Misuse of DeLong test to compare AUCs for nested models

Olga V. Demler[a,*,†], **Michael J. Pencina**[a], and **Ralph B. D'Agostino Sr**[b]

[a]Department of Biostatistics, Boston University School of Public Health Crosstown, 801 Massachusetts Ave Boston, MA 02118, U.S.A.

[b]Department of Mathematics and Statistics, Boston University, 111 Cummington Street, Boston, MA 02215, U.S.A.

Abstract

The area under the receiver operating characteristics curve (AUC of ROC) is a widely used measure of discrimination in risk prediction models. Routinely, the Mann–Whitney statistics is used as an estimator of AUC, while the change in AUC is tested by the DeLong test. However, very often, in settings where the model is developed and tested on the same dataset, the added predictor is statistically significantly associated with the outcome but fails to produce a significant improvement in the AUC. No conclusive resolution exists to explain this finding. In this paper, we will show that the reason lies in the inappropriate application of the DeLong test in the setting of nested models. Using numerical simulations and a theoretical argument based on generalized *U*-statistics, we show that if the added predictor is not statistically significantly associated with the outcome, the null distribution is non-normal, contrary to the assumption of DeLong test. Our simulations of different scenarios show that the loss of power because of such a misuse of the DeLong test leads to a conservative test for small and moderate effect sizes. This problem does not exist in cases of predictors that are associated with the outcome and for non-nested models. We suggest that for nested models, only the test of association be performed for the new predictors, and if the result is significant, change in AUC be estimated with an appropriate confidence interval, which can be based on the DeLong approach.

Keywords

AUC; DeLong test; logistic regression; *U*-statistics; discrimination; risk prediction

1. Introduction

Risk assessment based on statistical models is a useful tool in guiding clinicians towards reaching optimal treatment decisions. Such models have been developed in the diagnostic (disease has already occurred) or predictive setting (disease is yet to occur). Regression techniques with binary or survival outcomes are the most popular techniques employed for constructing such models [1, 2]. These modeling techniques allow researchers to combine medical test results into a single composite risk score [3]. Risk scores are usually defined as the model-based predicted probabilities of event and are a function of a linear combination of predictors estimated from these models. Higher risk scores correspond to higher

*Correspondence to: Olga V. Demler, Biostatistics , Boston University School of Public Health Crosstown, 801 Massachusetts Ave Boston, MA 02118, U.S.A..
†demler@bu.edu

probabilities of developing or having the adverse medical condition. When the composite risk score exceeds a certain threshold value, people can be assigned to the high risk group.

Appropriate methods for assessing the performance of these models are of fundamental importance. Generally, discrimination and calibration are of primary concern. Discrimination measures the model's ability to distinguish between subjects who have or will develop the disease of interest ('events') and those who did or will not ('nonevents'); calibration measures how close the model-based predictions agree with reality. Although several different tools have been suggested to evaluate quality of discrimination, the area under the receiver operating characteristic curve (AUC of ROC) remains the most widely used metric. It is defined as the probability that the risk score of a randomly picked nonevent is less than the risk score of an event. Models with perfect discrimination have AUC of 1.0 and ones with no discriminatory ability have AUC equal 0.5.

Ideally, a model developed on one sample is validated in a different sample from the same population. However, frequently the second sample is not available and researchers either split their original sample into training and validation or develop and validate their model on the same sample (referred to as 'direct resubstitution' in [4]). Because significance testing of the regression coefficient of the added predictor(s) occurs in the latter setting of 'direct resubstitution' and this is the approach of choice in the majority of practical applications, we will focus on this case in the remainder of this paper.

During the last decade a lot of biomedical research has been focused on identifying new phenotypic or genetic risk factors that may improve the performance of risk assessment models. For example, in the field of primary prevention of cardiovascular disease (CVD) many new biomarkers [5], measures of subclinical disease [6] and genetic risk factors [7] have been postulated as potential candidates to improve model performance beyond what is offered by standard risk factors (age, blood pressure, cholesterol levels, smoking, diabetes) [8]. In this context, the question of improvement in model performance has been frequently understood as the improvement in model discrimination, and hence, the increase in AUC has been considered as a method to quantify and test this improvement. A widely used test to compare the difference between two AUCs relies on the method developed in a seminal paper by DeLong *et al.* [9] (henceforth 'the DeLong test'). It provides a confidence interval and standard error of the difference between two (or more) correlated AUCs. This procedure has been frequently applied to test the incremental gain in model discrimination and is available as the default option in the logistic procedure in SAS 9.2, SAS Institute Inc., Cary, NC, USA [10].

It is important to observe that adding new risk factors to a model that contains 'standard' predictors results in two 'nested' models. Hence, the question of improvement in discrimination has often been addressed by comparing two nested AUCs [11, 12]. Yet quite frequently researchers observe that the statistical significance of the new risk factor did not translate into a statistically significant increase in the AUC. This repeatedly observed finding [7, 11, 13] has led to criticism of the increase in the AUC as the main measure of improvement in model performance [14] and raises the question of whether we understand the mechanism of discrimination correctly. It has been shown in [15] that this finding is not possible in the context of multivariate normal data, because in this situation, statistical significance of the new risk factor is equivalent to a statistically significant increase in the AUC.

In this paper we use numerical simulation and the result of [15] to show that the DeLong test should not be applied in a nested model using the same data that was used to develop the model and estimate the parameters (direct substitution). We present an explanation for a

general case of any distribution function and for a general set of models including logistic regression and LDA. We also show that our finding might be the reason behind numerous reports where statistical significance of a variable does not lead to a statistical significance of the AUC difference. The paper is organized as follows: in Section 2, we define the estimators of the AUC, introduce the DeLong test, and observe that for multivariate normal data analyzed using the LDA, two nested AUCs can be compared using an F-test, which is exact in this case. We simulate a large multivariate normal data set under the null hypothesis and show a clear discrepancy between p-values obtained using the F-test versus the DeLong test. Using numerical simulations under various scenarios, we further show that the DeLong test applied to nested models leads to a substantial loss of power for low to moderate conditional effect sizes. In Section 3, we provide a theoretical argument that can explain this discrepancy and show why it is likely to be present for any distribution of the data. In Section 4, we discuss possible solutions to this problem. We suggest that improvement in the AUC should only be quantified for variables that are statistically significantly associated with the outcome and hence argue against testing the null hypothesis of no difference for nested AUCs. Once the statistical significance of the added variable(s) has been established, the increase in AUCs can be reported as a confidence interval that can be constructed using the DeLong-based approach.

