Modern Stochastics: Theory and Applications 13 (1) (2026) 57–82
https://doi.org/10.15559/25-VMSTA285

A new confidence interval based on the theory of U-statistics for the area under the curve

Jürgen Kampf*, **Lukas H. Vogel, Iryna Dykun, Tienush Rassaf, Amir A. Mahabadi**

West German Heart and Vascular Center Essen, Department of Cardiology and Vascular Medicine, University Hospital Essen, Hufelandstraße 55, 45147 Essen, Germany

juergen.kampf@uk-essen.de (J. Kampf), lukas.vogel@uk-essen.de (L. Vogel), iryna.dykun@uk-essen.de (I. Dykun), tienush.rassaf@uk-essen.de (T. Rassaf), amir-abbas.mahabadi@uk-essen.de (A. Mahabadi)

Received: 8 April 2025, Revised: 12 September 2025, Accepted: 22 September 2025, Published online: 9 October 2025

Abstract The area under the receiver operating characteristic curve (AUC) is a suitable measure for the quality of classification algorithms. Here we use the theory of U-statistics in order to derive new confidence intervals for it. The new confidence intervals take into account that only the total sample size used to calculate the AUC can be controlled, while the number of members of the case group and the number of members of the control group are random. We show that the new confidence intervals can not only be used in order to evaluate the quality of the fitted model, but also to judge the quality of the classification algorithm itself. We would like to take this opportunity to show that two popular confidence intervals for the AUC, namely DeLong's interval and the Mann–Whitney intervals due to Sen, coincide.

Keywords Area under the curve (AUC), confidence interval, logistic regression, receiver operating characteristic (ROC), U-statistics
2010 MSC 62J12, 62P10

*Corresponding author.

1 Introduction

In medicine and other branches of science one frequently faces binarization problems, i.e. one wants to predict to which of two classes an item belongs based on characteristics or data of an individual item. To solve this kind of problems, a lot of competing algorithms—like neural nets, random forests or estimators for logistic regression models—are available. Despite all differences in details, the general approach of these algorithms is similar. First, a classification model is fit using some *training data* for which the true classification is known. Then this model can be used to classify new data for which the true classification is unknown. In order to evaluate the quality of a classification model, it is subsequently often applied to some *test dataset* for which the true classification is known and the true classification is compared with the prediction of the model. A frequently used method to quantify the result of this comparison is the area under the curve (AUC) that we are going to introduce in detail in Section 2.

There are several competing confidence intervals for the AUC. Most of them, like the popular DeLong algorithm [4], also those by Kottas, Kuss & Zapf [6], LeDell, Petersen & v. d. Laan [8], and all intervals compared in Qin & Hotilovac [12], assume that the number of members of the case and control groups are deterministic. However, in practice often only the total number of members of the test group can be controlled, while the assignment to the case and control groups is random. Using the theory of U-statistics we propose new confidence intervals that take this into account.

Moreover, the confidence intervals cited above are designed for the evaluation of the (fitted) classification model. It is assumed that this model is perfectly true and the randomness just comes from the fact that the test set is a random sample. This is alright, if one wants to assess the quality of the fitted model. However, often the real question is whether an algorithm is suitable for a certain kind of data. Then the uncertainty that arises due to the fitting of the model parameters based on a finite training set has to be considered as well. We will examine this for the logistic regression model. We will see in a simulation study that this uncertainty is of practical relevance and ignoring it can lead to a seriously too low coverage probability of the confidence intervals. On the theoretical side, however, we will see that this uncertainty is asymptotically neglectable.

A related question is, what happens, when all observations are used both for training and for testing. This situation was considered in [10].

We will take this opportunity to show that two well-known confidence intervals for the AUC, namely DeLong's intervals [4] and the Mann–Whitney intervals due to Sen [13], coincide. In the literature they are usually quoted under either name without noticing that they are the same, and in Qin & Hotilovac [12] they are even compared against each other.

This paper is organized as follows. In Section 2 we introduce the AUC and the logistic regression model in full detail. In Section 3 we derive the form of the confidence intervals from central limit theorems for the AUC. Section 4 is devoted to a simulation study and in Section 5 we apply the new confidence intervals to electrocardiogram (ECG) data. Finally, in Section 6 we discuss our results and point out directions for future research. The proofs of the theoretical results will be postponed to the Appendix.

2 Preliminaries

In this section we introduce the preliminaries we need on logistic regression models and the AUC.

A logistic regression model is a family of probability distributions on $\mathbb{R}^p \times \{0,1\}$ indexed by a parameter in $\mathbb{R}^p$. The random vectors (X, I) of a logistic regression model fulfill

$$\mathbb{P}(I = 1|X) = \frac{\exp\{\beta^T X\}}{\exp\{\beta^T X\} + 1} =: Y.$$

We assume that the observations $(X_{1,1}, I_{1,1}), \ldots, (X_{1,m}, I_{1,m})$ form an independent, identically distributed (i.i.d.) sample in which each member follows the logistic regression model. So we do not assume that the design points $X_{1,1}, \ldots, X_{1,m}$ are aligned on a grid, but we suppose that they are irregularly scattered. These data points form the training set. Logistic regression models are well studied in the literature; in particular the maximum-likelihood estimator $\hat{\beta}$ for β is known (see, e.g., [5]). It holds that $\hat{\beta} - \beta \to \mathcal{N}(0, F^{-1}(\beta))$ in distribution as $m \to \infty$, where $F(\beta)$ is the Fisher information matrix. A consistent estimator for $F(\beta)$ is given by

$$H(\beta) := \sum_{i=1}^{m} X_{1,i} X_{1,i}^T \pi\left(\beta^T X_{1,i}\right) \cdot \left(1 - \pi\left(\beta^T X_{1,i}\right)\right)$$

with $\pi(t) = 1/(\exp\{t\} + 1)$ (see [5, pp. 200–203]).

Once the model is fit, i.e. the parameter β is estimated, the probability $Y_{2,i} := \mathbb{P}(I_{2,i} = 1 \mid X_{2,i})$, $i = 1, \ldots, n$, can be estimated for new data points $X_{2,1}, \ldots, X_{2,n}$ that form the test set. In order to get a prediction for the class $I_{2,i}$ one can choose a threshold $c \in (0,1)$ and put $\hat{I}_{2,i} := \mathbf{1}_{[c,1]}(\hat{Y}_{2,i})$.

Notice that the behavior we saw above for the logistic regression model is typical for classification algorithms. At first, some $[0,1]$-valued score function $Y_{2,i}$ is derived—for logistic regression models this is the probability that $\{I_{2,i} = 1\}$, for neural nets it is the value of the nodes in the output layer and for random forests it is the ratio of all trees predicting $\{I_{2,i} = 1\}$. Then the predicted classification is obtained by thresholding $\hat{Y}_{2,i}$.

Now the prediction quality of the logistic regression model—or any other classification algorithm that works as indicated in the last paragraph—can be assessed using the area under the receiver operating curve (AUC). The empirical AUC is

$$\hat{A} = \frac{\sum_{i=1}^n \sum_{j=1}^n \mathbf{1}_{\{\hat{Y}_{2,i} < \hat{Y}_{2,j}\}} \mathbf{1}_{\{I_{2,i}=0\}} \mathbf{1}_{\{I_{2,j}=1\}}}{(\sum_{i=1}^n \mathbf{1}_{\{I_{2,i}=0\}}) \cdot (\sum_{j=1}^n \mathbf{1}_{\{I_{2,j}=1\}})} + \frac{1}{2} \cdot \frac{\sum_{i=1}^n \sum_{j=1}^n \mathbf{1}_{\{\hat{Y}_{2,i} = \hat{Y}_{2,j}\}} \mathbf{1}_{\{I_{2,i}=0\}} \mathbf{1}_{\{I_{2,j}=1\}}}{(\sum_{i=1}^n \mathbf{1}_{\{I_{2,i}=0\}}) \cdot (\sum_{j=1}^n \mathbf{1}_{\{I_{2,j}=1\}})}.$$

Its theoretical counterpart is

$$A = \mathbb{P}(Y_1 < Y_2 \mid I_1 = 0, I_2 = 1) + \frac{1}{2} \cdot \mathbb{P}(Y_1 = Y_2 \mid I_1 = 0, I_2 = 1).$$

The name of the AUC comes from the fact that it equals a certain area; see Figure 1. For further information on the AUC, see [11].

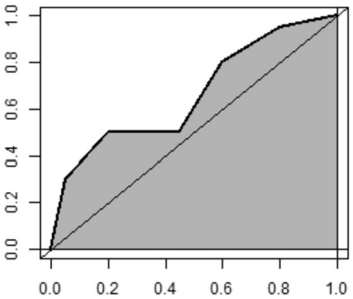

Fig. 1. Interpretation of the AUC as an area. If the ROC curve (see [11]) is the thick black line, then the AUC is the area of the gray polygon

3 Theoretical results

In this section, we present the mathematical theorems that are needed in order to establish the new confidence intervals.

We consider two models.

Model 1. Let $(\hat{Y}_{2,i}, I_{2,i}) = (Y_{2,i}, I_{2,i})$, $i = 1, \ldots, n$, be an i.i.d. sample of a real-valued variable Y and a binary variable I.

Model 2. Let $\hat{Y}_{2,i}$, $i = 1, \ldots, n$, be the fitted values of a logistic regression model (see Section 2) with true parameter $\beta_0 \in \mathbb{R}^p$, applied to i.i.d. data points $X_{2,i}$, $i = 1, \ldots, n$, and let $I_{2,i}$, $1, \ldots, n$, be the known true classification. We shall assume that the distribution of the points $X_{2,i}$, $i = 1, \ldots, n$, is absolutely continuous with the Lebesgue density q. For two points $X_{2,i}$ and $X_{2,j}$, $j \neq i$, the distribution of $(X_{2,i} - X_{2,j})/\|X_{2,i} - X_{2,j}\|$ has a bounded density with respect to the $(p-1)$-dimensional Hausdorff measure on the unit sphere. Moreover, the cardinality m of the training set and the cardinality n of the test set should fulfill

$$\liminf_{n\to\infty} \frac{m}{n} > 0, \qquad \limsup_{n\to\infty} \frac{m}{n} < \infty. \tag{1}$$

Model 1 is the classical model used most frequently in the literature so far (see DeLong et al. [4], Kottas et al. [6], Qin & Hotilovac [12] and Sen [13]). The idea is that $\hat{Y}_{2,i}$, $i = 1, \ldots, n$, are fitted values obtained by applying a completely known model to an i.i.d. sample of data points. As a consequence of this simplification $\hat{Y}_{2,i}$ and $Y_{2,i}$ always coincide under Model 1 and the "real" value of $Y_{2,i}$ is ignored. Notice that the first index 2 of the observations is not necessary if one only considers Model 1, since then there are no training observations $(X_{1,i}, I_{1,i})$. We just add this index in order to be able to treat Model 1 and Model 2 jointly.

