Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Full receiver operating characteristic curve estimation using two alternative forced choice studies

Francesc Massanes
Jovan G. Brankov

Full receiver operating characteristic curve estimation using two alternative forced choice studies

Francesc Massanes* and Jovan G. Brankov

Illinois Institute of Technology, Medical Imaging Research Center, Chicago, Illinois 60616, United States

Abstract. Task-based medical image quality is typically measured by the degree to which a human observer can perform a diagnostic task in a psychophysical human observer study. During a typical study, an observer is asked to provide a numerical score quantifying his confidence as to whether an image contains a diagnostic marker or not. Such scores are then used to measure the observers' diagnostic accuracy, summarized by the receiver operating characteristic (ROC) curve and the area under ROC curve. These types of human studies are difficult to arrange, costly, and time consuming. In addition, human observers involved in this type of study should be experts on the image genre to avoid inconsistent scoring through the lengthy study. In two-alternative forced choice (2AFC) studies, known to be faster, two images are compared simultaneously and a single indicator is given. Unfortunately, the 2AFC approach cannot lead to a full ROC curve or a set of image scores. The aim of this work is to propose a methodology in which multiple rounds of the 2AFC studies are used to re-estimate an image confidence score (a.k.a. rating, ranking) and generate the full ROC curve. In the proposed approach, we treat image confidence score as an unknown rating that needs to be estimated and 2AFC as a two-player match game. To achieve this, we use the ELO rating system, which is used for calculating the relative skill levels of players in competitor-versus-competitor games such as chess. The proposed methodology is not limited to ELO, and other rating methods such as TrueSkill™, Chessmetrics, or Glicko can be also used. The presented results, using simulated data, indicate that a full ROC curve can be recovered using several rounds of 2AFC studies and that the best pairing strategy starts with the first round of pairing abnormal versus normal images (as in the classical 2AFC approach) followed by a number of rounds using random pairing. In addition, the proposed method was tested in a pilot human observer study. These pilot results indicate that three to five rounds of 2AFC studies require less human observer time than a full scoring study and that the re-estimated ROC curves and associated area under ROC curve values have high statistical agreement with the full scoring study. © 2016 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: [10.1117/1.JMI.3.1.011010](https://doi.org/10.1117/1.JMI.3.1.011010)]

Keywords: image quality; image evaluation; reader studies; computer tomography.

Paper 15162SSPRR received Jul. 31, 2015; accepted for publication Jan. 13, 2016; published online Feb. 5, 2016.

1 Introduction

The task-based medical image quality assessment paradigm usually involves collecting data on a discrete, confidence scoring scale.¹ This means that the observer uses a discrete scoring scale to indicate his or her confidence about whether an image contains a diagnostic marker or not. For example, in a 1 to 5 scoring scale: “1” indicates high confidence that the decision marker is not present; and “5” indicates high confidence that the decision marker is present. Scores are summarized by the area under the receiver operating characteristic curve (AUC) that quantifies the observer's accuracy. In order to achieve a proper and smooth receiver operating characteristic (ROC) curve, the scoring should be evenly distributed,^{1,2} which is generally hard to achieve. Furthermore, the observers may have difficulties using the rating scale consistently throughout an experiment,³ and the observers should be familiar with the image genre to avoid inconsistent scoring within the lengthy study.¹

In this paper, we aim to reduce the complexity and duration of human reader studies using multiple rounds of a two-alternative forced choice (2AFC) paradigm while regenerating a full confidence-scoring scale. The classical 2AFC is a simple decision test

in which an observer is presented with two alternative images (one with and one without a diagnostic marker) and he/she is asked to choose the one that is more likely to contain the diagnostic marker, thus meriting a higher score. The average of all those outcomes allows one to calculate the empirical AUC value. For most human readers, a 2AFC study is less time consuming and more accurate, and adds less strain on the observer in comparison to providing confidence scoring for each image separately.^{4,5} In addition, this approach does not utilize memory recall and can lessen observer fatigue.^{4,6} For more details, see Ref. 5.