2. Motivation

2.1. General framework: notation and definitions

Let D be the outcome of interest, with $D = 1$ for events and $D = 0$ for nonevents. Our goal is to predict the event status using p test results, which we denote as $\mathbf{x} = x_1, \ldots, x_p$. Assume D and vector of $\mathbf{x}$ are available for N patients. The prediction based on the full set of p test results is to be compared with that based on a reduced number of tests, $p - k$. One of the risk-prediction models is employed to calculate the risk score, for example, with logistic regression or LDA. It produces linear coefficients estimates $\mathbf{a}' = (a_1, \ldots, a_p)$ for the full model and $\mathbf{a}'_R = (a_1^R, \ldots, a_{p-k}^R)$ for the reduced model. Corresponding risk scores are calculated as $\mathbf{a}'\mathbf{x}$ and $\mathbf{a}'_R\mathbf{x}_{p-k}$. We want to test whether the risk prediction model with p predictors discriminates between the two subgroups better than the model with only the first $p - k$ predictors. Using the AUC as a measure of model performance, we formulate the following hypothesis:

$$
\begin{aligned}
&H_0^{AUC} : AUC_p = AUC_{p-k} \\
&H_a^{AUC} : AUC_p \neq AUC_{p-k}
\end{aligned} , \quad (1)
$$

where AUC_p and the AUC_{p-k} are the AUCs of the full and reduced model, respectively. In general, there is no explicit formula for the AUC. Therefore, the AUC is frequently estimated by the Mann–Whitney statistic [3] — a nonparametric unbiased estimator, often referred to as the c-statistic [1]. To distinguish it from the parametric AUC estimator available in multivariate normal case, we denote it as eAUC (e stands for empirical). It is given as:

$$
eAUC = \frac{1}{n_0 n_1} \sum_{x_i \in D_0} \sum_{x_j \in D_1} I[\mathbf{a}'\mathbf{x}_i, \mathbf{a}'\mathbf{x}_j], \quad (2)
$$

where D_1 and D_0 are the sets of subjects with and without events, respectively, n_1 and n_0 are the sizes of these sets and I[.] is the indicator function adjusted for ties

$$I[\mathbf{a}'\mathbf{x}_i, \mathbf{a}'\mathbf{x}_j] = \begin{cases} 1, & \text{if } \mathbf{a}'\mathbf{x}_i < \mathbf{a}'\mathbf{x}_j \\ 0.5, & \text{if } \mathbf{a}'\mathbf{x}_i = \mathbf{a}'\mathbf{x}_j \\ 0, & \text{otherwise.} \end{cases}$$

In general there is no exact test of (1), so an asymptotic test developed by DeLong *et al.* [9] is usually applied to compare the two eAUCs. This test calculates the statistic

$$z = \frac{(\text{eAUC}_p - \text{eAUC}_{p-k} - (\text{AUC}_p - \text{AUC}_{p-k}))}{((1, -1)S\,(1, -1)')^{1/2}},$$

where S is the estimated variance–covariance matrix of $(\text{eAUC}_p, \text{eAUC}_{p-k})$ and can be found in [9]. Under the null hypothesis, the test statistic z has a standard normal distribution. The p-value is calculated as $2(1 - \Phi(|z|))$, where $\Phi(.)$ is the standard normal cumulative distribution function.

2.2. Testing under assumption of normality

If data are normally distributed in the event and nonevents subgroups, there exists a parametric estimator of the AUC (we denote it as pAUC) and, as shown in [15], the difference of two nested pAUCs can be tested by an F-test. As we outline in more detail below, the F-test for pAUC difference is based on the multiple partial F-test in discriminant analysis [16] and therefore it is an exact test. We can treat it as a gold standard test in this situation. Formally, the normality assumption means that the vector $\mathbf{x}$ of p predictor variables is conditionally normally distributed: $\mathbf{x}|D = 0 \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ and $\mathbf{x}|D = 1 \sim N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_0$ and $\mathbf{x}\boldsymbol{\mu}_1$ are vectors of the means for the p test results among the nonevents and events, respectively, and $\boldsymbol{\Sigma}$ is the variance–covariance matrix, which for simplicity of the presentation we assume to be the same in the two subgroups. In this situation several authors showed that LDA is the best method to estimate linear coefficients [3, 17] and presented the explicit formula for its AUC [3]

$$\text{AUC} = \Phi\left(\sqrt{\frac{(\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 - \mu_0)}{2}} \right), \quad (3)$$

where $\Phi(.)$ is the standard normal CDF. In practical situations we can replace the unknown population parameters with their consistent estimators to obtain a consistent estimator of the AUC [17]. This is the parametric AUC, or pAUC.

As demonstrated in (3) the AUC is also a function of the squared Mahalanobis distance, M^2 [18] defined as

$$M^2 \stackrel{\text{def}}{=} (\mu_1 - \mu_0)'\Sigma^{-1}(\mu_1 - \mu_0).$$

Therefore, as argued by Demler *et al.* [15], the two AUCs are equal, if and only if, the corresponding M^2s are equal. Likewise, under normality, to compare two AUCs we can use the test of significance for the difference of two nested M^2s, which is the well-known F-test [18].

Thus, it makes sense to compare the DeLong test to this gold standard when working with nested models and multivariate normal data. If the DeLong test is adequate for this

application, we would expect its size to be close to the size of the gold standard F-test. Hence, p-values obtained using the DeLong test plotted against p-values calculated with the F-test should form a 45 degree line or be close to it.

2.3. Example

To determine if the hypothesized relationship holds, we used numerical simulations on multivariate normal data with equal covariance matrices in the event and nonevent subgroups. To make these simulations mimic reality as much as possible, we used the means and correlation structure of an actual data set from the Framingham Heart Study [19–21]. A total of 8261 observations on people free of cardiovascular disease at baseline examination in the 1970s were available. Measurements of risk factors and results of medical tests were obtained, including age, total (TCL) and high-density lipoprotein (HDL) cholesterol, and systolic (SBP) and diastolic blood pressure (DBP). Participants in the study were followed for 12 years for the development of coronary heart disease (CHD) and were categorized as cases (621 observations) if they developed CHD or as noncases otherwise. The effect sizes (defined as difference of the means in two groups divided by standard deviation) of the logarithmically transformed predictors were 0.65, –0.47, 0.48, 0.62, and 0.42 for age, TCL, HDL, SBP, and DBP, respectively. For purposes of this example we used traditional LDA analysis, which assumes equal correlation structure among subgroups. The set of noncases is much larger, so we used its correlation matrix in our simulations for both cases and noncases. It is given below

$$\begin{pmatrix} 1.0 & .39 & .06 & .44 & .19 \\ & 1.0 & .08 & .25 & .18 \\ & & 1.0 & -.04 & -.11 \\ & & & 1.0 & .74 \\ & & & & 1.0 \end{pmatrix}$$

Of note, the simulations described here are robust to the correlation matrix structure. When we performed simulations with the correlation structure of cases and also run them using the results of Su and Liu for LDA [17], the results remained unchanged (data not shown).