Model 2 takes the more realistic point of view that the classification model is disturbed by random effects that arose in the model fitting procedure. However, under Model 2 we require that the used classification model is the logistic regression model, while under Model 1 we make no assumptions on the classification model.

We put

$$\sigma_A^2 = v^T \Sigma v,$$

where Σ is the asymptotic covariance matrix of

$$\begin{pmatrix} \sum_{i=1}^n \sum_{j=1}^n \mathbf{1}_{\{Y_{2,i}<Y_{2,j}\}}\mathbf{1}_{\{I_{2,i}=0\}}\mathbf{1}_{\{I_{2,j}=1\}} + \frac{1}{2}\cdot\sum_{i=1}^n \sum_{j=1}^n \mathbf{1}_{\{Y_{2,i}=Y_{2,j}\}}\mathbf{1}_{\{I_{2,i}=0\}}\mathbf{1}_{\{I_{2,j}=1\}} \\ \sum_{i=1}^n \mathbf{1}_{\{I_{2,i}=0\}} \\ \sum_{j=1}^n \mathbf{1}_{\{I_{2,j}=1\}} \end{pmatrix}$$

and where

$$v = \begin{pmatrix} \frac{1}{\mathbb{P}(I_1=0)\cdot\mathbb{P}(I_1=1)} \\ -\frac{\mathbb{P}(Y_1<Y_2, I_1=0, I_2=1)}{(\mathbb{P}(I_1=0)^2\cdot\mathbb{P}(I_2=1)} - \frac{1}{2}\cdot\frac{\mathbb{P}(Y_1=Y_2, I_1=0, I_2=1)}{\mathbb{P}(I_1=0)^2\cdot\mathbb{P}(I_2=1)} \\ -\frac{\mathbb{P}(Y_1<Y_2, I_1=0, I_2=1)}{\mathbb{P}(I_1=0)\cdot\mathbb{P}(I_2=1)^2} - \frac{1}{2}\cdot\frac{\mathbb{P}(Y_1=Y_2, I_1=0, I_2=1)}{\mathbb{P}(I_1=0)\cdot\mathbb{P}(I_2=1)^2} \end{pmatrix},$$

with (Y_1, I_1, Y_2, I_2) having the same distribution as $(Y_{2,i}, I_{2,i}, Y_{2,j}, I_{2,j})$, $j \neq i$. In the Appendix, we will show that σ_A^2 is the asymptotic variance of $\hat{A}$. Let S_A^2 be the plug-in estimator for σ_A^2, where all probabilities involved in the definition of v are estimated by their corresponding relative frequencies and where

$$\begin{aligned} \hat{\Sigma} = {} & \frac{1}{n\cdot(n-1)\cdot(n-2)} \sum_{i,j,k=1}^n \begin{pmatrix} a_{ij} \\ \mathbf{1}_{\{I_{2,i}=0\}} + \mathbf{1}_{\{I_{2,j}=0\}} \\ \mathbf{1}_{\{I_{2,i}=1\}} + \mathbf{1}_{\{I_{2,j}=1\}} \end{pmatrix} \\ & \times \begin{pmatrix} a_{ik} & \mathbf{1}_{\{I_{2,i}=0\}} + \mathbf{1}_{\{I_{2,k}=0\}} & \mathbf{1}_{\{I_{2,i}=1\}} + \mathbf{1}_{\{I_{2,k}=1\}} \end{pmatrix} \\ & - \left(\frac{1}{n\cdot(n-1)} \sum_{i,j=1}^n \begin{pmatrix} a_{ij} \\ \mathbf{1}_{\{I_{2,i}=0\}} + \mathbf{1}_{\{I_{2,j}=0\}} \\ \mathbf{1}_{\{I_{2,i}=1\}} + \mathbf{1}_{\{I_{2,j}=1\}} \end{pmatrix} \right) \\ & \times \left(\frac{1}{n\cdot(n-1)} \sum_{i,j=1}^n \begin{pmatrix} a_{ij} & \mathbf{1}_{\{I_{2,i}=0\}} + \mathbf{1}_{\{I_{2,j}=0\}} & \mathbf{1}_{\{I_{2,i}=1\}} + \mathbf{1}_{\{I_{2,j}=1\}} \end{pmatrix} \right) \end{aligned}$$

with

$$\begin{aligned} a_{ij} = {} & \mathbf{1}_{\{\hat{Y}_{2,i}<\hat{Y}_{2,j}\}}\cdot\mathbf{1}_{\{I_{2,i}=0\}}\cdot\mathbf{1}_{\{I_{2,j}=1\}} + \frac{1}{2}\cdot\mathbf{1}_{\{\hat{Y}_{2,i}=\hat{Y}_{2,j}\}}\cdot\mathbf{1}_{\{I_{2,i}=0\}}\cdot\mathbf{1}_{\{I_{2,j}=1\}} \\ & + \mathbf{1}_{\{\hat{Y}_{2,j}<\hat{Y}_{2,i}\}}\cdot\mathbf{1}_{\{I_{2,j}=0\}}\cdot\mathbf{1}_{\{I_{2,i}=1\}} + \frac{1}{2}\cdot\mathbf{1}_{\{\hat{Y}_{2,j}=\hat{Y}_{2,i}\}}\cdot\mathbf{1}_{\{I_{2,j}=0\}}\cdot\mathbf{1}_{\{I_{2,i}=1\}} \end{aligned}$$

is the estimator for Σ. The consistency of S_A^2 will be established in the course of the proof of Theorem 1. We remark that the calculation of $\hat{\Sigma}$ is indeed not $O(n^3)$, but $O(n^2)$, because

$$\begin{aligned} & \sum_{i,j,k=1}^n \begin{pmatrix} a_{ij} \\ \mathbf{1}_{\{I_{2,i}=0\}} + \mathbf{1}_{\{I_{2,j}=0\}} \\ \mathbf{1}_{\{I_{2,i}=1\}} + \mathbf{1}_{\{I_{2,j}=1\}} \end{pmatrix} \begin{pmatrix} a_{ik} & \mathbf{1}_{\{I_{2,i}=0\}} + \mathbf{1}_{\{I_{2,k}=0\}} & \mathbf{1}_{\{I_{2,i}=1\}} + \mathbf{1}_{\{I_{2,k}=1\}} \end{pmatrix} \\ & = \sum_{i=1}^n w_i w_i^T, \end{aligned}$$

where

$$w_i = \sum_{j=1}^n \begin{pmatrix} a_{ij} \\ \mathbf{1}_{\{I_{2,i}=0\}} + \mathbf{1}_{\{I_{2,j}=0\}} \\ \mathbf{1}_{\{I_{2,i}=1\}} + \mathbf{1}_{\{I_{2,j}=1\}} \end{pmatrix}.$$

Theorem 1. *Under Model 1 or under Model 2, with the notation introduced above,*

$$\sqrt{n} \cdot \frac{\hat{A} - A}{\sqrt{S_A^2}} \to \mathcal{N}(0,1)$$

in distribution as $n \to \infty$.

The proof of this theorem will be given in Appendix A.

Corollary 1. *Assume Model 1 or Model 2 with the notation introduced above. Let* $g : (0,1) \to \mathbb{R}$ *be a* C^1*-function with* $g'(x) \neq 0$ *for all* $x \in (0,1)$. *Then*

$$\sqrt{n} \cdot \frac{g(\hat{A}) - g(A)}{g'(\hat{A}) \cdot \sqrt{S_A^2}} \to \mathcal{N}(0,1)$$

in distribution as $n \to \infty$.

The proof of this corollary will be given in Appendix A.

Let z_α be the α-quantile of the $\mathcal{N}(0,1)$-distribution.

Corollary 2. *Under Model 1 or under Model 2, with the notation introduced above, the interval*

$$\left(\hat{A} + z_{\alpha/2} \cdot \sqrt{\frac{S_A^2}{n}}, \hat{A} + z_{1-\alpha/2} \cdot \sqrt{\frac{S_A^2}{n}} \right)$$

has asymptotically the coverage probability $1 - \alpha$ *for A.*

This corollary is immediate from the theorem.

Corollary 3. *Assume Model 1 or Model 2 with the notation introduced above. Let* $g : (0,1) \to \mathbb{R}$ *be a bijective* C^1*-function with* $g'(x) > 0$ *for all* $x \in (0,1)$. *Then the interval*

$$\left(g^{-1}\left(g(\hat{A}) + z_{\alpha/2} \cdot g'(\hat{A}) \cdot \sqrt{\frac{S_A^2}{n}} \right), g^{-1}\left(g(\hat{A}) + z_{1-\alpha/2} \cdot g'(\hat{A}) \cdot \sqrt{\frac{S_A^2}{n}} \right) \right)$$

has asymptotically the coverage probability $1 - \alpha$ *for A.*

This corollary is immediate from Corollary 1.

For practice, we recommand to let g be the logit-function,

$$g(x) := \log\left(\frac{x}{1-x} \right), \quad x \in (0,1).$$

4 Simulations

In this section, we compare the performance of the new proposed confidence intervals from Corollary 2 and Corollary 3 with the logit-function as g to DeLong's interval [4, 13] and the Modified Wald interval [6] based on simulations. We consider two different scenarios, namely the binormal model which is classical for the investigation of the AUC and the fitting of a logistic regression model.

In the binormal model the procedure of fitting the model is ignored and it is assumed that the fitted values of the control observations and the fitted values of the

Table 1. Estimated coverage probabilities in the binormal model. In the first row we give the number n of observations and in the second row we give the expected value μ_1 of the fitted values of the case observations

	20	200	2000	20	200	2000
	1	1	1	2	2	2
Corollary 2	0.6154	0.9359	0.9494	0.0038	0.8772	0.9462
Corollary 3	0.5999	0.9389	0.9494	0.0000	0.8864	0.9463
DeLong	0.9026	0.9446	0.9505	0.7910	0.9369	0.9499
Modified Wald	0.9225	0.9543	0.9590	0.8577	0.9709	0.9797

Table 2. Mean value of the interval lengths in the binormal model. The further details are the same as for Table 1

	20	200	2000	20	200	2000
	1	1	1	2	2	2
Corollary 2	0.1911	0.1261	0.0412	0.0126	0.0602	0.0225
Corollary 3	0.1859	0.1258	0.0412	0.0125	0.0612	0.0225
DeLong	0.4280	0.1315	0.0414	0.2208	0.0721	0.0228
Modified Wald	0.4251	0.1365	0.0432	0.2496	0.0858	0.0272

case observations are normally distributed with different means. We assume that the fitted values of the case observations follow the $\mathcal{N}(0,1)$-distribution and the fitted values of the control observations follow the $\mathcal{N}(\mu_1,1)$-distribution for $\mu_1 = 1$ or $\mu_1 = 2$. We use $n = 20, 200, 2000$ observations in the test set of which one half belongs to the case group and the other half belongs to the control group. For each parameter combination we determined the coverage probability and the mean length of the confidence intervals based on 10,000 simulation runs. The true AUC needed to calculate the coverage probability was determined analytically. The results are reported in Table 1 and Table 2.

