Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/46/20241317

Comparison of methods for calculating confidence intervals of AUC in ROC curve considering sampling error

Yuxuan She[1,4,7], Jiahao Cui[2,5] and Xinran Liu[3,6]

[1]School of Mathematical Science, Peking University, No.5 Yiheyuan Road, Beijing, People's Republic of China
[2]Department of Mathematics, Southeast University, No.2 Southeast University Road, Nanjing, Jiangsu, People's Republic of China
[3]Department of Land Economy, University of Cambridge,16-21 Silver St, Cambridge, United Kingdom

[4]pku-accelerator@pku.edu.cn
[5]213234115@seu.edu.cn
[6]xl548@cantab.ac.uk
[7]corresponding author

Abstract. The Receiver Operating Characteristic (ROC) curve is a crucial method for evaluating the effectiveness of diagnostic medical indicators and has found extensive applications. However, errors are inevitable in the data acquisition process. Therefore, discussions on error and various methods for improving and handling data have not only become the focus of academic discourse but also hold practical significance. Unlike general statistics, the diversity of error situations, ranges, and impacts in biostatistics often present unique challenges. In practical scenarios, such as drug experiments, limited sample sizes and variations in individual responses to the same drug necessitate the use of error models, data scales, and statistical processing based on historical data, biomedical knowledge, and experimental data. Furthermore, the choice of an appropriate method depends on the specific objectives of the experiment, which is essential for producing compelling conclusions. Importantly, the field of biology has introduced methods to address errors, such as cross-comparison experiments or repeated experiments, and data processing must adapt to changes in experimental designs. This paper presents a statistical approach based on the widely used practice of error reduction through repeated experiments in the context of assessing generic drug consistency. The paper first summarizes the common types of errors encountered in biostatistics and the corresponding analytical, control, and optimization measures. It explores several methods for calculating the Area Under the ROC Curve (AUC) when sampling error is introduced and applies error reduction through repeated experiments. Subsequently, the paper validates the methods under different error scenarios using simulated data, highlighting the suitability of different statistical models and their reasons for selection in cases where the difference between healthy and diseased populations is not substantial. This paper offers valuable insights into handling various types of real-world data to eliminate errors and obtain more accurate statistical conclusions.

Keywords: ROC curve, AUC, confidence interval, sampling error.

DOI: 10.54254/2755-2721/46/20241317

1. Introduction

The ROC curve, as an effective method for evaluating the effectiveness of using continuous medical indicators to determine health status, is of paramount importance in diagnostic medicine. The Area Under the ROC Curve (AUC) is considered a significant measure to assess the effectiveness of this method [1]. For such diagnostic methods, a threshold is chosen. Values above this threshold are considered indicative of disease, while values below it are indicative of non-disease [2]. For example, in the diagnosis of hypertension, if a patient's systolic blood pressure exceeds 140 mmHg and diastolic blood pressure exceeds 90 mmHg, they are diagnosed with hypertension. Similarly, in the diagnosis of coronary artery disease, when more than 50% of a patient's blood vessels are blocked during a cardiac angiogram, they are considered to have heart disease. These are examples of using continuous medical indicators for diagnosis. However, before determining the threshold, it is essential to verify the method's effectiveness, which involves showing that the values for healthy individuals are lower than those for patients. Bamber demonstrated that the AUC represents Pr(Y > X) [3], where X and Y are the measurements for healthy and unhealthy populations, respectively. AUC = 0.5 indicates that the method is no different from random chance in distinguishing healthy and unhealthy individuals, rendering the metric meaningless. AUC values closer to 1 signify a greater diagnostic effectiveness. Normally, under parametric assumptions, the normal distribution is widely applied, while in non-parametric cases, AUC is estimated using the Mann-Whitney statistic. Regardless of the scenario, AUC remains the most critical metric.