We used all five predictors for the full model ($p = 5$) and the first four ($p - k = 4$) for the reduced model. The first four predictors were simulated according to the means and correlation structure described above. However, because the focus here was on the size of the test, the data for the added fifth predictor (DBP) was simulated under the null. This was accomplished by making its means equal in the two subgroups and ensured that it was uncorrelated with the rest of the predictors by setting nondiagonal elements in the last column of the correlation matrix to zero. We repeated the simulations 1000 times.

Linear discriminant analysis was fitted to each simulated dataset, first for the reduced model and then for the full model. P-values for the difference of pAUCs were calculated using the F-test, and p-values comparing eAUCs were calculated with the DeLong test. All required parameters for the models and two tests were estimated with each simulated dataset. These pairs of p-values are plotted below (Figure 1).

It is clear that the p-values do not form a 45 degree line; they are very scattered. Because the F-test is the gold standard here, we conclude that the application of the DeLong test to nested eAUCs may not be adequate. If we look at the 0.05 significance lines, all points below the horizontal 0.05 line are significant based on the F-test. All points to the left of the vertical 0.05 line are significant based on the DeLong test. Discrepant tests lie in the upper left or lower right rectangles. We see that all but one discrepant simulation lies in the upper

left rectangle corresponding to the situation when the F-test is significant but the DeLong test is not. Moreover, the significance level for the F-test is close to the nominal 0.05 level, whereas it is 0.001 for the DeLong test. Dramatic as it seems, our simulation-based comparison of the DeLong test with the F-test suggests that the former biases the results towards the null. Therefore, our simulation suggests that the DeLong test applied to nested models tends to be overly conservative.

3. Practical problem with extending DeLong test to nested models

3.1. Distribution of change in empirical AUCs under null hypothesis

The finding based on the simulated example of the previous section can possibly be explained by arguing that the F-test calculated the difference of two pAUCs, whereas the DeLong test compared eAUCs. In this case, because both estimators are consistent, we would expect the two p-values to become closer as we increase the sample size. Our simulations, with sample sizes of 50,000 and 100,000 observations, respectively, show that this is not the case (results are very similar to Figure 1 and are not shown). In this section, we show that the true reason behind the inconsistency observed in Figure 1 lies in the null distribution utilized by the DeLong test. We first show that the empirical distribution of the change in eAUC under the null is quite different from the normal distribution utilized in the DeLong test. In the next section we provide the theoretical explanation.

We again revert to our numerical simulation described above to plot a histogram of the difference in eAUCs between the full and reduced models and superimpose a normal curve used by the DeLong test (mean zero and variance calculated by the DeLong formula [9]). The result is presented in Figure 2.

We observe that the empirical distribution of 1000 simulations of eAUC difference is skewed, nonsymmetrical and hence does not appear to be normal. Increasing the sample size does not improve the situation — the plot looks virtually the same when sample size of 100,000 observations with 50,000 cases is used (results not shown). We notice that the distribution of the difference in eAUC converges to zero mostly from the right and is always skewed.

3.2. Extent of power loss

In the previous two sections, we saw that the DeLong test appears to be overly conservative, which may result in a loss of power. Given common applications of the DeLong test in the literature and the above result, it makes sense to investigate the extent of the power loss incurred. Different scenarios defined by the sample size, the distribution of predictors, and the baseline AUC are analyzed. We considered data in two distributional settings: a mixture of 3 continuous and 2 binary predictors (smoking status and diabetes) for real life data in one setting and 5 normally distributed predictors in another setting, with the added predictor being continuous. Of note, when we repeated the simulations adding a categorical predictor the results remained similar. For each setting we used a large sample of 8261 subjects with 621 events, corresponding to the data used in our initial example, a smaller sample of 700 subjects but with 350 events and a sample of 700 subjects with 53 events. Finally, we considered cases of low, moderate, and high baseline AUC of 0.52, 0.76, and 0.90 for large data of 8261 sample size. The desired AUCs and effect sizes are achieved by changing means among cases for the continuous predictors and altering the corresponding prevalence of binary exposure. Power is assessed as a number of rejections (based on a simulated null distribution) in 1000 repetitions of the experiment. Linear coefficients and all necessary parameters were estimated with each bootstrap sample. The following tests were considered:

1. The F-test for normal data analyzed by the LDA or the Wald test for the logistic regression coefficient for non-normal data;

2. DeLong test of AUC difference;

3. A test based on a nonparametric bootstrap of the difference in the eAUCs with $B = 1000$ replications of the original data. We used the approach described in [22–25] by drawing random sample with replacement, 1000 times for each effect size. To perform a bootstrap test of eAUC difference, we fitted the full and reduced model to each bootstrap sample and estimated the corresponding eAUCs. Then, the bootstrap p-value of eAUC difference was obtained from the empirical distribution function under the null hypothesis as a probability of exceeding the observed value [23–25].

Given the striking similarities of the power curves for the 10 scenarios, we present only two of them in Figures 3(A) and (B): one for non-normal data and the large sample of 8261 subjects with baseline AUC of 0.76 (Figure 3(A)) and one for normal data but a small sample of 700 with the same baseline AUC and prevalence of the outcome (Figure 3(B))

On the basis of the simulations illustrated above, we conclude that the DeLong test has the lowest power of the three tests considered. The extent of power loss depends on the combination of the effect size of the added predictor, and the strength of the baseline model — but mostly on the number of cases and the sample size. For large samples with many cases and a weak baseline model, the power loss is observed only for predictors with weak effect size, that is, less than 0.2. For example, for real life data with 621 events out of 8261 observations, with a baseline AUC of 0.76 and the effect size of added predictor equals to 0.2, the power of the DeLong test is only 0.473 compared with 0.941 and 0.904 for the Wald and bootstrap tests, respectively. On the other hand, for small samples with few cases, we observe power loss for effect sizes as large as 0.5 (Figure 3(B)). For example, using normal data with 53 events among 700 observations, with a baseline AUC of 0.76 and an effect size of the added predictor of 0.5, the power of the DeLong test is only 0.243, compared with 0.927 and 0.890 for the F-test and the bootstrap tests, respectively.

