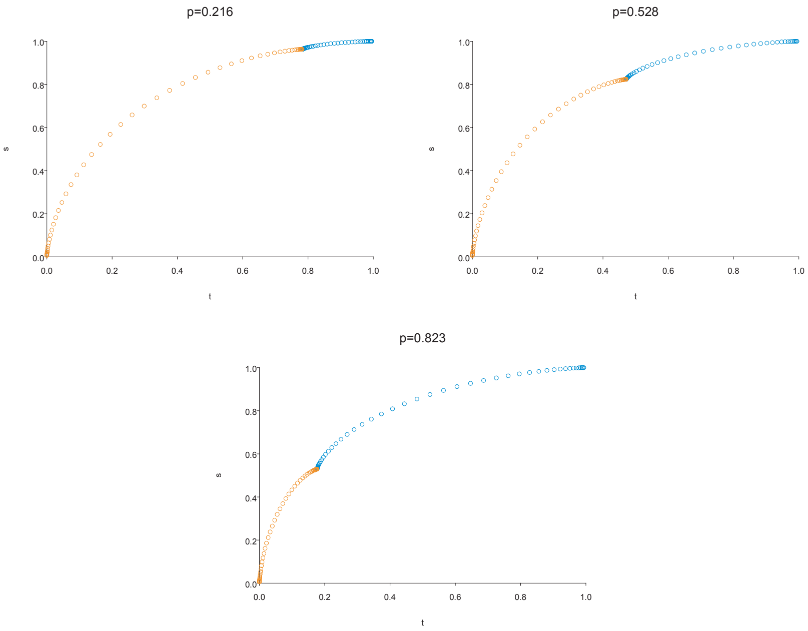

Figure 5: Three ROC curves (BP and BN together) from the ROC surface in Figure 2.

3.2 Choice between BP and BN

Since the performance of BP and BN depends on the parameter setting (Figure 2 and 3), one obvious way to choose between them is to use the maximum of $s_M^{BP}(t)$ and $s_M^{BN}(t)$:

$$s_M(t) = \max\{s_M^{BP}(t), S_M^{BN}(t)\}.$$

This is equivalent to locating the maximum s on the path across the two surfaces for fixed t (Figure 1). Then the maximum can be on the BP surface or on the BN surface. Since such a path also includes the sensitivities achieved at each individual test, $s_M(t)$ is greater than any one of them. We computed $s_M(t)$ for the distributions shown in Figure 3. The curves are shown in Figure 6 with orange indicating that s_M is on the BN surface and blue indicating that s_M is on the BP surface. It can be seen that one rule (BP or BN) is always better than the other when the correlation is of different sign. When the correlation sign is the same, BN outperforms BP at high specificity region and the opposite occurs at low specificity region. For example, one wants to use the BN rule to maximize the sensitivity when t is smaller than 0.3 (specificity greater than 0.7) and use BP rule when t is larger than 0.3 (specificity less than 0.7) under double-positive and double-negative correlation scenarios.

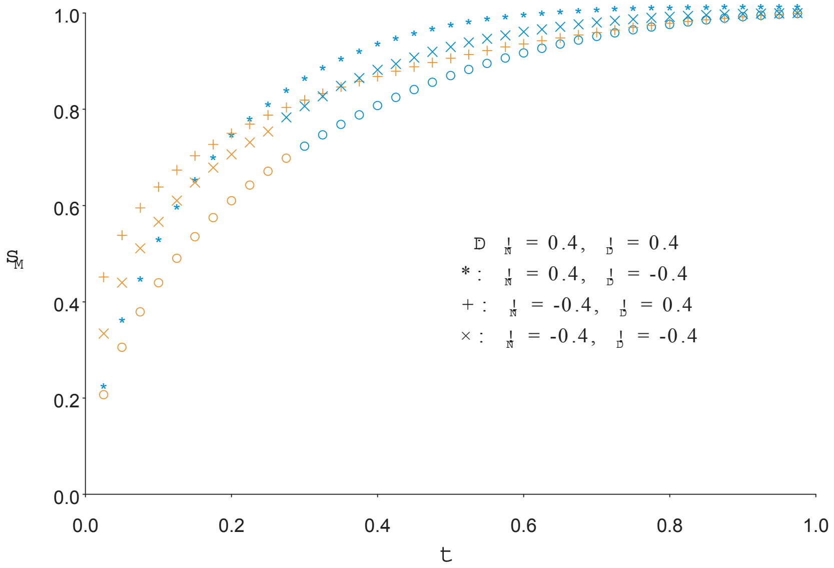

Figure 6: $S_M(t)$ under $\mu_N = (0,0)^T$ and $\sigma_{N1} = \sigma_{N2} = 1$ and $\mu_D = (1,1)^T$ (Orange: $s_M(t)$ from BN; Blue: $s_M(t)$ from BP.