The idea for the proposed method comes from round-robin competitive tournaments, such as chess, where the players are paired in some manner, compete in a head-to-head game, then are re-paired for the next round. In round-robin tournaments, the overall rating is estimated from all head-to-head games. Here we consider that two images are “competing” and that when an image is chosen as more likely to contain a diagnostic marker, it “wins” in a game. Therefore, the winning image should have a higher score (a higher likelihood of containing a diagnostic marker). We will refer to our method as a 2AFC with rounds where, after one round is finished, the ratings are updated and the images are repaired for a new game. This process is repeated

*Address all correspondence to: Francesc Massanes, E-mail: fmassane@iit.edu

for a desired number of times (rounds). Note that in the proposed method, we assume that a round is $N/2$ 2AFC experiments (paired games), where N is the number of images.

This setting has an additional benefit over the classical, one-round 2AFC, or “yes-no” setting: the relevance of a single game diminishes as the number of rounds increases; therefore, the proposed methodology should yield estimated scores with a smaller error and consequently help to minimize the impact of inconsistency of the observer.

To recover ratings (scores), we utilized the ELO rating system,^{7,8} named after its creator Arpad Elo, which the World Chess Federation adopted in 1970.^{1,2,9} One can find many rating algorithms for competitive games, such as TrueSkill^{TM},^{3,7} Chessmetrics,¹⁰ or Glicko.¹¹ We chose ELO as it is the oldest rating system and delivers good results, as we will show later. More complex methods such as TrueSkill^{TM}⁷ and Chessmetrics¹⁰ are adept at adding new players into an existing pool, accounting for the time span between games; however, this property is not necessary in our application.

The proposed methodology has several anticipated practical advantages. (1) It allows ROC curve estimation instead of only a percentage of correctness,¹² while still utilizing the fast and easy 2AFC paradigm. (2) It allows for partial area under ROC curve analysis,¹³ which may have better clinical relevance for a specific level of sensitivity and/or specificity. (3) In our current research on model observer development using a machine learning methodology,^{6,14–16} we observed that models perform better if trained on a human confidence rating.

Next, in Sec. 2, we will describe the ELO rating system followed by consideration of the pairing strategy in Sec. 3. In Sec. 4, we present evaluation and validation results using synthetic and human observer data.

2 Background

2.1 Receiver Operating Characteristic Curve

The ROC curve is used to visualize the performance of a binary classifier in differentiating between two classes, by varying the threshold used to perform the classification. For example, an image score histogram and its ROC curve are shown in Fig. 1. Each point of the ROC curve corresponds to a threshold, which selects the trade-off between true positives (TP, shown on the vertical axis of the ROC curve in Fig. 1) and false positives (FP, shown on the horizontal axis of the ROC curve).

The example in Fig. 1 uses a 1 to 5 scoring scale and the corresponding ROC curve for a discrete distribution. Therefore, the ROC curve has only four operating points. Note that for

binary scoring (0 to 1), the ROC curve will have only one operating point. If one desires to estimate the full ROC curve from a limited set of operating points, one can utilize various models and assumptions.^{2,12} However, in this article, we will utilize linear interpolation between the operating points. That is, for a given set of M operating points $[(FP_m, TP_m)]_{m=1, \dots, M}$, one can generate an interpolated ROC curve using the following equation:

$$\hat{r}(FP) = TP_m + (TP_{m+1} - TP_m) \frac{FP - FP_m}{FP_{m+1} - FP_m},$$

$$FP \in (FP_m, FP_{m+1}]. \quad (1)$$

2.2 ELO Rating

In 1960, the United States Chess Federation implemented a statistical rating system, known as ELO rating, developed by Elo.⁸ The central assumption of this system is that the performance of each player follows a normal or logistic distribution, where the player's true skill is represented by the distribution's mean. While it is likely that each player might have a different standard deviation to his performance, the ELO rating simplifies by assuming a uniform standard deviation. After each game, the player's rating is updated based on whether the game was a win, loss, or draw for that player. Although a player might perform significantly better or worse from one game to the next, the mean value of his performance changes slowly over time.

