Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Xiwen Zhao Date

Development and Evaluation of Concavity-Respecting ROC Curve Estimators

By

Xiwen Zhao

Master of Science in Public Health

Biostatistics and Bioinformatics

Yijian (Eugene) Huang, PhD

(Thesis Advisor)

Amita Manatunga, PhD

(Reader)

Development and Evaluation of Concavity-Respecting ROC Curve Estimators

By

Xiwen Zhao

B.S.

Shanghai Jiao Tong University

2013

Thesis Committee Chair: Yijian (Eugene) Huang, PhD

Reader: Amita Manatunga, PhD

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics and Bioinformatics

2018

Abstract

Development and Evaluation of Concavity-Respecting ROC Curve Estimators

By Xiwen Zhao

Biomarkers are widely used for the diagnosis of a specific disease. To evaluate the diagnostic accuracy of a biomarker, the Receiver Operating Characteristic (ROC) curve analysis is one of the best developed statistical tools. In clinical context, typically higher biomarker value indicates larger possibility of the disease and then, the corresponding ROC curve should be strictly concave. However, the empirical ROC curve, as often adopted, does not necessarily respect the concavity property. In this study, we developed two methods to modify the empirical ROC curve in order to restore the concavity. Extensive simulations were conducted, and they showed that our modified ROC estimators for the area under curve (AUC) and the specificity at a fixed sensitivity have performance comparable to the empirical estimators.

Development and Evaluation of Concavity-Respecting ROC Curve Estimators

By

Xiwen Zhao

B.S.

Shanghai Jiao Tong University

2013

Thesis Committee Chair: Yijian (Eugene) Huang, PhD

Reader: Amita Manatunga, PhD

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics and Bioinformatics

2018

Table of Contents

1.Introduction

In this era of precision medicine, many candidate biomarkers have been or are being discovered to assist in monitoring asymptomatic patients, disease detection and diagnosis, and prediction of treatment response. To validate these biomarkers, it is important to evaluate their accuracy to correctly classify one condition from another (i.e., diseased versus non-diseased) (Søreide, 2009). When the testing results of biomarkers are measured on ordinal or continuous scale, the sensitivity and specificity would be functions of a selected cut-off point "c" ranging over all the possible threshold values (Hajian-Tilaki, 2013). In such a context, the Receiver Operating Characteristic (ROC) curve analysis is currently the best-developed method to describe the performance of biomarkers (Pepe, 2003).

The ROC curve was developed during WWII for analyzing classification accuracy in radar detection, before its principles later were expanded to improve medical decision making. Over the years, it has implemented to many fields including atmospheric science, biology, experimental psychology, and sociology etc. (Goncalves,2014). ROC analysis has also been increasingly applied in machine learning and data mining recently. Its advantages include testing accuracy across the entire range of test scores without requiring a predetermined cut-off point. In addition, ROC analysis allows comparison across different diagnostic tests or classifiers, by simply plotting the curves, or comparing the values of summary index (e.g., area under the curve). Furthermore, ROC does not require the knowledge of disease prevalence in population, which is very important in biomedical research as case-control studies are often conducted.

In diagnostic test with dichotomous outcomes, the conventional approach of diagnostic test evaluation uses sensitivity (proportion of true positives that are correctly classified by the test) and specificity (proportion of true negatives that are correctly classified by the test) as measure of accuracy, in comparison with gold standard. In situations that the test results are measured on continuous scales, the sensitivity and specificity can be ranged over all the possible threshold values (Hajian-Tilaki,2013). Sensitivity is also known as True Positive Rate (TPR) and 1- Specificity is known as False Positive Rate (FPR). The plot of TPR versus FPR gives rise to the Receiver Operating Characteristic (ROC) curve. The ROC curve is a monotone increasing function mapping (0,1) onto (1,0). For an uninformative test, whose probability of detecting positive disease status is unrelated to truly getting disease, we have TPR(c)=FPR(c) given any threshold c. The ROC curve for such a test will be just a line with slope=1. On the other hand, a perfect test can perfectly separate diseased subjects from non-diseased. Therefore, the ROC curve can reach the point TPR(c)=1 and FPR(c)=0 at certain value of c (Pepe, 2003). Many parametric, semiparametric, and nonparametric estimation methods have been established to construct the ROC curve and its associated summary indices. Among these methods, the simplest and commonly used one is the empirical estimator, which we will further discuss in the following section.

When using biomarker for disease diagnosis, it is typically plausible that a higher level of test score always indicates a larger probability of the disease status. Then the corresponding ROC curve should be strictly concave, as will be shown in next section. However, the

empirical ROC curves do not respect concavity: see Figure1for illustration. Constructing strictly concave ROC curves has been studied in the machine learning literatures. Provost and Fawcett (2001) addressed this problem with an ROC convex hull algorithm, to obtain the convex hull of the empirical ROC curve. Flach and Wu (2005) further developed a concavity-respecting ROC estimator by approximately combining the Provost and Fawcett's estimator and the empirical ROC curve. However, the statistical properties of these concavity-respecting estimators have not been investigated.

In this study, we developed and evaluated two concavity-respecting estimators of the ROC curve. The empirical ROC is a step function (in the absence of tied biomarker observation between diseased and non-diseased). Each jump involves a top point and a bottom point. The first concavity-respecting ROC curve is the convex hull of these top points, and the second one is the convex hull of the bottom points; these two ROC estimators will be referred to as top concavity-respecting (TCR) estimator and bottom concavity-respecting (BCR) estimator, respectively. The TCR estimator coincides with the Provost-Fawcett's method. We also conducted extensive simulation studies to investigate the statistical properties of these estimators in comparison with the empirical ROC curve.

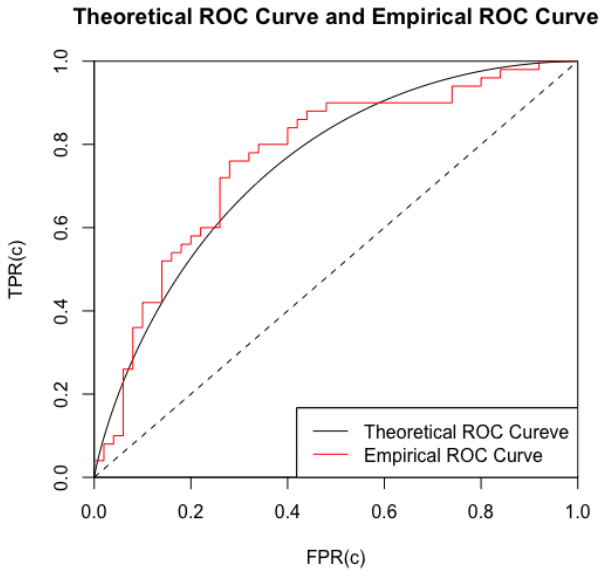


Figure 1: Illustration of an empirical ROC curve, which does not respect concavity.

2.Problem and Notation

Consider a continuous biomarker. Represent it in the diseased group (D=1) and the non-diseased group (D=0) by Y and X, respectively. For a given cut-off point c, the test result is diseased if it is greater than c and non-diseased otherwise. Let G and F be the distribution functions of the random variables Y and X. The sensitivity of the test is given by $Se(c) = 1 - G(c)$, and the specificity is defined as $Sp(c) = F(c)$. A schematic plot is presented as Figure 2 (Gonçalves, 2014).