All confidence intervals have a too low coverage probability at small sample size. Not surprisingly, this gets better as the number of observation grows. We see that for a small sample size, the new confidence intervals are shorter than the ones reported in the literature at the price of having a lower coverage probability. For a large sample size there is hardly a difference between the new confidence intervals and DeLong's confidence intervals.

Now we consider the AUC from fitting a logistic regression model. For these simulations we assumed that $m + n = 100, 1000, 10000$ independent design points are drawn from a multivariate standard normal distribution in $\mathbb{R}^p$ for $p = 10, 100$. However, we dropped the combination $m + n = 100$ and $p = 100$, since then we have more parameters than observations. We let 80% of the observations be training data and 20% be test data; so $n = 20, 200, 2000$ observations are used for testing and the results are comparable to the results for the binormal model. We considered two models for the true class: a logistic regression model with the first unit vector as true parameter and a logistic regression model whose true parameter satisfies

$$\langle \beta_0, e_j \rangle = \frac{j - p/2}{\sqrt{\sum_{i=1}^{p} (i - p/2)^2}}, \quad j = 1, \ldots, p.$$

It is easily seen that the (absolute) probability that an observation is assigned to the

case class is one half—however, unlike for the binormal model, now the choice is made independently for each observation. For determining the coverage probability one has to decide what should be the target parameter. The approach considered in the literature so far takes

$$A_1 = \mathbb{P}\big(\beta^T X_1 < \beta^T X_2 \mid I_1 = 0, I_2 = 1\big)_{|\beta=\hat{\beta}} + \frac{1}{2} \cdot \mathbb{P}\big(\beta^T X_1 = \beta^T X_2 \mid I_1 = 0, I_2 = 1\big)_{|\beta=\hat{\beta}},$$

which is alright if you are interested in the quality of the fitted model. If you are interested in the quality of the classification algorithm, it makes more sense to consider

$$A_2 = \mathbb{P}\big(\beta_0^T X_1 < \beta_0^T X_2 \mid I_1 = 0, I_2 = 1\big) + \frac{1}{2} \cdot \mathbb{P}\big(\beta_0^T X_1 = \beta_0^T X_2 \mid I_1 = 0, I_2 = 1\big),$$

where β_0 is the true parameter. For each parameter combination we determined the coverage probability for A_1, the coverage probability for A_2 and the mean length of the confidence intervals based on 10,000 simulation runs. For simulating the true value of A_2 we generated a sample of 10^8 observations and for simulating the true value of A_1 we generated in each simulation run a sample of 10^6 observations. The results are reported in Tables 3–8.

We obtain essentially the same results as for the binormal model. All confidence intervals have a too low coverage probability at small sample size, but this gets better as the number of observation grows. At small sample size the new confidence intervals have a lower coverage probability and a shorter length than DeLong's intervals or the modified Wald intervals, while at a large sample size there is not much difference between the intervals. Moreover, we see that when A_2 is the target, we have a curse of dimensionality, i.e. the coverage probability drops at high dimensions. In particular, it is seriously too low for $p = 100$ and $m + n = 1000$ and a bit too low for $p = 100$ and $m + n = 10{,}000$. The results for the first unit vector as β_0 are quite similar to the results for the "skew" vector β_0. This is not surprising, since the first unit vector is

Table 3. Estimated coverage probabilities for A_1 in the logistic regression model with the first unit vector as true parameter. In the first row we give the dimension p and in the second row we present the number $m + n$ of observations

	10	10	10	100	100
	100	1000	10000	1000	10000
Corollary 2	0.727	0.937	0.948	0.939	0.951
Corollary 3	0.737	0.943	0.948	0.946	0.951
DeLong	0.915	0.946	0.949	0.946	0.952
Modified Wald	0.912	0.953	0.955	0.951	0.960

Table 4. Estimated coverage probabilities for A_2 in the logistic regression model with the first unit vector as true parameter. The further details are the same as in Table 3

	10	10	10	100	100
	100	1000	10000	1000	10000
Corollary 2	0.718	0.936	0.949	0.665	0.903
Corollary 3	0.691	0.935	0.950	0.616	0.894
DeLong	0.919	0.946	0.950	0.684	0.904
Modified Wald	0.909	0.950	0.957	0.689	0.915

Table 5. Mean value of the interval length for the logistic regression model with the first unit vector as true parameter. The further details are the same as in Table 3

	10 100	10 1000	10 10000	100 1000	100 10000
Corollary 2	0.2926	0.1329	0.0428	0.1428	0.0432
Corollary 3	0.2806	0.1325	0.0428	0.1421	0.0432
DeLong	0.4921	0.1379	0.0429	0.1473	0.0434
Modified Wald	0.4646	0.1415	0.0445	0.1490	0.0448

Table 6. Coverage probability of A_1 for the logistic regression model with the "skew" true parameter. The further details are the same as in Table 3

	10 100	10 1000	10 10000	100 1000	100 10000
Corollary 2	0.724	0.937	0.948	0.939	0.948
Corollary 3	0.734	0.944	0.949	0.945	0.950
DeLong	0.907	0.947	0.949	0.947	0.950
Modified Wald	0.905	0.955	0.957	0.951	0.957

Table 7. Coverage probability of A_2 for the logistic regression model with the "skew" true parameter. The further details are the same as in Table 3

	10 100	10 1000	10 10000	100 1000	100 10000
Corollary 2	0.716	0.940	0.949	0.661	0.897
Corollary 3	0.686	0.938	0.949	0.616	0.890
DeLong	0.912	0.948	0.950	0.681	0.898
Modified Wald	0.903	0.953	0.958	0.687	0.910

Table 8. Mean length for the logistic regression model with the "skew" true parameter. The further details are the same as in Table 3

	10 100	10 1000	10 10000	100 1000	100 10000
Corollary 2	0.2918	0.1329	0.0428	0.1428	0.0432
Corollary 3	0.2800	0.1324	0.0428	0.1420	0.0432
DeLong	0.4903	0.1379	0.0429	0.1472	0.0434
Modified Wald	0.4638	0.1415	0.0445	0.1490	0.0448

mapped by a rotation on the "skew" vector and both the logistic regression model and the distribution of the design points are invariant under rotations.

What can be done against the curse of dimensionality? Dimension reduction techniques like LASSO have the potential to mitigate the problem. In Tables 9–14 we report the results of LASSO logistic regression with $\lambda = 0.05$. The logistic regresssion with LASSO is no longer rotation invariant. Indeed, when the true parameter β_0 is the first unit vector, LASSO can be expected to work quite well, since we have one quite large entry and many zero entries. Under this easy parameter setting, LASSO provides a satisfactory solution. For the "skew" true parameter, LASSO can be expected to have problems, since there are many entries which are close to zero, but nonzero. Under this difficult parameter setting, the result with LASSO are even worse than the results without LASSO.

Table 9. Coverage probability for A_1 for LASSO logistic regression with the first unit vector as true parameter

	100 1000	100 10000
Corollary 2	0.935	0.953
Corollary 3	0.943	0.953
DeLong	0.945	0.954
Modified Wald	0.954	0.961

Table 10. Coverage probability for A_2 for LASSO logistic regression with the first unit vector as true parameter

	100 1000	100 10000
Corollary 2	0.935	0.953
Corollary 3	0.943	0.954
DeLong	0.945	0.954
Modified Wald	0.953	0.961

Table 11. Mean interval length for LASSO logistic regression with the first unit vector as true parameter

	100 1000	100 10000
Corollary 2	0.1315	0.0427
Corollary 3	0.1311	0.0427
DeLong	0.1366	0.0429
Modified Wald	0.1404	0.0444

Table 12. Coverage probability for A_1 for LASSO logistic regression with "skew" true parameter

	100 1000	100 10000
Corollary 2	0.9379	0.0227
Corollary 3	0.9423	0.0228
DeLong	0.9437	0.0228
Modified Wald	0.9432	0.0228

Table 13. Coverage probability for A_2 for LASSO logistic regression with "skew" true parameter

	100 1000	100 10000
Corollary 2	0.0055	0.0000
Corollary 3	0.0032	0.0000
DeLong	0.0067	0.0000
Modified Wald	0.0068	0.0000

Table 14. Mean interval length for LASSO logistic regression with "skew" true parameter

	100 1000	100 10000
Corollary 2	0.15580	0.00124
Corollary 3	0.15458	0.00124
DeLong	0.15938	0.00124
Modified Wald	0.15881	0.05062

Table 15. Bias and standard deviation of the AUC in the binormal model

	20 1	200 1	2000 1	20 2	200 2	2000 2
bias	6.11e-04	9.10e-05	1.27e-04	2.09e-04	7.36e-05	5.69e-05
standard deviation	0.10805	0.03341	0.01050	0.06137	0.01853	0.00578

Table 16. Bias and standard deviation of the AUC in the logistic regression model

	10 100	10 1000	10 10000	100 1000	100 10000
mean of A	0.690	0.733	0.739	0.682	0.732
A1	0.684	0.733	0.739	0.682	0.732
A2	0.74	0.74	0.74	0.74	0.74
bias to target A1	6.22e-03	5.29e-05	1.57e-05	1.15e-04	7.12e-05
bias to target A2	0.049483	0.006789	0.000745	0.057935	0.007481
standard deviation	0.1171	0.0352	0.0110	0.0385	0.0110

It is a natural question, whether these confidence intervals can be further improved by bias reduction. In order to assess that, we investigated the bias and the standard deviation under the model assumptions explained above. The results are reported in Table 15 and Table 16. We see that, while the bias in the binormal model and the bias to the target A_1 in the logistic regression model are neglectable, there is a considerable bias to A_2 in the logistic regression model.

5 Real data application

In this section we apply the confidence intervals to medical data.

We want to predict the presence of an obstructive coronary artery disease from ECGs and from seven risk factors (age, sex, systolic blood pressure, LDL, diabetes, smoking status, family history). Of the ECGs we extracted 648 features using the MUSE(TM) (General Electrics, Boston, US) algorithm yielding 648 explanatory variables. The seven risk factors lead eight explanatory variables, since we decided to split the family history in two variables ("present vs. absent or unknown" and "unknown vs. present or absent"). Notice that four of these risk factors are binary and thus, strictly speaking, the assumptions of Model 2 are not fulfilled.

We had data from 283,897 ECGs conducted at the University Hospital of Essen. Since we need to know the true classification, we combined this data with the ECAD registry containing the results of 33,865 coronary angiographies. We found a matching coronary angiography for 13,538 ECGs. The patients, to which these ECGs belong, were assigned to the training group with probability 0.6 and to the test group with probability 0.4 independently of each other. This resulted in 8136 coronary angiographies being assigned to the training group and 5402 coronary angiographies being assigned to the test group.