3 Methodology

3.1 ELO Implementation

The ELO rating consists of two iterative steps: (1) calculation of the expected game outcome for each player based on his current rating and (2) a rating adjustment based on the game outcome.

1. The expected game outcome in the i 'th pairing (round) between players k and l is modeled using a logistic regression, as shown in Eq. (2), where $\mathbb{E}_{k \rightarrow l}^i$ is the expected game outcome of player k with rating R_k^i competing with player l with rating R_l^i ; similarly, $\mathbb{E}_{l \rightarrow k}^i$ is the expected game outcome of player l competing with player k , and s is an arbitrary scale parameter explained later. The scale parameter can be expressed in terms of the standard deviation σ as $s = \sqrt{3}/\pi\sigma$.

Fig. 1 Example score's histogram and estimated ROC curve.

$$\mathbb{E}_{k \rightarrow l}^i = \frac{1}{1 + 10^{\frac{R_l^i - R_k^i}{s}}} \quad \mathbb{E}_{l \rightarrow k}^i = 1 - \mathbb{E}_{k \rightarrow l}^i. \quad (2)$$

The expected game outcome is the player's probability of winning plus half his probability of drawing (Note: a draw is considered half of a win and half of a loss). Thus an expected game outcome of 0.75 could represent a 75% chance of winning, 25% chance of losing, and 0% chance of drawing. At the other extreme, it could represent a 50% chance of winning, 0% chance of losing, and 50% chance of drawing. Figure 2 shows the expected game outcome as a function of the difference in the player's rating.

2. Rating adjustment is based on the game outcome. After the game is played, the outcome of the game is used to modify the ratings of both players using the following equation:

$$\begin{aligned} R_k^{i+1} &= R_k^i + K(S_{k \rightarrow l}^i - \mathbb{E}_{k \rightarrow l}^i) \\ R_l^{i+1} &= R_l^i + K(S_{l \rightarrow k}^i - \mathbb{E}_{l \rightarrow k}^i) \\ &= R_l^i - K(S_{k \rightarrow l}^i - \mathbb{E}_{k \rightarrow l}^i), \end{aligned} \quad (3)$$

where R_k^i is the old rating, R_k^{i+1} is the new rating, $S_{k \rightarrow l}^i$ is the outcome of the game (1 if the k 'th player wins or 0 if the k 'th player loses), $\mathbb{E}_{k \rightarrow l}^i$ is the expected game outcome, and K is the step size constant. Note that $S_{l \rightarrow k}^i = 1 - S_{k \rightarrow l}^i$.

In our implementation, we choose an initial rating of zero for all players, i.e., $R_k^0 = 0$, $k = 1, \dots, N$. Further, we select the scale parameter to be $s = 0.5$ so that 98% of the cdf is in the interval $[-1, 1]$ and $K = 0.04$. From Fig. 2, we can see that if R_k^i is 0.25 points higher than R_l^i , (so $R_l^i - R_k^i = -0.25$), the expected game outcome $\mathbb{E}_{k \rightarrow l}^i$ is 0.76.

Note that we kept the same ratio of K/s as in the original ELO rating system, in which $s = 400$ and $K = 32$ is used. The values of these constants only change the scale and spread of the final ratings, and the actual values are not relevant to the ROC curve calculation.

Next we will describe pairing strategy and its effect on the results. As a reminder, our goal is to use the ELO rating system

Fig. 2 Expected game outcome $\mathbb{E}_{k \rightarrow l}^i$ as a function of the difference in rating $R_l^i - R_k^i$, modeled by a logistic cumulative density function with the scale parameter of $s = 0.5$.

to convert 2AFC tests results into a confidence-scoring scale and consequently estimate the full ROC curve.