Opposite correlation signs between the non-diseased and diseased populations tend to yield higher sensitivity in general. This is not surprising since different correlation direction provides another ”dimension” to discriminate non-diseased from diseased subjects. Particularly, negative correlation in the non-diseased and positive correlation in the diseased yield higher sensitivity than others at high specificity region, and positive correlation in the diseased and negative correlation in the non-diseased yield higher sensitivity at low specificity region. For the two situations where the two populations have the same sign for the correlation, double-negative-correlation yields higher sensitivity than double-positive-correlation uniformly in all specificity values. Therefore, one might want to choose a pair of tests that have opposite correlation signs between the diseased and non-diseased populations if possible. It appears that the double-positive-correlation is the least favorable situation, which, unfortunately, is often what we deal with.

In reality, we rarely know the true underlying distribution of the test measurements for the diseased and non-diseased population. Usually, we will have some data from subjects whose diseases status is known before we decide what should be the test procedure. Then one can estimate $s_M(t)$ (denoted by $\hat{s}_M(t)$) by fitting a parametric distribution to the two variables and standard errors of $\hat{s}_M(t)$ can be estimated by Bootstrap (Efron and Tibshirani, 1993). A tool to construct such parametric models is the copula functions (Nelsen, 1998). However, it is sometimes difficult to specify a parametric distribution model in practical application. Then an empirical approach by searching appropriate cut-off points on the observed data can be applied. The search allows one to maximize the sensitivity when the specificity value is restricted to be at least equal to a fixed value. Specifically, let $x_1 = (x_{11}, x_{21}, \ldots, x_{n1})$ and $x_2 = (x_{12}, x_{22}, \ldots, x_{n2})$ be the observed values of the two variables and $d = (d_1, d_2, \ldots, d_n)$ be the disease status indicator such that $d_i = 1$ indicates subject i has the disease and 0 otherwise. Define

$$\begin{aligned}
\hat{F}_D(c_1, c_2) &= \sum_{i=1}^{n} I(x_{i1} \le c_1, x_{i2} \le c_2, d_i = 1) / \sum_{i=1}^{n} d_i \\
\hat{F}_N(c_1, c_2) &= \sum_{i=1}^{n} I(x_{i1} \le c_1, x_{i2} \le c_2, d_i = 0) / \sum_{i=1}^{n} (1 - d_i) \\
\hat{S}_D(c_1, c_2) &= \sum_{i=1}^{n} I(x_{i1} \ge c_1, x_{i2} \ge c_2, d_i = 1) / \sum_{i=1}^{n} d_i \\
\hat{S}_D(c_1, c_2) &= \sum_{i=1}^{n} I(x_{i1} \ge c_1, x_{i2} \ge c_2, d_i = 0) / \sum_{i=1}^{n} (1 - d_i)
\end{aligned}$$

Then an empirical approach for fixed t seeks

$$\hat{s}_M^{BP}(t) = \max_{1-\hat{F}_N(c_1,c_2) \le t} 1 - \hat{F}_D(c_1, c_2) \tag{3.3}$$

$$\hat{s}_M^{BN}(t) = \max_{\hat{S}_D(c_1,c_2) \le t} \hat{S}_D(c_1, c_2) \tag{3.4}$$

Often $\hat{s}_M^{BP}(t)$ and $\hat{s}_M^{BN}(t)$ are obtained at different specificity values and in general not comparable. One can then decide which rule to choose based on relative "importance" between sensitivity and specificity (Kraemer, 1992). The standard errors of $\hat{s}_M^{BP}(t)$ and $\hat{s}_M^{BN}(t)$ can be easily estimated and are omitted here.