We fitted a logistic regression model based on the training group and we calculated the AUC together with 95%-confidence intervals to predict obstructive coronary artery disease as detected in subsequently preformed coronary angiography procedures. When the prediction was based on the ECGs, the AUC for the training group was 0.709 and the AUC for the test group was 0.578. We got an AUC for the training

Table 17. A comparison of different confidence intervals for the AUC for the diagnosis of an obstructive CAD for the full data of 13,538 coronary angiographies via logistic regression models

Method	ECG (training data)	ECG (test data)	Risk factors (training data)	Risk factors (test data)
AUC	0.709	0.578	0.595	0.581
Corollary 2	(0.697; 0.721)	(0.562; 0.594)	(0.582; 0.608)	(0.565; 0.597)
Corollary 3	(0.697; 0.721)	(0.561; 0.594)	(0.581; 0.608)	(0.565; 0.597)
DeLong	(0.697; 0.721)	(0.562; 0.594)	(0.582; 0.608)	(0.565; 0.597)
Modified Wald	(0.698; 0.721)	(0.563; 0.593)	(0.582; 0.607)	(0.566; 0.596)

Table 18. A comparison of different confidence intervals for the AUC for the diagnosis of an obstructive CAD for the reduced data of 100 coronary angiographies via logistic regression models

Method	ECG (training data)	ECG (test data)	Risk factors (training data)	Risk factors (test data)
AUC	1	0.378	0.751	0.543
Corollary 2	(1; 1)	(0.229; 0.527)	(0.602; 0.901)	(0.346; 0.739)
Corollary 3	(1; 1)	(0.244; 0.534)	(0.576; 0.870)	(0.350; 0.724)
DeLong	(1; 1)	(0.214; 0.542)	(0.575; 0.927)	(0.330; 0.755)
Modified Wald	(1; 1)	(0.225; 0.531)	(0.607; 0.896)	(0.386; 0.700)

group of 0.595 and for the test group of 0.581 for the prediction of an obstructive CAD from seven risk factors.

The results are reported in Table 17. Though strictly speaking outside the scope of this article, we added the results for the training group. In order to see how the confidence intervals behave on a smaller sample, we applied our methods to a subsample consisting of 100 coronary angiographies. The results are shown in Table 18.

For the whole sample all confidence intervals have approximately the same length—the new intervals have the same length as the ones from the literature and the intervals based on the ECGs have the same length as the ones based on the seven risk factors. Not surprisingly, as we reduce the number of observations, the intervals get longer. In particular, for all 13,538 coronary angiographies the logistic regression model is significantly better than a pure random choice (i.e. an AUC of 0.5), which is no longer true if we use only 100 coronary angiographies. For the subsample the new confidence intervals are slightly narrower than the ones from the literature.

In Tables 19–21 we look what happens, when one uses neural nets, random forests or support vector machines instead of logistic regression models. We see that

Table 19. Neural nets. The further details are the same as in Table 17

Method	ECG (training data)	ECG (test data)	Risk factors (training data)	Risk factors (test data)
AUC	0.725	0.587	0.635	0.622
Corollary 2	(0.713; 0.737)	(0.571; 0.604)	(0.622; 0.648)	(0.606; 0.637)
Corollary 3	(0.713; 0.737)	(0.571; 0.604)	(0.622; 0.648)	(0.606; 0.637)
DeLong	(0.713; 0.737)	(0.571; 0.604)	(0.622; 0.648)	(0.606; 0.637)
Modified Wald	(0.714; 0.736)	(0.572; 0.602)	(0.623; 0.647)	(0.607; 0.636)

Table 20. Random forests. The further details are the same as in Table 17

Method	ECG (training data)	ECG (test data)	Risk factors (training data)	Risk factors (test data)
AUC	0.999	0.599	1	0.572
Corollary 2	(0.999; 0.999)	(0.583; 0.615)	(1; 1)	(0.556; 0.588)
Corollary 3	(0.999; 0.999)	(0.582; 0.615)	(1; 1)	(0.556; 0.588)
DeLong	(0.999; 0.999)	(0.583; 0.615)	(1; 1)	(0.556; 0.588)
Modified Wald	(0.999; 1)	(0.584; 0.614)	(1; 1)	(0.557; 0.587)

Table 21. Support vector machines. The further details are the same as in Table 17

Method	ECG (training data)	ECG (test data)	Risk factors (training data)	Risk factors (test data)
AUC	0.997	0.571	0.618	0.608
Corollary 2	(0.996; 0.998)	(0.555; 0.588)	(0.605; 0.631)	(0.592; 0.624)
Corollary 3	(0.996; 0.998)	(0.555; 0.588)	(0.605; 0.631)	(0.592; 0.624)
DeLong	(0.996; 0.998)	(0.555; 0.588)	(0.605; 0.631)	(0.592; 0.624)
Modified Wald	(0.996; 0.999)	(0.556; 0.586)	(0.606; 0.63)	(0.593; 0.623)

the confidence intervals are slightly shifted due to the different values of the point estimates, but that they all have approximately the same length as the confidence intervals of the logistic regression model.

In order to evaluate the computation times for the confidence intervals, observe that their computation is a two-step procedure. First, the chosen model estimator is used to calculate the fitted values $\hat{Y}_{2,i}$, $i = 1, \ldots, n$, and in the second step the confidence intervals are calculated from these numbers. So the total computation time of a confidence interval is the sum of one component which does depend on the model estimator, but not on the confidence interval method, and one component which does depend on the confidence interval method, but not on the model estimator. The computation times are reported in Table 22 and Table 23. We see that the computation times for the new intervals are longer than for those from the literature, but that also the computation of the new confidence intervals is feasible. In particular, for random forests and support vector machines the difference between the new computation times and the old ones is neglectable compared to the time needed for the calculation of the fitted values $\hat{Y}_{2,i}$, $i = 1, \ldots, n$, anyway.

Table 22. Computation time (in seconds) for the whole sample (13,538 patients)

Method	ECG (training data)	ECG (test data)	Risk factors (training data)	Risk factors (test data)
logistic regression	26.23	26.13	0.56	0.49
neural net	4.50	4.68	2.52	2.09
random forest	164.74	165.06	3.78	3.84
support vector machines	856.41	857.25	181.66	177.82
AUC	0.01	0.00	0.01	0.01
Corollary 2	71.00	34.50	74.28	38.77
Corollary 3	73.74	35.30	81.97	38.60
DeLong	0.42	0.26	0.51	0.27
Modified Wald	0.01	0.00	0.01	0.00

Table 23. Computation time (in seconds) for the subsample of 100 patients

Method	ECG (training data)	ECG (test data)	Risk factors (training data)	Risk factors (test data)
logistic regression	0.44	0.42	0.01	0.02
AUC	0.00	0.00	0.00	0.00
Corollary 2	0.13	0.01	0.01	0.01
Corollary 3	0.16	0.01	0.01	0.02
DeLong	0.04	0.00	0.00	0.00
Modified Wald	0.01	0.00	0.00	0.00

6 Discussion and outlook

In this paper we have taken into account two facts that are usually ignored in the study of confidence intervals for the AUC. First, only the total size of the test cohort can be controlled, while its splitting into the case and control groups is random. Second, the fitted binarization model is itself subjected to random effects. The first fact brought new confidence intervals that are narrower than the ones in the literature, but have a too low coverage probability at a small sample size. The second fact did not bring new confidence intervals, since we saw that the confidence intervals we got from considering the first fact still had asymptotically the correct coverage probability under Model 2. All what we changed was that we had to add additional parts to the proofs (the ones that we have only for Model 2 and not for Model 1). It can be expected that in a similar manner the confidence intervals proposed in the literature have asymptotically the correct coverage probability under Model 2.

Can it be expected for other binarization algorithms as well that the old confidence intervals still have asymptotically the correct coverage probability when the model uncertainty is taken into account? The estimators in linear discriminant analysis and in quadratic discriminant analysis are combinations of standard estimators. Hence central limit theorems for these estimators are easily established and from there on it is straightforward to generalize the results of the present article. For quadratic discriminant analysis a certain challenge will be that the set of all test points $(X_1, X_2) \in \mathbb{R}^p \times \mathbb{R}^p$ for which $Y_1 < Y_2$, but $\hat{Y}_1 > \hat{Y}_2$, will be more complicated than for logistic regression models or for linear discriminant analysis. For algorithms from machine learning, like neural nets, random forests and support vector machines, a first problem already occurs in the definition of the theoretical AUC. Since there is only an algorithm and no underlying probability model, we cannot define the theoretical AUC as a probability like we have done for logistic regression models. One could define the theoretical AUC as the average of many independent realizations of the empirical AUC or as the limit of the empirical AUC as the sample size tends to infinity (provided one can show that this limit exists). Still with either of these definitions, the proof will be much harder. The set of all test points $(X_1, X_2) \in \mathbb{R}^p \times \mathbb{R}^p$ for which $Y_1 < Y_2$, but $\hat{Y}_1 > \hat{Y}_2$, will be much more complicated for a machine learning algorithm than it was in our proof. Moreover, we used a central limit theorem for the estimator in a logistic regression model, and central limit theorems are unknown for machine learning algorithms.

While the theoretical results tell that asymptotically the old confidence intervals work under Model 2 as well, our simulation results tell that at small sample size these confidence intervals may have a seriously too low coverage probability—recall in

particular the results in Table 4 for $p = 100$. Hence the construction of new confidence intervals is desirable. A tempting idea is to use the δ-method in the same way as we use it in the proof. However, the derivative in Lemma 3 is zero and hence one will end up with the old confidence intervals when using this approach. A solution would be to use the second-order δ-method (see, e.g., [1, Lemma 5]). However, this may appear to be inelegant. The second-order δ-method yields that the limiting distribution of the AUC is a sum of squares of Gaussian random variables, but since not all Gaussian random variables involved in that sum have the same variance, this sum will not be χ^2-distributed in general. It is not clear whether a closed-form expression for the variances of these Gaussian random variables can be derived even in the ideal situation when the design points are multivariate-normally distributed or distributed uniformly on the ball. In the realistic situation, when the distribution of the design points is unknown and has to be estimated, it will even be a challenge to propose an algorithm that gives a reasonable approximation for these variances in acceptable time. The results for the LASSO logistic regression ranged from providing a satisfactory solution to being even worse than the pure logistic regression depending on the unknown true model parameter. Bootstrap [3] is known to have good finite-sample properties in many instances and hence would be another approach worth trying. Finally, our simulations in Table 16 show that the estimator $\hat{A}$ is seriously biased for A_2. Hence one can think of constructing an estimator for the bias of $\hat{A}$ for A_2 and then applying bias reduction.

