Radiological Physics and Technology
https://doi.org/10.1007/s12194-018-0479-9

Verification of modified receiver-operating characteristic software using simulated rating data

Junji Shiraishi[1] · Daisuke Fukuoka[2] · Reimi Iha[3] · Haruka Inada[3] · Rie Tanaka[4] · Takeshi Hara[5]

Received: 11 April 2018 / Revised: 18 September 2018 / Accepted: 18 September 2018

Abstract
ROCKIT, which is a receiver-operating characteristic (ROC) curve-fitting software package, was developed by Metz et al. In the early 1990s, it is a very frequently used ROC software throughout the world. In addition to ROCKIT, DBM-MRMC software was developed for multi-reader multi-case analysis of the difference in average area under ROC curves (AUCs). Because this old software cannot run on a PC with Windows 7 or a more recent operating system, we developed new software that employs the same basic algorithms with minor modifications. In this study, we verified our modified software and tested the differences between the index of diagnostic accuracies using simulated rating data. In our simulation model, all data were generated using target AUCs and a binormal parameter b. In ROC curve fitting with simulated rating data, we varied four factors: the total number of case samples, the ratio of positive-to-negative cases, a binormal parameter b, and the preset AUC. To investigate the differences between the statistical test results obtained from our software and the existing software, we generated simulated rating data sets with three levels of case difficulty and three degrees of difference in AUCs obtained from two modalities. As a result of the simulation, the AUCs estimated by the new and existing software were highly correlated ($R > 0.98$), and there were high agreements (85% or more) in the statistical test results. In conclusion, we believe that our modified software is as capable as the existing software.

Keywords Receiver-operating characteristic analysis (ROC) · Observer study · Computer software · Simulation data · Binormal distribution · Multi-reader multi-case

1 Introduction

Receiver-operating characteristic (ROC) analysis was initially developed through statistical decision theory for evaluating radar systems [1, 2]. In the early 1960s, Lusted first suggested the potential usefulness of ROC analysis for evaluating medical decision-making [3, 4]. Over the last 50 years, ROC analysis has been widely recognized as the most meaningful tool for quantifying the accuracy of a broad variety of diagnostic medical procedures [5–8], and its advantages have been well established [9–12].

An ROC curve represents the relationship between a true positive fraction (TPF) and false positive fraction (FPF). TPF and FPF are equivalent to "sensitivity" and "1-specificity", respectively. Therefore, an ROC curve indicates the trade-off between sensitivity and specificity. In general, an ROC curve is estimated based on the assumption that an observer's responses have a binormal distribution for actually positive and actually negative case samples [7, 10, 11]. To estimate the parameters of binormal distributions, maximum-likelihood estimation has been employed in a number of ROC programs for both single ROC curve fitting [2, 13, 14] and multi-reader multi-case analysis of the index of accuracy [15, 16], which is called the area under the ROC curve (AUC) [17].

✉ Junji Shiraishi
j2s@kumamoto-u.ac.jp

[1] Faculty of Life Sciences, Kumamoto University, 4-24-1 Kuhonji, Kumamoto 862-0976, Japan

[2] Faculty of Education, Gifu University, 1-1 Yanagido, Gifu 501-1193, Japan

[3] School of Health Sciences, Kumamoto University, 4-24-1 Kuhonji, Kumamoto 862-0976, Japan

[4] College of Medical, Pharmaceutical and Health Sciences, Kanazawa University, 5-11-80 Kodatsuno, Kanazawa, Ishikawa 920-0942, Japan

[5] Faculty of Engineering, Gifu University, 1-1 Yanagido, Gifu 501-1193, Japan

Published online: 22 September 2018

One of the most frequently used ROC software packages is called ROCKIT [2]; it was developed by Metz et al. in the early 1990s and was last modified in 2006. To this day, this software has been distributed via a publicly accessible website.[1] ROCKIT assumes that the rating data obtained by the reader in an observer study for actually positive and actually negative case samples follows a normal distribution [2, 8]. Although ROCKIT allows us to analyze a statistically significant difference between two modalities for a single reader, software called DBM-MRMC was developed for multi-reader multi-case analysis of the difference in average AUCs estimated from a number of readers [15, 16].

In most ROC studies, the most difficult task is determining the experimental procedures to obtain rating data sets for testing the diagnostic performances of two medical systems. Therefore, we developed a computer interface for ROC observer studies, which includes a display of digital images and ratings data obtained by observers (radiologists) [17]. The output files of our computer interface were designed to analyze rating data with ROCKIT and DBM-MRMC instantly. However, the old versions of ROCKIT and DBM-MRMC occasionally cannot run on a modern PC running Windows 7 or any later operating system.

To solve this problem, we developed new software that employs the same basic algorithms but with minor modifications. Our new software, called JLABROC and JSRT-MRMC, was designed to replace ROCKIT and DBM-MRMC, respectively, where the "J" in JLABROC and "JSRT" in JSRT-MRMC both stand for the Japanese Society of Radiological Technology. The aim of this study is to verify the practical utility of the JLABROC and JSRT-MRMC software in terms of differences in AUCs and the agreement of statistically significant difference test results obtained from both our software and the existing software.

