NISTIR 8218

Bootstrap Method versus Analytical Approach for Estimating Uncertainties of Measures in ROC Analysis on Large Datasets

Jin Chu Wu
Alvin F. Martin
Gregory A. Sanders
Raghu N. Kacker

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8218

NIST
National Institute of
Standards and Technology
U.S. Department of Commerce

NISTIR 8218

Bootstrap Method versus Analytical Approach for Estimating Uncertainties of Measures in ROC Analysis on Large Datasets

Jin Chu Wu
Alvin F. Martin
Gregory A. Sanders
Raghu N. Kacker

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8218

July 2018

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Undersecretary of Commerce for Standards and Technology

Bootstrap method versus analytical approach for estimating uncertainties of measures in ROC analysis on large datasets

Jin Chu Wu, Alvin F. Martin, Gregory A. Sanders and Raghu N. Kacker
National Institute of Standards and Technology, Gaithersburg, MD 20899

Abstract – The nonparametric two-sample bootstrap is employed to estimate uncertainties of measures in ROC analysis on large datasets with/without data dependency due to multiple use of the same subjects in many disciplines, based on our studies of bootstrap variability. On the other hand, it would seem that the analytical approach might be used for the same purpose. The differences between these two methods are noteworthy. The bootstrap method can intrinsically take account of how genuine scores and impostor scores are distributed, deal with data dependency, and solve the issue of the covariance occurred while the statistic is a weighted sum of two probabilities derived from two sets of data, respectively, in ROC analysis. The analytical approach cannot. The analytical approach generally underestimates the uncertainties of measures as opposed to the bootstrap method. The comparison was carried out using the real data obtained from the speaker recognition evaluations and the biometric evaluations, as well as the simulated data with normal distributions and nonparametric distributions, respectively.

Keywords: Metrology, measurement uncertainty, ROC analysis, large datasets, bootstrap, data dependency, analytical method, biometrics, speaker recognition

1 Introduction

A measure without estimated uncertainty is incomplete, because it cannot be employed in the practice of evaluating and comparing different systems [1, 2]. When the performance level of a system is very close to a criterion or when the performance levels of two systems are very close, the evaluation and comparison cannot be done without carrying out hypothesis testing to determine the statistical significance of differences. Such a testing demands that the uncertainties, i.e., the standard errors (SE) of measures be estimated.

Receiver operating characteristic (ROC) analysis on large datasets is an important statistical method to analyze statistics of interest for decision making of a classifier in many disciplines. Generally speaking, genuine scores are created by comparing two different objects (e.g., images, speech segments, etc.) of the same subject (e.g., face, speaker, etc.), and impostor scores are generated by matching two objects of two different subjects. The two distributions of continuous scores are schematically depicted in Figure 1 (A).

The cumulative probability of genuine scores from the lowest score to a threshold t is defined as the probability of type I error (miss) α(t), and the one of impostor score from the highest score to a threshold t is defined as the probability of type II error (false alarm) β(t). Thus, these two probabilities are associated with a threshold and negatively traded-off against each other in general. In the coordinate system of (1 – α) versus β, as the threshold t moves from the highest score down to the lowest score, an ROC curve is constructed as depicted in Figure 1 (B).

In different disciplines related to ROC analysis, different measures may be of interest. For instance, in biometrics, the true accept rate (TAR), i.e., 1 – α, at a specified false accept rate (FAR), i.e., β, is of interest. In speaker recognition evaluation (SRE), the measure of interest is the detection cost function (DCF) defined as a weighted sum of the probabilities of α(t) and β(t) at a given threshold t.

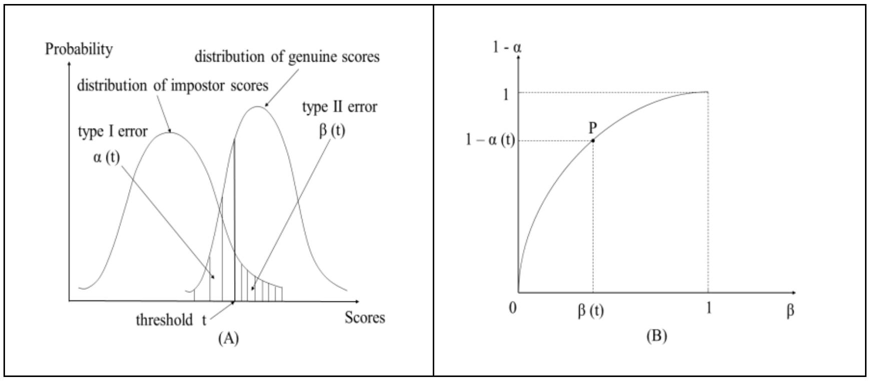

Figure 1 (A): A schematic diagram of two distributions of continuous genuine scores and impostor scores, showing three related variables: type I error α, type II error β, and threshold. (B): A schematic drawing of an ROC curve.

Then, a question arises: How to estimate the SEs of those measures? If datasets are independent and identically distributed (i.i.d.), the nonparametric two-sample bootstrap was employed to estimate uncertainties of measures in ROC analysis on large datasets [1, 3]. If datasets involve data dependency, the nonparametric two-sample two-layer bootstrap was used [4]. Both of them are based on our studies of bootstrap variability [1, 5-8]. On the other hand, it would seem that the analytical approach might be used [9]. The analytical approach in this article refers to using the formula for the sample proportion to compute the SE of an error rate (see Section 2.2.1). However, the differences between these two methods are quite noteworthy in terms of methodologies and results of estimation.

First, in real applications, no underlying parametric distribution functions can be applied to genuine and impostor scores; the score distributions may vary substantially from algorithm to algorithm in a way that differentiates algorithms in terms of matching accuracy [1, 10]. The nonparametric two-sample bootstrap algorithm randomly resamples with replacement (WR) the original genuine scores and impostor scores at each iteration [11-12]. In other words, the bootstrap method can intrinsically take account of how genuine scores and impostor scores are distributed.

Second, data dependency due to multiple use of the same subjects in order to generate more samples because of limited resources exits ubiquitously in many disciplines [4, 12]. If data dependency is involved, the datasets are constructed into a two-layer data structure based on probability theory [4]. Genuine scores are grouped into genuine sets, if they are created using the same subject, with equal number of scores, and likewise for impostor scores.

Further, from the perspective of the multinomial probabilities of selecting bootstrap samples from the original scores as well as the distributions of the bootstrap replications of the measure, in addition to the statistical analysis of the four examples, our research suggested that the nonparametric two-sample two-layer bootstrap algorithm be employed [4, 12]. The two-layer randomly resampling WR takes place not only on sets, but subsequently on scores within the sets. Scores in the same set are assumed to be conditionally independent, because they are generated by two sets of objects and objects in at least one set of the two sets are different.

Third, if the measure is a weighted sum of the probabilities of type I error and type II error at a threshold, a covariance can occur in the formation of the variance. The impact of the covariance on the estimated SE of the measure can be taken into account intrinsically by the bootstrap algorithm.