However, as a critical assessment of medical indicators, errors in the data must be considered. For most medical measurements, errors are introduced due to various factors such as external conditions and instrument limitations. Neglecting errors in the estimation of AUC significantly reduces reliability. Investigating the influence of errors on AUC contributes to a more accurate understanding of the method's effectiveness. Many articles have discussed various properties of AUC estimation, errors, and confidence intervals. Beyond the application of statistical methods to handle errors, methods involving multiple measurements have also been proposed to address error issues [3]. This paper explores several methods for estimating confidence intervals of AUC in the ROC curve and assesses their utility and characteristics using simulated data. It compares the performance of different methods under various data characteristics, offering recommendations for selecting AUC estimation methods in different situations where the difference between healthy and diseased populations is not substantial.

2. AUC Confidence Interval Estimation Based on David Faraggi's Method

Under the assumption that true values for measurements in the healthy populationUnder the assumption that true values for measurements in the healthy population, U_i, follow a normal distribution with parameters μ_X and ${\sigma_X}^2$, and true values for measurements in the diseased population, W_i, follow a normal distribution with parameters μ_Y and ${\sigma_Y}^2$, we define $A = Pr(X < Y) = \Phi(\delta)$, where $\delta = \frac{\mu_Y - \mu_X}{\sqrt{{\sigma_X}^2 + {\sigma_Y}^2}}$. Our observed values are $x_i = U_i + \varepsilon_i$ and $y_j = W_j + \eta_j$, where $\varepsilon_i \sim \mathrm{N}(0, {\sigma_\varepsilon}^2)$ and $\eta_j \sim N(0, {\sigma_\eta}^2)$, both following normal distributions. It is important to note that U, W, ε, η are mutually independent. Our confidence interval estimation is based on the assumption that ${\sigma_\varepsilon}^2 = {\sigma_\eta}^2$ and ${\sigma_U}^2 = {\sigma_W}^2 = \sigma^2$ [4].

When incorporating errors, the value of AUC, denoted as $A^* = Pr(y > x)$, can be expressed as $A^* = \Phi(\delta^*)$, where $\delta^* = \frac{\delta}{\sqrt{1+\theta^2}}$ and $\theta^2 = \frac{{\sigma_\varepsilon}^2}{\sigma^2}$. Consequently, the confidence interval estimation for A^* can be derived from the confidence interval for δ^*. To obtain both confidence intervals, we utilize the combined variance estimate ${S_p}^2$, as described by [5]:

${S_p}^2 = \frac{(m-1){S_x}^2 + (n-1){S_y}^2}{m+n-2}$ follows a χ^2 distribution with $m + n - 2$ degrees of freedom, where ${S_x}^2$ and ${S_y}^2$ are the sample variances for X and Y.

DOI: 10.54254/2755-2721/46/20241317

Additionally, $\frac{\bar{y}-\bar{x}}{\sigma\sqrt{1+\theta^2}\sqrt{\frac{1}{m}+\frac{1}{n}}}$ follows a normal distribution $N(\frac{\sqrt{2}\delta^*}{\sqrt{\frac{1}{m}+\frac{1}{n}}}, 1)$, where $\bar{x}$ and $\bar{y}$ are the sample means for X and Y. Since these two variables are independent, the ratio $t = \frac{\bar{y}-\bar{x}}{S_p\sqrt{\frac{1}{m}+\frac{1}{n}}}$ follows a t -distribution with $m+n-2$ degrees of freedom, denoted as $t_{m+n-2}(\lambda)$, where $\lambda = \frac{\sqrt{2}\delta^*}{\sqrt{\frac{1}{m}+\frac{1}{n}}}$.

The confidence interval for λ, with confidence level (1-α), is determined by the upper and lower limits, $\bar{\lambda}$ and $\underline{\lambda}$, which can be obtained using a non-central t-distribution with $m+n-2$ degrees of freedom:

$$Pr(t_{m+n-2}(\underline{\lambda}) \leq t) = 1-\frac{\alpha}{2}, \quad Pr(t_{m+n-2}(\bar{\lambda}) \leq t) = \frac{\alpha}{2}$$

$$(\underline{\delta^*}, \overline{\delta^*}) = \frac{\sqrt{\frac{1}{m}+\frac{1}{n}}}{\sqrt{2}}(\underline{\lambda}, \bar{\lambda})$$

The confidence interval for A^* can be calculated as $\left(\phi\left(\underline{\delta^*}\right), \phi(\overline{\delta^*})\right)$, where ϕ represents the standard normal distribution function.