Overall, these simulations provide empirical evidence supporting the claim that the application of the DeLong test to nested models will lead to inference that is more likely biased towards the null hypothesis of no effect. This result is in agreement with numerous reports presented in the literature, where the significance of the regression coefficient did not translate into a significant increase in the AUC [7, 11, 13, 26].

4. Theoretical problems with extending DeLong test to nested models

In the previous sections we saw that the empirical distribution of the difference in eAUCs is very different from the normal distribution assumed by the DeLong test. In this section we present a theoretical explanation of the problems encountered when this test is extended from non-nested to nested models.

We first note that DeLong *et al.* [9] used the eAUC as an estimator of the true AUC. To arrive at the test of the difference of two AUCs, the authors argued that eAUC belongs to a class of generalized U-statistics and used asymptotic distribution theory of generalized U-statistics. Thus, the key to the explanation of the phenomenon illustrated in Figures 1–3 lies in the application of the U-statistics theory.

By definition, a generalized U-statistic is an average of a certain class of random variables, which are (possibly) correlated (see comprehensive review in [27]). The distribution theory of generalized U-statistics was developed in the 1940s and 1950s by Lehmann [28], Mises

[29], and Hoeffding [30] and can be viewed as an extension of the central limit theorem for sums of correlated variables. Definition of a generalized U-statistic does not allow for estimated parameters in the formula. For example, the AUC estimated according to the established formula (2), assuming the true parameters (that do not change with the sample data), is a generalized U-statistic. Let us denote it as eAUC*. However, the AUC estimated according to the same formula (2) but assuming the model-based estimated parameters belongs to a class of statistics under the aegis of the generalized U-statistic with estimated parameters. We will keep it denoted as eAUC.

Generalized U-statistics with estimated parameters form a wider class, and therefore we cannot directly apply to them the asymptotic distribution theory developed for the generalized U-statistic. There is no standalone, self-contained asymptotic distribution theory for generalized U-statistics with estimated parameters. However, as shown in [31–33], under certain regularity conditions, such statistics often 'inherit' distributional properties of the underlying generalized U-statistics. To obtain the asymptotic distribution for a generalized U-statistic with estimated parameters, one could first apply the asymptotic theory to the underlying generalized U-statistic (assuming true model parameters are known) and then check certain technical conditions [31–33] to see whether the results can be carried over to the corresponding generalized U-statistic with estimated parameters [27].

We attempt to follow this logic, trying to extend the DeLong test to nested models. First, we note that the difference between two eAUCs is also a generalized U-statistic with estimated parameters. Hence, to obtain its asymptotic distribution, we need to apply the asymptotic theory to the difference of eAUC*$_{full}$–eAUC*$_{reduced}$ and then determine if it can be extended to eAUC$_{full}$– eAUC$_{reduced}$. The former difference is a generalized U-statistic. Under very general conditions, any U-statistic converges asymptotically to a normal distribution unless it is a *degenerate* generalized U-statistic [27]. In the degenerate case, the normal approximation theory does not apply. It is not difficult to see that if the models are nested, parameters known and the new predictor is noninformative, then the full and reduced models are the same. Therefore, the distribution of eAUC*$_{full}$– eAUC*$_{reduced}$ degenerates to a point mass at 0. This fact automatically implies that it is a degenerate generalized U-statistics. Hence, the asymptotic normality does not hold, even in the simpler case of known model parameters. Consequently, we do not have a result to extend to the difference in eAUCs with estimated parameters.

The above argument illustrates that for nested models, under the null hypothesis of no association of the added predictor with the outcome, the corresponding difference in eAUC*s is a degenerate U-statistic for any distribution function of the predictors and for very general types of models such as logistic regression or LDA. This means that we cannot use the normal theory to approximate the distribution of the AUC difference and hence the DeLong test cannot be applied. This can explain the discrepancies observed in the previous section.

We have argued that if nonimprovement in the AUC coincides with nonsignificance of the added predictor(s), then we should not use the DeLong test. Demler *et al.* [15] showed that nonimprovement in the AUC for nested models is equivalent to nonsignificance of the added predictor for normal data analyzed by LDA. Hence, for nested models in the case of normality, we have a strict proof of the fact that the DeLong test is inappropriate for the very hypothesis on which it is applied. It is unclear if a similar equivalence could be established for non-normal data and further research is required in this area.

Of note, whenever results from the generalized U-statistics theory are being used, it is always necessary to explicitly check for nondegeneracy. This overlooked fact has led several

researchers to propose tests based on asymptotic normality of the generalized *U*-statistics, which are in fact degenerate; see for example Refs. [34, 35].

In conclusion, we would like to point out two important facts. First, in their original paper, DeLong *et al.* [9] developed their theory for comparison of several different tests. They never developed or explicitly proposed their test to compare nested models. Hence, the difference in the eAUCs that they considered is always nondegenerate and the problems outlined here do not apply. Second, they derived their results only for the situation when model parameters are known. In fact, the first papers addressing generalized *U*-statistics with estimated parameters only appeared around the time of the DeLong publication. These two facts were overlooked by researchers, who extended, without proof, the DeLong test to nested models with estimated parameters. Fortunately, the result of this misapplication led to an overly conservative test and only for specific combinations of sample sizes, effect sizes, and strengths of baseline models.

5. No problems with DeLong approach for nested models under alternative hypothesis and for non-nested models

In the previous sections we have argued that the distribution of the difference of two nested eAUCs is non-normal under the null hypothesis. Here, we ask whether it is approximately normal under the alternative. The question is important, because if normality under the alternative holds, we can apply the DeLong approach to estimate standard errors and construct confidence intervals for the difference of two nested AUCs.

We first employ numerical simulations to visualize the situation. We stay in the simulation framework described in Section 2, but set the conditional effect size of the new predictor to 0.25. In Figure 4 we plot the histogram of change in eAUC and superimpose the plot of the normal distribution function used by the DeLong method.

The histogram is very different from the one presented in Figure 2 and appears approximately normal. This suggests that the method of DeLong *et al.* [9] can be used to construct a confidence interval for the difference in nested AUCs. This has been confirmed in a comprehensive study by Obuchowski *et al.* [36] who showed that a confidence interval based on the DeLong approach has the appropriate coverage. Moreover, these authors have demonstrated that this interval is superior to intervals based on the bootstrap or likelihood ratio methods.