On the other hand, the analytical approach cannot take account of the distributions of similarity scores, cannot deal with the issue of data dependency, and cannot solve the issue of the covariance because it is difficult to estimate analytically how much it should be.

In terms of estimated results, as shown in the examples of this article and many examples in our practice, the analytical approach generally underestimates the uncertainties of measures in ROC analysis on large datasets with/without data dependency as opposed to the bootstrap method.

The bootstrap is a stochastic process, meaning that different executions of the bootstrap algorithms usually produce different results. But the analytical approach is a deterministic process and thus provides a unique solution. Hence, in this article, the bootstrap estimated SEs of a measure are represented in terms of 95% confidence interval (CI). Such a CI is relatively narrow. The SE produced by a random execution of the bootstrap algorithm should have 95% probability to fall in such a CI. Then, the analytical result or its upper bound (if the negative covariance is hard to estimate) is compared with such a 95% CI of SEs rather than a single bootstrap estimated SE. The absolute relative errors related to 95% CIs are also computed.

In this article, four types of dataset resources were employed[1]: (1) datasets obtained from the SREs involving data dependency; (2) datasets obtained from biometrics evaluation, which are i.i.d.; (3) simulated data with normal distributions; (4) simulated data with nonparametric distributions.

The bootstrap method and the analytical approach are presented in Section 2. The former includes the nonparametric two-sample bootstrap for i.i.d. datasets, and the nonparametric two-sample two-layer bootstrap for datasets with dependency. The bootstrap results and the

[1] Specific hardware and software products identified in this paper were used in order to adequately support the development of technology to conduct the performance evaluations described in this document. In no case does such identification imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

analytical results derived from different datasets, and the comparisons are shown in Section 3. The conclusions and discussion can be found in Section 4.

2 Methods

2.1 The bootstrap methods

2.1.1 The nonparametric two-sample bootstrap algorithm for i.i.d. datasets

For convenient computation, all non-integer scores were converted into integers by shifting the decimal point. While converting, as many decimal places of scores as possible are kept so that it does not result in loss of precision. Thereafter, the scores are expressed using the integer scores $\{s\} = \{s_{min}, s_{min}+1, \ldots, s_{max}\}$. The genuine score set **G** and the impostor score set **I** are denoted as

$$\mathbf{G} = \{m_i \mid m_i \in \{s\} \text{ and } i = 1, \ldots, N_G\}\ , \qquad (1)$$

and

$$\mathbf{I} = \{n_i \mid n_i \in \{s\} \text{ and } i = 1, \ldots, N_I\}\ , \qquad (2)$$

where N_G and N_I are the total numbers of scores. Both **G** and **I** are all multisets meaning that members may appear more than once, and have discrete probability distribution functions of scores.

The nonparametric two-sample bootstrap [1, 11-12] is employed to estimate the SEs of measures in ROC analysis while the datasets can be assumed to be i.i.d. The algorithm is as follows.

Algorithm I **(Nonparametric two-sample bootstrap)**

```
1: for i = 1 to B do
2:       select $N_G$ scores randomly WR from G to form a set {new $N_G$ genuine scores}$_i$
3:       select $N_I$ scores randomly WR from I to form a set {new $N_I$ impostor scores}$_i$
4:       {new $N_G$ genuine scores}$_i$ & {new $N_I$ impostor scores}$_i$ => statistic $\hat{T}_i$
5: end for
6: $\{\hat{T}_i \mid i = 1, \ldots, B\} \Rightarrow \widehat{SE}_B$
7: end
```

where B is the number of two-sample bootstrap replications and WR stands for "with replacement". As shown from Step 1 to 5, this algorithm runs B times. In the i-th iteration, N_G (N_I) scores are randomly selected WR from the original genuine (impostor) score set **G** (**I**) to form a new set of N_G genuine (N_I impostor) scores, and then from these two new sets of scores the i-th bootstrap replication of a measure, $\hat{T}_i$, is generated. A measure can be any statistic of interest in ROC analysis. For instance, it can be the true accept rate (TAR), i.e., 1 – α at a specified false accept rate (FAR), i.e., β (see Figure 1 (A)) [1].

Finally, in Step 6, the $\widehat{SE}_B$ of a measure is estimated to be the sample standard deviation of the B replications $\{\hat{T}_i \mid i = 1, \ldots, B\}$ [1, 11-12]. Further, based on our bootstrap variability studies in ROC analysis on large datasets, the number of bootstrap replications B was determined to be 2,000 in order to reduce the bootstrap variance and ensure the accuracy of the computation [1, 5-8].

In order to compare with the unique analytical result, due to the stochastic nature of the bootstrap method, a distribution of $\widehat{\mathrm{SE}}$s of a measure needs to be generated by running Algorithm I multiple times. Based on our previous studies, in order to create a stable distribution, it was suggested that the algorithm be executed 500 times [1, 5-8]. Thus, a distribution $\{\widehat{SE}_{\mathrm{B\,i}} \mid \mathrm{i} = 1, \dots, 500\}$ was generated. Thereafter, the 95% CI of such a distribution can be estimated.

2.1.2 The nonparametric two-sample two-layer bootstrap algorithm for datasets with dependency

If the data dependency is involved due to multiple use of the same subjects, then the datasets should be re-structured into a two-layer data structure based on the probability theory, and the nonparametric two-sample two-layer bootstrap algorithm should be employed to estimate the SEs of measures in ROC analysis on large datasets from the perspective of the multinomial probabilities of selecting bootstrap samples from the original scores as well as the distributions of the bootstrap replications of the measure [4, 12].

2.1.2.1 The two-layer data structure

genuine $\boldsymbol{S}_\mathrm{G}$	sets	$\boldsymbol{S}_{\mathrm{G}\,1}$	$\boldsymbol{S}_{\mathrm{G}\,2}$	……	$\boldsymbol{S}_{\mathrm{G}\ \mathrm{m_G}}$
	scores	$\alpha_{\mathrm{G}\,1\,1}$, $\alpha_{\mathrm{G}\,1\,2}$, …, $\alpha_{\mathrm{G}\,1\,\mu_{\mathrm{G}1}}$	$\alpha_{\mathrm{G}\,2\,1}$, $\alpha_{\mathrm{G}\,2\,2}$, …, $\alpha_{\mathrm{G}\,2\,\mu_{\mathrm{G}2}}$	……	$\alpha_{\mathrm{G\,m_G}\,1}$, $\alpha_{\mathrm{G\,m_G}\,2}$, …, $\alpha_{\mathrm{G\,m_G}\,\mu_{\mathrm{G\,m_G}}}$

Table 1 The genuine sets, the number of which is mG, and the genuine scores contained in each set.

impostor $\boldsymbol{S}_\mathrm{I}$	sets	$\boldsymbol{S}_{\mathrm{I}\,1}$	$\boldsymbol{S}_{\mathrm{I}\,2}$	……	$\boldsymbol{S}_{\mathrm{I}\ \mathrm{m_I}}$
	scores	$\alpha_{\mathrm{I}\,1\,1}$, $\alpha_{\mathrm{I}\,1\,2}$, …, $\alpha_{\mathrm{I}\,1\,\mu_{\mathrm{I}1}}$	$\alpha_{\mathrm{I}\,2\,1}$, $\alpha_{\mathrm{I}\,2\,2}$, …, $\alpha_{\mathrm{I}\,2\,\mu_{\mathrm{I}2}}$	……	$\alpha_{\mathrm{I\,m_I}\,1}$, $\alpha_{\mathrm{I\,m_I}\,2}$, …, $\alpha_{\mathrm{I\,m_I}\,\mu_{\mathrm{I\,m_I}}}$

Table 2 The impostor sets, the number of which is $\mathrm{m_I}$, and the impostor scores contained in each set.