Without considering errors, the confidence interval for A, denoted as $\left(\phi\left(\sqrt{1+\theta^2}\underline{\delta^*}\right), \phi(\sqrt{1+\theta^2}\overline{\delta^*})\right)$, can be calculated. It is evident that, as the error proportion increases, its impact on the final confidence interval becomes significant.

In practical applications, obtaining variance data is not straightforward. In fact, to use this method, each individual must be measured n times to estimate ${\sigma_\varepsilon}^2$ from the variation in data obtained from individual experiments and then estimate σ^2 from the overall variance. Subsequently, the average of measurements for each individual is used to estimate the AUC confidence interval. Regarding measurement errors, since measurements are averaged, ${\sigma_\varepsilon}^2$ must be transformed to $\frac{{\sigma_\varepsilon}^2}{n}$ for calculations [6].

3. Confidence Interval Estimation for Repeated Experiments AUC Based on Yanhong Li et al.'s Calculation Method

Repeated experiments are an important method to reduce the impact of errors; however, data processing after repeated experiments becomes more complex. The following discusses data processing after conducting repeated experiments. It's worth noting that if the confidence interval calculation method described below is not used and data is directly processed using methods like the Delta-method by Thomas and Hultquist [6], the results may turn out to be poor [7].

3.1. Discussion on Confidence Interval Calculation

For existing confidence intervals (l_1, u_1) and (l_2, u_2) with confidence level (1-α) for θ_1 and θ_2, and their point estimates $\widehat{\theta_1}$ and $\widehat{\theta_2}$, under the assumption of mutual independence, we can directly calculate the confidence interval for $\theta_1-\theta_2$ (L, U) as follows [8]:

$$L = \widehat{\theta_1} - \widehat{\theta_2} - z\sqrt{var(\widehat{\theta_1}) + var(\widehat{\theta_2})} \quad U = \widehat{\theta_1} - \widehat{\theta_2} + z\sqrt{var(\widehat{\theta_1}) + var(\widehat{\theta_2})}$$

Here, z is the critical value corresponding to the (1-α) confidence interval in the standard normal distribution. However, the confidence interval obtained in this manner tends to be too wide. To optimize it, we examine the distance between l_1-u_2 and L, which can be calculated as:

$$z\left\|\sqrt{var(\widehat{\theta_1}) + var(\widehat{\theta_2})} - \sqrt{var(\widehat{\theta_1})} - \sqrt{var(\widehat{\theta_2})}\right\|$$

DOI: 10.54254/2755-2721/46/20241317

This distance is less than the distance between $\widehat{\theta_1} - \widehat{\theta_2}$ and L: $z\left\|\sqrt{var(\widehat{\theta_1}) + var(\widehat{\theta_2})}\right\|$. Similarly, the distance between $u_1 - l_2$ and U is also less than the distance between $\widehat{\theta_1} - \widehat{\theta_2}$ and U. Therefore, when estimating the variances $\widehat{var}(\widehat{\theta_i}) = \frac{(\hat{\theta}_i - \theta_i)^2}{z^2}$, where $\theta_1 = l_1$ for L and u_1 for U, and $\widehat{\theta_2}$ is u_2 for L and l_2 for U, we have:

$$L_1 = \widehat{\theta_1} - \widehat{\theta_2} - z\sqrt{\widehat{var}(\widehat{\theta_1}) + \widehat{var}(\widehat{\theta_2})} = \widehat{\theta_1} - \widehat{\theta_2} - \sqrt{(\widehat{\theta_1} - l_1)^2 + (\widehat{\theta_2} - u_2)^2}$$

$$U_1 = \widehat{\theta_1} - \widehat{\theta_2} + z\sqrt{\widehat{var}(\widehat{\theta_1}) + \widehat{var}(\widehat{\theta_2})} = \widehat{\theta_1} - \widehat{\theta_2} + \sqrt{(\widehat{\theta_1} - u_1)^2 + (\widehat{\theta_2} - l_2)^2}$$