2 Methods

2.1 JLABROC

The conventional software (ROCKIT [2]) for ROC curve fitting involves the LABROC5 algorithm, which was designed based on the assumption that an observer's responses are binormally distributed for actually negative and actually positive cases that are generally obtained in an observer study [2, 7]. The LABROC5 algorithm was developed for fitting binormal ROC curves to continuously distributed rating data (i.e., the observer's responses) [2]. Figure 1 shows an example of the binormal distribution model used in the estimation of a binormal ROC curve. Note that the mean and standard deviation of the normal distribution of the observer's responses for actually negative cases are denoted as μ_n and σ_n, respectively, and those of the actually positive cases are denoted as μ_s and σ_s, respectively. In the binormal ROC model, the ROC parameters a and b are employed for describing an ROC curve and its AUC. The ROC parameters a and b are defined in the following way [2]:

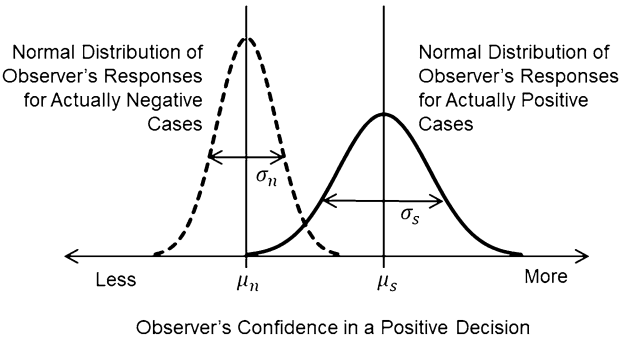

Fig. 1 Model of binormal distribution of observer's responses for actually negative cases and actually positive cases

$$a = \frac{(\mu_s - \mu_n)}{\sigma_s}, \tag{1}$$

$$b = \frac{\sigma_n}{\sigma_s}. \tag{2}$$

To estimate binormal ROC parameters a and b, LABROC5 implements a categorization process for continuously distributed data using a quasi-maximum-likelihood estimation [2]. The number of categories in the categorization process between 5 and 20 is determined by taking into account the number of case samples.

Because a binormal ROC curve was found empirically to be a straight line on normal-deviate axes, the vertical and horizontal coordinates of the ROC curve—true positive fraction (TPF) and false positive fraction (FPF), respectively—have the following relationship:

$$\mathrm{TPF} = \Phi\left(a + b * \Phi^{-1}(\mathrm{FPF})\right), \tag{3}$$

where $\Phi(z)$ represents the standard-normal cumulative distribution function.

In addition, the AUC, which is widely used as an index of diagnostic accuracy, is related to the ROC parameters a and b in the following way:

$$\mathrm{AUC} = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right). \tag{4}$$

[1] Imaging Section Website in JSRT: http://imgcom.jsrt.or.jp/rocGroup/.

The modified ROC software of JLABROC for ROC curve fitting uses the same assumption as LABROC5 for estimating a binormal ROC curve from continuously distributed data. In addition, JLABROC was designed to run with the same input data as ROCKIT uses. Therefore, we can run both ROCKIT and JLABROC with the same input file.

As opposed to LABROC5, JLABROC does not implement a categorization process for estimating binormal ROC parameters *a* and *b*. Because the LABROC5 algorithm was developed in the mid-1990 s, the developers had to implement a categorization process; otherwise, there would have been a substantial computational burden. However, optimal categorization was the most complicated task in the estimation of ROC curves [2]. In addition, the performance of central processing units in modern PCs are more than 1000 times faster than those of PCs in the 1990s [18]. Therefore, we adopted a categorization-free process in the estimation of binormal ROC parameters. In JLABROC, each negative and positive response was used individually as a single category rather than skipping the categorization process. For example, if 50 negative and 50 positive cases were used in an ROC study, JLABROC would estimate binormal distributions by regarding 100 responses as 100 data categories.

Although LABROC5 is just an algorithm included in the ROCKIT software, JLABROC was designed to be operable by itself as an independent piece of software.

The JLABROC software was written in the C programing language using Visual Studio 2017 (Microsoft, USA) and runs on the Windows 7 and Windows 10 operating systems.

2.2 JSRT-MRMC

The algorithm used in DBM-MRMC conducts an analysis of variance among multi-reader and multi-case variations using a pseudo-value matrix computed by jackknifing the AUC values [15]. The matrix data of the observers' responses across both actually negative and actually positive cases for each reader-modality combination are shown as follows [15]:

$$\begin{bmatrix} X_{111} & X_{112} & \cdots & X_{11k} \\ X_{121} & X_{122} & \cdots & X_{12k} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ X_{1j1} & X_{1j2} & \cdots & X_{1jk} \end{bmatrix} \begin{bmatrix} X_{211} & X_{212} & \cdots & X_{21k} \\ X_{221} & X_{222} & \cdots & X_{22k} \\ \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots \\ X_{2j1} & X_{2j2} & \cdots & X_{ijk} \end{bmatrix}, \tag{5}$$

where X_{ijk} denotes the observer's response (i.e., rating score) for the *k*th case of the *j*th reader on the *i*th modality. Using this matrix data, the pseudo-value Y_{ijk} of the AUC is calculated with the following equation:

$$Y_{ijk} = c\hat{A}_{ij} - (c-1)A_{ij(k)}, \tag{6}$$

where c is the total number of cases (negative + positive), and $A_{ij(k)}$ is the AUC value estimated with all of the observer's responses obtained from the *j*th reader on the *i*th modality except for the *k*th case, whereas $\hat{A}_{ij}$ is the AUC estimated from the responses of all cases. Simply stated, if an observer's response for the *k*th case is correct (i.e., a high confidence level for a positive case), the pseudo-value Y_{ijk} is likely to become higher than $\hat{A}_{ij}$, because $A_{ij(k)}$ is lower than $\hat{A}_{ij}$ in most cases.