To preserve the data dependency while the bootstrap resampling takes place, the genuine scores and impostor scores that were generated using the same subject are grouped into a genuine set and an impostor set, respectively. Thus, a two-layer data structure is constructed: The first layer consists of genuine sets and impostor sets, and the second layer consists of genuine scores and impostor scores within those sets.

In this Section 2.1.2, the first subscript index indicates genuine or impostor, and the second and third subscript indices numerate sets and scores, respectively. Suppose that the numbers of the genuine sets and impostor sets are $\mathrm{m_G}$ and $\mathrm{m_I}$. Thus, the set $\boldsymbol{S}_\mathrm{G}$ ($\boldsymbol{S}_\mathrm{I}$) of all genuine (impostor) sets is expressed

$$\boldsymbol{S}_\mathrm{i} = \{\boldsymbol{S}_{\mathrm{i\,j}} \mid \mathrm{j} = 1, \dots, \mathrm{m_i}\},\ \mathrm{i} \in \{\mathrm{G}, \mathrm{I}\}, \tag{3}$$

where $\boldsymbol{S}_{G\,j}$ ($\boldsymbol{S}_{I\,j}$) are genuine (impostor) sets. In terms of scores, each set can be expressed as

$$\boldsymbol{S}_{i\,j} = \{\alpha_{i\,j\,k} \mid k = 1, \ldots, \mu_{i\,j}\},\ j = 1, \ldots, m_i \text{ and } i \in \{G, I\}, \tag{4}$$

where $\alpha_{G\,j\,k}$ ($\alpha_{I\,j\,k}$) are genuine (impostor) scores, and $\mu_{i\,j}$ stands for the number of scores in the corresponding set. The sets $\boldsymbol{S}_{i\,j}$ are all multiset.

The genuine and impostor sets and scores in sets are explicitly listed in Table 1 and Table 2, respectively. There are m_G genuine sets and m_I impostor sets. The genuine sets $\boldsymbol{S}_{G\,1}$, $\boldsymbol{S}_{G\,2}$, …, $\boldsymbol{S}_{G\,m_G}$ contain $\mu_{G\,1}$, $\mu_{G\,2}$, …, $\mu_{G\,m_G}$ genuine scores, respectively; and the impostor sets $\boldsymbol{S}_{I1}$, $\boldsymbol{S}_{I\,2}$, …, $\boldsymbol{S}_{I\,m_I}$ have μ_{I1}, $\mu_{I\,2}$, …, $\mu_{I\,m_I}$ impostor scores, respectively. Finally, the total number of genuine scores N_G and the total number of impostor scores N_I are, respectively,

$$N_i = \sum_{j=1}^{m_i} \mu_{i\,j} \text{ , where } i \in \{G, I\}. \tag{5}$$

2.1.2.2 The probabilities for each score being selected

For the two-layer randomly resampling WR taking place first on score sets and subsequently on scores in sets as described in Section I, the probability for a score $\alpha_{i\,j\,k}$ in a set $\boldsymbol{S}_{i\,j}$ being selected is

$$P_{\text{2-layer}}(\alpha_{i\,j\,k}) = P(\boldsymbol{S}_{i\,j}) \text{ x } P(\alpha_{i\,j\,k} \mid \boldsymbol{S}_{i\,j}) = \frac{1}{m_i} \text{ x } \frac{1}{\mu_{i\,j}}, \tag{6}$$

where $k = 1, \ldots, \mu_{i\,j}$, $j = 1, \ldots, m_i$ and $i \in \{G, I\}$.

These probabilities are the same for all scores within a set, regardless of whether it is a genuine set or an impostor set. However, the probabilities for scores being selected are different from set to set due to different score numbers in different sets indicated by $\mu_{i\,j}$.

The unequal selection probabilities for two-layer resampling must be eliminated in order to reduce the variance of the computation. If the numbers of scores in genuine sets, i.e., $\mu_{G\,j}$, $j = 1, \ldots, m_G$, are all set to be equal to μ_G, then each genuine score can have equal probability to be selected, which is $1/N_G$ due to Eq. (5). By the same token, if all $\mu_{I\,j}$, $j = 1, \ldots, m_I$, are set to be equal to μ_I, then the probability for each impostor score being selected is $1/N_I$.

2.1.2.3 The nonparametric two-sample two-layer bootstrap algorithm

This bootstrap algorithm is employed to estimate the SEs of DCF (see Section 2.1.2.4) at a threshold t, in which datasets involve dependency and two score distributions have no parametric model to fit. From here on, the superscript indices are used to numerate the resampling iterations. Here is a function WR_Random_Sampling_Set that will be frequently used in the following algorithm,

```
function WR_Random_Sampling_Set ($N$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Theta}$)
for $i = 1$ to $N$ do
    select randomly WR an index $j \in \{1, \ldots, N\}$
    $\theta_i = \gamma_j$
end for
```

```
6: end function
```

where WR stands for “with replacement”. $\boldsymbol{\Gamma}$ and $\boldsymbol{\Theta}$ represent a set of sets or a set of scores accordingly with the same cardinality N. Thus, this function can be applied to either a set of sets or a set of scores. As shown from Step 2 to Step 5, it runs N iterations. In the i-th iteration, a member γ_j of the set $\boldsymbol{\Gamma}$ is randomly selected WR to become a member θ_i of a new set $\boldsymbol{\Theta}$. As a result, N members (sets or scores) are randomly selected WR from the set $\boldsymbol{\Gamma}$ to form a new set $\boldsymbol{\Theta}$.

Then, the nonparametric two-sample two-layer bootstrap algorithm is described as follows.