7 The equality of the Mann–Whitney intervals and DeLong's intervals

Here we prove that the Mann–Whitney intervals due to Sen [13] coincide with DeLong's intervals [4]. For any real-valued sample $a_1, a_2, \ldots, a_N$, let $a_{(1)}, a_{(2)}, \ldots, a_{(N)}$ denote the ordered sample, i.e. the sample containing the same elements (with the same multiplicity), such that $a_{(1)} \leq a_{(2)} \leq \cdots \leq a_{(N)}$. We let $n_0 := |\{i \in \{1, \ldots, n\} \mid I_i = 0\}|$ denote the number of observations in the control group and $n_1 := |\{i \in \{1, \ldots, n\} \mid I_i = 1\}|$ the number of observations in the case group. Since Sen [13] and DeLong et al. [4] do not consider the training group, it is needless to say that we only mean observations of the test group here. We let $X_i, i = 1, \ldots, n_0$, be the observations of the control group—not to be confused with the design points of the logistic regression model, for which we used the same symbol—and $Y_j, j = 1, \ldots, n_1$, the observations of the case group. The Mann–Whitney intervals are defined as follows. We let

$$\begin{aligned} R_i &:= \left|\left\{k \in \{1, \ldots, n_0\} \mid X_k \leq X_{(i)}\right\}\right| + \left|\left\{j \in \{1, \ldots, n_1\} \mid Y_j \leq X_{(i)}\right\}\right|, \\ &\quad i = 1, \ldots, n_0, \\ S_j &:= \left|\left\{i \in \{1, \ldots, n_0\} \mid X_i \leq Y_{(j)}\right\}\right| + \left|\left\{k \in \{1, \ldots, n_1\} \mid Y_k \leq Y_{(j)}\right\}\right|, \\ &\quad j = 1, \ldots, n_1, \end{aligned}$$

denote the rank of $X_{(i)}$ or $Y_{(j)}$ respectively within the joint sample of control and case observations. Put

$$\bar{R} := \frac{1}{n_0} \sum_{i=1}^{n_0} R_i, \qquad \bar{S} := \frac{1}{n_1} \sum_{j=1}^{n_1} S_j,$$

$$S_{10}^2 = \frac{1}{(n_0-1)\cdot n_1^2}\cdot\left(\sum_{i=1}^{n_0}(R_i-i)^2 - n_0\cdot\left(\bar{R}-\frac{n_0+1}{2}\right)^2\right),$$
$$S_{01}^2 = \frac{1}{(n_1-1)\cdot n_0^2}\cdot\left(\sum_{j=1}^{n_1}(S_j-j)^2 - n_1\cdot\left(\bar{S}-\frac{n_1+1}{2}\right)^2\right),$$
$$\hat{\sigma}_M^2 = \frac{n_1\cdot S_{10}^2 + n_0\cdot S_{01}^2}{n_0\cdot n_1}.$$

Let z_α be the α-quantile of the standard normal distribution for $\alpha\in(0,1)$. Then

$$\left(\hat{A}+z_{\alpha/2}\cdot\sqrt{\hat{\sigma}_M^2}, \hat{A}+z_{1-\alpha/2}\cdot\sqrt{\hat{\sigma}_M^2}\right)$$

is the Mann–Whitney confidence interval. In order to define DeLong's intervals, put

$$V_{10}(y) := \frac{1}{n_1}\cdot\sum_{j=1}^{n_1}\left(\mathbf{1}_{\{y<Y_j\}}+\frac{1}{2}\cdot\mathbf{1}_{\{y=Y_j\}}\right),$$
$$V_{01}(y) := \frac{1}{n_0}\cdot\sum_{i=1}^{n_0}\left(\mathbf{1}_{\{X_i<y\}}+\frac{1}{2}\cdot\mathbf{1}_{\{X_i=y\}}\right),$$
$$\hat{\sigma}_D^2 = \frac{1}{n_0\cdot(n_0-1)}\sum_{i=1}^{n_0}\left(V_{10}(X_i)-\hat{A}\right)^2 + \frac{1}{n_1\cdot(n_1-1)}\sum_{j=1}^{n_1}\left(V_{01}(Y_j)-\hat{A}\right)^2.$$

Then DeLong's interval is

$$\left(\hat{A}+z_{\alpha/2}\cdot\sqrt{\hat{\sigma}_D^2}, \hat{A}+z_{1-\alpha/2}\cdot\sqrt{\hat{\sigma}_D^2}\right).$$

Theorem 2. *For any real-valued sample $(X_1,\dots,X_{n_0},Y_1,\dots,Y_{n_1})$ that does not contain ties, it holds that*

$$\hat{\sigma}_M^2 = \hat{\sigma}_D^2$$

and, in particular, the Mann–Whitney interval and DeLong's interval coincide.

This theorem will be proven in Appendix B.

A Proofs for Section 3

Corollary 2 is immediate from Theorem 1 and Corollary 3 is immediate from Corollary 1.

We start with the proof of Corollary 1 (taking Theorem 1 for granted) and then prove Theorem 1.

In order to prove Corollary 1, we need the following slight extension of the δ-method.

Lemma 1. *Let $X_n, n\in\mathbb{N}$, be a sequence of $\mathbb{R}^d$-valued random vectors, $C_n, n\in\mathbb{N}$, be a sequence of random numbers with $C_n\to\infty$ as $n\to\infty$ in probability, $\mu\in\mathbb{R}^d$ and X be an $\mathbb{R}^d$-valued random vector such that*

$$C_n(X_n-\mu)\to X$$

in distribution as $n \to \infty$. Let $g : \mathbb{R}^d \to \mathbb{R}^D$ be a C^1-function. Then

$$C_n \cdot \big(g(X_n) - g(\mu)\big) \to g'(\mu)X.$$

Proof of Corollary 1. By Theorem 1 the assumptions of Lemma 1 are fulfilled with $C_n := \sqrt{n/S_A^2}$ and $X \sim \mathcal{N}(0,1)$. Hence

$$\sqrt{n} \cdot \frac{g(\hat{A}) - g(A)}{g'(A) \cdot \sqrt{S_A^2}} \to \mathcal{N}(0,1)$$

in distribution as $n \to \infty$, which is the desired result except that we have $g'(A)$ instead of $g'(\hat{A})$ in the denominator. However, Theorem 1 in particular implies

$$\hat{A} - A = \sqrt{\frac{S_A^2}{n}} \cdot \sqrt{\frac{n}{S_A^2}}(\hat{A} - A) \to 0 \cdot X = 0$$

in probability as $n \to \infty$. So $g'(\hat{A}) \to g'(A)$ in probability as $n \to \infty$ by the continuous mapping theorem and Slutzky's theorem gives

$$\sqrt{n} \cdot \frac{g(\hat{A}) - g(A)}{g'(\hat{A}) \cdot \sqrt{S_A^2}} = \frac{g'(A)}{g'(\hat{A})} \cdot \sqrt{n} \cdot \frac{g(\hat{A}) - g(A)}{g'(A) \cdot \sqrt{S_A^2}} \to \mathcal{N}(0,1). \qquad \square$$

Proof of Lemma 1. By the definition of differentiability there is a function $r : \mathbb{R}^d \to \mathbb{R}^D$ with $\lim_{x \to \mu} r(x)/\|x - \mu\| = 0$ such that

$$g(x) - g(\mu) = g'(\mu)(x - \mu) + r(x).$$

Hence

$$C_n \cdot \big(g(X_n) - g(\mu)\big) = g'(\mu)\big(C_n \cdot (X_n - \mu)\big) + C_n \cdot r(X_n).$$

Now the definition of convergence in probability implies $1/C_n \to 0$ as $n \to \infty$ and therefore Slutzky's theorem yields

$$X_n = \frac{1}{C_n} \cdot C_n \cdot (X_n - \mu) + \mu \to 0 \cdot X + \mu = \mu$$

in probability as $n \to \infty$. By a sharp version of the continuous mapping theorem [2, Theorem 2.7] we get

$$\frac{r(X_n)}{\|X_n - \mu\|} \to 0$$

in probability as $n \to \infty$. So

$$C_n \cdot r(X_n) = C_n \cdot \|X_n - \mu\| \cdot \frac{r(X_n)}{\|X_n - \mu\|} \to \|X\| \cdot 0 = 0$$

in probability as $n \to \infty$ and Slutzky's theorem implies

$$C_n \cdot \big(g(X_n) - g(\mu)\big) \to g'(\mu)X. \qquad \square$$

Now then Corollary 1 is proven, we turn to the proof of Theorem 1. We start with the proof of two lemmata.

Let $\mathcal{H}^j$ denote the j-dimensional Hausdorff measure. Intuitively, $\mathcal{H}^1(A)$ is the length of a 1-dimensional set $A \subseteq \mathbb{R}^d$, $\mathcal{H}^2(A)$ is the area of a two-dimensional set $A \subseteq \mathbb{R}^d$, and so on. See [9] for a rigorous introduction.

Lemma 2. *Let $\beta_0, \beta_1 \in \mathbb{R}^d$, $\beta_0 \neq 0$, and put*

$$D := \left\{x \in S^{d-1} \mid \beta_0^T x < 0, \beta_1^T x > 0\right\} \cup \left\{x \in S^{d-1} \mid \beta_0^T x > 0, \beta_1^T x < 0\right\}.$$

Then

$$\mathcal{H}^{d-1}(D) \leq \omega_{d-2} \cdot \min\left\{\pi \cdot \frac{\|\beta_1 - \beta_0\|}{\|\beta_0\|}, \pi\right\},$$

where

$$\omega_{d-2} = \mathcal{H}^{d-2}\left(S^{d-2}\right) = \frac{2\pi^{(d-1)/2}}{\Gamma(\frac{d-1}{2})}.$$

Proof of Lemma 2. If β_1 is a nonnegative multiple of β_0, then $D = \emptyset$ and the assertion is fulfilled. If β_0 is a negative multiple of β_0, then $\|\beta_1 - \beta_0\|/\|\beta_0\| > 1$ and hence the assertion is fulfilled. Otherwise, put

$$\beta_t := \cos(t) \cdot \frac{\beta_0}{\|\beta_0\|} + \sin(t) \cdot \frac{\beta_1 - \langle \beta_1, \beta_0 \rangle \cdot \frac{\beta_0}{\|\beta_0\|^2}}{\|\beta_1 - \langle \beta_1, \beta_0 \rangle \cdot \frac{\beta_0}{\|\beta_0\|^2}\|}.$$

Then

$$D \subseteq \bigcup_{t \in [0, \arccos(\langle \frac{\beta_1}{\|\beta_1\|}, \frac{\beta_0}{\|\beta_0\|} \rangle)]} \left\{x \in S^{d-1} \mid \beta_t^T x = 0\right\}.$$

Hence the area formula ([9, Theorem 3.7]) yields

$$\begin{aligned}\mathcal{H}^{d-1}(D) &\leq \int_0^{\arccos(\langle \frac{\beta_1}{\|\beta_1\|}, \frac{\beta_0}{\|\beta_0\|} \rangle)} \mathcal{H}^{d-2}\left(\left\{x \in S^{d-1} \mid \beta_t^T x = 0\right\}\right) dt \\ &= \arccos\left(\left\langle \frac{\beta_1}{\|\beta_1\|}, \frac{\beta_0}{\|\beta_0\|} \right\rangle\right) \cdot \omega_{d-2}.\end{aligned}$$

