On the statistical analysis of image quality metrics based on alternative forced choice experiments

Frédéric Noo, Adam Wunderlich, Dominic J Heuscher, Katharina Schmitt, Zhicong Yu

***Abstract*—Task-based image quality assessment is a valuable methodology for development, optimization and evaluation of new image formation processes in CT. Such an assessment can be performed by building a receiver-operating characteristic (ROC) curve, or variants of it, such as the localization ROC (LROC) curve or the free-response ROC (FROC) curve. For comparisons, it is common to reduce the entire curve to a single scalar that is generally chosen as the area under the curve. In this setting, building the entire curve is not necessary: a two alternative forced choice (AFC) experiment can be performed to directly obtain the desired scalar. In this work, we discuss statistical inference for comparisons of image formation processes using multiple AFC studies.**

I. Introduction

Significant effort is currently spent on the development of statistical iterative reconstruction methods for CT imaging, particularly for the aim of enabling CT exams with a lower dose. To be successful, this effort needs to be accompanied by a careful methodology for assessment of image quality. Such an assessment should be task-based [1], particularly because the algorithms that are under development are non-linear, so that resolution, contrast and anatomical background effects are tangled and thus cannot be analyzed each on their own.

A popular methodology for task-based image quality assessment is the construction of a receiver operating characteristic (ROC) curve [1], [2]. The main idea behind this approach is to evaluate how well an observer (also called a reader) can differentiate images from two separate classes. Typically, these two classes are chosen as sets of images with signal (lesion) either present or absent, but the theory is not limited to such a type of classes. For example, the ROC curve can be used to evaluate the ability of an observer to distinguish lesions with fuzzy boundaries from lesions with sharp boundaries.

Two other popular methodologies for image quality assessment are the localization ROC (LROC) curve, and the free-response ROC (FROC) curve. When the task is defined as that of detecting a signal with unkown location, these two methodologies are often preferred over the classical ROC curve. This preference is due to the fact that, unlike the ROC methodology, the LROC and FROC curves do account for the visual search process. (When the lesion location is not specified, the ROC approach suffers from the fact that an observer may rate an image as containing a lesion and be correct while basing its decision on a reconstruction artifact.) Note that the FROC curve is more powerful than the LROC curve as it does not require telling the observer how many instances of the signal are present in an image.

Whether ROC, LROC or FROC curves are used, it is typical to reduce all information brought by the curve to a single scalar. For ROC and LROC studies, this scalar is generally chosen as the area under the curve. For FROC studies, the area under the curve is not defined, and no single metric has yet been universally accepted.

Interestingly, the area under the ROC or LROC curve has a clear probabilistic meaning: it is the probability of correct decision. In the ROC case, correct decision means correct classification. In the LROC case, correct decision means correct classification together with correct localization [3]. Given this probabilistic meaning, it was noted that the area under the ROC or LROC curve can be estimated without seeking the curve, using the concept of Bernoulli trials. In the context of image quality assessment, this trial is often referred to as a two alternative forced choice (2-AFC) experiment. Whereas the ROC or LROC curve involves only two choices, AFC experiments do not need to be limited to two choices. Multiple AFC experiments (MAFC) can be as easily implemented, and they can be advantageous over a 2-AFC experiment by allowing more stringent testing of image formation processes. However, note that the MAFC experiment does not have an ROC-curve interpretation.

The primary aim of a 2-AFC or MAFC experiment is to evaluate a proportion that serves as an estimate of the probability of correct decision. To achieve this aim, the experimentalist creates a number n of independent trials (cases), present these cases one after the other to an observer and records the number of times when the observer succeed to make a correct decision; this number divided by n is the sought proportion.

As presented above, the probability of correct decision in an MAFC experiment is a quantity that depends on the observer. To reduce this dependence, the mean probability of correct decision over a set of observers is often preferred as a figure-of-merit. Moreover, the proportion obtained for a given observer in an MAFC experiment depends on the selected cases as well as their number. The larger the number of cases, the closer the proportion is to the desired probability of correct decision. However, there are practical limits on the number of trials an observer can be subjected to. Hence, it is important to realize that image quality assessment results based on MAFC experiments include variability due to randomness in cases as well as in the reader pool. An MAFC experiment is inherently a so-called multi-reader multi-case (MRMC) study.

There are four different ways of reporting results from an

The authors are with the Department of Radiology, University of Utah, Salt Lake City, Utah, USA. E-mail: noo@ucair.med.utah.edu

This work was partially supported by NIH grants R01 EB007236 and R21 EB009168.