Similarly, for the confidence interval for $\theta_1 + \theta_2$ (L_2, U_2):

$$L_2 = \widehat{\theta_1} + \widehat{\theta_2} - z\sqrt{\widehat{var}(\widehat{\theta_1}) + \widehat{var}(\widehat{\theta_2})} = \widehat{\theta_1} + \widehat{\theta_2} - \sqrt{(\widehat{\theta_1} - l_1)^2 + (\widehat{\theta_2} - l_2)^2}$$

$$U_2 = \widehat{\theta_1} + \widehat{\theta_2} + z\sqrt{var(\widehat{\theta_1}) + var(\widehat{\theta_2})} = \widehat{\theta_1} + \widehat{\theta_2} + \sqrt{(\widehat{\theta_1} - u_1)^2 + (\widehat{\theta_2} - u_2)^2}$$

Finally, to calculate the confidence interval for $\frac{\theta_1}{\theta_2}$ (L_3, U_3), where $R = \frac{\theta_1}{\theta_2}$, we examine $\theta_1 - R\theta_2 = 0$. The lower and upper limits of the confidence interval (L_3, U_3) are determined as:

$$L_3 = \widehat{\theta_1} - R\,\widehat{\theta_2} - \sqrt{(\widehat{\theta_1} - l_1)^2 + R^2(\widehat{\theta_2} - u_2)^2}$$

$$U_3 = \widehat{\theta_1} - R\,\widehat{\theta_2} - \sqrt{(\widehat{\theta_1} - u_1)^2 + R^2(\widehat{\theta_2} - l_2)^2}$$

Thus, the confidence interval for R is obtained by solving $L_3 = 0$ and $U_3 = 0$, providing the smaller and larger roots as the confidence interval for R.

$$L_4 = \frac{\widehat{\theta_1\theta_2} - \sqrt{(\widehat{\theta_1\theta_2})^2 - l_1u_2(2\widehat{\theta_1} - l_1)(2\widehat{\theta_2} - u_2)}}{u_2(2\widehat{\theta_2} - u_2)}$$

$$U_4 = \frac{\widehat{\theta_1\theta_2} + \sqrt{(\widehat{\theta_1\theta_2})^2 - l_2u_1(2\widehat{\theta_1} - u_1)(2\widehat{\theta_2} - l_2)}}{l_2(2\widehat{\theta_2} - l_2)}$$

3.2. Applying Confidence Interval Calculation to Determine AUC Confidence Interval for Multiple Repeated Experiments

As mentioned in Section 2, the confidence interval for AUC is obtained using the $\left(\phi(\underline{\delta}), \phi(\overline{\delta})\right)$ method. Therefore, the calculation of the confidence interval is still based on determining the confidence interval for δ^*, where $\delta = \frac{\mu_Y - \mu_X}{\sqrt{\sigma_X{}^2 + \sigma_Y{}^2}}$.

Here, ω_{ij} represents the j-th observation for the i-th individual in the healthy group (similar calculations apply to the diseased group), where $j = 1, \cdots, k_i$,$i = 1, \cdots, n$.

$$\omega_{ij} = X_i + \varepsilon_{ij} \ , \ \bar{\omega}_{i\cdot} = \sum_j \frac{\omega_{ij}}{k_i} \ , \bar{\omega}_{\cdot\cdot} = \sum_i \frac{\omega_{i\cdot}}{n}, \ k_h = \frac{n}{\sum_i \frac{1}{k_i}}$$

Following the results of Thomas and Hultquist [6]:

DOI: 10.54254/2755-2721/46/20241317

$$\frac{k_h \sum_i (\bar{\omega}_{i.} - \bar{\omega}_{..})^2}{k_h {\sigma_x}^2 + {\sigma_\varepsilon}^2} \sim {\chi^2}_{n-1}, \quad \frac{\sum_i \sum_j (\bar{\omega}_{i.} - \omega_{ij})^2}{{\sigma_\varepsilon}^2} \sim {\chi^2}_{N-n}, \quad \bar{\omega}_{..} \sim N(\mu_{x,} \frac{\sum_i (\bar{\omega}_{i.} - \bar{\omega}_{..})^2}{n(n-1)})$$

Specific calculation method: Calculate the confidence interval for $\theta_1 = \mu_Y - \mu_X$ using the method from Section 3.2. Calculate the confidence interval for $k_h {\sigma_x}^2 + {\sigma_\varepsilon}^2$ and ${\sigma_\varepsilon}^2$ using the method from Section 3.1. Calculate the confidence interval for ${\sigma_x}^2$ in the same manner. Finally, using the method from Section 3.1, calculate the confidence interval for $\delta = \frac{\mu_Y - \mu_X}{\sqrt{{\sigma_x}^2 + {\sigma_Y}^2}}$ and, consequently, obtain the AUC confidence interval [9]. This method offers significant advantages compared to the Delta-Method.

4. Simulation Verification

Through fitting and calculations using simulated data, we will explore the strengths and weaknesses of these methods under different scenarios. We primarily investigate the performance of the same method under different datasets.

4.1. First Set of Simulated Data

Data for the healthy group is generated from $N(80,900)$, and data for the diseased group is generated from $N(160,900)$. Two sets of 51 data points are generated for each group. The errors are generated from $N(0,225)$. The method used is based on Yanhong Li et al.'s calculation method for estimating AUC confidence intervals with multiple repeated experiments. The results obtained are shown in the following table:

Table 1. Simulation data result 1

Prediction	Confidence Interval	
0.9785969	0.4681322	0.999987

It can be observed that the estimation for the lower bound of the confidence interval is notably poor. This is due to the large variance in calculating the confidence interval for $\theta_1 = \mu_Y - \mu_X$, where $\theta_1 \sim N(80,1800)$, leading to a wide confidence interval span. If this proportion is reduced, better confidence interval estimates may be obtained. Therefore, we proceed with a second set of simulated data, where we increase $\mu_Y - \mu_X$.

4.2. Second Set of Simulated Data

Data for the healthy group is generated from $N(80,900)$, and data for the diseased group is generated from $N(180,900)$. Two sets of 51 data points are generated for each group. The errors are generated from $N(0,225)$. The method used is based on Yanhong Li et al.'s calculation method for estimating AUC confidence intervals with multiple repeated experiments. The results obtained are shown in the following table:

Table 2. Simulation data result 2

Prediction	Confidence Interval	
0.9877686	0.6470907	0.9999865

It can be observed that only slightly increasing the difference between the healthy group and diseased group has no significant impact on the prediction and the upper bound of the confidence interval (less than 1%). However, it significantly improves the lower bound of the confidence interval. This indicates that if the value of $\theta_1 = \mu_Y - \mu_X$ is further increased, the estimation of the lower bound of the confidence interval will significantly improve.

DOI: 10.54254/2755-2721/46/20241317

4.3. Third Set of Simulated Data

Data for the healthy group is generated from $N(80,900)$, and data for the diseased group is generated from $N(200,900)$. Two sets of 51 data points are generated for each group. The errors are generated from N(0, 225). The method used is based on Yanhong Li et al.'s calculation method for estimating AUC confidence intervals with multiple repeated experiments. The results obtained are shown in the following table:

Table 3. Simulation data result 3

Prediction	Confidence Interval	
0.9973801	0.8023178	0.9999993

After further increasing the difference between the healthy group and diseased group, the obtained confidence interval span is highly satisfactory. It can be seen that the lower bound is most sensitive to $\theta_1 = \mu_Y - \mu_X$. This is because the region of the lower bound corresponds to the peak region of the standard normal distribution density function. Fluctuations in this range have a significant impact on the final lower bound of the confidence interval, while predictions and the upper bound are more stable in the presence of fluctuations. The fourth set of data is used to confirm this point.