We also mention that in the case of significant association between the new predictor and outcome or for non-nested models we do not encounter the degeneracy observed in Section 4. The difference of two eAUC*s in this case is nondegenerate and therefore it does have asymptotically normal distribution. Hence, it remains to show that the result for the eAUC difference with known parameters can be extended to the difference of eAUCs with estimated parameters. This goes beyond the scope of this paper; however, the plot presented in Figure 4 and the results in [36] can serve as empirical evidence supporting that claim.

6. Practical example

Now we return to the example introduced in Section 2 to illustrate the performance of the DeLong method applied to nested models developed on Framingham data [19]. We consider the five continuous predictors given in Section 2. Data was analyzed using logistic regression. Our full model includes age, TCL, HDL, SBP, and DBP as predictors of CHD. All five predictors are significant based on the Wald test for the logistic regression coefficient. Five reduced models are considered, omitting one of the predictors each time.

Empirical AUCs are calculated and p-values of the test of significance of AUC change based on the DeLong test and the Wald test of significance for the logistic regression coefficient. The results are presented in Table I.

This real life data example illustrates our findings. We see in Table I that the DeLong test is consistent with theWald test of the coefficient for those predictors that have high conditional effect sizes (age, HDL and total cholesterol and to some extent systolic blood pressure). However, diastolic blood pressure is significant in logistic regression but has conditional effect size of only 0.02, which puts it on that part of the power plot (see Figure 3(A)) where the DeLong test has very low power as compared with the Wald test. Thus, the DeLong test fails to reject the null of no change in AUC for diastolic blood pressure (p-value=0.47), whereas diastolic blood pressure is significantly related to the outcome of incident CHD based on the Wald test of the coefficient (p-value=0.01). Of more practical importance, however, we observe that the change in AUC is minimal in this case: 0.001 with a 95% confidence interval of (0.0001, 0.0031).

7. Conclusions and recommendations

On the basis of the numerical simulations supported by a theoretical explanation, we conclude that the test proposed by DeLong [9] for comparison of two AUCs should not be extended to compare two nested AUCs for models developed and validated on the same data. This misuse of the DeLong test leads to a conservative test with low power for a number of practical scenarios. Vickers *et al.* [26] arrived at a similar conclusion.

Our explanation of the invalidity of the DeLong test might provide some insight into possible ways to remedy the problem. In this paper we showed that the difference of two eAUCs of nested models under null is a degenerate U-statistic. This causes invalidity of the DeLong test when applied to nested models. We referred to standard U-statistics theory to argue that to extend the results of DeLong *et al.* beyond their original framework [9], we need first to check for nondegeneracy, and once nondegeneracy is established we need to account for estimated parameters. Luckily, the underlying U-statistic is degenerate only for nested models under the null. It is nondegenerate for non-nested models (therefore, we can use DeLong theory to test and construct a CI). It is also nondegenerate in nested models with significant new predictor (therefore, we still can use DeLong theory to construct CI for the difference in AUCs). In these situations only the question of the adjustment for estimated parameters remains. Results of extensive simulations by Obuchowski *et al.* [36] and our simulations in Figure 4 indicate that the DeLong method (which does not adjust for estimated parameters) works very well in nondegenerate cases despite estimated parameters. This indicates that adjustment for estimated parameters might not be needed in these situations after all.

To find possible ways to remedy the problem with testing AUC improvement for nested models discussed in this paper, let us first consider whether a test of significance for the difference of two nested AUCs is really needed. There are two steps to building the model: first, to select the best set of predictors associated with the outcome without over-fitting, and second, to evaluate the quality of discrimination of the resulting model. Comparisons of two nested models follow the same two steps: we can ask whether the new predictor(s) are associated with the outcome and whether they improve discrimination. Because AUC is primarily a measure of the quality of discrimination, its usefulness lies in the interpretation of its magnitude or the magnitude of the increase in model performance. Thus, the following strategy can be adopted. First, test the statistical significance of the added predictor(s) by usual methods. If not significant, stop and conclude that the relationship of the added variable(s) with outcome does not go beyond chance finding. If significant, proceed to

calculate the two nested AUCs and their difference with the corresponding confidence interval. This recommendation is in agreement with Vickers *et al.* [26]. In this setting, as shown in [36] and illustrated here, the approach of DeLong *et al.* [9] is the best method to arrive at this interval.

References

1. Harrell, FE, Jr. Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression and Survival Analysis. New York: Springer-Verlag; 2001.

2. Steyerberg, EW. Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating (Statistics for Biology and Health). New York: Springer Science+Business Media; 2009.

3. Pepe, MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. New York: Oxford University Press; 2004.

4. Stone B. Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society. Series B (Methodological). 1974; 36(2):111–147.

5. Ridker PM, Rifai N, Rose L, Buring JE, Cook NR. Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. NEJM. 2003; 348:1059–1061.

6. Polonsky TS, McClelland RL, Jorgensen NW, Bild DE, Burke GL, Guerci AD, Greenland P. Coronary artery calcium score and risk classification for coronary heart disease prediction. JAMA. 2010; 303(16):1610–1616. [PubMed: 20424251]

7. Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PWF, D'Agostino RB Sr, Cupples A. Genotype score in addition to common risk factors for prediction of type II diabetes. New England Journal of Medicine. 2008; 359(21):2208–2219. [PubMed: 19020323]

8. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham heart study. Circulation. 2008; 117:743–753. [PubMed: 18212285]

9. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing areas under two or more correlated reciever operating characteristics curves: a nonparamentric approach. Biometrics. 1988; 44(3):837–845. [PubMed: 3203132]

10. SAS/STAT software. Version 9.1(TS1M3) of the SAS System. Copyright © 2002–2003 SAS Institute Inc. SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. Cary, NC, USA:

11. Wang TJ, Gona P, Larson MG, Tofler GH, Levy D, Newton-Cheh C, Jacques PF, Rifai N, Selhub J, Robins SJ, Benjamin EJ, D'Agostino RB, Vasan RS. Multiple biomarkers for the prediction of first major cardiovascular events and death. The New England Journal of Medicine. 2006; 355(25):2631–2639. [PubMed: 17182988]

12. Ridker PM, Buring JE, Rifai N, Cook NR. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women. JAMA. 2007; 297(6):611–619. [PubMed: 17299196]

13. Cook NR. Use and misuse of the receiver operating characteristics curve in risk prediction. Circulation. 2007; 115:928–935. [PubMed: 17309939]

14. Tzoulaki I, Liberopoulos G, Ioannisis JPA. Assessment off claims of improved prediction beyond the Framnigham risk score. JAMA. 2009; 302:2345–2352. [PubMed: 19952321]

15. Demler OV, Pencina MJ, D'Agostino R. Equivalence of improvement in area under ROC curve and linear discriminant analysis coefficient under assumption of normality. Statistics in Medicine. 2011; 30(12):1410–1418. [PubMed: 21337594]

16. Rao, CR. Linear Statistical Inference and its Applications. New York: Wiley; 1973.

17. Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. JASA. 1993; 88(424):1350–1355.