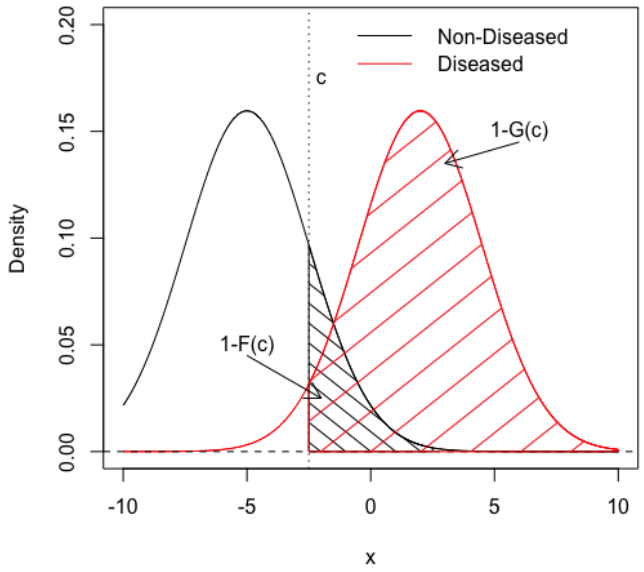


Figure 2: Distribution of the diagnostic test measures

for the diseased and non-diseased populations.

Therefore, the corresponding true and false positive factions at the threshold c are

$$\begin{aligned} TPR(c) &= Pr(Y > c) &= 1 - G(c) \\ FPR(c) &= Pr(X > c) &= 1 - F(c) \end{aligned}$$

The ROC curve is the entire set of possible true and false positive fractions computed by using different thresholds. It is easily to see that both FPR and TPR are a monotone decreasing function of c. Thus, the ROC curve is a monotone increasing curve that lies in the positive quadrant. We can write the ROC curve as:

$$ROC(t) = 1 - G(F^{-1}(1-t)) \ , \ t \in [0,1]$$

with $c = F^{-1}(1-t)$ being the threshold values such that $t = 1 - F(c) = FPR(c)$; see Figure 3.

Thus, the slope of $ROC(t)$ at t is

$$\frac{\partial ROC(t)}{\partial t} = \frac{f_Y(F^{-1}(1-t))}{f_X(F^{-1}(1-t))} = r(c)|_{c=F^{-1}(1-t)},$$

where f_Y, f_X denote the probability density functions of Y (Diseased Group) and X (Non-Diseased Group) respectively. In the biomedical settings that higher level of biomarker indicates larger probability of the disease status, the slope of ROC curve r(c) is a monotone increasing function of c. Since c is a function of $t = Pr(X \geq c)$, the slope of the ROC curve is monotone decreasing as t increases. Therefore, in such a context, the true ROC curve should be concave. It can also be explained by the second order derivative of ROR(t) with respect to t:

$$\frac{\partial ROC(t)^2}{\partial^2 t} = \frac{\partial r(c)}{\partial c}\frac{\partial}{\partial(1-t)}(F^{-1}(1-t))(-1)$$

Obviously the second order derivative is negative, ROC(t) is concave respecting to t.

The empirical estimator of the ROC curve is given by

$$\hat{ROC}(t) = 1 - \hat{G}\,(\hat{F}^{-1}(1-t))$$

where $\hat{F}^{-1}$ and $\hat{G}$ denote the empirical quantile function and the empirical distribution function associated to non-diseased and diseased populations respectively. The empirical distribution function is the percentage of subjects' test scores smaller or equal to t at any given t, and the empirical quantile function is its inverse (Gonçalves, 2014).

To restore concavity, we developed two algorithms to modify the empirical ROC curve.

3.Algorithm

Constructing Empirical ROC Curve

The empirical ROC curve preserves many properties of the empirical distribution function and it's uniformly convergent to the theoretical curve (Hsieh and Turnbull, 1996). Nevertheless, it has some drawbacks and may suffer from large variability, particularly for small sample size. In addition, the empirical ROC curve is not continuous, but a step function.

Input:	List of tuples (score, label), where:
	Score—numeric test results of each observation. Label—the true class of each observation.
Output:	Stepwise constant function
	1. Sort testing set instances increasing by score. 2. Starting from the lowest score, calculate TPR(c) and FPR(c) by setting c equals to chosen score. 3. Repeat step 2, use every score as threshold successively. 4. Plot TPR(c)s vs. FPR(c)s.

Identifying and Restoring Concavity

To restore the concavity, firstly we should locate all the points candidates on the empirical ROC curve to act as bases for the modified ROC curve. Two categories of special points on the empirical curves first come to mind: 1) the points on top of each step, which would be on the boundary of convex hull (marked in red); 2) the points at the bottom of each step (marked in green), as Figure 4 shown below.

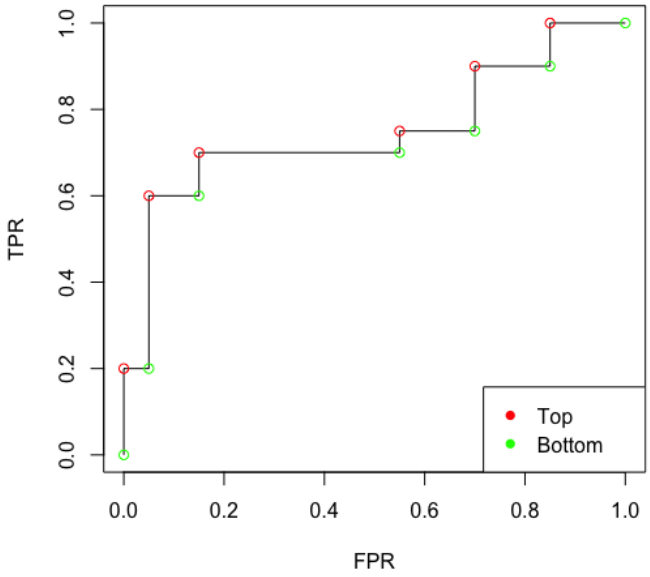


Figure 4: The candidate points of the two modification methods.

TCR: chose from the points having highest TPR level among all the points with same FPR levels, which locate on the top of each step. (Red Circles)

BCR: chose from the points having lowest TPR level among all the points with same FPR levels, which locate at the bottom of each step. (Green Circles)

Input:	A candidate list, containing points at every top/bottom of the steps, whose locations are expressed as ($FPR(c), TPR(c)$)
Output:	A list of points used to construct the modified ROC curve.
	1. A true ROC curve always contains (0,0) and (1,1), so (0,0) is the first point on modified ROC curve.
	2. Define the first point as "prior point" and derive the slopes between the prior point and all the following points in our candidate list.
	3. Among all the slopes, find out the maximum. Adding the corresponding point into modified ROC curve list. Thereafter, let this new point take place the prior point, then repeat step 2.
	4. Keep doing step 2 and 3, until the newest selected point reaches TPR=1 or FPR=1.
	5. Use the points in modified ROC list to draw a ROC curve by connecting them with line segments.

In the process of generating modified ROC curve estimators, we observed that the area under TCR curve always contains a positive bias. Such a phenomenon let us question TCR's asymptotic properties. However, for BCR we did not have this concern. In the

following simulation studies, we expected to observe that the performance of BCR would be better than TCR.

4.Simulation Study

To evaluate the statistical properties of the two modified estimators of ROC curve, we investigated two commonly used summary indexes of ROC curves, the area under ROC curve and specificity at fixed level of sensitivity. Simulation studies have been conducted using different population distributions and sample sizes.

Area Under Curve

Numerical indices for ROC curves are often used to summarize the curves. When it is not feasible to plot the ROC curve itself, such summary measures convey important information about the curve (Pepe, 2003). The area under the ROC curve (AUC) is a single number summary of an ROC curve which can be used to compare the effectiveness of two separate diagnostic tests or procedures. It is easier to compare a single number than to compare both the sensitivities and specificities of the two tests (Zhou,2005). Also it may come out that the ROC curves of two tests are very similar making it hard to detect which is better. Therefore, instead of comparing two ROC curves visually, the AUC for the two ROC curves are compared. As such, the AUC is "the most common quantitative index describing an ROC curve" (Hanley, 1997). It is defined as

$$AUC = \int_0^1 R\,OC(t)dt$$

A perfect test has the AUC value of 1. On the other hand, an uninformative test has AUC=0.5. The AUC is interpreted as "the probability that test results from a randomly

selected pair of diseased and non-diseased subjects are correctly ordered", that is, $Pr(Y > X)$ (Pepe, 2003).