MRMC study: the cases can either be seen as a fixed or a random effect, and the observers can also either be seen as a fixed or a random effect. Naturally, treating the readers and also the cases as a random rather than a fixed effect enables more general conclusions. However, it is important to realize that generality comes with a cost: error bars are increased. If only 3 or 4 observers are available, there is virtually no hope to make any useful conclusion between image formation processes while treating the readers as a random effect. In this paper, we are interested in the statistical analysis of MRMC results obtained with alternative forced choice experiments under the condition that the readers are seen as a fixed effect, and the cases as a random effect. Our results are primarily relevant for image quality assessment studies related to development and optimization of image formation processes, for which generating a large number of cases is typically easy whereas readers are scarce due to limited availability and high cost.

Technically speaking, the statistical analysis we are interested in amounts to making inferences based on a set of correlated proportions. Proper handling of correlations is where the complexity lies. In practice, correlations can be induced through a number of mechanisms. For example, a study involving two observers that read the exact same cases from one image formation process yields two correlated proportions. Similarly, comparing two image reconstruction algorithms using the exact same data sets with a single observer will yield two correlated proportions. Hypothesis tests have been developed for comparing two [4] or more [5] correlated proportions. Here, we extend on these results in two ways: first, we enable comparisons using confidence intervals rather than hypothesis testing, and second, we enable these comparisons to be performed between linear combinations of proportions, instead of proportions, which is crucially needed to compare reader-averaged proportions.

II. Theory

The problem we consider is that of drawing statistical inferences from K correlated proportions that are each the result of one MAFC experiment. The difference from one experiment to another may either be a change in the observer, or a change in the image formation process used to define the cases. In this section, we first give a mathematical formulation for this problem. Then, we derive the covariance matrix for the vector of correlated proportions, and we introduce a robust estimator for this matrix. Together with properties of asymptotic normality, this covariance matrix estimator is essentially all that we need to build confidence intervals for any function of the K correlated proportions.

A. Mathematical formulation

Let θ_k with $k = 1, \ldots K$ denote the probability of correct decision associated with the k-th MAFC experiment, and let $\hat{\theta}_k$ be the proportion used as estimate of this probability.

As discussed earlier, each $\hat{\theta}_k$ is obtained from a number n of independent Bernoulli trials. Let X_{ik} be the outcome of the i-th trial in the k-th MAFC experiment. This outcome is equal to one in case of success, and equal to zero otherwise. By definition,

$$\hat{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} X_{ik} \,. \tag{1}$$

Also, the expected value of X_{ik}, denoted as $E(X_{ik})$, is θ_k, and consequently, $E(\hat{\theta}_k) = \theta_k$.

Now, let $\underline{\theta}$ and $\underline{\hat{\theta}}$ be the two vectors in the K-dimensional Cartesian space that have the θ_k and $\hat{\theta}_k$ values as their components, respectively, and let $\underline{u}_i$ be the vector that has the X_{ik} as components for any fixed value of i. Using this vectorial notation, we can write $E(\underline{\hat{\theta}}) = \underline{\theta}$ and

$$\underline{\hat{\theta}} = \frac{1}{n} \sum_{i=1}^{n} \underline{u}_i \,. \tag{2}$$

In our setting, vector $\underline{\hat{\theta}}$ is a multivariate random variable with covariance matrix C. If there were no correlations between the MAFC experiments, C would be a diagonal matrix. However, here, we consider that correlations are present and thus C is not diagonal. In any case, the diagonal elements of C are each given by the variance expression for a proportion based on n Bernoulli trials, i.e.,

$$C(k,k) = \frac{\theta_k\,(1-\theta_k)}{n} \,. \tag{3}$$

B. Covariance matrix

Theorem 1. *Let p_{rs} be the probability of jointly reaching a correct decision in the experiments of indices r and s with $r \neq s$. Then,*

$$C(r,s) = \frac{p_{rs} - \theta_r\,\theta_s}{n} \,. \tag{4}$$

This theorem is proved as follows. First, recall that, by definition

$$\begin{aligned} C(r,s) &= E((\hat{\theta}_r - \theta_r)(\hat{\theta}_s - \theta_s)) \\ &= E(\hat{\theta}_r\,\hat{\theta}_s) - \theta_r\,\theta_s \,. \end{aligned} \tag{5}$$

From (1), we get

$$\begin{aligned} E(\hat{\theta}_r\,\hat{\theta}_s) &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{l=1}^{n} E(X_{ir}\,X_{ls}) \\ &= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{l \neq i} E(X_{ir}\,X_{ls}) + \frac{1}{n^2} \sum_{i=1}^{n} E(X_{ir}\,X_{is}) \,. \end{aligned} \tag{6}$$