18. Mardia, KV.; Kent, JT.; Bibby, JM. Multivariate analysis. San Diego: Academic Press; 1979.

19. Executive Summary of the Third Report of the National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults (Adult

Treatment Panel III). Journal of the American Medical Association. 2001; 285:2486–2497. [PubMed: 11368702]

20. Wilson P, D'Agostino R, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97:1837–1847. [PubMed: 9603539]

21. Anderson KM, Odell PM, Wilson PWF, Kannel WB. Cardiovascular disease risk profiles. American Heart Journal. 1991; 121:293–298. [PubMed: 1985385]

22. Steyerberg EW, Harrell FE, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. Journal of Clinical Epidemiology. 2001; 54:774–781. [PubMed: 11470385]

23. Efron, B.; Tibshirani, R. An Introduction to Bootstrap. Boca Raton: Chapman and Hall/CRC; 1993.

24. Tibshirani R, Hall P, Wilson SR. Bootstrap Hypothesis Testing. Biometrics. 1992; 48(3):969–970.

25. Hall P, Wilson SR. Two guidelines for bootstrap hypothesis testing. Biometrics. 1991; 47:757–762.

26. Vickers AJ, Cronin AM, Begg CB. One statistical test is sufficient for assessing new predictive markers. BMC Medical Research Methodology. 2011; 11(13):1–7. [PubMed: 21208427]

27. Lee, AJ. U-Statistics: Theory and Practice. New York and Basel: Marcel Dekkel; 1990.

28. Lehmann EL. Consistency and unbiasedness of certain nonparametric tests. The Annals of Mathematical Statistics. 1951; 22(2):165–179.

29. Mises RV. On the asymptotic distribution of differentiable statistical functions. Annals of Mathematical Statistics. 1947; 18:309–348.

30. Hoeffding W. A class of statistics with asymptotically normal distributions. Annals of Mathematical Statistics. 1948; 19(3):293–325.

31. De Wet T, Randles RH. On the effect of substituting parameter estimators in limiting $\chi 2$ U and V statistics. The Annals of Statistics. 1987; 15(1):398–441.

32. Iverson HK, Randles RH. The effects on convergence of substituting parameter estimates into U-statistics and other families of statistic. Probability Theory and Related Fields. 1989; 81:453–471.

33. Randles RH. On the asymptotic normality of statistics with estimated parameters. The Annals of Statistics. 1982; 10(2):462–474.

34. Antolini L, Namb B-H, D'Agostino RB. Inference on correlated discrimination measures in survival analysis: a nonparametric approach. Communications in Statistics Theory and Methods. 2004; 33(9):2117–2135.

35. Lee M-LT, Rosner BA. The average area under correlated receiver operating characteristic curves: a nonparametric approach based on generalized two-sample Wilcoxon statistics. Applied Statistics. 2001; 50(3):337–344.