AUC has been used as a summary measure in this study. We calculated the mean squared error (MSE), bias, and variance of AUCs of empirical estimator and the two modified ones, as well as the coverage accuracy of their bootstrap intervals (percentile intervals based on 1000 bootstrapping resamples), using different population distributions and sample sizes. In every simulation studies, we generated 1000 random samples of size m from the distribution for test responses of diseased patients, and another independent random samples of the sample size n from the distribution for test responses of non-diseased patients. We considered (m, n) = (20, 20), (50, 50) and (100, 100) to represent small to moderate sample size settings. In addition, we added (m, n) = (40, 20) as an unequal sample size setting. The population distribution settings in these simulation studies, were chosen to represent normally distributed data with different populations means, and non-normally distributed data. The parameters of the population distributions are determined to represent four level of distribution overlapping, which is to say 4 levels of theoretical AUC values. The settings of the simulation studies were following the publication of Zhou et al (2005). The detailed distribution parameters and the theoretical AUC values for the true ROC curves corresponding to the settings are shown in Table 1.

$$\text{Empirical AUC}(\hat{AUC}_{emp}) = \sum_{j=1}^{m}\sum_{i=1}^{n}\frac{I(Y_j>X_i)+1/2I(Y_j=X_i)}{mn}$$

$$\begin{aligned} Theoretical\,AUC(AUC_0) &= \int_0^1 R\,OC(t)dt \\ &= Pr(Y > X) \end{aligned}$$

$$\widehat{MSE}(\widehat{AUC}) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{AUC}_i - AUC_0)^2$$

$$\widehat{Bias}(\widehat{AUC}) = \frac{1}{n}\sum_{i=1}^{n}\widehat{AUC}_i - AUC_0$$

$$\widehat{Variance}(\widehat{AUC}) = \frac{1}{n-1}\sum_{i=1}^{n}(\widehat{AUC}_i - \overline{\widehat{AUC}})^2$$

Table 1: Distribution Setting for the simulation study on AUCs estimated by 3 methods, and the theoretical AUCs are calculated as below.

Index	Distribution of Diseased Group	Distribution of non-Diseased Group	Theoretical AUC
1	N (2, 1)	N (0, 1)	0.9214
2	Beta (2, 1)	Beta (1, 2)	0.8332
3	N (1, 1)	N (0, 1)	0.7603
4	N (0.5, 1)	N (0, 1)	0.6382

We display MSEs, Bias, Variances and Bootstrap Interval Coverage Probabilities for the empirical AUC and two modified AUCs, TCR for modified estimator based on convex hull, BCR for modified estimator constructing by the points at the bottom of each step, in Table 2 with (m,n)=(20,20), in Table 3 with (m,n)=(50,50), in Table 4 with (m,n)=(100,100), and in Table 5 with (m,n)=(40,20). With this series of comparison, we expected to see the advantages of our modified estimators under certain situations. From the results in Table 2-5, we find that for small sample size ((m,n)=(20,20)), TCR has the lowest MSE and variance than the empirical estimator when the theoretical AUC is close to 1, which means the diseased population distribution does not overlap with the non-

diseased population distribution very much. Generally, BCR has the best performance among all 3 estimators, especially when the theoretical AUC is between 0.7 to 0.9, which means the overlap between population distributions of the diseased and non-diseased groups are comparatively large. When the sample size increases, estimator BCR outperformed the other two in most situations, since it has a better balance of bias and variance, and the bias/standard deviation ratio converge to zero very quickly. However, though the bias of TCR does not shrink as much as the other two, it still works well when sample sizes are small, and most important, it always has the lowest variance.

Table 2: MSE, Bias, Variance, and Bootstrap Quantile Interval coverage probability of the AUCs estimated by 3 methods. Sample size (m, n) = (20, 20)

Theoretical AUC	Estimator	MSE	Bias	Variance	Coverage Probability
0.9214	Emp	0.00175	0.00046	0.00175	0.898
	TCR	0.00162	0.02362	0.00106	0.726
	BCR	0.00585	-0.05893	0.00238	0.711
0.8332	Emp	0.00249	0.00137	0.00249	0.919
	TCR	0.00307	0.03505	0.00184	0.786
	BCR	0.00279	-0.02242	0.00230	0.906
0.7603	Emp	0.00574	-0.00248	0.00574	0.950
	TCR	0.00664	0.05152	0.004	0.745
	BCR	0.00547	-0.03297	0.00439	0.927
0.6382	Emp	0.00791	0.00065	0.00791	0.937
	TCR	0.01087	0.0744	0.01758	0.676
	BCR	0.02032	-0.05238	0.0214	0.953

Table 3: MSE, Bias, Variance, and Bootstrap Quantile Interval coverage probability of the AUCs estimated by 3 methods. Sample size (m, n) = (50, 50)

Theoretical AUC	Estimator	MSE	Bias	Variance	Coverage Probability
0.9214	Emp	0.00069	-0.00015	0.00069	0.919
	TCR	0.00076	0.01571	0.00052	0.746
	BCR	0.00087	-0.01651	0.0006	0.947
0.8332	Emp	0.00152	0.00116	0.00152	0.933
	TCR	0.00197	0.02722	0.00123	0.771
	BCR	0.00134	-0.008	0.00128	0.949
0.7603	Emp	0.00219	0.00123	0.00219	0.931
	TCR	0.003	0.03463	0.0018	0.747
	BCR	0.00189	-0.00132	0.00189	0.953
0.6382	Emp	0.003	0.00055	0.00305	0.933
	TCR	0.00461	0.04739	0.00238	0.749
	BCR	0.00494	-0.00393	0.00493	0.955

Table 4: MSE, Bias, Variance, and Bootstrap Quantile Interval coverage probability of the AUCs estimated by 3 methods. Sample size (m, n) = (100, 100)

Theoretical AUC	Estimator	MSE	Bias	Variance	Coverage Probability
0.9214	Emp	0.00034	-0.00018	0.00034	0.932
	TCR	0.0004	0.01105	0.00028	0.788
	BCR	0.00033	-0.00564	0.0003	0.932
0.8332	Emp	0.00083	0.0016	0.00083	0.940
	TCR	0.0011	0.0193	0.00073	0.789
	BCR	0.00074	0.00119	0.00074	0.949
0.7603	Emp	0.0011	0.00178	0.0011	0.943
	TCR	0.00158	0.0245	0.00098	0.765
	BCR	0.00102	0.0058	0.00099	0.942
0.6382	Emp	0.00148	0.00006	0.00148	0.932
	TCR	0.00227	0.0319	0.00125	0.756
	BCR	0.0018	-0.0074	0.00175	0.930

Table 5: MSE, Bias, Variance, and Bootstrap Quantile Interval coverage probability of the AUCs estimated by 3 methods. Sample size (m, n) = (40, 20)

Theoretical AUC	Estimator	MSE	Bias	Variance	Coverage Probability
0.9214	Emp	0.00134	-0.00038	0.00134	0.891
	TCR	0.00129	0.02045	0.00087	0.718
	BCR	0.00284	-0.03932	0.00129	0.677
0.8332	Emp	0.00295	0.00156	0.00295	0.917
	TCR	0.0035	0.03726	0.00211	0.768
	BCR	0.00309	-0.02525	0.00245	0.906
0.7603	Emp	0.00452	0.00403	0.00451	0.930
	TCR	0.00582	0.04985	0.00334	0.751
	BCR	0.00374	-0.01452	0.00354	0.932
0.6382	Emp	0.00016	0.00093	0.00016	0.924
	TCR	0.00012	0.00675	0.00008	0.659
	BCR	0.01814	-0.08191	0.01144	0.943

Specificity at Fixed Level of Sensitivity

In clinical research, the fundamental measure of diagnostic accuracy is sensitivity (i.e. True Positive Rate) and specificity (i.e. True Negative Rate). For the continuous scaled diagnostic tests, there is an inherent trade-off between sensitivity and specificity which can be demonstrated by varying the cut-off points. In practice, the cut-off point is usually chosen to achieve a fixed level of sensitivity or specificity. The motivation to do so is that, depending on clinical context physicians may desire for maintaining high sensitivity or specificity. For example, in aggressive prostate cancer research, the cost of false negative diagnosis is much higher than false positive diagnosis. In that case, we would fix the test sensitivity level at 95 per cent, and then look at the corresponding specificity level. Hence it is sometimes of interest to only focus on a small part of the entire ROC curve where the diagnostic test is intended to operate in practice. Particularly, we are interested in the specificities at a fixed high sensitivity, for example, 95 percent.