JSRT-MRMC uses the same method as DBM-MRMC for testing whether there is a statistically significant difference between two average ROC curves using an analysis of variance. Therefore, it takes into account multi-reader and multi-case variations. In addition, JSRT-MRMC was designed to run using the same input data implemented for DBM-MRMC. Therefore, we can run both DBM-MRMC and JSRT-MRMC on the same input file. The only way that JSRT-MRMC differs from DBM-MRMC is in the core algorithm used for ROC curve fitting. JSRT-MRMC employs JLABROC only, whereas DBM-MRMC involves LABROC5, RSCORE for a discrete rating scale, and ProPROC, which employs a different curve-fitting algorithm based on a proper constant-shape bigamma model [16].

Although DBM-MRMC can test the difference between average AUCs, TPFs at any FPFs, and the partial area of two ROC curves [19], JSRT-MRMC has only one option for testing a statistically significant difference between average AUCs, which simplifies the use of the software.

2.3 Simulation of ROC observer study data

To verify the practical utility of JLABROC and JSRT-MRMC in terms of differences in AUCs and statistically significant difference test results between two average AUCs obtained from two ROC parameter settings, we created simulated rating data using a simple binormal model. Although there have been a number of reports on creating simulation data for ROC software [16, 20–27], we developed a simple model based on a formal multivariate model of variation proposed by Roe and Metz (the RM model) [21, 22]. Because the verification of ROCKIT and DBM-MRMC for employing these software packages in ROC studies has been done previously [2, 16], we focused on verifying that there was no difference between the AUCs and the statistical results obtained from the conventional and modified software.

In the RM model, an estimate of an ROC accuracy index (i.e., AUC), obtained from *r* observers, *c* case samples, and with modality *m* is given by the following:

$$\hat{\Theta}_{ijkn} = \mu_i + r_j + c_k + (mr)_{ij} + (mc)_{ik} + (rc)_{jk} + (mrc)_{ijk} + e_{ijkn}, \tag{7}$$

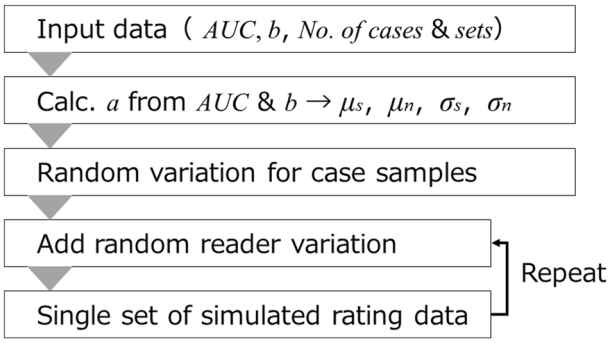

Fig. 2 Illustration of outline of observer's rating data simulation

where *i, j, k*, and *n* denote the parameters of imaging modality, image reader, case sample, and replication of the experiment, respectively. Thus, $\hat{\Theta}_{ijkn}$ is a particular estimate of Θ for the *i*th imaging modality that is obtained from the *j*th reader and the *k*th case sample in the *n*th replication [22]. Therefore, if variations of all factors, except for imaging modality, are considered to be random, then the total variation of the estimate Θ is given by the following:

$$\mathrm{Var}\left\{\hat{\Theta}_{RC|M}\right\} = \sigma_r^2 + \sigma_c^2 + \sigma_{mr}^2 + \sigma_{mc}^2 + \sigma_{rc}^2 + \sigma_{mrc}^2 + \sigma_e^2, \tag{8}$$

where *R, C*, and *M* indicate the observer, case sample, and modality, respectively [22].

In addition, if the estimate is normally distributed, the 95% confidence interval for the true value of Θ in the sampled population of readers and case samples is given by the following:

$$\left[\hat{\Theta} - 1.96\sqrt{\mathrm{Var}\left\{\hat{\Theta}_{RC|M}\right\}}, \hat{\Theta} + 1.96\sqrt{\mathrm{Var}\left\{\hat{\Theta}_{RC|M}\right\}}\right]. \tag{9}$$

In our simulation model, all data were produced using a target AUC and binormal parameter *b*. If particular values of AUC and *b* are provided, a binormal parameter *a* can be calculated using Eq. (4). By setting the mean and standard deviation of the normal distribution of an observer's responses for actually negative cases, those of the actually positive cases can be calculated using Eqs. (1) and (2).

We simulated the observer's rating score using normal distributions for actually positive and actually negative cases, of which the means and standard deviations were predetermined from the target values of AUC and *b*, and variations were determined by Eq. (9). Figure 2 illustrates an outline of an observer's rating data simulation. Because the case variation was assumed to be consistent for all the observers, only the reader variation was changed for the individual rating data set created from the same case samples.

To verify the equivalency between DBM-MRMC and JSRT-MRMC in terms of the statistically significant difference test results, we also created simulated rating data with group data of observers for evaluating paired case samples obtained from two modalities. In the RM model, the correlation of estimates Θ, which were obtained with the same case samples imaged with different modalities, is given by the following:

$$\mathrm{Corr}\left\{\hat{\Theta}_{R|MC}, \hat{\Theta}_{R|M'C'}\right\} = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_{mr}^2 + \sigma_{rc}^2 + \sigma_{mrc}^2 + \sigma_e^2}. \tag{10}$$

To estimate the degree of this correlation, we calculated actual correlation coefficient values for all the combinations of 20 rating data sets, which were obtained in our previous ROC study [28]. As a result, we obtained a 95% confidence interval (0.458 and 0.849) of correlation coefficients for simulating a rating data set obtained with two different modalities and observers but the same cases. Using this 95% confidence interval of correlation coefficients, we simulated paired rating data with a predetermined difference in AUCs.