3.2 Pairing Strategies

In a 2AFC study, it is customary to show an image pair with an image from the condition-absent dataset (denoted H_0) alongside an image from the condition-present dataset (denoted H_1). However, we have observed (as we will show later) that this strategy overestimates the AUC value after the initial round. Therefore, we will also explore a random pairing strategy and pairing of closely ranked images. In summary, we will explore the following pairing strategies:

- **Strategy A:** Pair an image from the H_0 (condition-absent) dataset with an image from the H_1 (condition-present) dataset.
- **Strategy B:** Randomly pair (without replacement) any two images.
- **Strategy C:** Randomly pair (without replacement) images that have similar rating values R_k^i and R_l^i .

Note that strategy C is used in online gaming communities where the players are paired so that only players with a similar skill level, or rating, are entered into a game. This is done in order to provide a more interesting gaming experience.⁷

In addition, we will also combine the proposed pairing strategies; these combinations will be described and justified in the next section, in which we also present the evaluation results.

4 Results

4.1 Simulated Data

To test the utility of the proposed ELO rating system, as well as the pairing strategies, we will start using simulated data. Here we used three different probability density functions (PDF) to draw random samples: Gaussian (a common assumption in medical diagnostic performance evaluation^{1,2}), Beta (appropriate for modeling highly polarized scores like the detection of pneumothoraxes³), and Uniform (providing a piecewise constant ROC curve). All distribution parameters were adjusted so that an AUC of 0.8 was achieved (see Fig. 3); the theoretical ROC curves are given in Fig. 4.

Next, from each distribution we generated $N = 2000$ random samples (1000 under H_0 hypothesis and 1000 under H_1 hypothesis). The values of these samples, x_k , $k = 1, \dots, N$, are used as true latent scores that we aim to estimate. In the i 'th pairing (round) between players k and l , we use the true latent scores to decide which sample is the winner, i.e.,

$$S_{k \rightarrow l}^i = \begin{cases} 1 & x_k > x_l \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad S_{l \rightarrow k}^i = 1 - S_{k \rightarrow l}^i. \quad (4)$$

To evaluate estimation variability and error bars, we repeated this process $P = 1000$ times so that in total one million samples are generated under each hypothesis and distribution.

4.2 Pairing Strategies Evaluation

Results of the nonparametric AUC estimation^{1,17} using the recovered ratings (scores) by the proposed methods as a function of the number of rounds are reported in Fig. 5 for the Gaussian dataset. The error bars represent a 95% confidence interval

Fig. 3 PDFs under H_0 and H_1 hypotheses for (a) Gaussian PDF, (b) Beta PDF, and (c) Uniform PDF.

Fig. 4 Theoretical ROC curves generated by the statistical distributions shown in Fig. 3. The distribution parameters are adjusted so that an AUC of 0.8 is achieved.

Fig. 5 Estimated AUC evolution of the different pairing strategies. It is obvious that strategy A + B has the most accurate AUC over any number of rounds. Error bars represent a 95% confidence interval estimated using 1000 independent replications of the simulated data.

estimated using 1000 realizations. It is interesting to observe that strategy A achieves an accurate AUC estimate of 0.8 in the first round,^{4,18} but in the following rounds AUC is overestimated. The overestimation of the AUC is an artifact of the proposed iterative approach and rating growth (from all zeros) at early

rounds. It depends on the initial ratings and selection of the step size K in a similar fashion as the step size controls optimization path in a gradient descended optimization problem.

Strategies B and C underestimate the AUC value at first, but they converge quickly to the correct value.

Fig. 6 Estimated AUC evolution of the different pairing strategies: (a) Beta distribution and (b) Uniform distribution. It is obvious that strategy A + B has the most accurate AUC estimate over any number of rounds. Error bars represent a 95% confidence interval estimated using 1000 independent replications of the simulated data.