In this part of the study, we fixed the sensitivity at 95 per cent, and evaluate the corresponding estimated specificity values based on empirical ROC curves and modified ROC curves, under population distribution and sample size settings similar to the previous section. Table 6 lists the detailed settings and theoretical specificities corresponding to 95 per cent level of sensitivities.

Table 6: Distribution Setting for the simulation study on Specificity given Sensitivity fixed at 0.95.

Index	Distribution of Diseased Group	Distribution of non-Diseased Group	Specificity (Sensitivity=0.95)
1	N (3, 1)	N (0, 1)	0.9123
2	N (2, 1)	N (0, 1)	0.6388
3	Beta (2, 1)	Beta (1, 2)	0.3972
4	N (1, 1)	N (0, 1)	0.2595

The empirical ROC curve is a step function. Therefore, multiple specificity values correspond to the same sensitivity. We took the minimum specificity within identical sensitivity levels. In Tables 7,8,9, and 10, we display the MSEs, Bias, Variances and Bootstrap Interval Coverage Probabilities for the empirical Specificities and two modified Specificities when Sensitivities fixed at 95 percent. The properties of TCR and BCR are similar to what we observed from comparing AUCs. As we can see, when sample size is relatively large, the BCR ROC curve would be very close to the empirical ROC curve. Therefore, the point estimates of Specificity at 95 per cent level of Sensitivity generated from empirical ROC curve and BCR ROC curve are similar to each other. And generally, the Bootstrap coverage probability of BCR estimator is better than the empirical estimator.

Table 7: MSE, Bias, Variance, and Bootstrap Quantile Interval coverage probability of Specificity estimated by 3 methods. Sample size (m, n) = (20, 20)

Theoretical Specificity	Estimator	MSE	Bias	Variance	Coverage Probability
0.9123	Emp	0.00645	0.0137	0.00627	0.921
	TCR	0.00516	0.02518	0.00453	0.775
	BCR	0.2352	-0.37595	0.09395	0.648
0.6388	Emp	0.03178	0.07109	0.02676	0.933
	TCR	0.02978	0.09747	0.0203	0.770
	BCR	0.0656	-0.1689	0.03711	0.912
0.3972	Emp	0.03702	0.08875	0.02917	0.901
	TCR	0.03874	0.12501	0.02314	0.727
	BCR	0.03187	-0.09482	0.0229	0.967
0.2595	Emp	0.03741	0.0949	0.02842	0.916
	TCR	0.03952	0.13481	0.02136	0.689
	BCR	0.02451	-0.03895	0.02302	0.995

Table 8: MSE, Bias, Variance, and Bootstrap Quantile Interval coverage probability of Specificity estimated by 3 methods. Sample size (m, n) = (50, 50)

Theoretical Specificity	Estimator	MSE	Bias	Variance	Coverage Probability
0.9123	Emp	0.00384	-0.0027	0.00383	0.944
	TCR	0.00255	0.01962	0.00217	0.826
	BCR	0.01766	-0.06648	0.01325	0.897
0.6388	Emp	0.01547	0.00172	0.01548	0.940
	TCR	0.01353	0.05588	0.01042	0.811
	BCR	0.01538	-0.04114	0.0137	0.955
0.3972	Emp	0.01501	0.00946	0.01494	0.949
	TCR	0.01612	0.07125	0.01105	0.773
	BCR	0.01165	-0.01787	0.01135	0.967
0.2595	Emp	0.01181	0.01056	0.01171	0.958
	TCR	0.0143	0.07356	0.0089	0.748
	BCR	0.00794	-0.00766	0.00789	0.988

Table 9: MSE, Bias, Variance, and Bootstrap Quantile Interval coverage probability of Specificity estimated by 3 methods. Sample size (m, n) = (100, 100)

Theoretical Specificity	Estimator	MSE	Bias	Variance	Coverage Probability
0.9123	Emp	0.00196	0.00075	0.00196	0.949
	TCR	0.00155	0.01035	0.00144	0.890
	BCR	0.0023	-0.01761	0.00199	0.966
0.6388	Emp	0.00753	0.0157	0.00729	0.924
	TCR	0.00711	0.0373	0.00572	0.837
	BCR	0.00665	-0.00757	0.0066	0.962
0.3972	Emp	0.00807	0.02142	0.00761	0.933
	TCR	0.00834	0.04928	0.00591	0.794
	BCR	0.00623	0.00343	0.00622	0.950
0.2595	Emp	0.00707	0.0251	0.00645	0.932
	TCR	0.00784	0.05329	0.00501	0.781
	BCR	0.00496	0.01194	0.00482	0.954

Table 10: MSE, Bias, Variance, and Bootstrap Quantile Interval coverage probability of Specificity estimated by 3 methods. Sample size (m, n) = (40, 20)

Theoretical Specificity	Estimator	MSE	Bias	Variance	Coverage Probability
0.9123	Emp	0.00578	0.0116	0.00565	0.910
	TCR	0.00442	0.02682	0.0037	0.838
	BCR	0.1008	-0.19753	0.06188	0.722
0.6388	Emp	0.02293	0.04434	0.02098	0.914
	TCR	0.02148	0.07705	0.01556	0.788
	BCR	0.03065	-0.08137	0.02405	0.872
0.3972	Emp	0.02468	0.05445	0.02174	0.954
	TCR	0.02577	0.096	0.01657	0.818
	BCR	0.02041	-0.04704	0.01822	0.956
0.2595	Emp	0.02074	0.04175	0.01901	0.927
	TCR	0.02238	0.08546	0.01509	0.745
	BCR	0.02018	-0.03092	0.01925	0.956

The research of Hsieh and Turnbull showed that the empirical ROC curve converges to the sum of 2 independent Brownian bridges. And the asymptotic normality of summary measures of the empirical ROC curve, such as AUC and specificity at fixed level of sensitivity, can be derived from their work (Hsieh and Turnbull, 1996). In this study, we plotted TCR and BCR versus standard normal distribution and modified AUCs versus empirical AUC.

In Figure 5, 6, 7, and 8, the red line is a 45° reference line, which implies that the modified estimator asymptotically equals to the empirical estimator if the points fall along the reference line. As we can see, the difference between empirical estimates and BCR estimates decreased faster than TCR estimates with sample size went to large, even though TCR seems to be more robust when sample size is relatively small. This finding agrees with what we observed from the previous simulations. Referring the trend shown in these scatter plots, we speculate that BCR is asymptotically equivalent to the empirical estimator, and it is possible that TCR is also asymptotically equivalent to the empirical estimator. Figure 9 and 10 are the Q-Q plots of AUCs of the 2 modified ROC curves versus standard normal quantiles, under different sample sizes. These plots show that both TCR AUC and BCR AUC are approximately normal.