36. Obuchowski NA, Lieber ML. Confidence intervals for the receiver operating characteristic area in studies with small samples. Academic Radiology. 1998; 5:561–571. [PubMed: 9702267]
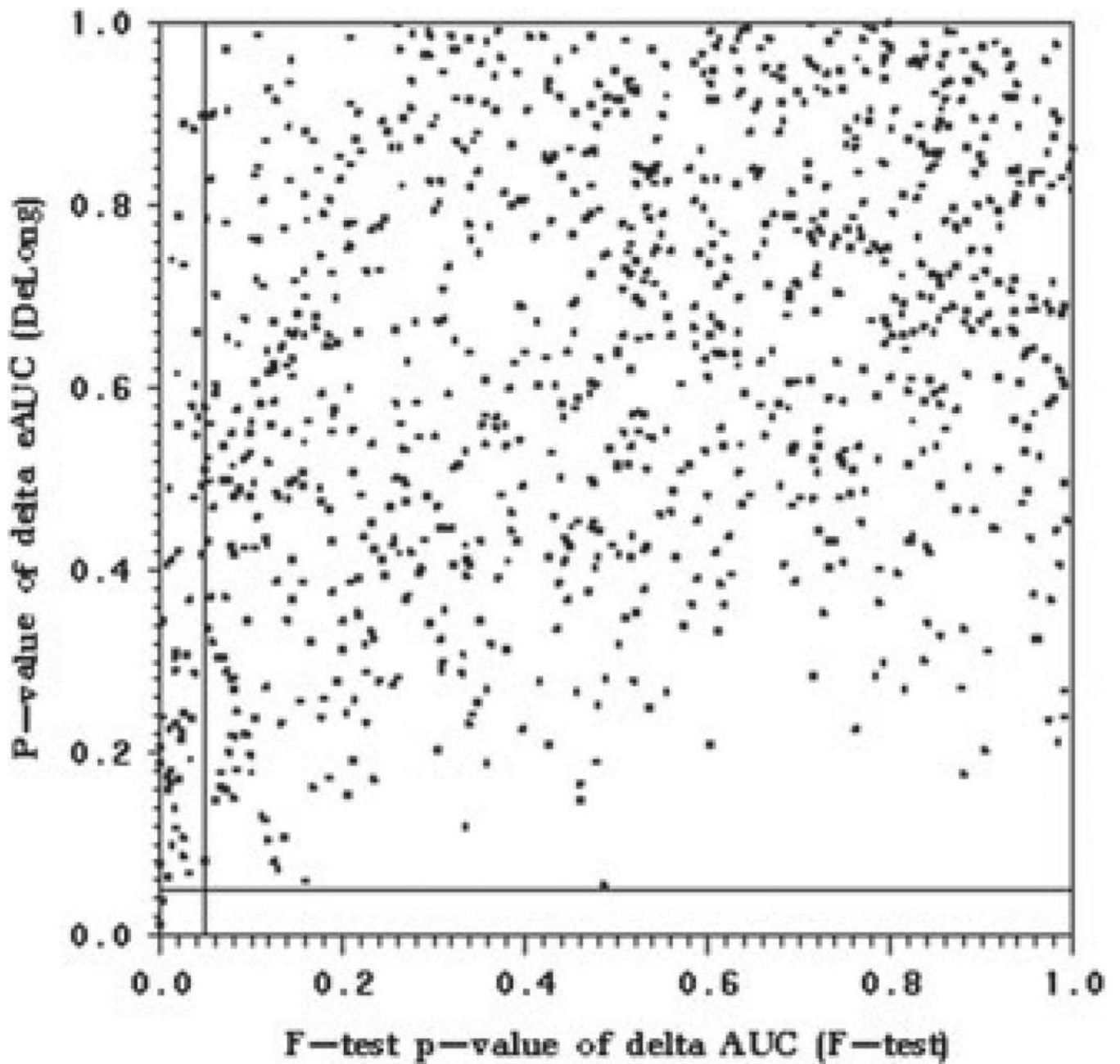
Figure 1.
Scatterplot of *p*-values produced by *F*-test versus corresponding *p*-values produced by DeLong test. 1000 simulations of multivariate normal data with sample size of 8261.
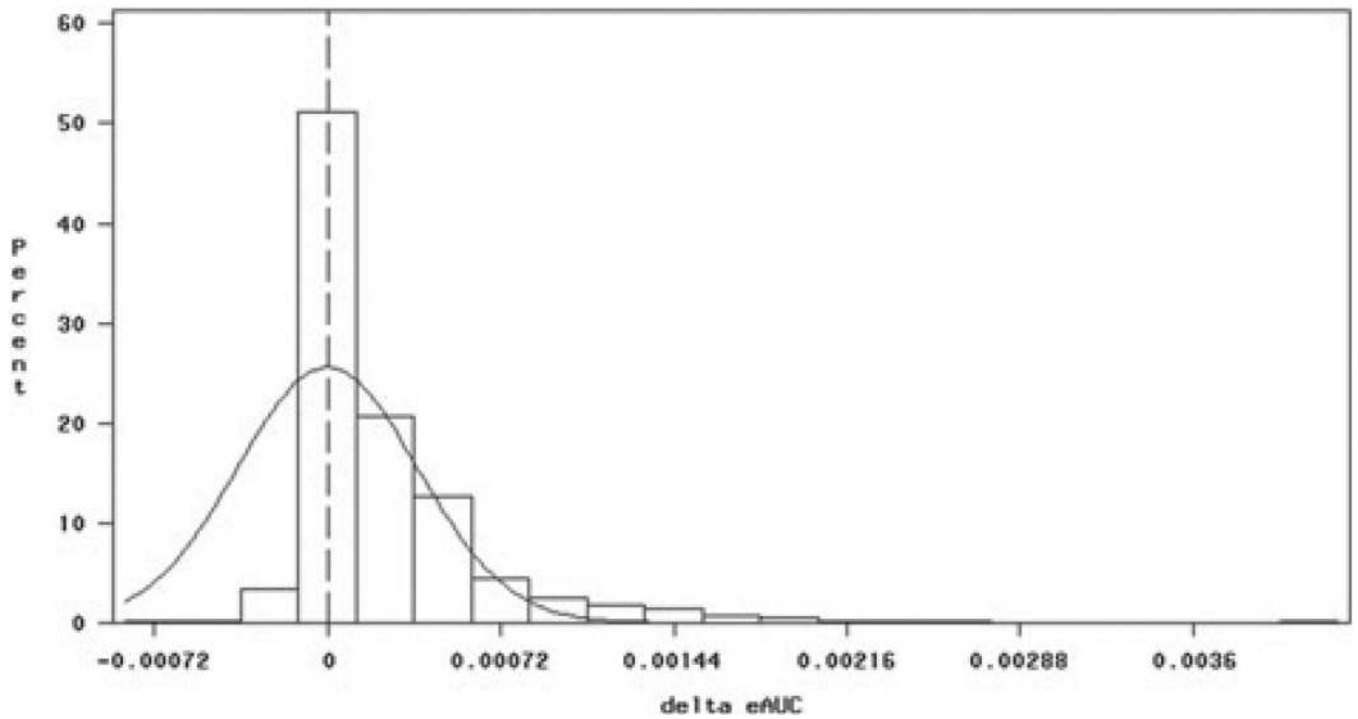
Figure 2.
Histogram of change in eAUC under null hypothesis for multivariate normal data and sample size of 8365 with superimposed plot of corresponding distribution function used by DeLong test.
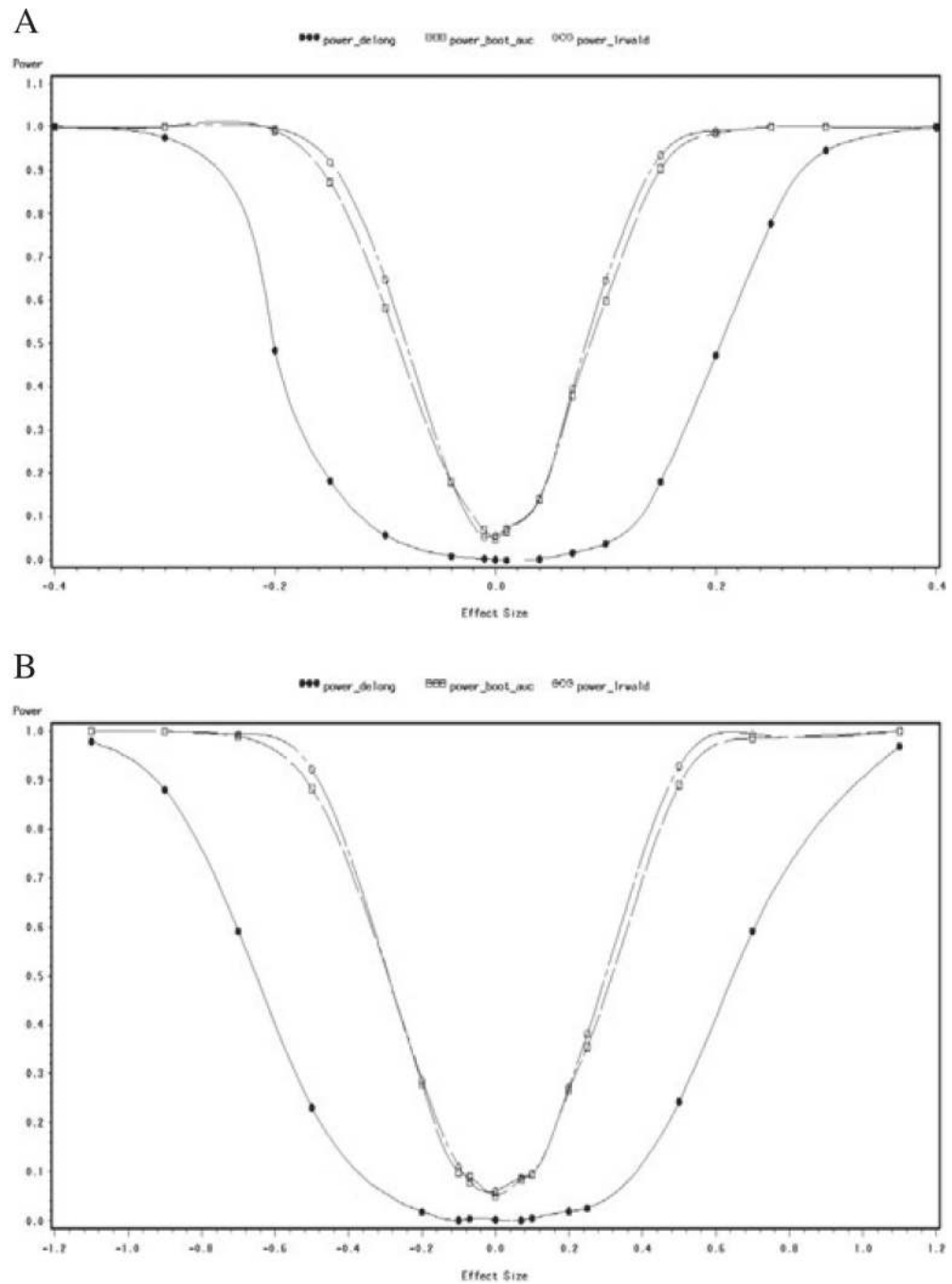
A