After analyzing the convergence of different strategies [see Fig. 5(a)], we added two additional options:

- **Strategy A + B:** In the first round of pairings, we utilize strategy A, meaning that we pair images from the H_0 (condition-absent) dataset against images from the H_1 (condition-present) dataset. The following rounds, however, obey strategy B, in which images are randomly paired regardless of the dataset.
- **Strategy A + C:** In the first round, we utilize strategy A, meaning that we pair images from the H_0 (condition-absent)

dataset against images from the H_1 (condition-present) dataset. In the following rounds, we utilize strategy C, meaning that we randomly pair images that have similar rating values.

Both strategies are motivated by the fact that strategy A has an accurate starting point, but requires a large number of rounds to converge to the actual AUC, while strategies B and C have a fast convergence rate but a less-accurate starting point. The results for these two strategies are shown in Fig. 5(b). It is obvious that strategy A + B has the most accurate AUC over any number of rounds.

Fig. 7 (a) MSE between the theoretical and predicted ROC curve for Gaussian, Beta, and Uniform distributions, using strategy A + B, as a function of the number of rounds. Error bars represent a 95% confidence interval estimated using 1000 independent replications of the simulated data; (b) bias^2 and var for the Gaussian distribution as a function of the number of rounds. Graphs are shown for two different sample sizes: $N = 2000$ and $N = 200$.

Finally, we report similar results for Beta and Uniform distribution utilizing A, B, and A + B strategies in Fig. 6.

In the following results, only the A + B pairing strategy was utilized.

4.3 Receiver Operating Characteristic Curve Convergence

Next we will evaluate whether the proposed methodology allows recovery of the theoretical (population) ROC curve as shown in Fig. 4. To quantitatively measure the proposed ROC curve estimation accuracy, we utilize the mean square error (MSE), as given in Eq. (5). MSE measures the difference between

estimated ROC curves in the i 'th round $\hat{r}^i(\text{FP})$, given in Eq. (1), against the theoretical ROC curve $R(\text{FP})$ (shown in Fig. 4). This integral was numerically approximated by discretization with a small step size of $\Delta\text{FP} = 10^{-4}$.

$$\text{MSE}(i) = E \left\{ \int [\hat{r}^i(\text{FP}) - R(\text{FP})]^2 d\text{FP} \right\}, \quad (5)$$

where $E[\cdot]$ denotes the expected value estimated using $P = 1000$ independent replications of the simulated data. In Fig. 7(a), we report the MSE between the estimated and the theoretical ROC curves for a given distribution as a function of the number of rounds.

Fig. 8 Averaged ROC curves after 1, 2, 4, and 20 rounds [estimated ROC curve (dashed line), true ROC curve (solid black line)], using strategy A + B. Dotted lines, in the top left figure, represent $\pm$ one standard deviation estimated using 1000 independent replications of the simulated data.

Fig. 9 Rank correlation coefficient between the estimated ratings R_k^i and the true scores (ranks) x_k as a function of the number of rounds for Gaussian, Beta, and Uniform distributions. Error bars represent a 95% confidence interval estimated using 1000 independent replications of the simulated data.

It is worth noting that MSE decreases quickly, especially for the Gaussian and Beta distributions. The Uniform distribution has a notably higher error than the other two distributions. The origin of this difference is clear after inspecting Fig. 8, which shows $\bar{r}^i(\text{FP})$, the estimated mean ROC curves, averaged over 1000 realizations with different numbers of rounds. For each distribution, once the first round is finished, the estimated rating is binary, having a value of $-K/2$ or $+K/2$, so the ROC curve has only one operating point. In the Gaussian and Beta cases, with three rounds of pairings, the estimated ROC curve (dashed line) closely matches the true ROC curve (solid black line). For the case of the uniform distribution (which is a discontinuous distribution), more rounds are needed to achieve a more accurate ROC curve.