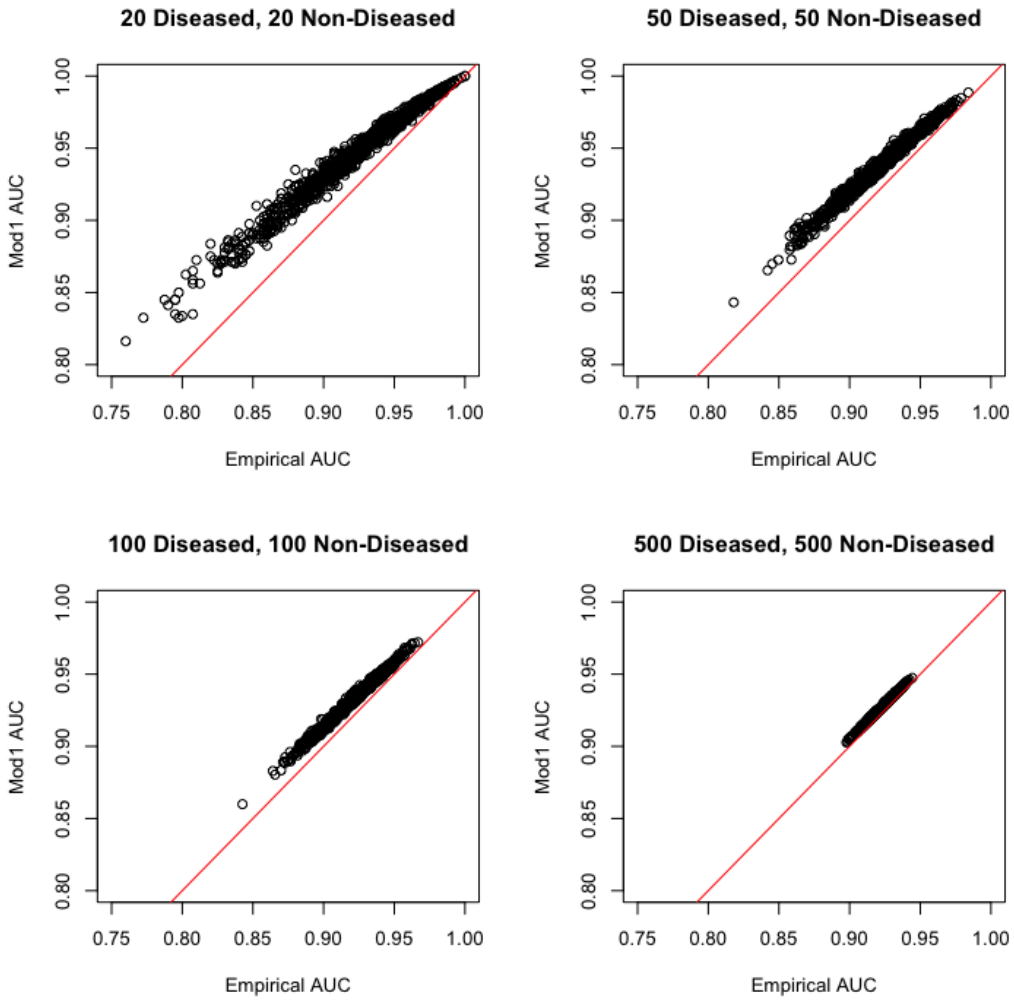


Figure 5: Diseased Group- N(2,1); Non-Diseased Group- N(0,1)

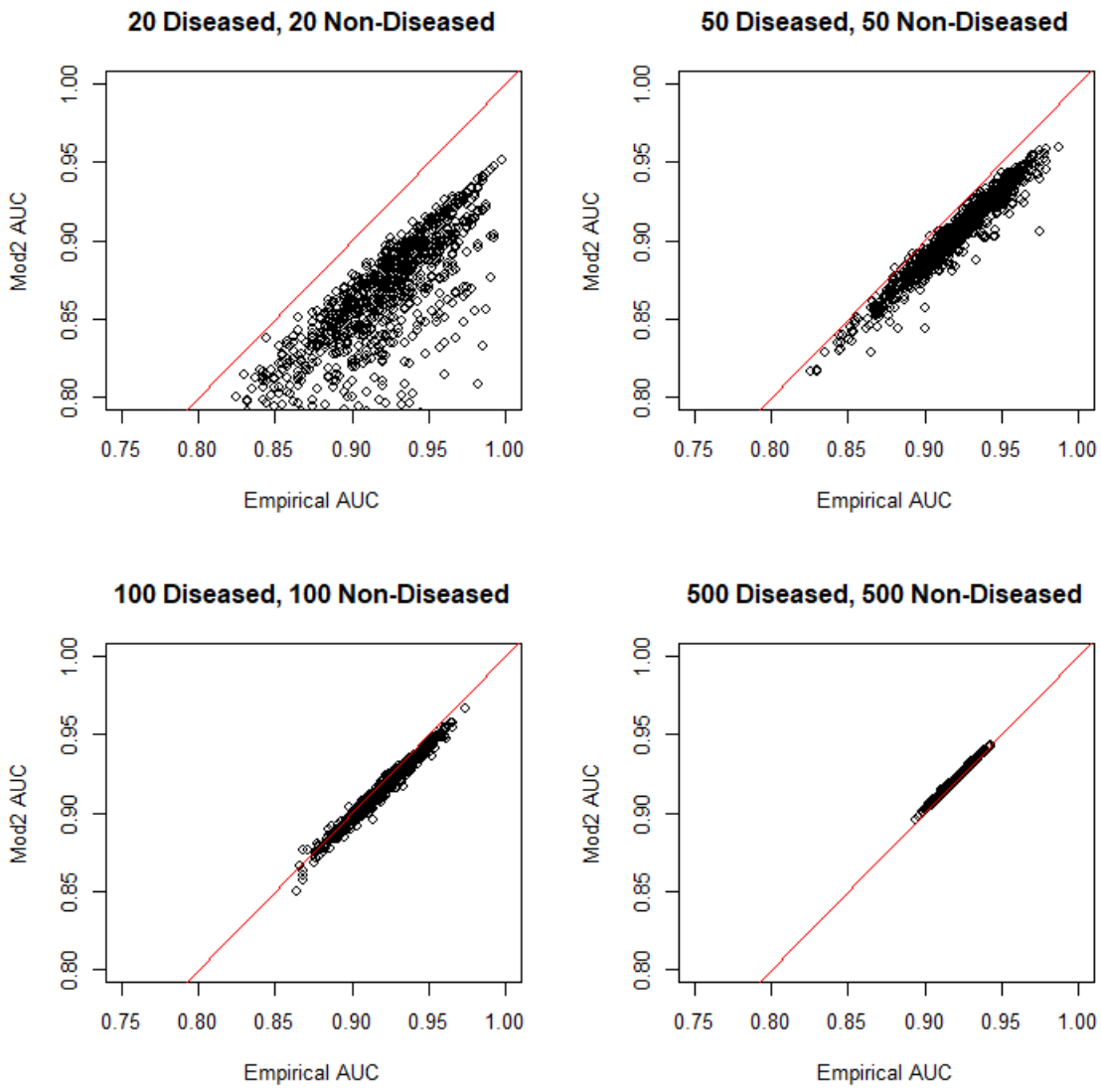


Figure 6: Diseased Group - N(2,1); Non-Diseased Group - N(0,1)