4.4. Fourth Set of Simulated Data

Data for the healthy group is generated from $N(80,900)$, and data for the diseased group is generated from $N(120,900)$. Two sets of 51 data points are generated for each group. The errors are generated from $N(0,225)$. The method used is based on Yanhong Li et al.'s calculation method for estimating AUC confidence intervals with multiple repeated experiments. The results obtained are shown in the following table:

Table 4. Simulation data result 4

Prediction	Confidence Interval	
0.8284095	0.1507986	0.9984336

It can be seen that after $\theta_1 = \mu_Y - \mu_X$ is reduced, the estimation of the lower bound becomes poor, while the prediction value remains high. This confirms our analysis from the third set of simulated data. Therefore, in scenarios where $\mu_Y - \mu_X$ is small (less than double the true variance), the results obtained using this method are not satisfactory. We will adjust the proportion of sampling variance to true variance to observe its impact on confidence interval estimation.

4.5. Fifth Set of Simulated Data

Data for the healthy group is generated from $N(80,900)$, and data for the diseased group is generated from $N(160,900)$. Two sets of 51 data points are generated for each group. The errors are generated from $N(0,900)$. The method used is based on Yanhong Li et al.'s calculation method for estimating AUC confidence intervals with multiple repeated experiments. The results obtained are shown in the following table:

Table 5. Simulation data result 5

Prediction	Confidence Interval	
0.9352721	0.4761308	0.9992275

It can be observed that enlarging the sampling error relative to the true error has no significant impact on the prediction value and the confidence interval data. $\mu_Y - \mu_X$ still remains the primary factor influencing the final results.

DOI: 10.54254/2755-2721/46/20241317

4.6. Sixth Set of Simulated Data

Data for the healthy group is generated from $N(80,900)$, and data for the diseased group is generated from $N(160,900)$. Two sets of 51 data points are generated for each group. The errors are generated from $N(0,25)$. The method used is based on Yanhong Li et al.'s calculation method for estimating AUC confidence intervals with multiple repeated experiments. The results obtained are shown in the following table:

Table 6. Simulation data result 6

Prediction	Confidence Interval	
0.9698731	0.4704401	0.9999481

It can be observed that reducing the sampling error relative to the true error has no significant impact on the prediction value and the confidence interval data. $\mu_Y - \mu_X$ still remains the primary factor influencing the final results.

We will now use the method of David Faraggi to estimate the confidence interval of AUC. Alongside comparing its results to those obtained in the previous six sets of experiments, we will determine the strengths and weaknesses of these methods in different scenarios.

4.7. Seventh Set of Simulated Data

Data for the healthy group is generated from $N(80,900)$, and data for the diseased group is generated from $N(160,900)$. Two sets of 51 data points are generated for each group. The errors are generated from $N(0,225)$. The method used is based on David Faraggi's AUC confidence interval estimation method. It's essential to note that the data used for processing represents the average of data obtained twice. According to the properties of the normal distribution, the calculation requires transforming ${\sigma_\varepsilon}^2$ into $\frac{{\sigma_\varepsilon}^2}{n}$ and substituting it for the solution.

We first calculate the t-value as required for the data and then use software to solve the corresponding estimates and upper and lower bounds for λ, which are subsequently substituted into the equation to obtain the results.

Table 7. Simulation data result 7

Prediction	Confidence Interval	
0.863176	0.7782511	0.9375467

Compared to the first set of data, although the prediction value has decreased, it's evident that this method's ability to provide confidence intervals is significantly superior to Yanhong Li's method. This is because David Faraggi's method estimates the upper and lower bounds based on the t-distribution, which means that the obtained lower bound does not significantly affect the result when applied to the standard normal distribution. This is a significant characteristic that sets it apart from other methods. We will continue to explore the changes in numerical mean values for healthy and diseased populations.