2.4 Verification of modified ROC software using simulated ROC observer study data

2.4.1 Verification of JLABROC

To investigate the equivalency of AUC estimations obtained by LABROC5 and JLABROC, we conducted a series of simulation experiments in which we generated continuous rating data from the binormal model described earlier. In the ROC curve fitting with the simulated rating data, we varied four factors: the total number of case samples (positive + negative = 50, 100, 125, and 200), the ratio of positive-to-negative cases (positive:negative = 1:1 and 1:4), the binormal parameter (b = 0.75, 1.00, and 1.25), and a preset AUC (0.60 up to 0.90 in increments of 0.10). For each possible combination of factors, we simulated 50 (for 100 case samples) or 20 (for the other case samples) data sets. After we analyzed each data set with both ROCKIT and JLABROC, we recorded the resulting AUC estimates along with their average and differences in the results obtained.

2.4.2 Verification of JSRT-MRMC

To investigate the difference between statistical test results obtained from DBM-MRMC and JSRT-MRMC, we generated simulated rating data sets with three levels of case difficulty and three degrees of difference in AUC (ΔAUC) obtained from two modalities. Figure 3 shows ROC curves with three levels of AUC: high (detection is very easy: AUC 0.980), middle (detection is relatively easy: AUC 0.873), and low (detection is difficult: AUC 0.711) [2]. Figure 4

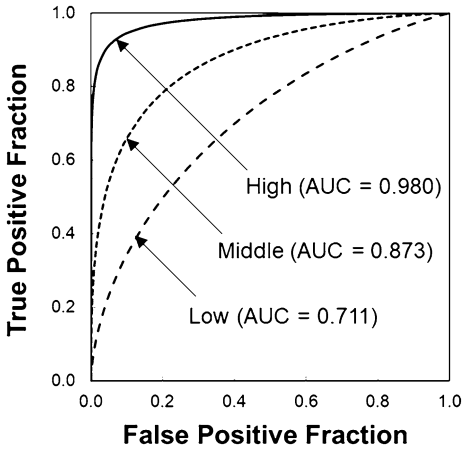

Fig. 3 Binormal ROC curves with three levels of AUCs in simulation studies. Low ($a=0.75$, $b=0.90$; AUC 0.711) and high ($a=0.50$, $b=0.70$; AUC 0.980) curves represent bounds on range of ROCs usually encountered, whereas middle curve ($a=1.50$, $b=0.85$; AUC 0.873) represents more typical ROC [2]

measure of statistical significance, the resulting estimates of AUC, and their average and differences in the results obtained from each.

3 Results

3.1 Verification of JLABROC software

Table 1a–d shows estimated average AUCs obtained using the LABROC algorithm (ROCKIT) and JLABROC and the difference in the two AUCs for the same data sets. All simulation data sets were created for the resulting designated AUCs (Preset AUC). When the preset AUC was adjusted to fit the estimated average AUCs obtained by JLABROC, the calculated AUCs obtained by LABROC5 were slightly higher than those obtained by JLABROC. The estimated AUCs obtained from LABROC5 and JLABROC were highly correlated ($R=0.9875$–0.9897), whereas the differences in the two AUCs ranged from 0.003 to 0.013 when the number of case samples was 50 (25 positives and 25 negatives). There were no visible relationships with the value of b. When the number of case samples was high ($n=100$, 200), the correlation coefficients between average AUCs obtained by both ROCKIT and JLABROC increased, and the differences became small. As shown in Table 1d, an asymmetric combination of positive and negative case samples (25:100) also produced a very high correlation between the two estimated average AUCs ($R=0.9899$–0.9966), whereas the range of differences in the two AUCs (0.002 to 0.011) was equivalent to that of 50 case samples (Table 1a).

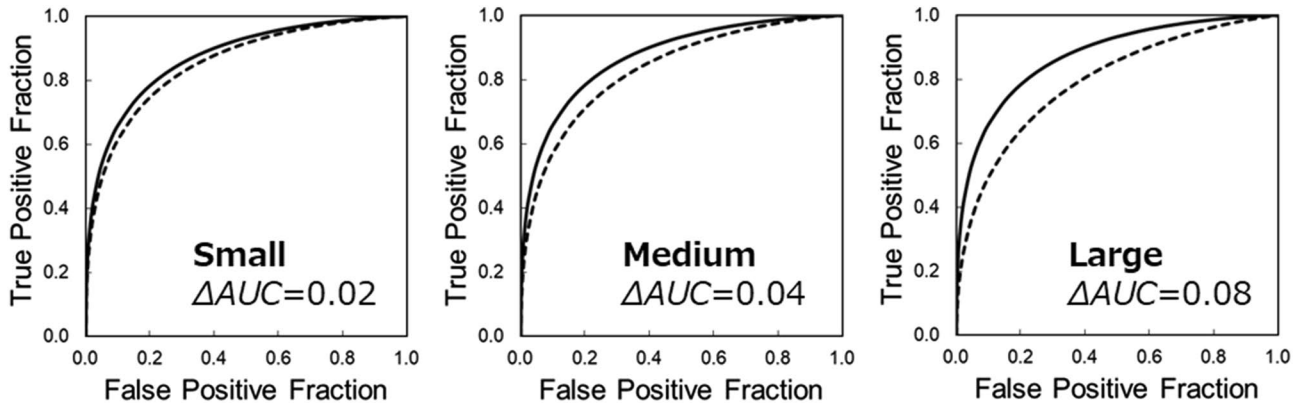

Fig. 4 Pairs of binormal ROC curves with three levels of differences in AUCs. Small (ΔAUC 0.02), medium (ΔAUC 0.02), and large (ΔAUC 0.02) represent degrees of difficulties for demonstrating statistically significant differences