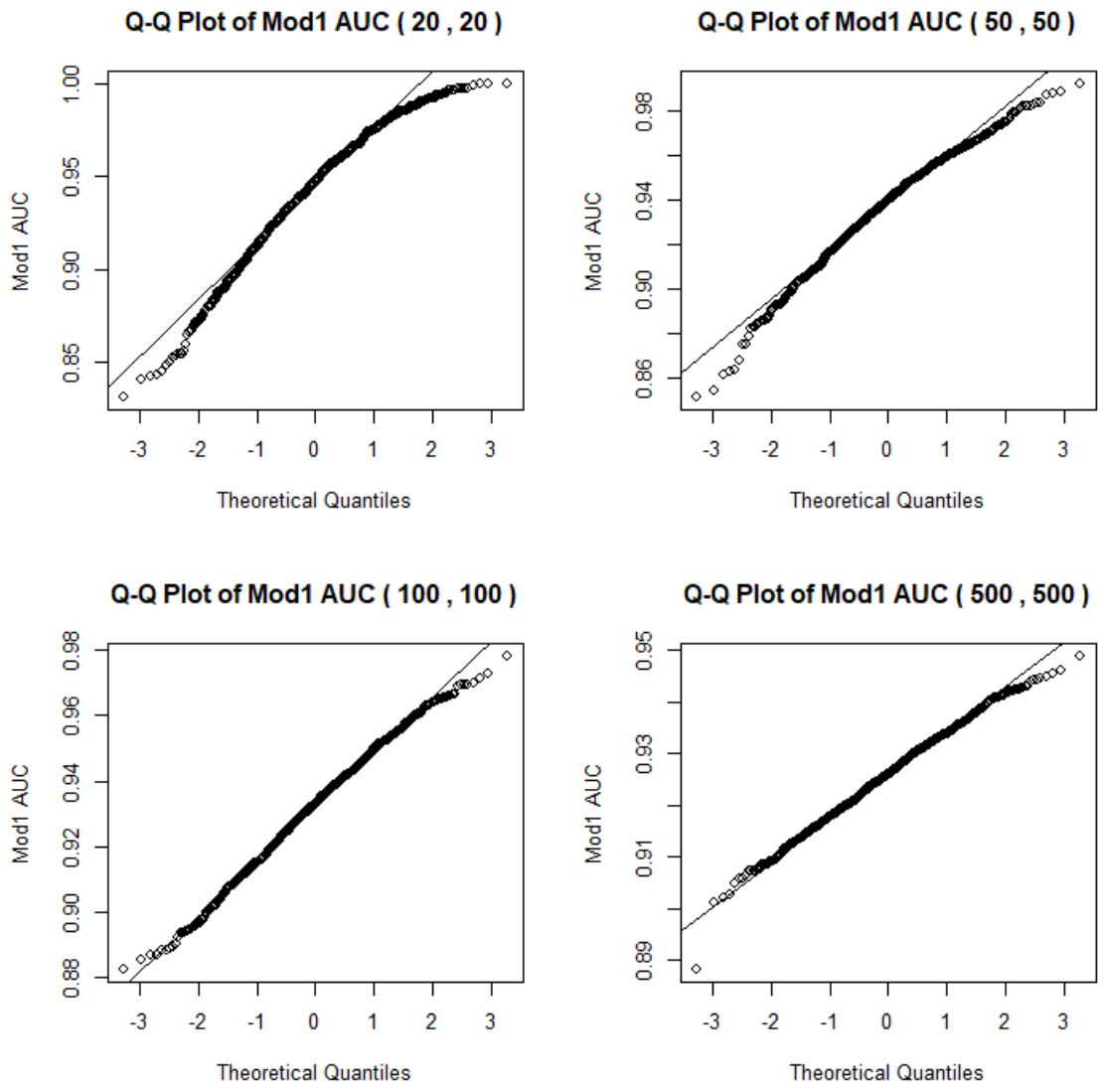


Figure 9: Diseased Group – N(2,1); Non-Diseased Group – N(0,1)

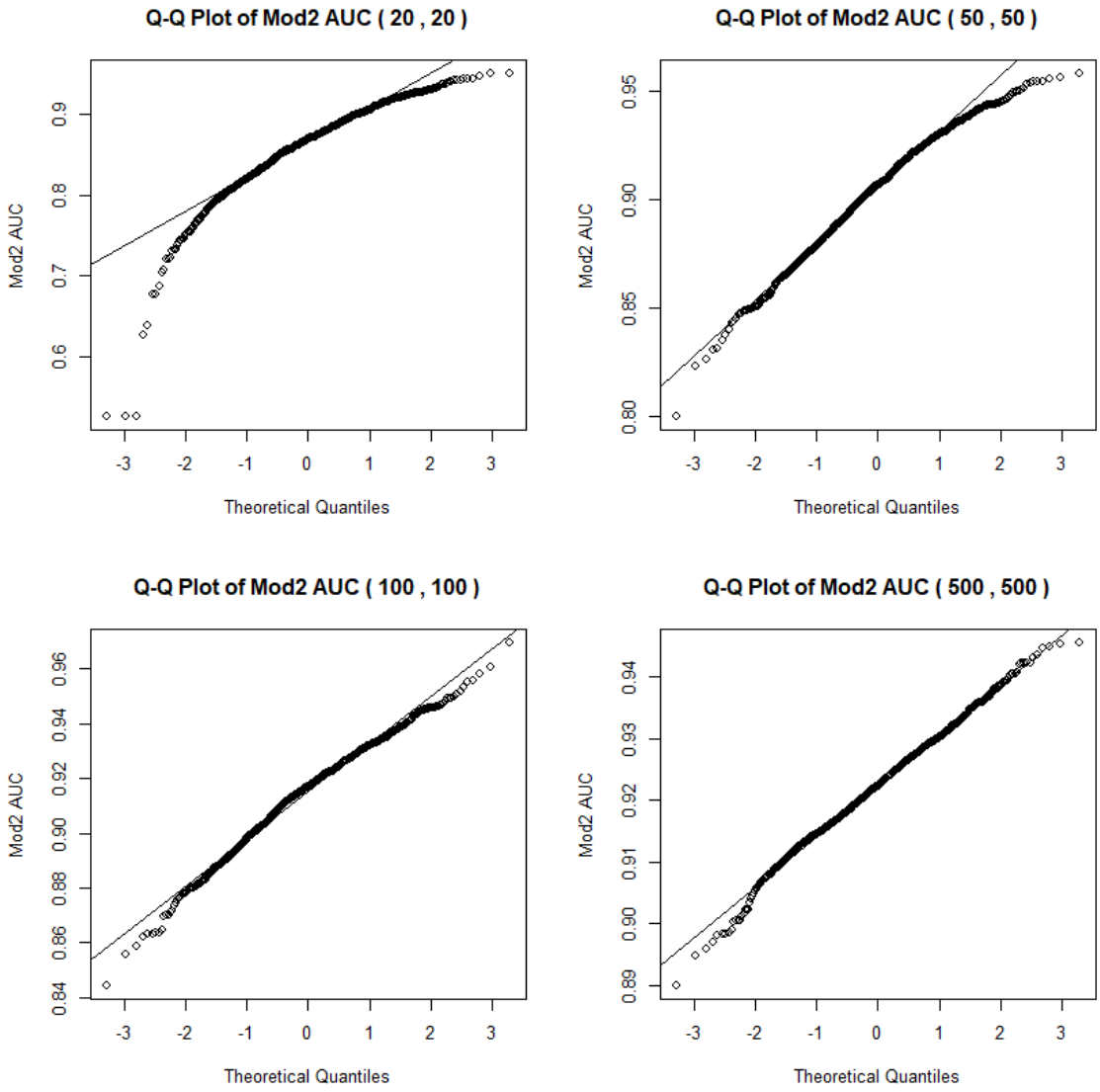


Figure 10: Diseased Group – N(2,1); Non-Diseased Group – N(0,1)

5.Real Data Application

We illustrate the application of these two modified ROC curves in a real data study. The following data come from the ovarian cancer gene expression array study (Pepe, 2003). Ovarian cancer claims more lives than any other gynecological cancer in this century. It is the fifth most common cancer in American women and the fifth most common cause of cancer death. Though the 5-year survival for women diagnosed in early-stage is 90%, but the majority of patients are diagnosed with late-stage disease and have a 5-year survival of less than 30% (Lu, 2004). Furthermore, a considerable increase in the risk of catching ovarian cancer is observed in patients with a family history. Most cases of familial ovarian cancer are based on mutations in the *BRCA1* and *BRCA2* genes (Lux, 2006). Therefore, the early stage diagnosis of ovarian cancer has been very important and gene expression array data became critical indicator of early stage detection. The relative gene expression intensities of a particular gene are displayed below for 23 non-diseased ovarian tissues and 30 ovarian tumor tissues (Pepe, 2003).