B

Figure 3.
(A) Power of Wald test, DeLong test, and test based on bootstrap for different conditional effect sizes. On the basis of real-life data sample size 8261 (with 621 cases) baseline AUC is 0.76. (B) Power of F-test, DeLong test, and test based on bootstrap conditional effect sizes. On the basis of simulated multivariate normal data sample size 700 (with 53 cases) baseline AUC is 0.76.
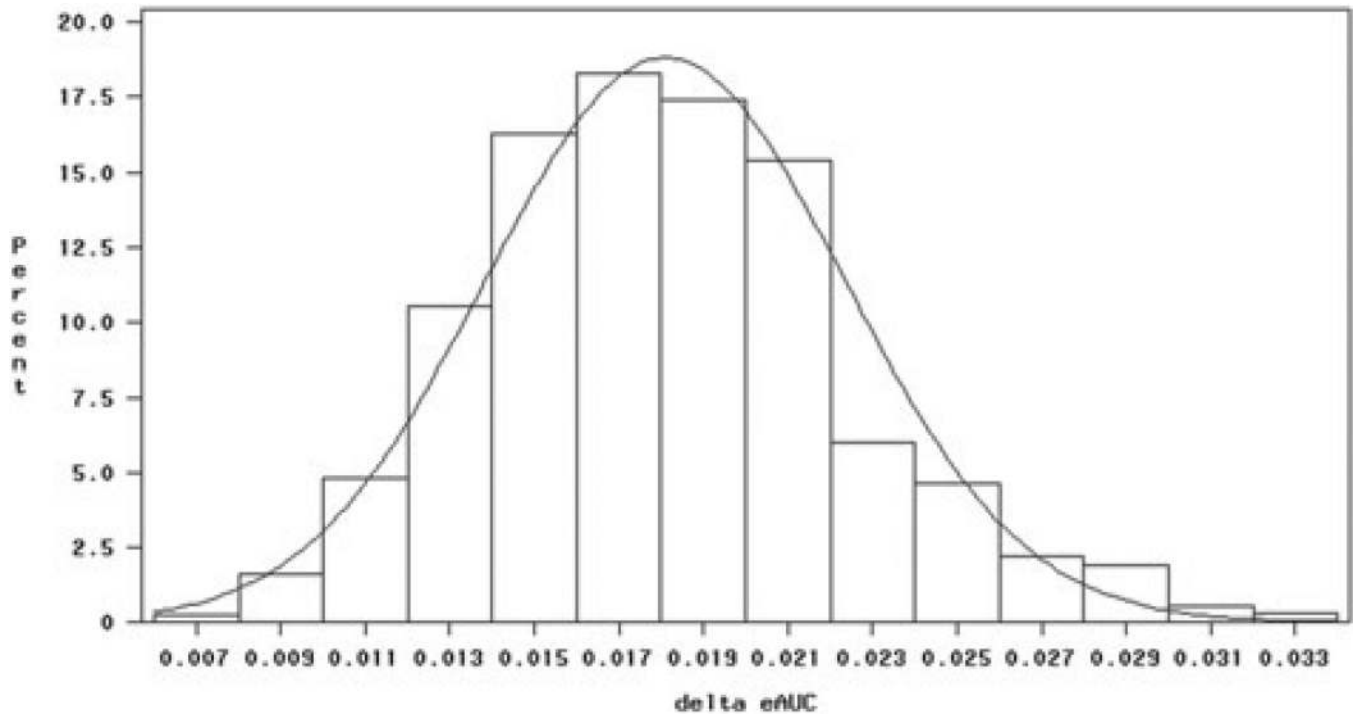
Figure 4.
Histogram of distribution of change in eAUC under alternative hypothesis. Simulations were performed for conditional effect size of 0.25 of multivariate normal data with sample size of 8261.

Table I

Significance of change in empirical AUC tested by DeLong versus significance of regression coefficient using Wald test; $N = 8261$.

Excluded variable (conditional effect size)	eAUC		*P*-value	
	Full model	**Reduced model**	**DeLong**	**Wald test**
Age (0.39)	0.758	0.735	< 0:01	< 0:01
HDL cholesterol (–0.46)	0.758	0.726	< 0:01	< 0:01
Total cholesterol (0.29)	0.758	0.746	< 0:01	< 0:01
Systolic blood pressure (0.18)	0.758	0.756	0.02	0.01
Diastolic blood pressure (0.02)	0.758	0.757	0.47	0.01