Given that the cases correspond to independent trials, $E(X_{ir}\,X_{ls}) = \theta_r\,\theta_s$ when $l \neq i$. Moreover, from the definition of p_{rs}, we have $E(X_{ir}\,X_{is}) = p_{rs}$ for any value of i. Therefore,

$$\begin{aligned} E(\hat{\theta}_r\,\hat{\theta}_s) &= \frac{n\,(n-1)}{n^2}\,\theta_r\theta_s + \frac{1}{n}\,p_{rs} \\ &= \theta_r\,\theta_s + \frac{1}{n}\,(p_{rs} - \theta_r\,\theta_s) \,. \end{aligned} \tag{7}$$

Direct combination of this last result with (5) yields the announced result.

Note, as expected, that (3) and (4) are fully consistent with each other, since $p_{rs} = \theta_r$ when $r = s$. In addition, when the proportions are independent, $p_{rs} = \theta_r\,\theta_s$ and thus $C(r,s) = 0$.

C. Estimator for the covariance matrix

Theorem 2. *Let*

$$\hat{p}_{rs} = \frac{1}{n} \sum_{i=1}^{n} X_{ir}\, X_{is}\,. \tag{8}$$

Then,

$$\hat{C}(r,s) = \frac{1}{n-1}\left(\hat{p}_{rs} - \hat{\theta}_r\,\hat{\theta}_S\right) \tag{9}$$

is an unbiased and consistent estimator of C*. Furthermore,* $\hat{C}$ *is definite positive with probability one.*

The unbiasedness of C is proved as follows. First, we note, from its definition, that $E(\hat{p}_{rs}) = p_{rs}$. Second, we observe that

$$E(\hat{\theta}_r\,\hat{\theta}_s) = C(r,s) + \theta_r\,\theta_s\,. \tag{10}$$

Consequently,

$$\begin{aligned} E(\hat{C}(r,s)) &= \frac{1}{n-1}\,E(\hat{p}_{rs}) - \frac{1}{n-1}\,E(\hat{\theta}_r\,\hat{\theta}_s) \\ &= \frac{1}{n-1}\,p_{rs} - \frac{1}{n-1}\,(C(r,s) + \theta_r\,\theta_s) \\ &= \frac{1}{n-1}\,(p_{rs} - \theta_r\,\theta_s) - \frac{1}{n-1}\,C(r,s) \\ &= C(r,s) \end{aligned} \tag{11}$$

where the last equality comes from (4).

To prove consistency, we need to evaluate the behavior of $\hat{C}(r,s)$ as a function n. Given that $\hat{C}(r,s)$ is expressed as the sum of two random variables, we have

$$\begin{aligned} \mathrm{Var}\left(\hat{C}(r,s)\right) &\le \left(\sqrt{\mathrm{Var}\left(\frac{\hat{p}_{rs}}{n-1}\right)} + \sqrt{\mathrm{Var}\left(\frac{\hat{\theta}_r\hat{\theta}_s}{n-1}\right)}\right)^2 \\ &\le \frac{1}{(n-1)^2}\left(\sqrt{\mathrm{Var}(\hat{p}_{rs})} + \sqrt{\mathrm{Var}(\hat{\theta}_r\hat{\theta}_s)}\right) \\ &\le \frac{1}{(n-1)^2}\left(\sqrt{\frac{\hat{p}_{rs}\,(1-\hat{p}_{rs})}{n}} + \sqrt{\mathrm{Var}(\hat{\theta}_r\hat{\theta}_s)}\right)\,. \end{aligned} \tag{12}$$

Also, by the delta method, we know that

$$\mathrm{Var}(\hat{\theta}_r\hat{\theta}_s) \simeq \frac{W}{n} \tag{13}$$

for n large where W is a constant. Therefore, $\mathrm{Var}(\hat{C}(r,s))$ decays at least as $n^{-5/2}$ with n, which proves consistency.