indicates three levels of difference in the AUCs of two ROC curves. As shown in Fig. 4, we generated a paired data set by keeping the AUC with a target value for modality A, whereas the AUC value for modality B was varied by subtracting the difference in AUC from the target AUC. In addition to varying the case difficulty and degree of difference, we varied the number of observers (3, 6, and 9) for each combination of simulation data set. The total number of case samples (positive 50, negative 50) and the value of binormal parameter b (1.0) were consistent for all the simulation data sets. For each possible combination of paired data sets, we generated and tested 100 data sets. Thus, we simulated 2700 pairs of simulated rating data sets ($3\times3\times3\times100$).

After we analyzed each paired data set with both DBM-MRMC and JSRT-MRMC, we calculated a p value as a

Figure 5 illustrates the relationship between the average difference in two AUCs estimated by ROCKIT and JLABROC and a number of case samples. As described above,

Table 1 Average AUCs and the differences in AUCs between LABROC5 and JLABROC estimated using simulated rating data sets

Preset AUC	LABROC5			JLABROC			ΔAUC		
	$b=0.75$	$b=1.00$	$b=1.25$	$b=0.75$	$b=1.00$	$b=1.25$	$b=0.75$	$b=1.00$	$b=1.25$
(a) 50 case samples (25 positives and 25 negatives)									
0.60	0.613	0.605	0.606	0.601	0.601	0.600	0.012	0.004	0.006
0.70	0.708	0.709	0.705	0.699	0.702	0.700	0.003	0.008	0.005
0.80	0.810	0.805	0.809	0.798	0.795	0.800	0.011	0.010	0.006
0.90	0.906	0.912	0.910	0.898	0.898	0.898	0.009	0.013	0.012
*CC				0.9875	0.9875	0.9897			
(b) 100 case samples (50 positives and 50 negatives)									
0.60	0.605	0.602	0.601	0.601	0.600	0.600	0.004	0.002	0.002
0.70	0.704	0.707	0.707	0.701	0.702	0.704	0.003	0.005	0.004
0.80	0.803	0.808	0.807	0.797	0.801	0.801	0.006	0.007	0.006
0.90	0.903	0.910	0.905	0.896	0.904	0.898	0.007	0.007	0.007
*CC				0.9961	0.9970	0.9963			
(c) 200 case samples (100 positives and 100 negatives)									
0.60	0.600	0.599	0.601	0.599	0.596	0.598	0.001	0.002	0.002
0.70	0.700	0.705	0.700	0.698	0.705	0.698	0.002	0.000	0.002
0.80	0.807	0.808	0.804	0.805	0.803	0.800	0.002	0.004	0.004
0.90	0.905	0.903	0.904	0.901	0.897	0.899	0.004	0.006	0.004
*CC				0.9981	0.9971	0.9978			
(d) 125 case samples (25 positives and 100 negatives)									
0.60	0.602	0.605	0.606	0.600	0.600	0.601	0.005	0.002	0.005
0.70	0.708	0.705	0.709	0.702	0.698	0.704	0.006	0.005	0.002
0.80	0.809	0.803	0.807	0.799	0.794	0.801	0.010	0.009	0.006
0.90	0.910	0.903	0.907	0.901	0.897	0.899	0.011	0.006	0.008
*CC				0.9899	0.9952	0.9966			

**CC* cross correlation between AUCs estimated from LABROC5 and JLABROC

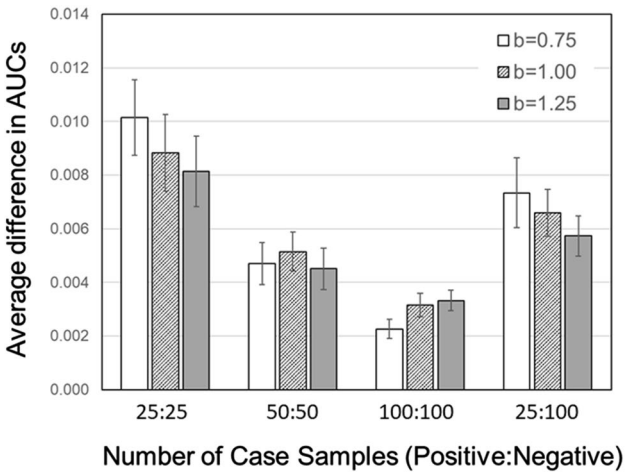

Fig. 5 Relationship between average difference in two *AUC*s estimated by two software packages (ROCKIT and JLABROC) and a number of case samples

the average difference in the two AUCs became small when the number of case samples increased, except when there was an asymmetric combination of case samples (25:100).

3.2 Verification of JSRT-MRMC

Table 2 demonstrates the agreement of statistically significant test results for the difference in two average AUCs estimated by DBM-MRMC and JSRT-MRMC in each of 100 data sets. For instance, the test results were considered to be in agreement when both p values obtained by both DBM-MRMC and JSRT-MRMC were lower than 0.05, equal, or higher than 0.05. Although the average agreement for all combinations was relatively high (94.2%), there were no trends between the three levels of case difficulty, three degrees of differences in AUCs, and the number of observers.