Algorithm II **(Nonparametric two-sample two-layer bootstrap)**

```
1: for i = 1 to B do
2:     WR_Random_Sampling_Set ($m_G$, $\boldsymbol{S}_G$, $\boldsymbol{S}'^{\,i}_{G} = \{\boldsymbol{S}'^{\,i}_{G\,j} \mid j = 1, \dots, m_G\}$)
3:     for k = 1 to $m_G$ do
4:         WR_Random_Sampling_Set ($\mu_G$, $\boldsymbol{S}'^{\,i}_{G\,k}$, $\boldsymbol{S}''^{\,i}_{G\,k}$)
5:     end for

6:     WR_Random_Sampling_Set ($m_I$, $\boldsymbol{S}_I$, $\boldsymbol{S}'^{\,i}_{I} = \{\boldsymbol{S}'^{\,i}_{I\,j} \mid j = 1, \dots, m_I\}$)
7:     for k = 1 to $m_I$ do
8:         WR_Random_Sampling_Set ($\mu_I$, $\boldsymbol{S}'^{\,i}_{I\,k}$, $\boldsymbol{S}''^{\,i}_{I\,k}$)
9:     end for

10:    $\boldsymbol{S}''^{\,i}_{G} = \{\boldsymbol{S}''^{\,i}_{G\,j} \mid j = 1, \dots, m_G\}$ and $\boldsymbol{S}''^{\,i}_{I} = \{\boldsymbol{S}''^{\,i}_{I\,j} \mid j = 1, \dots, m_I\}$ => statistic $\hat{C}^i$
11: end for
12: $\{\hat{C}^i \mid i = 1, \dots, B\} \Rightarrow \widehat{SE}_B$
13: end
```

where B is the number of bootstrap replications. The sets $\boldsymbol{S}_G$ and $\boldsymbol{S}_I$ refer to Eq. (3). This algorithm runs B times as shown from Step 1 to 11. In Steps 2 and 6, the function WR_Random_Sampling_Set is applied twice to the first layer of datasets, i.e., sets. That is, m_G genuine sets are randomly selected WR from the set $\boldsymbol{S}_G$ to constitute a new set $\boldsymbol{S}'^{\,i}_{G} = \{\boldsymbol{S}'^{\,i}_{G\,j} \mid j = 1, \dots, m_G\}$, and m_I impostor sets are randomly selected WR from the set $\boldsymbol{S}_I$ to form a new set $\boldsymbol{S}'^{\,i}_{I} = \{\boldsymbol{S}'^{\,i}_{I\,j} \mid j = 1, \dots, m_I\}$.

Subsequently, the same function is applied to the second layer of datasets, i.e., the scores in sets. From Step 3 to 5, m_G iterations take place. In the k-th iteration, μ_G genuine scores are randomly selected WR from the genuine set $\boldsymbol{S}'^{\,i}_{G\,k}$, which is the k-th new genuine set from the first-layer resampling, to form the k-th new genuine set $\boldsymbol{S}''^{\,i}_{G\,k}$ of the second-layer resampling. The analogous interpretation can be applied to impostor scores in the impostor set $\boldsymbol{S}'^{\,i}_{I\,k}$ as shown from Step 7 to 9.

In Step 10, all genuine scores in the new set $\boldsymbol{S}''^{\,i}_{G} = \{\boldsymbol{S}''^{\,i}_{G\,j} \mid j = 1, \dots, m_G\}$ and all impostor scores in the new set $\boldsymbol{S}''^{\,i}_{I} = \{\boldsymbol{S}''^{\,i}_{I\,j} \mid j = 1, \dots, m_I\}$ are used to generate the i-th bootstrap replication of the measure $\hat{C}^{i}$. Finally in Step 12, from the set { $\hat{C}^{i}$ | i = 1, …, B}, the standard error $\widehat{SE}$ of the measure is estimated using the sample standard deviation of the B replications.

With the new data structure described in Section 2.1.2.2, not only does each genuine (impostor) score have the same probability of being selected, but also the same number of genuine scores and the same number of impostor scores at different iterations are obtained in Step 10 to estimate the bootstrap replications of the measure. This can reduce the computational variance.

By the same token discussed in Section 2.1.1, in order to reduce the bootstrap variance and ensure the accuracy of computation, based on our bootstrap variability studies, the appropriate number of bootstrap replications was determined to be 2,000 [1, 5-8]. And also, because of the stochastic nature of the bootstrap, the 95% CI of the bootstrap estimated SEs of a measure is generated from the distribution of 500 bootstrap estimated SEs of the measure [1, 5-8].

2.1.2.4 The measure in SRE

The measure in this study for the datasets involving data dependency is the DCF in SRE. Let $f_i(s)$, $i \in \{G, I\}$, denote the continuous probability density functions of genuine scores and impostor scores. The two corresponding discrete probability distribution functions, denoted by $P_i(s)$ where $s \in \{s\}$ and $i \in \{G, I\}$, are expressed as

$$\boldsymbol{P}_{\mathbf{i}} = \{P_i(s) \mid \forall\ s \in \{s\} \text{ and } \sum_{s=S_{min}}^{S_{max}} P_i(s) = 1\},\ i \in \{G, I\}. \qquad (7)$$

Hence, the probabilities of type I error α (t) and type II error β (t) related to a threshold $t \in \{s\}$ can be expressed by

$$\alpha(t) = \int_{-\infty}^{t} f_G(s)\,ds = \sum_{s=S_{min}}^{t} P_G(s) = 1 - \sum_{s=t+1}^{S_{max}} P_G(s), \qquad (8)$$

and

$$\beta(t) = \int_{t}^{+\infty} f_I(s)\,ds = \sum_{s=t}^{S_{max}} P_I(s), \qquad (9)$$

where $\sum_{s=S_{max}+1}^{S_{max}} P_G(s) = 0$ is assumed and thus the normalization in Eq. (7) is preserved [1, 10]. For discrete probability distribution, while computing α(t) and β(t) at a threshold t, the probabilities of genuine scores and impostor scores at this threshold t must be taken into account [13].

The DCF is defined as a weighted sum of the probabilities of type I error α(t) and type II error β(t) at a given threshold t [14] and can be expressed in terms of the cumulative discrete probability distribution functions of genuine scores and impostor scores using Eqs. (8) and (9),

$$\begin{aligned} C(t) &= \eta_1 \times \alpha(t) + \eta_2 \times \beta(t) \\ &= \eta_1 \times \left[1 - \sum_{s=t+1}^{S_{max}} P_G(s)\right] + \eta_2 \times \left[\sum_{s=t}^{S_{max}} P_I(s)\right], \end{aligned} \qquad (10)$$

where $\eta_1 = C_{Miss} \times P_{Genuine}$, $\eta_2 = C_{FalseAlarm} \times (1 - P_{Genuine})$, and the parameters C_{Miss}, $C_{FalseAlarm}$, and $P_{Genuine}$ were set to be 10, 1, and 0.01, respectively [14].

2.1.3 The stochastic nature and the absolute relative errors

Because of the stochastic nature of the bootstrap algorithms, the bootstrap estimated SE of a measure γ is represented using the 95% CI of SE, i.e., $[\xi_1, \xi_2]$, meaning that the SE of the measure γ estimated from a random execution of the bootstrap algorithm may fall with 95% confidence in this interval of SE. Indeed, this is what was encountered in our practice [1, 2, 4].