Last, to prove that $\hat{C}$ is definite positive with probability one, we first note that the following equality holds:

$$n\,\hat{C} = \frac{1}{n}\sum_{i=1}^{n} \underline{u}_i\,\underline{u}_i^T - \underline{\hat{\theta}}\,\underline{\hat{\theta}}^T\,. \tag{14}$$

Thus, for any vector $\underline{x}$, we have

$$n\,\underline{x}^T\hat{C}\underline{x} = \frac{1}{n}\sum_{i=1}^{n}(\underline{x}^T\,\underline{u}_i)^2 - (\underline{x}^T\underline{\hat{\theta}})^2\,. \tag{15}$$

Furthermore, since $\underline{\hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\underline{u}_i$, this last equality is equivalent to

$$n^2\,\underline{x}^T\hat{C}\underline{x} = \sum_{i=1}^{n}(\alpha_i)^2 - \frac{1}{n}\left(\sum_{i=1}^{n}\alpha_i\right)^2 \tag{16}$$

with $\alpha_i = \underline{x}^T\,\underline{u}_i$. However, Cauchy-Schwartz's inequality implies that

$$\left(\sum_{i=1}^{n}\alpha_i\right)^2 \le n\,\sum_{i=1}^{n}(\alpha_i)^2\,. \tag{17}$$

Therefore,

$$n^2\,\underline{x}^T\hat{C}\underline{x} \ge 0\,, \tag{18}$$

which demonstrates that $\hat{C}$ is semi-definite positive.

Last, we examine the condition under which the equality in (18) can hold. Because the inequality in (18) was found using Cauchy-Schwartz's inequality, the condition is simple: equality only holds only when $\underline{x}^T\,\underline{u}_i$ is equal to a constant for all i. Since this constraint corresponds to a set of measure zero for any given $\underline{x}$, the strict inequality holds with probability one.

D. Asymptotic properties

Theorem 3. *The random vector* $\hat{C}^{-1/2}(\underline{\hat{\theta}} - \underline{\theta})$ *converges in distribution to a multivariate normal vector with mean zero and identity covariance matrix.*

This theorem is a direct consequence of the following two results. First, $\hat{C}$ converges towards C with probability one, because $\hat{C}$ is a consistent estimator of C with a converging rate of $n^{-5/2}$. Second, equation (2) and the central limit theorem for multivariate random variables imply together that $\hat{C}^{-1/2}(\underline{\hat{\theta}} - \underline{\theta})$ converges in distribution to a multivariate normal vector with mean zero and identity covariance matrix.

E. Summary

Thanks to the asymptotic properties of Theorem 3, the covariance matrix estimator defined by (9) can be used to build confidence intervals (or regions) for any linear combination of components of $\hat{\theta}$. More precisely, let $\underline{\hat{d}} = F\underline{\hat{\theta}}$ where F is a matrix of non-random coefficients, and let $\Omega = FCF^T$ be the covariance matrix of $\underline{\hat{d}}$. Our results imply that $\hat{\Omega} = F\hat{C}F^T$ is a consistent unbiased estimator of Ω and that $\hat{\Omega}^{-1/2}(\underline{\hat{d}} - E(\underline{\hat{d}}))$ is asymptotically distributed as a multivariate normal vector with mean zero and identity covariance matrix.

III. Example of utilization

In this section, we illustrate how the results of the previous section can be utilized for comparison between image reconstruction algorithms using results from 2-AFC experiments.

A. Reconstruction algorithms

The algorithms selected for our example perform image reconstruction from fan-beam data collected in two dimensions. The first two algorithms, called algorithms A and B, use a full-scan of data, whereas the third algorithm, called algorithm C, only uses a short-scan of 240 degrees. Algorithms A and C are both implementations of the fan-beam filtered-backprojection (FBP) formula with different weighting schemes: algorithm A weights all measurements with a factor of 1/2, whereas algorithm C invokes a Parker weighting so that only data over a short-scan are needed. Algorithm B is an implementation of the parallel-beam FBP formula that is applied after rebinning

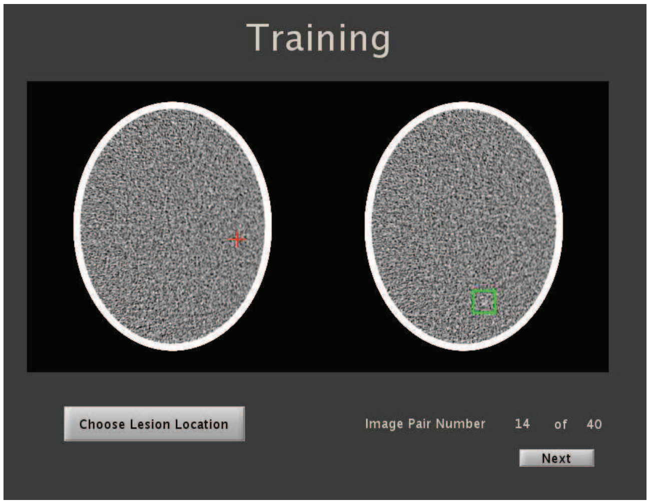

Fig. 1. Image display for a 2-AFC experiment that assesses the area under an LROC curve.

the fan-beam data to the parallel-beam geometry. Like algorithm A, algorithm C assigns a weight of 1/2 to handle all data redundancy.