Now

$$\cos(x) \leq 1 - \frac{2}{\pi^2} \cdot x^2, \quad x \in [0, \pi],$$

yields

$$\arccos\left(\left\langle \frac{\beta_1}{\|\beta_1\|}, \frac{\beta_0}{\|\beta_0\|} \right\rangle\right) \leq \sqrt{\frac{\pi^2}{2} \cdot \left(1 - \left\langle \frac{\beta_1}{\|\beta_1\|}, \frac{\beta_0}{\|\beta_0\|} \right\rangle\right)}.$$

Further

$$\begin{aligned}&1 - \left\langle \frac{\beta_1}{\|\beta_1\|}, \frac{\beta_0}{\|\beta_0\|} \right\rangle \\ &= \frac{1}{2} \cdot \left(\left\| \frac{\beta_1}{\|\beta_1\|} \right\|^2 + \left\| \frac{\beta_0}{\|\beta_0\|} \right\|^2 - 2 \cdot \left\langle \frac{\beta_1}{\|\beta_1\|}, \frac{\beta_0}{\|\beta_0\|} \right\rangle\right)\end{aligned}$$

$$= \frac{1}{2} \cdot \left\| \frac{\beta_1}{\|\beta_1\|} - \frac{\beta_0}{\|\beta_0\|} \right\|^2 \leq \frac{1}{2} \cdot \left(\frac{\|\beta_1 - \beta_0\|}{\|\beta_0\|} + \|\beta_1\| \cdot \frac{|\|\beta_0\| - \|\beta_1\||}{\|\beta_1\| \cdot \|\beta_0\|} \right)^2$$
$$\leq 2 \cdot \frac{\|\beta_1 - \beta_0\|^2}{\|\beta_0|^2}.$$

Hence

$$\mathcal{H}^{d-1}(D) \leq \omega_{d-2} \cdot \min\left\{ \pi \cdot \frac{\|\beta_1 - \beta_0\|}{\|\beta_0\|}, \pi \right\}. \qquad \square$$

Lemma 3. *Put*

$$z(\beta) := \mathbb{P}\big(\beta^T X_1 < \beta^T X_2, I_1 = 0, I_2 = 1\big).$$

Then, under Model 2, z is differentiable on $\mathbb{R}^p \setminus \{0\}$ with

$$\frac{d}{d\beta} z(\beta)_{|\beta=\beta_0} = 0.$$

Proof of Lemma 3. Recall that q is the Lebesgue density of a design point X_1 and that $\pi(t) = 1/(\exp\{t\} + 1)$. For any $\beta \in \mathbb{R}^p$ it holds that

$$\begin{aligned} z(\beta) &= \mathbb{P}\big(\beta^T X_1 < \beta^T X_2, I_1 = 0, I_2 = 1\big) \\ &= \int_{\mathbb{R}^p} \int_{\mathbb{R}^p} \mathbf{1}_{\{0<\beta^T(x_2-x_1)\}} \cdot \pi\big(\beta_0^T x_1\big) \cdot q(x_1) \cdot \big(1 - \pi\big(\beta_0^T x_2\big)\big) \cdot q(x_2)\, dx_1\, dx_2 \\ &= \int_{\mathbb{R}^p} \mathbf{1}_{\{0<\beta^T v\}} \int_{\mathbb{R}^p} \pi\big(\beta_0^T w\big) \cdot q(w) \cdot \big(1 - \pi\big(\beta_0^T (v+w)\big)\big) \cdot q(v+w)\, dw\, dv. \end{aligned}$$

Now consider $\beta_t = \beta + tu$ with $u \in S^{p-1}$ and $t > 0$. If $u = \pm\beta/\|\beta\|$, then $z(\beta_t) = z(\beta)$ for $|t| < 1/\|\beta\|$. So assume that u and β are linearly independent now. Put

$$u' := \frac{u - \frac{\langle u,\beta\rangle}{\|\beta\|^2}\beta}{\|u - \frac{\langle u,\beta\rangle}{\|\beta\|^2}\beta\|}.$$

Let M be an orthogonal matrix mapping $\beta/\|\beta\|$ to the first unit vector and u' to the second unit vector. Then

$$\begin{aligned} &\int_{\mathbb{R}^p} \mathbf{1}_{\{0<\beta_t^T v\}} \int_{\mathbb{R}^p} \pi\big(\beta_0^T w\big) \cdot q(w) \cdot \big(1 - \pi\big(\beta_0^T (v+w)\big)\big) \cdot q(v+w)\, dw\, dv \\ &\quad = \int_{\mathbb{R}^p} \mathbf{1}_{\{0<\|\beta\| \cdot x_1 + t \cdot \langle u,u'\rangle \cdot x_2\}} \int_{\mathbb{R}^p} \pi\big(\beta_0^T w\big) \cdot q(w) \\ &\qquad \cdot \big(1 - \pi\big(\beta_0^T (Mx+w)\big)\big) \cdot q(Mx+w)\, dw\, dx. \end{aligned}$$

Now the fundamental theorem of calculus implies

$$\begin{aligned} &\frac{d}{dt} \int_{\mathbb{R}^p} \mathbf{1}_{\{0<\|\beta\| \cdot x_1 + t \cdot \langle u,u'\rangle \cdot x_2\}} \int_{\mathbb{R}^p} \pi\big(\beta_0^T w\big) \cdot q(w) \\ &\qquad \cdot \big(1 - \pi\big(\beta_0^T (Mx+w)\big)\big) \cdot q(Mx+w)\, dw\, dx_{|t=0} \\ &\quad = \int_{\mathbb{R}^{p-1}} \frac{\langle u,u'\rangle \cdot x_2'}{\|\beta\|} \int_{\mathbb{R}^p} \pi\big(\beta_0^T w\big) \cdot q(w) \cdot \big(1 - \pi\big(\beta_0^T \big(M\big(0,x'\big)+w\big)\big)\big) \end{aligned}$$

$$\cdot\, q\big(M\,(0,x')+w\big)\,dw\,dx'$$
$$= \int_{\{v\in\mathbb{R}^p\,|\,\beta^T v=0\}} \frac{\langle v,u\rangle}{\|\beta\|} \int_{\mathbb{R}^p} q(w)\cdot\pi\big(\beta_0^T w\big)\cdot q(v+w)\cdot\big(1-\pi\big(\beta_0^T(v+w)\big)\big)dw\,dv,$$

because $\langle u,u'\rangle\cdot\langle v,u'\rangle=\langle v,u\rangle$. Since this expression is continuous in β, we conclude that z is differentiable with

$$\frac{d}{d\beta}z(\beta)=\int_{\{v\in\mathbb{R}^p\,|\,\beta^T v=0\}}\frac{v}{\|\beta\|}\int_{\mathbb{R}^p} q(w)\cdot\pi\big(\beta_0^T w\big)\cdot q(v+w)\cdot\big(1-\pi\big(\beta_0^T(v+w)\big)\big)\,dw\,dv.$$

Since for v with $\beta_0^T v=0$ it holds that

$$\int_{\mathbb{R}^p} q(w)\cdot\pi\big(\beta_0^T w\big)\cdot q(v+w)\cdot\big(1-\pi\big(\beta_0^T(v+w)\big)\big)\,dw$$
$$=\int_{\mathbb{R}^p} q(w)\cdot\pi\big(\beta_0^T w\big)\cdot q(-v+w)\cdot\big(1-\pi\big(\beta_0^T(-v+w)\big)\big)\,dw,$$

we conclude that $d/d\beta\, z(\beta)_{|\beta=\beta_0}=0$. □

Proof of Theorem 1. We start by proving Theorem 1 under Model 1.

Put

$$g\big((y_1,i_1),(y_2,i_2)\big):=\begin{pmatrix}(\mathbf{1}_{\{y_1<y_2\}}+\frac{1}{2}\mathbf{1}_{\{y_1=y_2\}})\cdot(1-i_1)\cdot i_2\\ 1-i_1\\ i_2\end{pmatrix},$$
$$h\big((y_1,i_1),(y_2,i_2)\big)=\frac{1}{2}\cdot g\big((y_1,i_1),(y_2,i_2)\big)+\frac{1}{2}\cdot g\big((y_2,i_2),(y_1,i_1)\big)$$

and

$$U=\frac{2}{n\cdot(n-1)}\sum_{i=1}^{n}\sum_{j=i+1}^{n} h\big((Y_{2,i},I_{2,i}),(Y_{2,j},I_{2,j})\big).$$

Now [7, Section 5.1.1, Theorem 1] implies

$$\sqrt{n}\cdot(U-\theta)\to\mathcal{N}(0,\Sigma)$$

in distribution as $n\to\infty$, where

$$\theta=\begin{pmatrix}\mathbb{P}(Y_1<Y_2,I_1=0,I_2=1)+\frac{1}{2}\cdot\mathbb{P}(Y_1=Y_2,I_1=0,I_2=1)\\ \mathbb{P}(I_1=0)\\ \mathbb{P}(I_1=1)\end{pmatrix}$$

is the expected value vector and where Σ denotes the covariance matrix of $\mathbb{E}[h((Y_1,I_1),(Y_2,I_2))|Y_1,I_1]$—recall that (Y_1,I_1,Y_2,I_2) has the same distribution as $(Y_{2,i},I_{2,i},Y_{2,j},I_{2,j})$ for $i\neq j$.

Put

$$f:\mathbb{R}^3\to\mathbb{R},\quad (x_1,x_2,x_3)\mapsto\frac{x_1}{x_2\cdot x_3}$$

so that $\hat{A}=n/(n-1)\cdot f(U)$. Now the δ-method implies

$$\sqrt{n}\cdot(\hat{A}-A)\to\mathcal{N}\big(0,\sigma_A^2\big),$$

where

$$\sigma_A^2 = f'(\theta)^T \Sigma f'(\theta) = v^T \Sigma v.$$

It follows from [7, pp. 163, 164] that $\hat{\Sigma}$ (defined on page 61) is a consistent estimator for Σ. Hence we get that S_A^2 is a consistent estimator for σ_A^2. So Slutzky's theorem implies the assertion.