Assume that the distribution of γ is normal [1, 2, 4]. Thus, the measure γ is within $[\gamma - 1.96\ \xi_1, \gamma + 1.96\ \xi_1]$ with 95% probability due to the bootstrap estimated SE ξ_1, and within $[\gamma - 1.96\ \xi_2, \gamma + 1.96\ \xi_2]$ due to the SE ξ_2. Hence, the measure γ caused by different bootstrap estimated SEs in $[\xi_1, \xi_2]$ may occur in $[\gamma - 1.96\ \xi_2, \gamma - 1.96\ \xi_1]$ and $[\gamma + 1.96\ \xi_1, \gamma + 1.96\ \xi_2]$. Indeed, these two regions show the impact of the stochastic nature of the bootstrap algorithm while estimating SEs on the measure γ.

So, the absolute relative errors caused by such a stochastic nature on the measure γ vary from $\varepsilon_1 = 1.96\ \xi_1/\gamma$ to $\varepsilon_2 = 1.96\ \xi_2/\gamma$. In Section 3, it will be shown that the variation of the absolute relative errors $\varepsilon_2 - \varepsilon_1$ is very small. Hence, in real practice, while using the bootstrap method to estimate the SE of a measure, the bootstrap algorithm only needs to be carried out once.

2.2 The analytical approach

2.2.1 Standard error of a single error rate

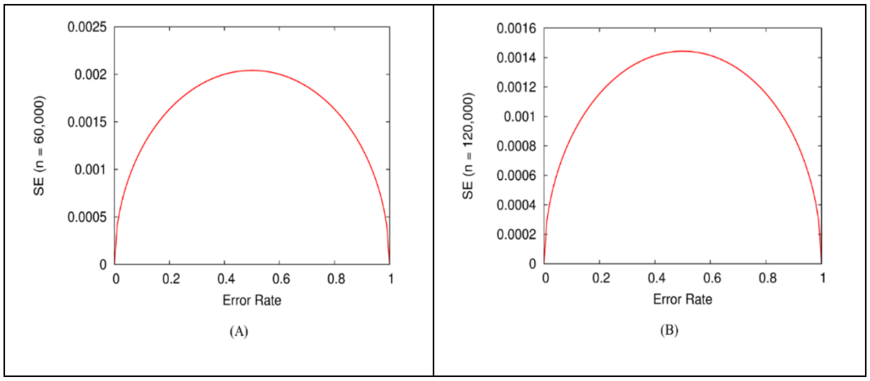

Figure 2 The two curves of the analytically estimated SEs of an error rate for 60,000 (A) and 120,000 (B) scores, respectively.

A classifier in the matching process may be simplified as making a dichotomous decision (i.e., yes or no) with respect to a specified criterion, albeit different genuine scores and impostor scores are generated. This indicates that the populations consist of only two classes, i.e., success and failure [15]. Therefore, the SE of the sample proportion p may be estimated by

$$\mathrm{SE} = \sqrt{\frac{\mathrm{p}\ (1 - \mathrm{p})}{n}} \qquad (11)$$

where n is the total number of scores. While computing the SEs of type I error α(t) and type II error β(t) at an operational threshold t_o in Eq. (10), p = α(t_o) as n = N_G, and p = β(t_o) as n = N_I.

It is easy to prove that the analytically computed SE in Eq. (11) is symmetric with respect to p = 0.5, reaches the maximum at that point, and then decreases and approaches 0. All these can also be seen in Figure 2, while n is 60,000 (A) and 120,000 (B), respectively.

The score distributions vary substantially from classifier to classifier in a way that differentiates classifiers in terms of matching accuracy. However, Eq. (11) does not take account of how the scores are distributed, the combined impact of the two score distributions, and data dependency if it is involved in the datasets [16]. The analytically computed SEs for any two classifiers in the same test remain the same if their two corresponding error rates are equal. The analytical approach generally underestimates the SE of a measure as opposed to the bootstrap method (see below).

Regarding the total numbers of genuine scores and impostor scores, there are tens of thousands of scores in the datasets used in this article [17]. In the SREs' datasets, N_G = 12,672 and N_I = 31,720; in the biometric datasets, N_G = 61,531 and N_I = 121,994; and in the simulated datasets, N_G = 60,000 and N_I = 120,000.

2.2.2 Standard error of a weighted sum of two rates

The analytically estimated SE of DCF can be derived from Eq. (10) [9, 15]

$$SE(t)^2 = \eta_1^2\, SE^2_{\alpha(t)} + \eta_2^2\, SE^2_{\beta(t)} + 2\, \eta_1\, \eta_2\, Cov(\alpha(t), \beta(t)), \qquad (12)$$

where Cov(α(t), β(t)) is the covariance between α(t) and β(t) [15]. Since α(t) and β(t) are all single-error rates, $SE_{\alpha(t)}$ and $SE_{\beta(t)}$ may be estimated using Eq. (11), but its drawback is stated in Section 2.2.1 [9, 15-16].

It is difficult to estimate the covariance Cov(α(t), β(t)). However, as pointed out in Section I, α(t) and β(t) are generally traded off and thus negatively correlated as the threshold t varies. Hence, this covariance must be negative [15]. As a result, even though it is difficult to estimate this covariance, the upper bound of the analytically estimated SEs of DCF can be obtained by setting this negative Cov(α(t), β(t)) in Eq. (12) to zero, and thereafter is used to compare with the 95% CI of the bootstrap estimated SEs.

It is obvious that the analytically computed SE(t) must be smaller than its upper bound. In this article, such upper bounds will be compared with the 95% CIs of the bootstrap estimated SEs due to the stochastic nature of the bootstrap method.

Again, Eq. (12) does not take account of how similarity scores are distributed, and data dependency if it is involved in the datasets. Moreover, it is hard to estimate the covariance in Eq. (12).

3 Results

The datasets from the SREs with data dependency and from the biometrics evaluations with i.i.d. assumption, respectively, are employed. And the simulated datasets with normal distributions and with nonparametric distributions, respectively, are also used.

3.1 Speaker recognition evaluations

As described in Section 2.1.2, the datasets generated from the SRE involve data dependency due to multiple use of the same subjects. Hence, the datasets are reconstructed into a two-layer structure, and the nonparametric two-sample two-layer bootstrap algorithm is employed to estimate the SE of DCF in terms of a 95% CI of SEs of DCF due to the stochastic nature of the bootstrap algorithms [4]. And the upper bound of the analytically computed SE of DCF, in which the negative covariance is set to be zero, is compared with such a 95% CI.

The related results of 12 speaker recognition systems presented in Ref. [4] are shown in Table 3. The smaller the DCFs are, the more accurate the systems are. Table 3 shows that for all 12 speaker recognition systems with different matching accuracies, the upper bounds of the analytically estimated SEs of DCF are correspondingly smaller than the lower bounds of the 95% CIs of the bootstrap estimated SEs of DCF. It indicates that the analytical approach underestimates the SEs of DCF in SREs.