3.3 An example

The Indianapolis Study of Health and Aging is an on-going longitudinal study of dementia and Alzheimer's disease in the elderly starting 1992 (Hendrie, Ogunniyi, Hall, Baiyewu, Unverzagt, Gureje, Gao, Evans, Ogunseyinde, Adeyinka, Musick and Hui, 2001). The study participants are 2212 African Americans age 65 and older living in Indianapolis. A population-based two-phase survey (Pickles, Dunn and Vazquez-Barquero, 1995) was conducted at each data collection wave. There was first an in-home screening followed by a full clinical assessment for a subsample of participants selected based on the performance of the screening test. The screen in the first phase is intended to select subjects with a high chance to have dementia in a cost-efficient way. Currently, the Community Screening Interview for Dementia (CSID) (Hall, Gao, Emsley, Ogunniyi, Morgan and Hendrie, 2000) is used as the screen test, which consists of a cognitive assessment of the study participants and an interview with a close relative (informant) evaluating the daily functioning of the participants.

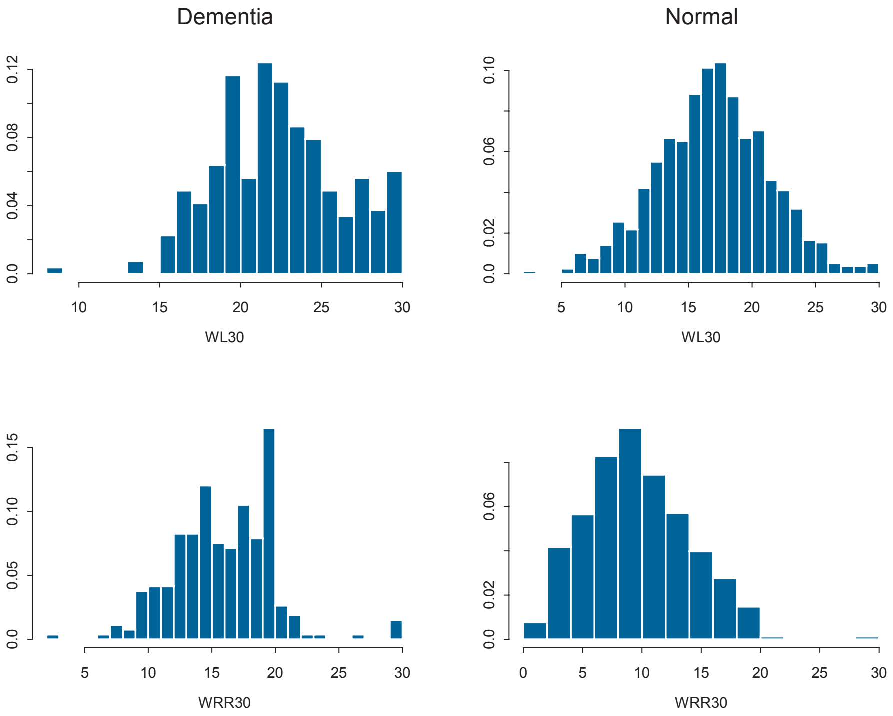

Figure 7: Histogram of $WL30$ $(30-WL)$ and $WRR30$ $(30-WRR)$ for subjects with dementia and normal subjects.

Here we consider using two continuous neuropsychological tests to discriminate subjects with dementia from normal subjects. Since most, if not all, neuropsychological tests are positive correlated, we do not have much choice in selecting pair of tests based on their correlations. We choose two tests that are used to measure the learning and memory function domains because dementia is always accompanied by loss in these two domains. The first one is called Word List Learning (WL) and the second one is called Word List Delayed Recall and Recognition (WRR). For Word List Learning, 10 words are read to the participant in three different trials (same 10 words, different orders) and the participant is asked to repeat the words at the end of each trial. For World List Delayed Recall and Recognition, the participant is asked to repeat the 10 words after a little while and recognize them from another list of 20 words read to them. The maximum score of both X_1 and X_2 is 30. Since high-risk group tends to have lower scores on both variables, two new variables are created for the implementation of our approach: $WL30 = 30 - WL$ and $WRR30 = 30 - WRR$. We select 1046 subjects who had at least one clinical assessment, among which 266 were diagnosed as dementia. WL and WRR scores right before the time of diagnosis are selected for diseased subjects and the same scores at last neuropsychological evaluation are selected for normal subjects. In Figure 7, we show the histogram of $WL30$ and $WRR30$ for the two groups of subjects. It demonstrates that the distributions of the two variables for subjects with dementia shifts towards the right from the distributions of the normal subjects. The normality assumption for normal subjects seems to be reasonable, though somehow questionable for the other group. In Figure 8 we show $\hat{s}_M(t)$ based on bivariate normal approximation (blue and orange circle), $\hat{s}_M^{BP}(t)$ (blue circle) and $\hat{s}_M^{BN}(t)$ (orange circle) based on (3.3) and (3.4). For comparison, we also included in the graph the ROC curves from best linear combination of the two tests (bivariate normal approach and distribution-free approach) (Pepe, 2000, 2003). It can be seen that all curves are quite close to each other. In Table 1 we show the thresholds for $WL30$ and $WRR30$ and associated sensitivity (S.E.) and specificity (S.E.) for BP and BN based on numerical search. We use the first row as an example to illustrate how to read the table. For BP, when the test positive is defined as $WL30$ greater or equal to 27 (WL less or equal to 3) OR WRR30 greater or equal to 19 (WR less or equal to 11), the sensitivity is 0.383 and specificity is 0.953; for BN, when the test positive is defined as $WL30$ greater or equal to 19 (WL less or equal to 11) AND $WRR30$ greater or equal to 18 (WR less or equal to 12), the sensitivity is 0.402 and specificity is 0.950. The BP and BN seem to perform quite similarly in most cutoff points, though at certain situations one is better than the other one (e.g. 6th and 7th row, BP is better than BN; 4th row, BN is better than BP).