Note that MSE can be rewritten in terms of cumulative bias and variance as

$$\begin{aligned} \text{MSE}(i) &= E \left\{ \int [\hat{r}^i(\text{FP}) - \bar{r}^i(\text{FP})]^2 d\text{FP} \right\} \\ &\quad + \int [\bar{r}^i(\text{FP}) - R(\text{FP})]^2 d\text{FP} \\ &= \overline{\text{var}}^i + \overline{\text{bias}}^{2i}, \end{aligned} \quad (6)$$

where $\bar{r}^i(\text{FP})$ is defined as $\bar{r}^i(\text{FP}) = E[\hat{r}^i(\text{FP})]$. In Fig. 7(b), we report $\overline{\text{var}}^i$ and $\overline{\text{bias}}^{2i}$ for the Gaussian distribution as a function of the number of rounds and for two different dataset sizes: $N = 2000$ and $N = 200$.

Inspecting bias and variance curves in Fig. 7(b), a similar conclusion can be made as in the preceding paragraph. In addition, the results indicate that for a large dataset (i.e., $N = 2000$), five rounds should be sufficient so that low bias, in comparison to almost constant variance, is achieved. In addition, for a smaller dataset (i.e., $N = 200$), the presented results indicate that three rounds may be sufficient.

4.4 Correlation Between True and Estimated Ratings (Scores)

In addition to the ROC curve, the output of the proposed methodology is a rating of all the images. Therefore, we can also compare the true (original) rating x_k with the estimated rating R_k^i . Since we cannot recover the true scale of the ratings, we will use Kendall's τ rank correlation coefficient,¹⁹ which measures the similarity of the orderings of the data when ranked by each of the quantities. Specifically, we used the $\tau - b$ method, which makes adjustments for ties;²⁰ the results are shown in Fig. 9. Note that after the coefficient value exceeds 0.80, the data correlation can be considered strong.²¹ One can observe that the proposed methodology improves the rank correlation at each step, and this is independent of the data distribution type.

Fig. 10 (a) MSE between the theoretical and estimated ROC curves as a function of the number of rounds for Gaussian-distributed data and different values of AUC; (b) rank correlation coefficient between the estimated ratings and the true ranks as a function of number of rounds for Gaussian-distributed data and different values of AUC.

Fig. 11 Examples of the cardiac SPECT images used in Ref. 22: (a) has the lesion present while (b) does not. (c) The difference between images; therefore, the darkest part is the location of the defect.

4.5 Different AUC Values

Previous results evaluated the proposed methodology using data that had AUC of 0.8; in this section, we investigate the effect of varying the AUC value. The simulation data were generated using only the Gaussian PDF, with different AUC values. As before, we generated 2000 samples (1000 under hypothesis H_0 and 1000 under hypothesis H_1) and, to estimate variability and report error bars, we generated 1000 realizations. MSE and rank correlation results are shown in Fig. 10. Almost identical conclusions and observations can be made as in the previous section.

Table 1 Time required by the readers to perform the full scoring study and three or five rounds of 2AFC comparisons.

	Reader A	Reader B
Full scoring	20.13 min	15.49 min
ELO rating with three rounds of 2AFC matches	7.15 min	8.5 min
ELO rating with five rounds of 2AFC matches	11.91 min	14.15 min

4.6 Comparison Using Human Observer Study Data

In this section, we apply and test the proposed methodology to a human observer study. In this experiment, we used imaging data from a previously published human observer study that evaluated various reconstruct algorithms for cardiac SPECT imaging.²² Examples of the images from this dataset are shown in Fig. 11.

For the human observer study, we recruited two readers with experience reading cardiac SPECT images. Each reader was asked to perform two studies:

1. A typical scoring study wherein each observer is asked to provide a numerical score (1 to 6) quantifying the observer’s confidence as to whether the images contain a defect or not. In this study, 200 images were scored (100 images without a perfusion defect and 100 images with the perfusion defect).
2. The proposed ELO rating with three and five rounds of 2AFC matches utilizing the A + B pairing strategy, as detailed in the previous sections. In this study, 500 image pairs were scored.