To directly investigate differences in p values obtained by DBM-MRMC and JSRT-MRMC, we examined histograms of the difference in p values for changes in the number of readers (Fig. 6), three levels of case difficulty (Fig. 7), and three degrees of differences in AUCs (Fig. 8). The kurtosis of the histograms increased when the number of readers was increased. In the same way, the kurtosis increased when the case difficulty decreased, and also when the degrees of differences in AUCs increased.

Table 2 Agreement (%) of statistically significant test results ($p < 0.05$) for the difference in AUCs estimated by DBM-MRMC and JSRT-MRMC in each of 100 data sets

Number of readers	Preset difference in AUCs	Preset AUC		
		High (0.980)	Middle (0.873)	Low (0.711)
3	Small (0.02)	92.0	100.0	100.0
	Medium (0.04)	87.0	95.0	97.0
	Large (0.08)	87.0	91.0	86.0
6	Small (0.02)	85.0	96.0	100.0
	Medium (0.04)	96.0	87.0	96.0
	Large (0.08)	100.0	99.0	92.0
9	Small (0.02)	90.0	96.0	100.0
	Medium (0.04)	97.0	88.0	90.0
	Large (0.08)	100.0	100.0	97.0

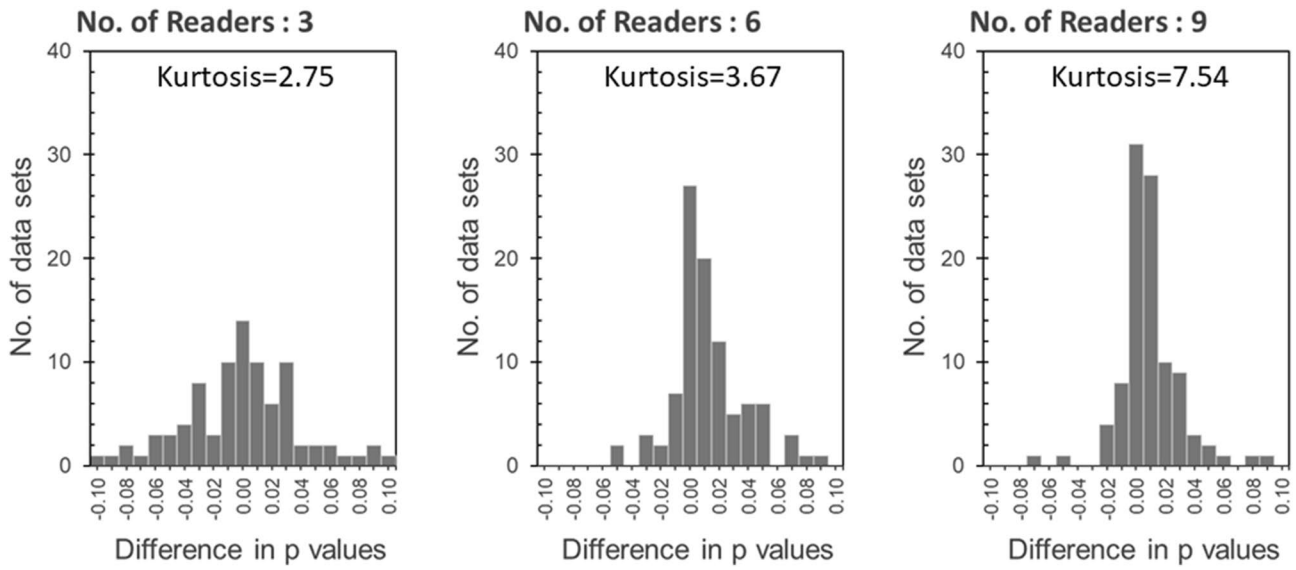

Fig. 6 Histograms of difference in p values for changes in number of readers

4 Discussion

As shown in Table 1, JLABROC had a tendency to underestimate the AUC value slightly compared to the LABROC5 algorithm. A typical example for indicating a difference between two ROC curves estimated by LABROC5 and JLABROC with the same data set is shown in Fig. 9 with a reference ROC curve without curve fitting. Although an ROC curve estimated by LABROC5 slightly exceeded that of JLABROC for the entire curve, the AUC value estimated from JLABROC was closer to that of the reference curve compared to LABROC5.

When LABROC5 implemented a categorization process in its binormal parameter estimation, the number of categories was increased (5–20) corresponding to the number of case samples. In other words, the effects of the categorization process can be reduced when the number of case samples is increased in LABROC5. Therefore, we assumed that the differences of AUC between LABROC5 and JLABROC became small when the number of case samples was increased, and thus, the difference in the number of categories for each became small.

In general, the confidence of p values obtained from the observer study was statistically higher when the number of readers increased [12]. Therefore, the kurtosis of the histograms for the difference in p values obtained from DBM-MRMC and JSRT-MRMC became high when the number of readers increased.