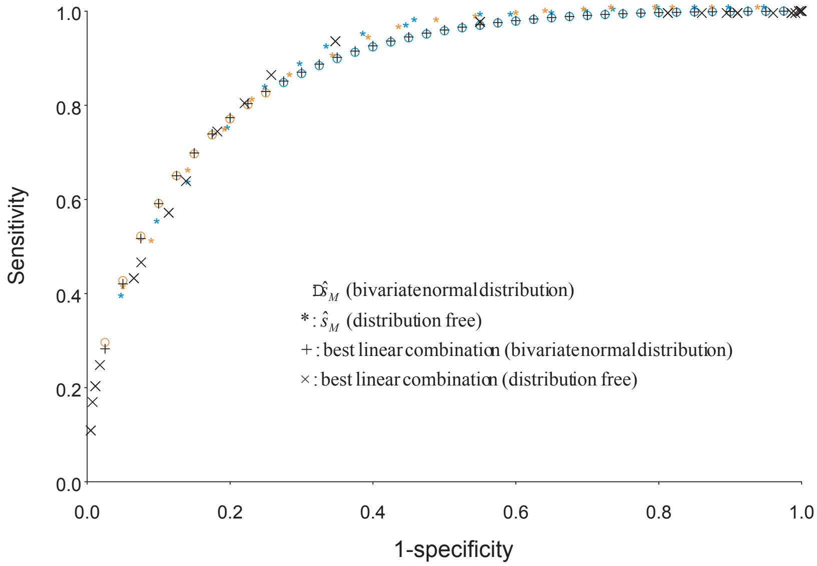

Figure 8: Different approaches based on WL30 and WRR30: $\hat{s}_M$ from bivariate normal distribution approximation (blue and orange circle), $\hat{s}_M^{BP}$ from empirical search (blue star) and $\hat{s}_M^{BN}$ from empirical search (orange star), ROC curve of best linear combination by a distribution-free approach (black mark sign).

4. Discussion

We have demonstrated that one of BP and BN principles, as a simple rule for diagnosis based on two continuous tests, has favorable power in detecting diseased subjects under most scenarios. It is our belief that application of the optimal one (BP or BN, depending on the underlying distribution) is of great practical feasibility and cost-effectiveness due to their simplicity and potential to improve prediction accuracy as compared with each single test. In many clinical settings, we want to balance the prediction accuracy and the cost. For example, in Section 3.3, first phase screen is conducted for every subject available at the corresponding data collection wave to select demented subjects. We hope to include as many demented subjects as possible at the price of some false positives. Hence, the test in the screen phase does not need to be highly specific, yet it should be sensitive enough to be cost-effective since every subject needs to go through it. Currently, the selection process based on Community Screening Interview for Dementia (CSID) involves categorizing the subjects into three performance groups and a random sample is drawn from each group with different sampling rates for the second phase assessment. The BP and BN rules can also be extended to fit this framework and is subject to further investigation.