Now we turn to Model 2. Put

$$\Psi_n(\beta) := \sqrt{n} \cdot \Big(Z_n(\beta) - \mathbb{P}\big(\beta^T X_1 < \beta^T X_2 \mid I_1 = 0, I_2 = 1\big) - \frac{1}{2} \cdot \mathbb{P}\big(\beta^T X_1 = \beta^T X_2 \mid I_1 = 0, I_2 = 1\big) \Big),$$

where

$$Z_n(\beta) := \frac{1}{|\{i \in \{1, \ldots, n\} \mid I_{2,i} = 0\}|} \cdot \frac{1}{|\{j \in \{1, \ldots, n\} \mid I_{2,j} = 1\}|} \cdot \sum_{i,j=1}^{n} \big(\mathbf{1}_{\{\beta^T X_{2,i} < \beta^T X_{2,j}, I_{2,i}=0, I_{2,j}=1\}} + 1/2 \cdot \mathbf{1}_{\{\beta^T X_{2,i} = \beta^T X_{2,j}, I_{2,i}=0, I_{2,j}=1\}} \big).$$

Then $Z_n(\hat{\beta})$ is the empirical AUC. Put

$$z(\beta) = \mathbb{P}\big(\beta^T X_1 < \beta^T X_2 \mid I_1 = 0, I_2 = 1\big) + 1/2 \cdot \mathbb{P}\big(\beta^T X_1 = \beta^T X_2 \mid I_1 = 0, I_2 = 1\big)$$

and

$$R := \sqrt{n} \cdot \big(z(\hat{\beta}) - z(\beta_0) \big).$$

The aim is to derive a central limit theorem for

$$\Psi_n(\hat{\beta}) + R.$$

We treat the two summands separately. Put

$$g\big((x_1, i_1), (x_2, i_2); \beta\big) := \begin{pmatrix} (\mathbf{1}_{\{\beta^T x_1 < \beta^T x_2\}} + \frac{1}{2} \mathbf{1}_{\{\beta^T x_1 = \beta^T x_2\}}) \cdot (1 - i_1) \cdot i_2 \\ 1 - i_1 \\ i_2 \end{pmatrix},$$

$$h\big((x_1, i_1), (x_2, i_2); \beta\big) = \frac{1}{2} \cdot g\big((x_1, i_1), (x_2, i_2); \beta\big) + \frac{1}{2} \cdot g\big((x_2, i_2), (x_1, i_1); \beta\big)$$

and

$$U(\beta) = \frac{2}{n \cdot (n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} h\big((X_{2,i}, I_{2,i}), (X_{2,j}, I_{2,j}); \beta\big).$$

Now [7, Section 5.1.1, Theorem 1] implies

$$\sqrt{n} \cdot \big(U(\beta_0) - \theta \big) \to \mathcal{N}(0, \Sigma)$$

in distribution as $n \to \infty$, where

$$\theta = \begin{pmatrix} z(\beta_0) \\ \mathbb{P}(I_1 = 0) \\ \mathbb{P}(I_1 = 1) \end{pmatrix}$$

is the expected value vector and Σ is the covariance matrix of $\mathbb{E}[h((X_1, I_1), (X_2, I_2); \beta_0) \mid X_1, I_1]$. We would like to show that also

$$\sqrt{n} \cdot \big(U(\hat{\beta}) - \hat{\theta}\big) \to \mathcal{N}(0, \Sigma), \tag{2}$$

where

$$\hat{\theta} = \begin{pmatrix} z(\hat{\beta}) \\ \mathbb{P}(I_1 = 0) \\ \mathbb{P}(I_1 = 1) \end{pmatrix}.$$

For this, by Slutzky's theorem it suffices to show that

$$\sqrt{n} \cdot \big(U_1(\hat{\beta}) - \mathbb{E}\big[U_1(\hat{\beta})|\hat{\beta}\big] - U_1(\beta_0) + \mathbb{E}\big[U_1(\beta_0)\big]\big) \to 0 \tag{3}$$

in probability as $n \to \infty$, where $U_1(\beta)$ denotes the first component of $U(\beta)$—notice the second and third components do not depend on β. At first, however, consider $\sqrt{n}(U_1(\beta_1) - \mathbb{E}[U_1(\beta_1) - U_1(\beta_0) + \mathbb{E}U_1(\beta_0))$ for a deterministic point $\beta_1 \in \mathbb{R}^p$ which is not a multiple of β_0. For $\epsilon > 0$ we have by the Chebychev inequality that

$$\mathbb{P}\big(\sqrt{n} \cdot \big|\big(U_1(\beta_1) - \mathbb{E}U_1(\beta_1)\big) - \big(U_1(\beta_0) - \mathbb{E}U_1(\beta_0)\big)\big| > \epsilon\big) \le \frac{n}{\epsilon^2} \cdot \mathsf{Var}\big(U_1(\beta_1) - U_1(\beta_0)\big).$$

Similar to (3.71) in [7] one gets

$$\begin{aligned}\mathsf{Var}\big(U_1(\beta_1) - U_1(\beta_0)\big) \le \frac{8}{n} \cdot \mathsf{Var}\Bigg(&\left(\mathbf{1}_{\{\beta_1^T X_1 < \beta_1^T X_2\}} + \frac{1}{2}\mathbf{1}_{\{\beta_1^T X_1 = \beta_1^T X_2\}}\right)\mathbf{1}_{\{I_1=0\}}\mathbf{1}_{\{I_2=1\}} \\ &- \left(\mathbf{1}_{\{\beta_0^T X_1 < \beta_0^T X_2\}} + \frac{1}{2}\mathbf{1}_{\{\beta_0^T X_1 = \beta_0^T X_2\}}\right)\mathbf{1}_{\{I_1=0\}}\mathbf{1}_{\{I_2=1\}}\Bigg) + \frac{8}{n^2}.\end{aligned}$$

Put

$$D := \big\{x \in S^{p-1} \mid \beta_1^T x < 0, \beta_0^T x > 0\big\} \cup \big\{x \in S^{p-1} \mid \beta_1^T x > 0, \beta_0^T x < 0\big\}$$

and denote by $\bar{D}$ the closure of D and by $\mathsf{relbd}\, D$ the boundary of D relative to S^{p-1} as surrounding topological space. Neglecting the event $\{X_1 = X_2\}$ which occurs with probability 0, we have

$$\begin{aligned}&\left|\mathbf{1}_{\{\beta_1^T X_1 < \beta_1^T X_2\}} + \frac{1}{2}\mathbf{1}_{\{\beta_1^T X_1 = \beta_1^T X_2\}} - \mathbf{1}_{\{\beta_0^T X_1 < \beta_0^T X_2\}} - \frac{1}{2}\mathbf{1}_{\{\beta_0^T X_1 = \beta_0^T X_2\}}\right| \\ &\quad = \begin{cases} 1 & \text{if } \frac{X_2 - X_1}{\|X_2 - X_1\|} \in D \\ \frac{1}{2} & \text{if } \frac{X_2 - X_1}{\|X_2 - X_1\|} \in \mathsf{relbd}\, D \\ 0 & \text{if } \frac{X_2 - X_1}{\|X_2 - X_1\|} \notin \bar{D} \end{cases}.\end{aligned}$$

Hence

$$\begin{aligned}\mathsf{Var}\Bigg(&\left(\mathbf{1}_{\{\beta_1^T X_1 < \beta_1^T X_2\}} + \frac{1}{2}\mathbf{1}_{\{\beta_1^T X_1 = \beta_1^T X_2\}}\right)\mathbf{1}_{\{I_1=0\}}\mathbf{1}_{\{I_2=1\}} \\ &- \left(\mathbf{1}_{\{\beta_0^T X_1 < \beta_0^T X_2\}} + \frac{1}{2}\mathbf{1}_{\{\beta_0^T X_1 = \beta_0^T X_2\}}\right)\mathbf{1}_{\{I_1=0\}}\mathbf{1}_{\{I_2=1\}}\Bigg)\end{aligned}$$

$$\leq \mathbb{E}\left[\mathbf{1}_{\bar{D}}\left(\frac{X_2 - X_1}{\|X_2 - X_1\|}\right)\right] \leq M \cdot \mathcal{H}^{p-1}(D),$$

where $\mathcal{H}^{p-1}(D)$ denotes the $(p-1)$-dimensional Hausdorff measure of D and where M is the upper bound of the density of $(X_2 - X_1)/\|X_2 - X_1\|$ with respect to the $(p-1)$-dimensional Hausdorff measure (recall that by the assumptions of Model 2 such a bound exists).

Therefore Lemma 2 implies

$$\begin{aligned}&\mathbb{P}\left(\sqrt{n} \cdot \left|U_1(\beta_1) - \mathbb{E}U_1(\beta_1) - U_1(\beta_0) + \mathbb{E}U_1(\beta_0)\right| > \epsilon\right)\\ &\quad\leq \frac{8}{\epsilon^2} \cdot M \cdot \omega_{p-2} \cdot \min\left\{\pi \cdot \frac{\|\beta_1 - \beta_0\|}{\|\beta_0\|}, \pi\right\} + \frac{8}{n \cdot \epsilon^2}.\end{aligned}$$

$$\begin{aligned}&\mathbb{P}\left(\sqrt{n} \cdot \left|U_1(\hat{\beta}) - \mathbb{E}U_1(\hat{\beta}) - U_1(\beta_0) + \mathbb{E}U_1(\beta_0)\right| > \epsilon\right)\\ &\quad= \mathbb{E}\left[\mathbb{P}\left(\sqrt{n} \cdot \left|U_1(\hat{\beta}) - \mathbb{E}\left[U_1(\hat{\beta})|\hat{\beta}\right] - U_1(\beta_0) + \mathbb{E}U_1(\beta_0)\right| > \epsilon \mid \hat{\beta}\right)\right]\\ &\quad\leq \mathbb{E}\left[\frac{8}{\epsilon^2} \cdot M \cdot \omega_{p-2} \cdot \min\left\{\pi \cdot \frac{\|\hat{\beta} - \beta_0\|}{\|\beta_0\|}, \pi\right\}\right] + \frac{8}{n \cdot \epsilon^2} \to 0 \qquad (4)\end{aligned}$$

as $n \to \infty$. So (3) holds, which concludes the proof of (2). Now the δ-method implies

$$\Psi_n(\hat{\beta}) \to \mathcal{N}\left(0, \sigma_A^2\right),$$

where σ_A^2 is as in the proof for Model 1. Let us turn to R. Recall that by [5, p. 203] $\hat{\beta}$ obeys a central limit theorem as $m \to \infty$. Under Model 2 we have

$$\mathbb{P}\left(\beta^T X_1 = \beta^T X_2 | I_1 = 0, I_2 = 1\right) = 0$$

for all $\beta \in \mathbb{R}^p \setminus \{0\}$ and thus

$$\begin{aligned}z(\beta) &= \mathbb{P}\left(\beta^T X_1 < \beta^T X_2 \mid I_1 = 0, I_2 = 1\right) + \frac{1}{2} \cdot \mathbb{P}\left(\beta^T X_1 = \beta^T X_2 \mid I_1 = 0, I_2 = 1\right)\\ &= \mathbb{P}\left(\beta^T X_1 < \beta^T X_2 \mid I_1 = 0, I_2 = 1\right).\end{aligned}$$

Hence the δ-method together with Lemma 3 and assumption (1) gives

$$R = \sqrt{n} \cdot \left(z(\hat{\beta}) - z(\beta_0)\right) \to 0$$

in probability as $n \to \infty$. This shows

$$\sqrt{n} \cdot \left(Z_n(\hat{\beta}) - z(\beta_0)\right) \to \mathcal{N}\left(0, \sigma_A^2\right).$$