System	DCF	SE of DCF		
		Analytical	Bootstrap	
		(Upper bound)	95% CI	Abs. relative error
EL	0.022199	0.000686	(0.001916, 0.002043)	(16.92%, 18.04%)
UJ	0.028996	0.000502	(0.001961, 0.002093)	(13.25%, 14.15%)
BK	0.031588	0.000520	(0.001818, 0.001934)	(11.28%, 12.00%)
LZ	0.040098	0.000888	(0.002781, 0.002956)	(13.59%, 14.45%)
DL	0.040880	0.000571	(0.001756, 0.001878)	(8.42%, 9.00%)
AF	0.073500	0.000502	(0.001683, 0.001788)	(4.49%, 4.77%)
FI	0.096988	0.000346	(0.000805, 0.000856)	(1.63%, 1.73%)
PB	0.098744	0.001118	(0.004149, 0.004420)	(8.24%, 8.77%)
PM	0.161254	0.001886	(0.005055, 0.005381)	(6.14%, 6.54%)
CO	0.223263	0.002194	(0.006623, 0.007026)	(5.81%, 6.17%)
CH	0.236771	0.002294	(0.005020, 0.005345)	(4.16%, 4.42%)
DG	0.455384	0.002777	(0.009339, 0.009926)	(4.02%, 4.27%)

Table 3 The DCF and its SE estimated using the upper bound of the analytical result, and the nonparametric two-sample two-layer bootstrap method in terms of 95% CI of SEs and absolute relative error of 12 speaker recognition systems. The datasets involve data dependency.

Furthermore, the impact of the stochastic nature of the bootstrap algorithm on the estimates of SEs of DCF is explored in Table 3. Based on the measures proposed in Section 2.1.3, the absolute relative errors due to the bootstrap estimated SEs of DCF vary just about 1%. For instance, for the most accurate System EL, the absolute relative error changes from 16.95% to 18.04%. It suggests that such an impact be little. In other words, a random execution of the nonparametric two-sample two-layer bootstrap algorithm provide a stable estimate of the SE of DCF with 95% confidence in spite of its stochastic nature. As a result, the SE of DCF derived from a random execution of the bootstrap algorithm can be treated as the bootstrap estimated SE of DCF.

3.2 Biometrics evaluations

Algorithm	TAR at FAR = 0.001	SE of TAR		
		Analytical	Bootstrap	
			95% CI	Abs. relative error
A	0.993255	0.000330	(0.000320, 0.000341)	(0.06%, 0.07%)
B	0.796753	0.001622	(0.003362, 0.003618)	(0.83%, 0.89%)

Table 4 The TAR at FAR = 0.001 and its SE estimated using the analytical approach, and the nonparametric two-sample bootstrap method in terms of 95% CI and absolute relative error for the high-accuracy Algorithm A and the low-accuracy Algorithm B. The biometric datasets are assumed to be i.i.d.

The biometric datasets are assumed to be i.i.d. The measure is TAR at FAR = 0.001 [1]. The SE of TAR was computed using the analytical formula Eq. (11), and was also estimated using the nonparametric two-sample bootstrap algorithm shown in Section 2.1.1. The results for the high-accuracy Algorithm A and the low-accuracy Algorithm B are shown in Table 4.

Indeed, the analytical SE of TAR can be roughly found in Figure 2 (A) as well, which has no relationship with how the genuine scores and the impostor scores are distributed. If the matcher is of very high accuracy, i.e., the TAR at FAR = 0.001 is very close to 1, then it is possible that the analytically estimated SE is quite close to the bootstrap estimated SE. This is what happens to Algorithm A. However, for Algorithm B, the analytically estimated SE of TAR is smaller than the lower bound of the 95% CI of SEs of the TAR at FAR = 0.001.

Again, the range of the absolute relative errors caused by the stochastic nature of the bootstrap algorithm is very narrow as shown in Table 4, which is 0.01% for Algorithm A and 0.06% for Algorithm B. In other words, the bootstrap algorithm provides a stable estimate of SE of a measure with 95% confidence.

3.3 The simulated datasets with normal distributions

The matching ability of an image matcher determines how genuine scores and impostor scores are distributed [1, 10]. Thus, the SEs of a measure should not be unrelated to such distributions. In this simulation study, the accuracies of the two image matchers are intentionally designed to be almost the same, but their score distributions are totally different except they are all normally distributed. Under such circumstances, as shown and elaborated below, the analytically computed SEs of the measure stay almost the same, but those bootstrap estimated SEs of the measure vary.

As presented in Table 5 and depicted in Figure 3, in Case (A), the mean and standard deviation of the normal distribution of the impostor scores are 14.0 and 3.0, and those of the genuine scores are 26.0 and 2.0, respectively. In Case (B), the mean and standard deviation of the normal distribution of the impostor scores are 20.0 and 1.5, and those of the genuine scores are 26.0 and 1.0, respectively.

The measure is TAR at FAR = 0.001. As stated in Section 2.2.1, N_I = 120,000. So, if the FAR is set to be 0.001, the number of the falsely accepted instances would be about 120, which is reasonably large in operational practice [1, 10].

Case	Impostor Scores		Genuine Scores		TAR at FAR = 0.001	SE of TAR		
							Bootstrap	
	mean	sd	mean	sd		Analytical	95% CI	Abs. relative error
A	14.0	3.0	26.0	2.0	0.906000	0.001191	(0.005239, 0.005622)	(1.13%, 1.22%)
B	20.0	1.5	26.0	1.0	0.908200	0.001179	(0.009311, 0.009732)	(2.01%, 2.10%)

Table 5 The assigned means and standard deviations for normal probability distributions using simulated data for two Cases (A) and (B) with almost the same TAR at FAR = 0.001, in which the SEs of TAR are computed using the analytical approach and the nonparametric two-sample bootstrap method in terms of 95% CI and absolute relative error, respectively.

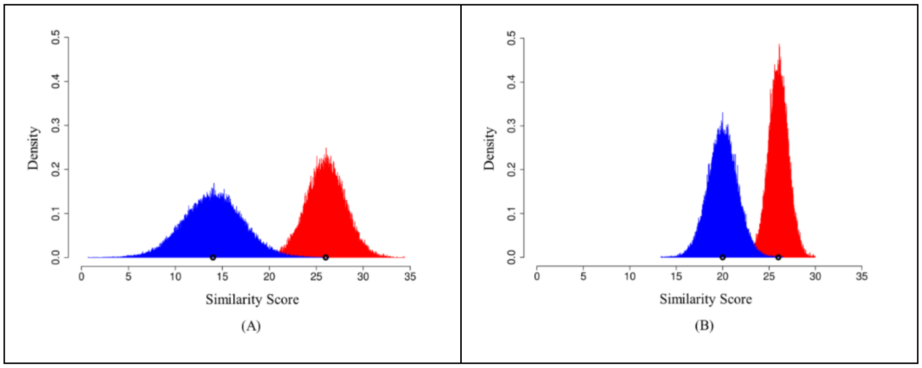

Figure 3 The normal probability distributions of the simulated genuine scores (red) and impostor scores (blue) for two Cases (A) and (B) with different means and standard deviations.