4.8. Eighth Set of Simulated Data

Data for the healthy group is generated from $N(80,900)$, and data for the diseased group is generated from $N(180,900)$. Two sets of 51 data points are generated for each group. The errors are generated from $N(0,225)$. The method used is based on David Faraggi's AUC confidence interval estimation method. The results obtained are shown in the following table:

Table 8. Simulation data result 8

Prediction	Confidence Interval	
0.9503237	0.914344	0.9890933

DOI: 10.54254/2755-2721/46/20241317

It can be seen that as the gap between the diseased population and healthy population data widens, the prediction value rises quickly, approaching the prediction made using Yanhong Li's method. Simultaneously, the width of the confidence interval significantly narrows, indicating a marked improvement in the prediction. Therefore, in this scenario, this method is considered superior to Yanhong Li's method.

Continuing to increase$\mu_Y - \mu_X$ has limited research value for this method. We will now reduce this difference and observe its impact on the final results.

4.9. Ninth Set of Simulated Data

Data for the healthy group is generated from $N(80,900)$, and data for the diseased group is generated from $N(120,900)$. Two sets of 51 data points are generated for each group. The errors are generated from $N(0,225)$. The method used is based on David Faraggi's AUC confidence interval estimation method. The results obtained are shown in the following table:

Table 9. Simulation data result 9

Prediction	Confidence Interval	
0.7250927	0.6212518	0.8249914

In this scenario, Yanhong Li's method, although providing a higher prediction value, fails to offer an effective confidence interval. While this method provides a smaller prediction value, it offers a more reliable confidence interval. It's worth noting that Yanhong Li's method provides a prediction value that falls outside the 95% confidence interval of this method, indicating potential overestimation. Additionally, as $\mu_Y - \mu_X$ decreases, the width of the confidence interval increases, confirming our earlier speculation. We will now adjust the ratio of sampling variance to true variance to observe its impact on confidence interval estimation.

4.10. Tenth Set of Simulated Data

Data for the healthy group is generated from $N(80,900)$, and data for the diseased group is generated from $N(160,900)$. Two sets of 51 data points are generated for each group. The errors are generated from $N(0,900)$. The method used is based on David Faraggi's AUC confidence interval estimation method. The results obtained are shown in the following table:

Table 10. Simulation data result 10

Prediction	Confidence Interval	
0.857375	0.7712289	0.9333739

It can be observed that increasing the sampling error relative to the true error has no significant impact on the prediction value and the confidence interval data. The prediction value slightly decreases, and the confidence interval width slightly widens. At the same time, the confidence interval maintains a significant advantage compared to the fifth set of simulated data.

4.11. Eleventh Set of Simulated Data

Data for the healthy group is generated from $N(80,900)$, and data for the diseased group is generated from $N(160,900)$. Two sets of 51 data points are generated for each group. The errors are generated from $N(0,25)$. The method used is based on David Faraggi's AUC confidence interval estimation method. The results obtained are shown in the following table:

Table 11. Simulation data result 11

Prediction	Confidence Interval	
0.9154539	0.8444226	0.9716023

DOI: 10.54254/2755-2721/46/20241317

It can be observed that reducing the sampling error relative to the true error results in a significant increase in the prediction value and a noticeable decrease in the confidence interval width. This indicates that in scenarios where $\theta^2 = \frac{{\sigma_\varepsilon}^2}{\sigma^2}$ is small, this method is sensitive to this parameter. As the parameter further increases, its sensitivity gradually decreases.