Note that sufficient time was allowed between the studies to minimize memorization of the cases.

Table 2 The estimated AUC values and p -value of rejecting the null hypothesis for three rounds of the proposed method.

	Reader A		Reader B	
	AUC	σ	AUC	σ
Full scoring	0.69	0.04	0.82	0.03
ELO rating with three rounds of 2AFC matches	0.74	0.03	0.84	0.03
Test for H_0 : Both methods have the same AUC	F value	$Pr > F$	F value	$Pr > F$
	1.50	0.223	0.48	0.491
Mean difference	−0.0523		−0.02120	
Standard error	0.04272		0.03070	
Confidence interval	(−0.13654, 0.03194)		(−0.08175, 0.03935)	
	F -test power for 0.05 difference: 1.00		F -test power for 0.02 difference: 1.00	

Table 3 The estimated AUC values and p -value of rejecting the null hypothesis for five rounds of the proposed method.

	Reader A		Reader B	
	AUC	σ	AUC	σ
Full scoring	0.69	0.04	0.82	0.03
ELO rating with five rounds of 2AFC matches	0.76	0.03	0.83	0.03
Test for H_0 : both methods have the same AUC	F value	$\Pr > F$	F value	$\Pr > F$
	2.65	0.105	0.10	0.751
Mean difference	−0.06615		−0.01020	
Standard error	0.04065		0.03210	
Confidence interval	(−0.14630, 0.01400)		(−0.07351, 0.05311)	
	F -test power for 0.07 difference: 1.00		F -test power for 0.01 difference: 0.91	

The first interesting result, shown in Table 1, is that the necessary time to perform a full scoring study is longer than for three or five rounds of 2AFC comparisons. This is in agreement with results shown in Ref. 23.

Next, in Tables 2 and 3, we show the AUC, standard deviation, p -value of rejecting the null hypothesis, and the power of the test that both methods provide the same AUC for each reader, as well as for the whole study. These values were estimated using DBM MRMC 2.2²¹ and the fixed readers and random cases scenario.

Note that for each reader or method we could not reject the null hypothesis that full scoring or ELO rating with three or five rounds of 2AFC matches provide the same AUC, although the agreement is much higher for the second reader. Finally, the estimated ROC curves are shown in Fig. 12, where one can visually confirm higher agreement for the second reader.

5 Conclusions

The aim of this work was to recover the full ROC curve using 2AFC studies. In the presented approach, we treat image confidence score as an unknown rating that must be estimated, and 2AFC as a two-player game. To estimate the unknown image rating, we utilize the ELO algorithm, which was originally developed to calculate the relative skill levels of players in competitor-versus-competitor games such as chess. The presented simulation results indicate that a full ROC curve can be recovered using several rounds of 2AFC studies and that the best pairing strategy starts with a first round of matching abnormal versus normal images followed by random pairing. In addition, the ratings estimated by the proposed methodology have high $\tau - b$ rank correlation with true scores. Finally, the proposed methodology was tested in a small human observer study. The presented results indicate that three or five rounds of 2AFC

Fig. 12 ROC curves for readers: (a) for three and five rounds with MSE of 0.0053 and 0.0068, respectively; and (b) with MSE of 0.0031 and 0.0040 for the full scoring and the ELO rating.

studies require less human observer time than a full scoring study and that the re-estimated ROC curves and associated AUC values have high statistical agreement with the full scoring study.

Acknowledgments

This work was supported by National Institutes of Health under Grant Nos. HL0654425, HL 091017, and HL122484.