On the other hand, the categorization procedure in ROC curve fitting became difficult when the degrees of difficulty for the detection of actually positive cases were likely to be high. Thus, as shown in Fig. 7, the differences in p values obtained from DBM-MRMC and JSRT-MRMC became small when the case difficulty was low (detection is very easy: AUC 0.980). Similarly, as shown in Fig. 8, the

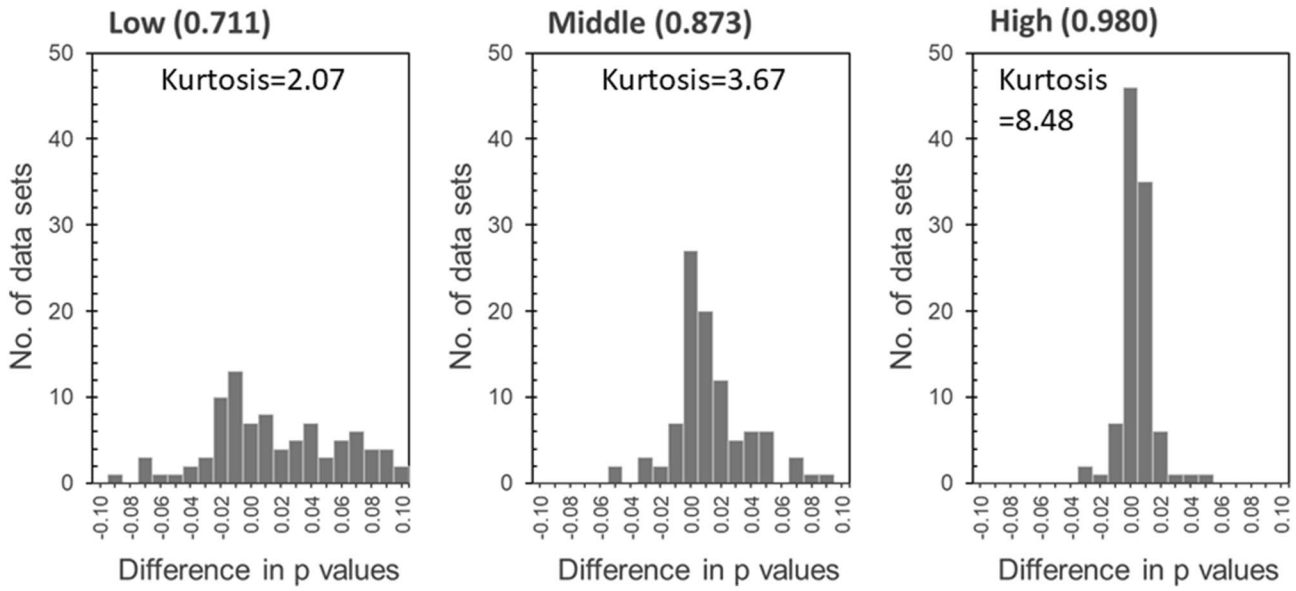

Fig. 7 Histograms of difference in *p* values for changes in three levels of case difficulty

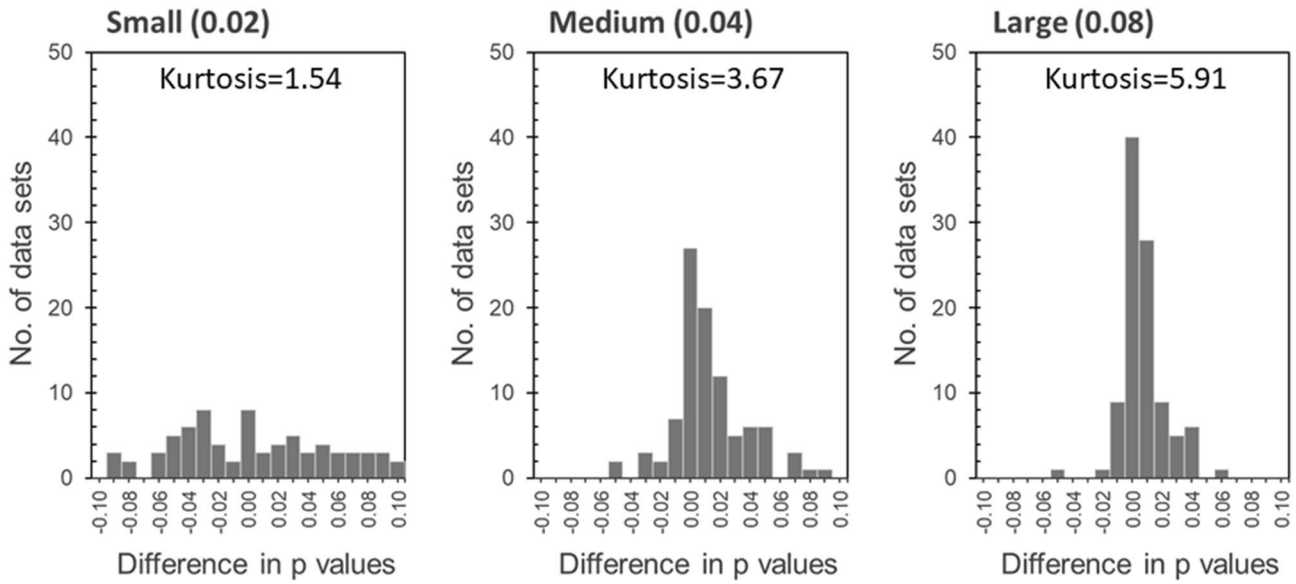

Fig. 8 Histograms of difference in *p* values for changes in three degrees of difference in AUCs

differences in p values also became small when the levels of differences in AUCs became large.

5 Conclusion

We developed and verified modified ROC software to replace the existing outdated software of ROCKIT and DBM-MRMC. In this simulation study, we used a simplified binormal model that was proposed in a previous report. In conclusion, we demonstrated that our modified software for ROC curve fitting and that for testing the difference between the index of diagnostic accuracies obtained in a multi-reader and multi-case manner were equally as capable as the existing software in terms of differences in the estimated AUC and high agreement in a statistically significant difference test.