Normal Tissues:	0.442, 0.500, 0.510, 0.568, 0.571, 0.574, 0.588, 0.595,0.595, 0.595, 0.598, 0.606, 0.617, 0.628, 0.641, 0.641,0.680, 0.699, 0.746, 0.793, 0.884, 1.149, 1.785
Cancer Tissues:	0.543, 0.571, 0.602, 0.609, 0.628, 0.641, 0.666, 0.694,0.769, 0.800, 0.800, 0.847, 0.877, 0.892, 0.925, 0.943,1.041, 1.075, 1.086, 1.123, 1.136, 1.190, 1.234, 1.315,1.428, 1.562, 1.612, 1.666, 1.666, 2.127

The empirical and modified ROC curves have been plotted using algorithm described before. (Figure 11) The sample size of diseased group is 30 and the sample size of non-diseased group is 23. According to the simulation study results, it suggested that BCR may

have as good as or even better performance compared to empirical estimator under given observations. The point estimates and Bootstrap percentile confidence interval of estimated AUC and Specificity at 95 percent level of Sensitivity using the three methods are shown in Table 11.

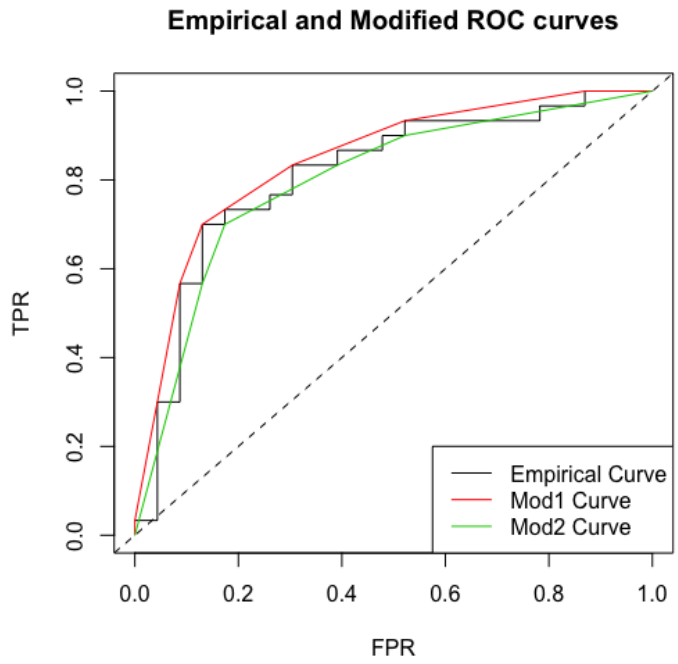


Figure 11: Empirical and Modified ROC Curves of Ovarian Cancer Gene Expression Data

Table 11 Point estimates and Bootstrap percentile confidence interval using Ovarian Cancer gene expression data

	Estimated AUC	Bootstrap Interval		Estimated Specificity	Bootstrap Interval	
		Lower	Upper		Lower	Upper
Empirical	0.8121	0.6884	0.9275	0.3295	0.0434	0.7391
TCR	0.8619	0.7615	0.9514	0.4598	0.2064	0.7826
BCR	0.7761	0.6580	0.8783	0.2616	0.0978	0.6089

6.Discussion

It is often the case that higher level of certain biomarker value indicates larger probability of a disease. Under this situation, the corresponding ROC curve should be always concave. In this study, we developed and evaluated two modification methods on empirical estimator of ROC curve, TCR and BCR, in order to restore concavity of the empirical ROC curve. As shown in the simulation studies, for bi-normal population distributions, when the difference between 2 normal means are relatively small, BCR estimator has better performance over TCR and empirical estimator. The MSE and variance of BCR AUCs are smaller and have fairly good bootstrap interval coverage. In clinical practice, a lot of biomarkers do not have perfect classification accuracy, and the empirical AUC would be far from 1. The BCR estimator has a promising potential in these situations. Also, referring to the Specificity at fixed level of Sensitivity, BCR's Bootstrap percentile confidence interval generally has better coverage probability even when the mean difference is big, as long as the sample size is not too small. Sometimes the entire area under the curve may not be very informative, so people would use partial AUC instead. The specificity at fixed level of sensitivity (or sensitivity at fixed level of specificity) can be viewed as the ultimate partial AUC, so our modified estimator may also provide better estimates for partial AUC.

One existing problem of our method is that, the bootstrap interval coverage probabilities of TCR were not as good as expected. We observed that the bias of TCR relative to its standard deviation (SD) was larger than the other two estimators. Nevertheless, the bias/SD ratio still decreased as sample size increased. That suggests that this might be a finite sample issue, rather than a large-sample one. Further investigation is warranted.

Reference

Søreide, K. (2009). Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research. *Journal of clinical pathology*, *62*(1), 1-5.

Pepe, M. S. (2003). *The statistical evaluation of medical tests for classification and prediction*. Medicine.

Choi, B. C. (1998). Slopes of a receiver operating characteristic curve and likelihood ratios for a diagnostic test. *American Journal of Epidemiology*, *148*(11), 1127-1132.

Hsieh, F., & Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The annals of statistics*, 25-40.

Provost, F., & Fawcett, T. (2001). Robust classification for imprecise environments. *Machine learning*, *42*(3), 203-231.

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation. *Caspian journal of internal medicine*, *4*(2), 627.

Gonçalves, L., Subtil, A., Oliveira, M. R., & Bermudez, P. Z. (2014). ROC curve estimation: An overview. *REVSTAT–Statistical Journal*, *12*(1), 1-20.

Zhou, X. H., & Qin, G. (2005). Improved confidence intervals for the sensitivity at a fixed level of specificity of a continuous-scale diagnostic test. *Statistics in medicine*, *24*(3), 465-477.

Qin, G., & Zhou, X. H. (2006). Empirical likelihood inference for the area under the ROC curve. *Biometrics*, *62*(2), 613-622.

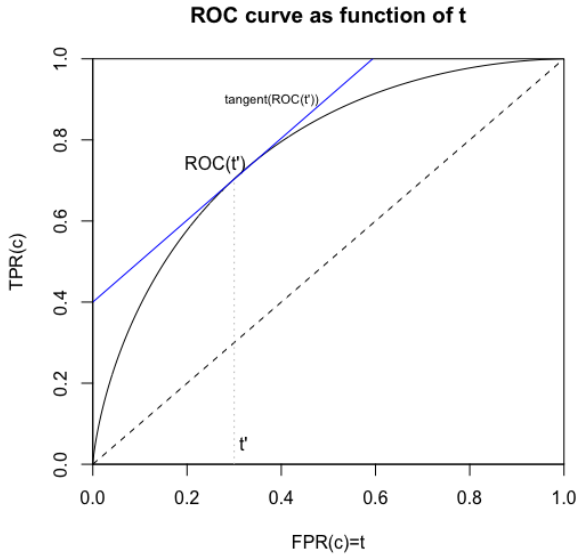
Flach, P. A., & Wu, S. (2005, July). Repairing Concavities in ROC Curves. In *IJCAI* (pp. 702-707).

Hanley, J. A., & Hajian-Tilaki, K. O. (1997). Sampling variability of nonparametric estimates of the areas under receiver operating characteristic curves: an update. *Academic radiology*, *4*(1), 49-58.

Lu, K. H., Patterson, A. P., Wang, L., Marquez, R. T., Atkinson, E. N., Baggerly, K. A., ... & Smith, D. (2004). Selection of potential markers for epithelial ovarian cancer with gene expression arrays and recursive descent partition analysis. *Clinical cancer research*, *10*(10), 3291-3300.

Lux, M. P., Fasching, P. A., & Beckmann, M. W. (2006). Hereditary breast and ovarian cancer: review and future perspectives. *Journal of molecular medicine*, *84*(1), 16-28.