B. Task description

Image quality was assessed using two-AFC experiments corresponding to LROC analysis. The LROC task was to detect a small lesion within a uniform brain phantom. Both the position of the lesion and the contrast of the lesion were random ($[25, 35]$ HU), whereas the lesion size was fixed (5 mm diameter). The lesion was always within the gray-matter area of the brain, and was not allowed to overlap with the skull.

In our context, the Bernoulli trial corresponded to presenting the observer with a pair of images as shown in Figure 1. One of the images always contained exactly one lesion whereas the other image did not contain any lesion. The observer was asked to insert a mark within one of the two images (see the red cross). A success was recorded when the mark identified the lesion within 10 pixels, otherwise a failure was recorded. In Fig.1, the lesion is indicated with a green square, showing that the mark was inserted at the wrong location.

C. Study design

We decided to assess performance using four observers reading each 250 pair of images (in two sessions of 125 images, with 40 training images before each session). To optimize statistical power, the exact same data sets were used for all three reconstruction algorithms, and different cases were used from one reader to another. Hence, the computed proportions were only correlated between algorithms.

Denote the proportions for algorithms A, B and C and reader j as A_j, B_j and C_j, and let $\hat{C}_j$ be the 3×3 covariance matrix for these three proportions. This matrix was estimated for each reader using (9). Next, define the reader-averaged proportions for the three algorithms as $\overline{A} = (A_1 + A_2 + A_3 + A_4)/4$, $\overline{B} = (B_1 + B_2 + B_3 + B_4)/4$, $\overline{C} = (C_1 + C_2 + C_3 + C_4)/4$. Given that the cases were independent from one reader to another, the covariance matrix for these reader-average proportions was

$$\Omega = \frac{1}{16} \sum_{j=1}^{4} C_j \,. \tag{19}$$

Confidence intervals were estimated for $\overline{A}$, $\overline{A} - \overline{B}$ and $\overline{A} - \overline{C}$; $\overline{A}$ was included to provide a reference value. Let $\underline{\hat{d}} = [\overline{A},\ \overline{A} - \overline{B},\ \overline{A} - \overline{C}]$. The covariance matrix for $\underline{\hat{d}}$ was obtained from Ω and the diagonal elements of this matrix were used to find a 98.33% confidence interval for each entry of $\underline{\hat{d}}$, by relying on asymptotic normality. The confidence intervals found for $\overline{A}$, $\overline{A} - \overline{B}$ and $\overline{A} - \overline{C}$ were $[0.7909, 0.8491]$, $[-0.0199, 0.0359]$ and $[0.1141, 0.1819]$ respectively. By Bonferroni's inequality, the joint probability for the three intervals together is at least 95%. As expected, we observe that Algorithm A significantly performs better than Algorithm C, due in particular to the extra amount of data involved in the reconstruction process. On the other hand, the difference between algorithms A and B is relatively small, and no conclusion can be made in favor of one method versus the other.

IV. Conclusion

We have presented a nonparametric methodology to evaluate the statistical variability of image quality assessment results based on MAFC experiments with multiple readers and cases. Our methodology views the readers as a fixed effect and the cases as a random effect. This setting is ideal for development and optimization of image formation processes, where using a large number of readers is impractical. For studies that invoke many readers, we recommend evaluating the variability due to the reader pool as well as that due to the cases, which may be done using the results in [6].

Although not discussed here, it can be shown that our theory also enables simple sample size calculations. The procedure to follow is very similar to that presented in [3] for LROC studies. Moreover, it turns out that there exist interesting links between our covariance matrix estimator, Jack-knifing techniques, and maximum likelihood estimation. These links will be discussed in the future.

References

[1] H. H. Barrett and K. J. Myers, *Foundations of Image Science*. Wiley, 2004.

[2] M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Univ. Press, 2003.

[3] A. Wunderlich and F. Noo, "A nonparametric procedure for comparing the areas under correlated lroc curves," *IEEE Trans. Med. Imaging*, In Press. 2012.

[4] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 17, pp. 153–157, 1947.

[5] B. Bennett, "Tests of hypotheses concerning matched samples," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 29, no. 3, pp. 468–474, 1967.

[6] P. G. Gallas, B.D. and K. Myers, "Multireader multicase variance analysis for binary data," *J. Opt. Soc. Am. A*, vol. 24, no. 12, pp. B70–B80, Dec. 2007.