5. Summary

In summary, for all scenarios, the AUC predictions obtained using Yanhong Li et al.'s method for estimating AUC confidence intervals with multiple repeated experiments are greater than the predictions obtained using David Faraggi's AUC confidence interval estimation method. However, the confidence intervals provided by David Faraggi's AUC confidence interval estimation method are often significantly narrower than those provided by Yanhong Li et al.'s method for estimating AUC confidence intervals with multiple repeated experiments. We also found that the confidence interval width given by Yanhong Li et al.'s method for estimating AUC confidence intervals with multiple repeated experiments is more sensitive to $\mu_Y - \mu_X$, but less so to changes in $\frac{{\sigma_\varepsilon}^2}{\sigma^2}$. In contrast, the confidence intervals provided by David Faraggi's AUC confidence interval estimation method are sensitive to both $\mu_Y - \mu_X$ and $\frac{{\sigma_\varepsilon}^2}{\sigma^2}$. Therefore, if a higher AUC prediction is desired, Yanhong Li et al.'s method for estimating AUC confidence intervals with multiple repeated experiments should be used. If wider confidence intervals are sought, then David Faraggi's AUC confidence interval estimation method should be employed. In cases where $\mu_Y - \mu_X$ is substantial, the estimates from both methods are similar.

While Yanhong Li et al.'s method for estimating AUC confidence intervals with multiple repeated experiments may have relatively weaker overall performance, it has a broader range of applications. It does not require every experimental subject to participate in the same number of trials and does not demand an equivalent sampling error variance and true variance for both the healthy and diseased populations. Therefore, it still holds important practical value.

6. Conclusion

The area under the ROC curve (AUC) serves as the most crucial diagnostic method effectiveness metric, and its wide range of applications means that it requires different data processing approaches for various data characteristics. Obtaining more reliable data based on data characteristics is of great significance. However, there are various methods for estimating AUC confidence intervals, and each method naturally comes with its assumptions, limitations, and applicable scenarios. In addition to the two methods introduced, improved, and validated in this paper for estimating AUC under the assumption of a known parameter and a normal distribution, there are various other methods. These include methods for estimating AUC confidence intervals when data is affected by the instrument's measurable range using Maximum Likelihood Estimation (MLE) and methods for estimating AUC confidence intervals for data that follows an exponential random variable distribution, among others. The simulated experiments in this paper also highlight that choosing an estimation method that closely aligns with the existing experimental data conditions leads to better conclusions regarding confidence intervals. When dealing with real data, one should conduct preliminary data preprocessing based on knowledge of the relevant information, data sources, and inherent characteristics. By doing so, the corresponding confidence interval estimation method can be determined. When necessary, various methods can be used for small-scale simulations to determine the optimal estimation method.

References

[1] Zou,K.H, Hall,W.J. and Shapiro,D.E.‘Smooth non-parametric receiver operating characteristic(ROC) curves for continuous diagnostic tests’, Statistics in Medicine, 16, 2143-2156 (1997).

DOI: 10.54254/2755-2721/46/20241317

[2] Wieand, S., Gail M. H., James, B.R. and James K. L.'A family of non-parametric statistics for comparing diagnostic markers with paired or unpaired data',Biometrika, 76, 585-592 (1989).
[3] Bamber, D.C. 'The area above the ordinal dominance graph and the area below the receiver operating characteristic graph',Journal of Mathematical Psychology, 12, 387-415 (1975).
[4] David F.'The effect of random measurment error on receiver operating characteristic (ROC) curves',Statictics in Medicine, 19, 61-70 (2000).
[5] Owen,D.B., Craswell,K.J. and Hanson,D.L.'Non-parametric upper confidence bound for P(Y<X) and confidence limits for P(Y<X) when X and Y are normal'Journal of the American Statistical Association, 59, 906-924 (1964).
[6] Thomas J.D., Hultquist R.A. 'Interval estimation for the unbalanced case of the one-way random effect model', Annals of Statistics, 6, 582-587 (1978)
[7] Yanhong L., John J.K., Allan D. and Zou G.Y.'Interval estimation for the area under the receiver operating characteristic curve when data are subject to error' Statistics in Medicine, 29, 2521-2531 (2010)
[8] Howe WG. 'Approximate confidence limits on the mean of X+Y where X and Y are two tabled independent random variable',Journal of the American Statistical Association, 69, 789-794 (1974)
[9] Graybill F.A. and Wang C.M. 'Confidence intervals on nonnegative linear combination of variances', Journal of the American Statistics Association. 75, 869-873 (1980)