The two normal distributions in Case (B) are narrower than the corresponding ones in Case (A). However, as far as the TAR at FAR = 0.001 is concerned, it is intended to design in such a way that they are almost the same for both Cases (A) and (B): the former is 0.906000, and the latter is 0.908200, as shown in Table 5.

Hence, based on N_G = 60,000 provided in Section 2.2.1, the SE of TAR calculated using the analytical formula Eq. (11) is 0.001191 in Case (A), and 0.001179 in Case (B), which can also be found in Figure 2 (A) based on the corresponding values of the TAR at FAR = 0.001. These two SEs of TAR are almost the same. It indicates that the analytically computed SE of TAR is immune to how the genuine scores and the impostor score are distributed.

All simulated scores are i.i.d. Hence, the nonparametric two-sample bootstrap algorithm presented in Section 2.1.1 is employed to compute the SE of TAR at FAR = 0.001 [1]. The 95% CI of the bootstrap estimated SEs of TAR is (0.005239, 0.005622) in Case (A), but (0.009311, 0.009732) in Case (B), which are very different.

The rationale for this observation is that genuine scores and impostor scores are distributed closer to the point where FAR = 0.001 in Case (B) than in Case (A). This indicates that more similarity scores can affect the area, where the TAR at FAR = 0.001 is determined, in Case (B) than in Case (A). As a result, more uncertainty, i.e., larger SE of measure can occur in Case (B) than in Case (A). This is why the SE of TAR at FAR = 0.001 should be larger in Case (B) than in Case (A), which is exactly what the bootstrap method provides. Hence, the bootstrap algorithm takes account of how similarity score are distributed, i.e., how the matchers are functioning.

The analytically computed SE of TAR is fixed, but smaller than the lower bound of the 95% CI of the bootstrap estimated SEs of TAR. It indicates that the analytical approach underestimates the SE of TAR. Furthermore, the two ranges of the bootstrap estimated absolute relative errors shown in Table 5 are all about 0.09%. This suggests with 95% confidence that the impact of the stochastic nature of the bootstrap algorithm on estimating SE of TAR be negligible.

3.4 The simulated datasets with nonparametric distributions

Case	Impostor Scores		Genuine Scores		TAR at FAR = 0.001	SE of TAR		
							Bootstrap	
	Region	Prob.	Region	Prob.		Analytical	95% CI	Abs. relative error
A	0.0	0.98	[0.0, 50.0)	0.02	0.980783	0.000560	(0.001398, 0.001647)	(0.28%, 0.33%)
	(0.0, 50.0)	0.019	[50.0, 1000.0)	0.88				
	[50.0, 100.0)	0.001	1000.0	0.10				
B	20.0	0.98	[20.0, 50.0)	0.02	0.979217	0.000582	(0.001745, 0.001867)	(0.35%, 0.37%)
	(20.0, 50.0)	0.019	[50.0, 700.0)	0.88				
	[50.0, 70.0)	0.001	700.0	0.10				

Table 6 The assigned probabilities at different regions for nonparametric probability distributions using simulated data for two Cases (A) and (B) with almost the same TAR at FAR = 0.001, in which the SEs of TAR are computed using the analytical approach and the nonparametric two-sample bootstrap method in terms of 95% CI and absolute relative error, respectively.

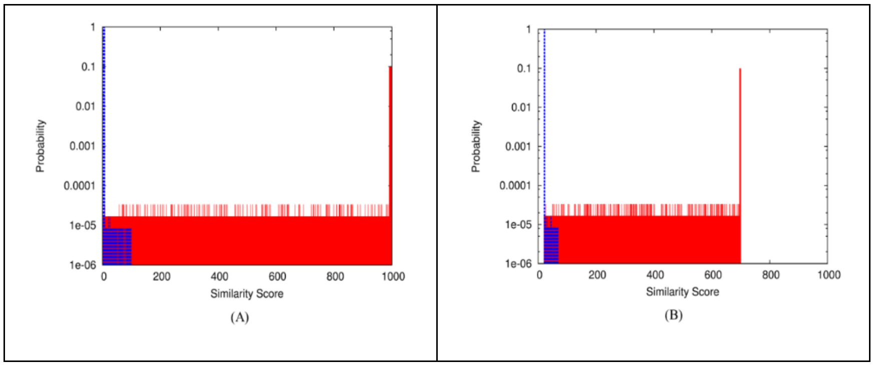

Figure 4 The nonparametric probability distributions of the simulated genuine scores (red) and impostor scores (blue) for two Cases (A) and (B).

This simulation study is different from the previous one in Section 3.3 only in terms of the score distributions, but the same conclusions can be reached. Here, the score distributions are all nonparametric with a high score peak at one end of the distribution, which is a way to increase the matching accuracy [1, 10]. The same comments on the measure TAR at FAR = 0.001 stated in Section 3.3 are also valid here.

As presented in Table 6 and depicted in Figure 4, in Case (A), the impostor scores with probabilities 0.98, 0.019, and 0.001 are uniformly distributed in the regions 0.0, (0.0, 50.0), and [50.0, 100.0); so do the genuine scores with probabilities 0.02, 0.88, and 0.10 in the regions [0.0, 50.0), [50.0, 1000.0), and 1000.0; and the TAR at FAR = 0.001 is 0.980783. In Case (B), the impostor scores with probabilities 0.98, 0.019, and 0.001 are uniformly distributed in the regions 20.0, (20.0, 50.0), and [50.0, 70.0); so do the genuine scores with probabilities 0.02, 0.88, and 0.10 in the regions [20.0, 50.0), [50.0, 700.0), and 700.0; and the TAR at FAR = 0.001 is 0.979217.

Based on Eq. (11) and N_G = 60,000 shown in Section 2.2.1, the analytically estimated SEs of TAR in Cases (A) and (B) are 0.000560 and 0.000582, respectively, which are almost the same and can also be seen in Figure 2 (A). Again, the analytical SE of TAR is independent of how the similarity scores are distributed.

However, the score distributions in Case (A) are quite different from those in Case (B), as shown in Figure 4. This should have impact on the SEs of TAR. Since the simulated data are i.i.d., the nonparametric two-sample bootstrap algorithm shown in Section 2.1.1 is employed to estimate the SE of TAR at FAR = 0.001 [1].

In Case (A) the 95% CI of the bootstrap estimated SEs of TAR is (0.001398, 0.001647), and in Case (B) it is (0.001745, 0.001867). Thus, the bootstrap estimated SEs of TAR at FAR = 0.001 are larger in Case (B) than in Case (A). Regarding this issue, the same comments made in Section 3.3 can be applied here as well. Again, it suggests that the SE of a measure can be affected by how the similarity scores are distributed.