The focus in our work is the comparison and selection of the BP and BN due to their close connection so that prediction accuracy can be improved. Furthermore, the BP and BN discriminate non-diseased from diseased via the difference in the correlation in addition to the difference in the location of the two distributions. This property can be useful in guiding us in the search for candidate tests for diagnosis. For example, tests for cancer aimed at the same anatomic or morphologic features of a tumor, such as palpation and mammography can be quite differently correlated for subjects with tumors than subjects without tumors (Marshall, 1989). From another perspective, Marshall (Marshall, 1989) discussed relationship between predictive value and asymmetry and strength of the correlation of two binary tests.

Table 1: Thresholds of $WL30$ and $WRR30$ for BP and BN and associated sensitivity (S.E.) and specificity (S.E.) based on numerical search on the observed data points

BP				BN			
Sen	Spe	$WL30$	$WRR30$	Sen	Spe	$WL30$	$WRR30$
(S.E.)	(S.E.)			(S.E.)	(S.E.)		
0.383	0.953	27	19	0.402	0.950	19	18
(0.017)	(0.008)			(0.018)	(0.008)		
0.541	0.903	27	17	0.500	0.910	22	15
(0.018)	(0.010)			(0.018)	(0.010)		
0.624	0.859	25	16	0.650	0.859	20	14
(0.017)	(0.012)			(0.017)	(0.012)		
0.741	0.804	24	15	0.737	0.808	20	12
(0.016)	(0.014)			(0.016)	(0.014)		
0.827	0.751	22	15	0.801	0.769	19	12
(0.014)	(0.015)			(0.014)	(0.015)		
0.876	0.703	24	13	0.853	0.717	2	13
(0.013)	(0.016)			(0.013)	(0.016)		
0.914	0.665	22	13	0.895	0.656	2	12
(0.011)	(0.017)			(0.011)	(0.017)		
0.940	0.614	22	13	0.932	0.606	16	11
(0.009)	(0.017)			(0.009)	(0.017)		
0.959	0.554	23	11	0.955	0.564	17	10
(0.007)	(0.018)			(0.007)	(0.018)		
0.970	0.542	20	12	0.970	0.512	15	10
(0.006)	(0.018)			(0.006)	(0.018)		

Sen = Sensitivity, Spe = Specificity for short

In summary, the specialty of choosing between BP and BN for two continuous tests lies in its balance in simplicity and discriminating power. Performance of such a procedure is most favorable when the correlation of the two tests has different signs between the diseased and non-diseased population under the bivariate normal model. Although it is straight forward to generalize the idea to more than two continuous tests, the intensive computational burden and the ”curse of dimensionality” tremendously limit its practical implementation. Many alternative approaches might be more efficient such as CART, adaptive thresholds (Thompson, 2003) and so on for more than two continuous tests.

Acknowledgment

This research was supported by NIH grants: R01 AG15813, R01 AG09956 and P30 AG10133. We would like to thank Dr. Sujuan Gao for critical comments on the manuscript.

References

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall/CRC.

Hall, K. S., Gao, S., Emsley, C. L., Ogunniyi, A. O., Morgan, O. and Hendrie, H. C. (2000). Community screening interview for dementia (CSI’D’); performance in five disparate study sites. *International Journal of Geriatric Psychiatry* **15**, 521-531.

Hendrie, H. C., Ogunniyi, A., Hall, K. S., Baiyewu, O., Unverzagt, F. W., Gureje, O., Gao, S., Evans, R. M., Ogunseyinde, A. O., Adeyinka, A. O., Musick, B. and Hui, S. L. (2001). Incidence of dementia and Alzheimer disease in two communities. *Journal of the American Medical Association* **285**, 739-747.

Kraemer, H. C. (1992). *Evaluating Medical Tests*. Sage Publications.

Marshall, R. J. (1989). The predictive value of simple rules for combining two diagnostic tests. *Biometrics* **45**, 1213-1222.

Nelsen, R. B. (1998). *An Introduction for Copulas*. Springer.

Pepe, M. S. (2003). *Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press.

Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123-140.

Pickles, A., Dunn, G. and Vazquez-Barquero, J. L. (1995). Scre ening for stratification in two-phase (”two-stage”) epidemiological survey. *Statistical Methods in Medical Research* **4**, 73-89.

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350-1355.

Thompson, M. L. (2003). Assessing the diagnostic accuracy of a sequence of tests. *Biostatistics* **4**, 341-351.

Received January 11, 2007; accepted March 25, 2007.

Changyu Shen
Division of Biostatistics
School of Medicine
Indiana University
1050 Wishard Boulevard RG R4101
Indianapolis, IN 46202, USA
chashen@iupui.edu