Figures



Picture 3: How ROC Curves formed

and the tangent of a point on curve

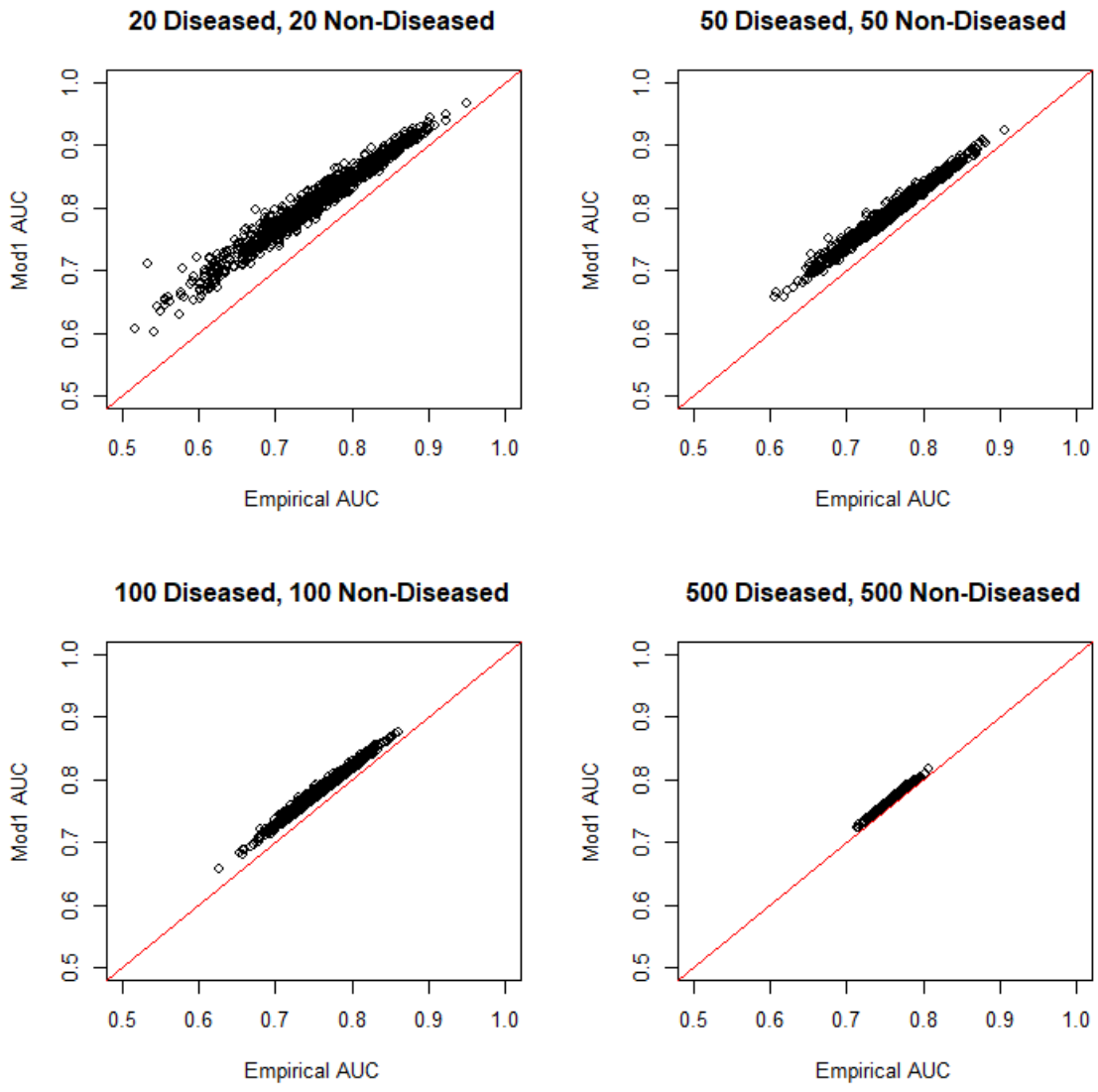


Figure 7: Diseased Group - N(1,1); Non-Diseased Group - N(0,1)

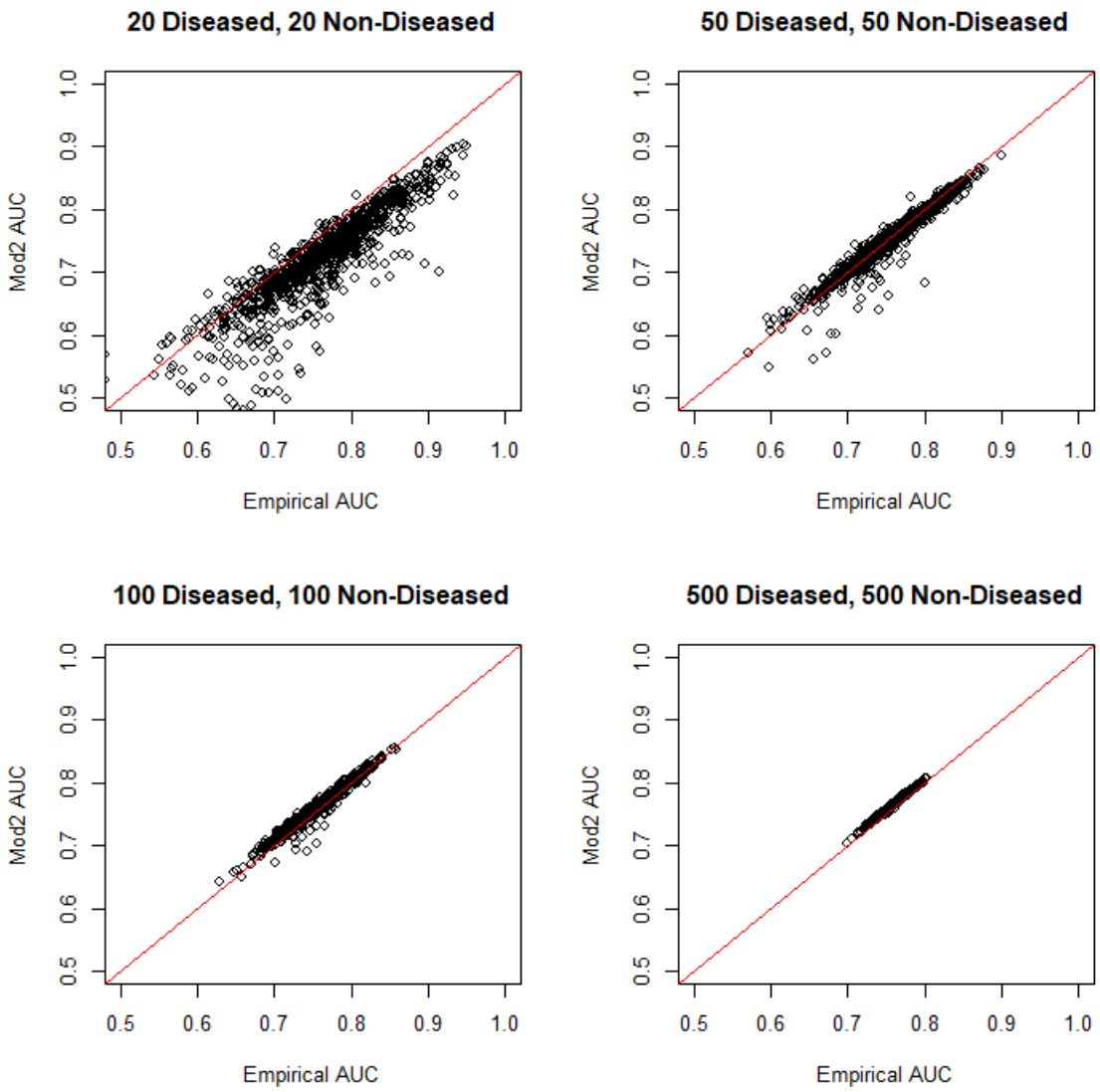


Figure 8: Diseased Group - N(1,1); Non-Diseased Group - N(0,1)