It remains to estimate σ_A^2. A consistent pseudo-estimator for Σ is given by

$$\begin{aligned}\tilde{\Sigma} = &\frac{1}{n \cdot (n-1) \cdot (n-2)} \sum_{i,j,k=1}^{n} \begin{pmatrix} \tilde{a}_{ij} \\ \mathbf{1}_{\{I_{2,i}=0\}} + \mathbf{1}_{\{I_{2,j}=0\}} \\ \mathbf{1}_{\{I_{2,i}=1\}} + \mathbf{1}_{\{I_{2,j}=1\}} \end{pmatrix}\\ &\times \begin{pmatrix} \tilde{a}_{ik} & \mathbf{1}_{\{I_{2,i}=0\}} + \mathbf{1}_{\{I_{2,k}=0\}} & \mathbf{1}_{\{I_{2,i}=1\}} + \mathbf{1}_{\{I_{2,k}=1\}} \end{pmatrix}\end{aligned}$$

$$- \left(\frac{1}{n \cdot (n-1)} \sum_{i,j=1}^{n} \begin{pmatrix} \tilde{a}_{ij} \\ \mathbf{1}_{\{I_{2,i}=0\}} + \mathbf{1}_{\{I_{2,j}=0\}} \\ \mathbf{1}_{\{I_{2,i}=1\}} + \mathbf{1}_{\{I_{2,j}=1\}} \end{pmatrix} \right)$$
$$\times \left(\frac{1}{n \cdot (n-1)} \sum_{i,j=1}^{n} \begin{pmatrix} \tilde{a}_{ij} & \mathbf{1}_{\{I_{2,i}=0\}} + \mathbf{1}_{\{I_{2,j}=0\}} & \mathbf{1}_{\{I_{2,i}=1\}} + \mathbf{1}_{\{I_{2,j}=1\}} \end{pmatrix} \right)$$

with

$$\tilde{a}_{ij} = \mathbf{1}_{\{\beta_0^T X_{2,i} < \beta_0^T X_{2,j}\}} \cdot \mathbf{1}_{\{I_{2,i}=0\}} \cdot \mathbf{1}_{\{I_{2,j}=1\}} + \frac{1}{2} \cdot \mathbf{1}_{\{\beta_0^T X_{2,i} = \beta_0^T X_{2,j}\}} \cdot \mathbf{1}_{\{I_{2,i}=0\}} \cdot \mathbf{1}_{\{I_{2,j}=1\}}$$
$$+ \mathbf{1}_{\{\beta_0^T X_{2,j} < \beta_0^T X_{2,i}\}} \cdot \mathbf{1}_{\{I_{2,j}=0\}} \cdot \mathbf{1}_{\{I_{2,i}=1\}} + \frac{1}{2} \cdot \mathbf{1}_{\{\beta_0^T X_{2,j} = \beta_0^T X_{2,i}\}} \cdot \mathbf{1}_{\{I_{2,j}=0\}} \cdot \mathbf{1}_{\{I_{2,i}=1\}}$$

(see [7, pp. 163, 164]). However, $\tilde{\Sigma}$ is only a pseudo-estimator, since it depends on β_0 which is unknown when working with real data. So we have to use the estimator $\hat{\Sigma}$ in which β_0 is replaced with $\hat{\beta}$. By the Markov inequality we have for all $\epsilon > 0$ and all coordinates $r, s = 1, 2, 3$ that

$$\mathbb{P}\left(|\tilde{\Sigma}_{rs} - \hat{\Sigma}_{rs}| > \epsilon\right)$$
$$\leq \frac{1}{\epsilon} \cdot \mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{\{\hat{\beta}^T X_1 < \hat{\beta}^T X_2\}} + \frac{1}{2} \mathbf{1}_{\{\hat{\beta}^T X_1 = \hat{\beta}^T X_2\}} - \mathbf{1}_{\{\beta_0^T X_1 < \beta_0^T X_2\}} - \frac{1}{2} \mathbf{1}_{\{\beta_0^T X_1 = \beta_0^T X_2\}} \mid \hat{\beta} \right] \right]$$

and hence

$$\lim_{n \to \infty} \mathbb{P}\left(|\tilde{\Sigma}_{rs} - \hat{\Sigma}_{rs}| > \epsilon\right) = 0$$

holds in the same way as (4). So $\hat{\Sigma}$ is a consistent estimator for Σ. Therefore $S_A^2 \to \sigma_A^2$ in probability, and Slutzky's theorem implies

$$\sqrt{n} \cdot \frac{\hat{A} - A}{\sqrt{S_A^2}} \to \mathcal{N}(0, 1). \qquad \square$$

B Proof for Section 7

Proof of Theorem 2. It holds that

$$R_i - i = \sum_{j=1}^{n_1} \mathbf{1}_{\{Y_j \leq X_{(i)}\}} = n_1 - n_1 \cdot V_{10}(X_{(i)})$$

and

$$S_j - j = \sum_{i=1}^{n_0} \mathbf{1}_{\{X_i \leq Y_{(j)}\}} = n_0 \cdot V_{01}(Y_{(j)}).$$

Hence

$$\bar{R} - \frac{n_0 + 1}{2} = \frac{1}{n_0} \cdot \sum_{i=1}^{n_0} (R_i - i) = \frac{1}{n_0} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} \mathbf{1}_{\{Y_j \leq X_i\}} = n_1 - n_1 \cdot \hat{A}$$

and

$$\bar{S} - \frac{n_1+1}{2} = \frac{1}{n_1} \cdot \sum_{j=1}^{n_1} (S_j - j) = \frac{1}{n_1} \sum_{j=1}^{n_1} \sum_{i=1}^{n_0} \mathbf{1}_{\{X_i \le Y_j\}} = n_0 \cdot \hat{A}.$$

So

$$\begin{aligned}
S_{10}^2 &= \frac{1}{(n_0-1)\cdot n_1^2} \left[\sum_{i=1}^{n_0} (R_i - i)^2 - n_0 \cdot \left(\bar{R} - \frac{n_0+1}{2} \right)^2 \right] \\
&= \frac{1}{(n_0-1)\cdot n_1^2} \left[\sum_{i=1}^{n_0} \left(n_1 - n_1 \cdot V_{10}(X_{(i)}) \right)^2 \right. \\
&\quad \left. - 2 \left(\sum_{i=1}^{n_0} \left(n_1 - n_1 \cdot V_{10}(X_{(i)}) \right) \right) \cdot (n_1 - n_1 \cdot \hat{A}) + n_0 \cdot (n_1 - n_1 \cdot \hat{A})^2 \right] \\
&= \frac{1}{(n_0-1)\cdot n_1^2} \left[\sum_{i=1}^{n_0} \left(\left(n_1 - n_1 \cdot V_{10}(X_{(i)}) \right) - (n_1 - n_1 \cdot \hat{A}) \right)^2 \right] \\
&= \frac{1}{n_0 - 1} \sum_{i=1}^{n_0} \left(V_{10}(X_i) - \hat{A} \right)^2
\end{aligned}$$

and

$$\begin{aligned}
S_{01}^2 &= \frac{1}{(n_1-1)\cdot n_0^2} \left[\sum_{j=1}^{n_1} (S_j - j)^2 - n_1 \cdot \left(\bar{S} - \frac{n_1+1}{2} \right)^2 \right] \\
&= \frac{1}{(n_1-1)\cdot n_0^2} \left[\sum_{j=1}^{n_1} \left(n_0 \cdot V_{01}(Y_{(j)}) \right)^2 \right. \\
&\quad \left. - 2 \cdot \sum_{j=1}^{n_1} \left(n_0 \cdot V_{01}(Y_{(j)}) \right) \cdot n_0 \cdot \hat{A} + n_1 \cdot (n_0 \cdot \hat{A})^2 \right] \\
&= \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} \left(V_{01}(Y_j) - \hat{A} \right)^2.
\end{aligned}$$

This yields

$$\begin{aligned}
\hat{\sigma}_M^2 &= \frac{n_0 \cdot S_{01}^2 + n_1 \cdot S_{10}^2}{n_0 \cdot n_1} \\
&= \frac{1}{n_1 \cdot (n_1-1)} \sum_{j=1}^{n_1} \left(V_{01}(Y_j) - \hat{A} \right)^2 + \frac{1}{n_0 \cdot (n_0-1)} \sum_{i=1}^{n_0} \left(V_{10}(X_i) - \hat{A} \right)^2 = \hat{\sigma}_D^2.
\end{aligned}$$

In particular, the two confidence intervals are equal. □

Supplementary Material

The file AUC_CI.R contains all confidence intervals mentioned in this article—the new ones proposed here and the ones used in the simulation study for comparison.

The files Simulation_binormal.R, Simulation_logistic.R, Siumlation_logistic_2_fast.R, Simulation_LASSO.R, Simulation_LASSO_2_fast.R, Simulation_binormal_bias.R and Simulation_logistic_bias.R contain the source code for the simulations reported in this article.

References

[1] Afendras, G., Papadatos, N., Piperigou, V.: On the limiting distribution of sample central moments. Ann. Inst. Stat. Math. **72**, 399–425 (2020). MR4067230. https://doi.org/10.1007/s10463-018-0695-4

[2] Billingsley, P.: Covergence of Probability Measures. Wiley (1999). MR1700749. https://doi.org/10.1002/9780470316962

[3] Carpenter, J., Bithell, J.: Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. Stat. Med. **19**, 1141–1164 (2000). https://doi.org/10.1002/(SICI)1097-0258(20000515)19:9<1141::AID-SIM479>3.0.CO;2-F

[4] DeLong, E., DeLong, D., Clarke-Pearson, D.: Comparing the areas under two or more operating characteristic curves: A nonparametric approach. Biometrics **44**, 837–845 (1988). https://doi.org/10.2307/2531595

[5] Fahrmeir, L., Kneib, T., Lang, S.: Regression. Modelle, Methoden und Anwendungen. Springer (2007)

[6] Kottas, M., Kuss, O., Zapf, A.: A modified Wald interval for the area under the ROC curve (AUC) in diagnostic case-control studies. BMC Med. Res. Methodol. **14**, 26 (2014) (9 pages). https://doi.org/10.1186/1471-2288-14-26

[7] Kowalski, J., Tu, X.: Modern Applied U-Statistics. John Wiley & Sons (2007). MR2368050

[8] LeDell, E., Peterson, M., v. d. Laan, M.: Computationally efficient confidence intervals for cross-validated area under the ROC curve estimates. Electron. J. Stat. **9**, 1583–1607 (2015). MR3376118. https://doi.org/10.1214/15-EJS1035

[9] Morgan, F.: Geometric Measure Theory—A Beginner's Guide. Elsevier (2016). MR3497381

[10] Noma, H., Shinozaki, T., Iba, K., Teramukai, S., Furukawa, T.: Confidence intervals of prediction accuracy measures for multivariable prediction models based on the bootstrap-based optimism correction methods. Stat. Med. **40**, 5691–5701 (2021). MR4330574. https://doi.org/10.1002/sim.9148

[11] Pepe, M.: The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press (2003). MR2260483

[12] Qin, G., Hotilovac, L.: Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. Stat. Methods Med. Res. **17**, 207–221 (2008). MR2432389. https://doi.org/10.1177/0962280207087173

[13] Sen, P.K.: A note on asymptotically distribution-free confidence bounds $P(X < Y)$, based on two samples. Sankhya **29**, 95–102 (1967). MR0226772