In these two cases, the analytically computed SE of TAR is smaller than the lower bound of the 95% CI of the bootstrap estimated SEs of TAR, which indicates that the analytical approach underestimates the SE of TAR. Moreover, in these two cases, the ranges of the bootstrap estimated absolute relative errors are 0.05% and 0.02%, respectively. It indicates that with 95% confidence the stochastic nature of the bootstrap algorithm has little impact on the bootstrap estimated SEs of TAR.

4 Conclusions and discussion

Evaluation and comparison of performance levels of matching systems in ROC analysis on large datasets cannot be fulfilled without estimating uncertainties of measures. The estimation is usually done using the analytical approach or the bootstrap method. However, there are substantial differences between these two methods, in terms of the methodologies, the conditions of the methods, and the estimates of SEs, etc. The comparisons between these two methods were carried out using the real data obtained from the SREs and the biometric evaluations, as well as the simulated data with normal distributions and nonparametric distributions, respectively.

The analytical approach simply uses a formula without any input regarding how similarity scores are distributed. The bootstrap algorithm resamples the original similarity scores at each iteration and thus takes account of the distributions of similarity scores, which are associated with how the matcher works.

If the datasets involve data dependency due to multiple use of the same subjects like the data in SRE, the analytical approach cannot deal with such a situation, whereas the nonparametric two-sample two-layer bootstrap algorithm based on the two-layer data structure can.

Moreover, if a measure is defined as a weighted sum of the probabilities of type I error and type II error like DCF in SRE and thus a covariance can be derived while estimating its uncertainty analytically, then it is very hard to estimate it. However, because of the way of estimating the SE of DCF, the bootstrap algorithm can intrinsically take the covariance into consideration. And more similar cases can be found in Ref. [1].

The analytical approach generally underestimates the SEs of measures as opposed to the bootstrap method. Of the 18 datasets employed in this article, there is only one case, shown in Section 3.2, in which the analytically estimated SE falls in the middle of the 95% CI of the bootstrap estimated SEs. In all others, the analytically estimated SEs are smaller than the lower bounds of the 95% CIs of the bootstrap estimated SEs.

The analytical approach is a deterministic process and provides a unique value. The nature of the bootstrap algorithm is stochastic, and thus different executions of the bootstrap algorithm may generate slightly different estimates of SE of a measure in ROC analysis. Hence, the bootstrap estimated SEs are represented in terms of its 95% CI, meaning that the estimated SE of a measure derived from a random execution of the bootstrap algorithm should fall with 95% confidence in this interval of SEs. Then, the analytical SE is compared with such a 95% CI of the bootstrap SEs.

Further, after converting to absolute relative errors, it is worth noting that the 95% CI of the bootstrap estimated SEs is generally very narrow. It suggests that a random execution of the bootstrap algorithm provide a stable estimate of the SE of a measure with 95% confidence in spite of its stochastic nature. As a result, the SE of a measure generated from a random execution of the bootstrap algorithm can be treated as the bootstrap estimated SE of the measure.

As pointed out in Ref. [1], in some literature [18] the false non-match rate (FNMR) was employed, which is defined to be 1 – TAR. It is trivial to prove that as far as the estimated SE is concerned, there is no difference between TAR and FNMR. However, they have different coefficient of variation of SE [1].

References

1. Wu, J.C., Martin, A.F., Kacker, R.N. (2011). Measures, uncertainties, and significance test in operational ROC analysis. J. Res. Natl. Inst. Stand. Technol. 116 (1):517-537.
2. Wu, J.C., Martin, A.F., Greenberg, C.S., Kacker, R.N., Stanford, V.M. (2013). Significance test with data dependency in speaker recognition evaluation. In: Active and Passive Signatures IV, Proceedings of SPIE. 8734:87340I.

3. Wu, J.C., Halter, M., Kacker, R.N., Elliott, J.T., Plant, A.L. (2017). A novel measure and significance testing in data analysis of cell image segmentation. BMC Bioinformatics 18:168.
4. Wu, J.C., Martin, A.F., Greenberg, C.S., Kacker, R.N. (2017). The impact of data dependence on speaker recognition evaluation. IEEE/ACM Trans. Audio, Speech, Lang. Process. 25 (1): 5-18.
5. Wu, J.C., Martin, A.F., Kacker, R.N. (in press). Monte Carlo studies of bootstrap variability in ROC analysis with data dependency. Communications in Statistics – Simulation and Computation.
6. Wu, J.C., Martin, A.F., Kacker, R.N. (2016). Validation of nonparametric two-sample bootstrap in ROC analysis on large datasets. Communications in Statistics – Simulation and Computation 45 (5):1689-1703.
7. Wu, J.C., Martin, A.F., Kacker, R.N. (2014). Bootstrap variability studies in ROC analysis on large datasets. Communications in Statistics – Simulation and Computation 43 (1):225-236.
8. Wu, J.C. (2007). Studies of operational measurement of ROC curve on large fingerprint data sets using two-sample bootstrap. NISTIR 7449, National Institute of Standards and Technology.
9. van Leeuwen, D.A., Martin, A.F., Przybocki, M.A., Bouten, J.S. (2006). NIST and NFI-TNO evaluations of automatic speaker recognition. Computer Speech & Language 20:128-158.
10. Wu, J.C., Wilson, C.L. (2007). Nonparametric analysis of fingerprint data on large data sets. Pattern Recognition 40 (9):2574-2584.
11. Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. Ann. Statistics 7:1-26.
12. Efron, B., Tibshirani, R.J. (1993). An Introduction to the Bootstrap. New York: Chapman & Hall.
13. Ostle, B., Malone, L.C. (1988). Statistics in Research: Basic Concepts and Techniques for Research Workers, fourth ed. Ames, Iowa: Iowa State University Press.
14. Doddington, G.R., Przybocki, M.A., Martin, A.F., Reynolds, D.A. (2000). The NIST speaker recognition evaluation – Overview, methodology, systems, results, perspective. Speech Communication 31:225–254.
15. Snedecor, G.W., Cochran, W.G. (1989). Statistical Methods, 8th ed. Ames, Iowa: Iowa State University Press.
16. Linnet, K. (1987). Comparison of quantitative diagnostic tests: type I error, power, and sample size. Statistics in Medicine 6:147-158.
17. Wu, J.C., Wilson, C.L. (2006). An empirical study of sample size in ROC-curve analysis of fingerprint data. In: Biometric Technology for Human Identification III, Proceedings of SPIE. 6202:620207.
18. Cappelli, R., Maio, D., Maltoni, D., Wayman, J.L., Jain, A.K. (2006). Performance evaluation of fingerprint verification systems. IEEE Trans. Pattern Analysis and Machine Intelligence 28 (1):3-18.