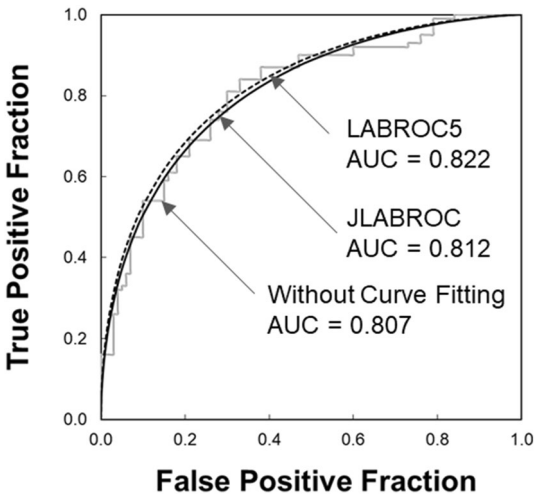

Fig. 9 Typical example of estimated ROC curves fitted using the LABROC5 algorithm (ROCKIT: dash) and JLABROC (solid), and reference ROC curve (gray) without curve fittings obtained from the same original rating data

Acknowledgements We gratefully acknowledge the support of a Japanese Society of Radiological Technology (JSRT) research grant (2016 and 2017). This work was also partially supported by JSPS KAKENHI Grant number 15K09898.

Compliance with ethical standards

Ethical approval This article does not contain any studies with human participants performed, and thus, we have no informed consent from any individuals. In addition, this article does not contain any studies with animals performed.

Conflict of interest The authors declare that they have no conflict of interest about this article.

References

1. Green DM, Swets JA. Signal detection theory and psychophysics. New York: Wiley; 1966 **(reprinted with updated topical bibliographies by Kreiger, New York, 1974)**.
2. Metz CE, Herman BA, Shen J-H. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat Med. 1998;17:1033–53.
3. Lusted LB. Logical analysis in Roentgen diagnosis. Radiology. 1960;74:178–93.
4. Lusted LB. Introduction to medical decision making. Springfield: Charles C Thomas; 1968.
5. Swets JA. Measuring the accuracy of diagnostic systems. Science. 1988;240:1285–93.
6. Goodenough DJ, Rossmann K, Lusted LB. Radiographic applications of receiver operating characteristic (ROC) curves. Radiology. 1974;110:89–95.
7. Metz CE. ROC methodology in radiologic imaging. Invest Radiol. 1986;21:720–33.
8. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978;8:283–98.
9. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology. 2003;229:3–8.
10. ICRU Report 79. Receiver operating characteristic analysis in medical imaging, vol. 8, No.1. Oxford: Oxford University Press; 2008 **(J. of the ICRU)**.
11. Metz CE. ROC analysis in medical imaging: a tutorial review of the literature. Radiol Phys Technol. 2008;1:2–12.
12. Shiraishi J, Pesce L, Metz CE, Doi K. Experimental design and data analysis in receiver operating characteristic studies: lessons learned from reports in Radiology from 1997 to 2006. Radiology. 2009;253:822–30.
13. Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data. J Math Psychol. 1969;6:487–96.
14. Metz CE, Pan X. "Proper" binormal ROC curves: theory and maximum-likelihood estimation. J Math Psychol. 1999;43(1):1–33.
15. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest Radiol. 1992;27:723–31.
16. Metz CE, Roe CA. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: validation with computer simulation. Acad Radiol. 1997;4(4):298–303.
17. Shiraishi J, Fukuoka D, Hara T, Abe H. Basic concepts and development of an all-purpose computer interface for ROC/FROC observer study. Radiol Phys Technol. 2013;6(1):35–41.
18. Waldrop MM. More than Moore. Nature. 2016;530:144–7.
19. Metz CE. Receiver operating characteristic analysis: a tool for the quantitative evaluation of observer performance and imaging systems. J Am Coll Radiol. 2006;3:413–22.
20. Dorfman DD, Berbaum KS, Metz CE, Lenth RV, Hanley JA, Dagga HA. Proper receiver operating characteristic analysis: the Bigamma model. Acad Radiol. 1996;4:138–49.
21. Roe CA, Metz CE. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic data: validation with computer simulation. Acad Radiol. 1997;4:298–303.
22. Roe CA, Metz CE. Variance-component modeling in the analysis of receiver operating characteristic index estimates. Acad Radiol. 1997;4:587–600.
23. Pan X, Metz CE. The "Proper" binormal model: parametric receiver operating characteristic curve estimation with degenerate data. Acad Radiol. 1997;4:380–9.
24. Wagner RF, Beiden SV, Metz CE. Continuous versus categorical data for ROC analysis: some quantitative considerations. Acad Radiol. 2001;8(4):328–34.
25. Pesce LL, Horsch K, Drukker K, Metz CE. Semiparametric estimation of the relationship between ROC operating points and the test-result scale: application to the proper binormal model. Acad Radiol. 2011;18:1537–48.
26. Hillis SL, Berbaum KS. Monte Carlo validation of the Dorfman-Berbaum-Metz method using normalized pseudo values and less data-based model simplification. Acad Radiol. 2005;12:1534–41.
27. Hillis SL, Berbaum KS, Metz CE. Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis. Acad Radiol. 2008;15:647–61.
28. Shiraishi J, Katsuragawa S, Ikezoe J, Matsumoto T, Kobayashi T, Komatsu K, Matsui M, Fujita H, Kodera Y, Doi K. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. AJR. 2000;174:71–4.