References

1. H. E. Rockette, D. Gur, and C. E. Metz, "The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques," *Invest. Radiol.* **27**(2), 169 (1992).
2. C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," *Invest. Radiol.* **24**(3), 234 (1989).
3. D. Gur et al., "Practical issues of experimental ROC analysis selection of controls," *Invest. Radiol.* **25**(5), 583 (1990).
4. Y. Jang, J. T. Wixted, and D. E. Huber, "Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory," *J. Exp. Psychol. Gen.* **138**(2), 291–306 (2009).
5. A. E. Burgess, "Comparison of receiver operating characteristic and forced choice observer performance measurement methods," *Med. Phys.* **22**(5), 643–655 (1995).
6. T. Marin et al., "Numerical surrogates for human observers in myocardial motion evaluation from SPECT images," *IEEE Trans. Med. Imaging* **33**, 38 (2014).
7. R. Herbrich, T. Minka, and T. Graepel, "Trueskill™: a Bayesian skill rating system," in *Advances in Neural Information Processing Systems 20*, pp. 569–576, MIT Press (2006).
8. A. E. Elo, *The Rating of Chess Players: Past and Present*, Arco Publishing, New York (1978).
9. M. E. Glickman, "A comprehensive guide to chess ratings," *Am. Chess J.* **3**, 59–102 (1995).
10. J. Sonas, Ed., *Chessmetrics*, <http://www.chessmetrics.com/> (2005).
11. M. E. Glickman, "Parameter estimation in large dynamic paired comparison experiments," *J. R. Statist. Soc. C* **48**(3), 377–394 (1999).
12. C. E. Metz, B. A. Herman, and J.-H. Shen, "Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data," *Stat. Med.* **17**, 1033–1053 (1998).
13. S. D. Walter, "The partial area under the summary ROC curve," *Stat. Med.* **24**(13), 2025–2040 (2005).
14. J. G. Brankov et al., "Learning a channelized observer for image quality assessment," *IEEE Trans. Med. Imaging* **28**(7), 991–999 (2009).
15. M. N. Wernick et al., "Machine learning in medical imaging," *IEEE Signal Process Mag.* **27**(4), 25–38 (2010).
16. M. M. Kalayeh, T. Marin, and J. G. Brankov, "Generalization evaluation of machine learning numerical observers for image quality assessment," *IEEE Trans. Nucl. Sci.* **60**(3), 1609–1618 (2013).
17. J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," *Radiology* **143**(1), 29–36 (1982).
18. S. Park et al., "Efficiency of the human observer detecting random signals in random backgrounds," *J. Opt. Soc. Am. A* **22**(1), 3 (2005).
19. M. G. Kendall, "A new measure of rank correlation," *Biometrika* **30**, 81–93 (1938).
20. M. G. Kendall, *Rank Correlation Methods*. Griffin, Oxford, England (1948).
21. J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics* **33**, 159–174 (1977).
22. M. V. Narayanan et al., "Optimization of iterative reconstructions of ^{99m}Tc cardiac SPECT studies using numerical observers," *IEEE Trans. Nucl. Sci.* **49**(5), 2355–2360 (2002).
23. A. E. Burgess, "Comparison of receiver operating characteristic and forced choice observer performance measurement methods," *Med. Phys.* **22**(5), 643–655 (1995).

Francesc Massanes received two degrees, mathematics in 2009 and computer science in 2010, from the Center for Interdisciplinary Studies at UPC-Barcelona Tech with a final thesis titled "Emulation of Human Motion Perception." In 2011, he received his MSEE degree from the Department of Electrical Engineering, Illinois Institute of Technology, where now he is a PhD student with a fellowship grant for postgraduate studies from "La Caixa," Barcelona, Spain.

Jovan G. Brankov is an associate professor of electrical and computer engineering at Illinois Institute of Technology. His current research includes medical image quality assessment based on a human-observer model using machine learning, analyzer-based phase contrast x-ray imaging, 4-D and 5-D tomographic image reconstruction for cardiac SPECT, and quantitative molecular imaging using dual-tracer kinetic modeling. He established the Advanced X-Ray Imaging Laboratory at IIT, which is currently developing a phase-sensitive x-ray device.