**JafrocR: an R Package for Multiple Reader Multiple Case ROC Data Analysis**

# 1. Abstract

TBA

# 1. Introduction

In an imaging department the radiologist interprets images that could be from non-diseased or diseased patients, the true disease status is a-priori unknown, and the task is to correctly diagnose the patients. If the radiologist responds "diseased" to a diseased patient image, that is a true positive (TP) outcome, and if the radiologist responds "diseased" to a non-diseased patient image, that is a false positive (FP). False negatives (FN) and true negatives (TN) are the complements of true and false positives, respectively. Outcomes preceded by "true" represent correct decisions and those preceded by "false" represent incorrect decisions. An objective measure of performance metric, or figure-of-merit (FOM), would credit correct decisions and penalize incorrect ones. The FOM can be used to compare imaging systems, a common problem in medical imaging, given the rapid evolution of imaging technology and the need to maximize image quality and minimize radiation dose.

Receiver Operating Characteristic (ROC) studies are simplistic, but practical, laboratory simulations of what actually happens in the clinic. They are widely used to objectively compare different imaging modalities. As an example of two modalities, in the past there has been great interest in comparing conventional analog x-rays to digital images of the chest for the detection of lung nodules (refs) - that argument has been settled, as a result of several   
ROC studies, in favor of digital. In an ROC study the experimenter selects a set of patients on which the true-disease status for each patient is known, independently of the modalities to be compared. Typically about a hundred cases are selected, about half of which are actually diseased. The patients are imaged in each of the modalities to be evaluated. The images are shown one-at-a-time to a number (about five or more) of radiologists and their individual confidence levels in presence of disease is recorded using an agreed-upon rating scale. Commonly the rating (frequently recorded as a number, but it is really an ordered label) takes one of 5 values, e.g., the integers 1 through 5, where a 1-rating denotes high confidence in absence of disease and 5-rating denotes high confidence in presence of disease, and intermediate values denote intermediate confidence levels. Some experimenters prefer a finer scale, e.g., the integers 1 through 100, or even a floating point numeric scale (ref).

By cumulating the numbers of TP and FP events rated at or above a selected threshold, and dividing by the total number of diseased and non-diseased patients, respectively, one obtains the fraction of TP and FP events at or above the selected threshold, termed true positive fraction (TPF) and false positive fraction (FPF), respectively. These define an operating point on the ROC plot: the latter is defined as a plot of TPF (ordinate) vs. FPF. For example, cumulating only the events rated 5 yields the lowest ROC operating point, cumulating the 5s and 4s yields the next higher operating point, etc. The trivial operating points (0,0), obtained by cumulating no ratings, and (1,1), obtained by cumulating all the ratings, belong on every ROC plot. Statistical methods (refs) exist to fit the ROC operating points to parametric models; the fit can be used to infer the area under the fitted ROC curve. Alternatively, a non-parameteric measure of the area under the curve (AUC) can be obtained by calculating the trapezoidal area under the ROC, which can be shown to be equivalent to the Wilcoxon statistic applied to the ratings. When every reader rates every image in all modalities to be compared, that is defined as the multiple reader multiple case (MRMC) paradigm.

The fundamental idea behind ROC analysis is to compute a FOM metric that rewards correct decisions and penalizes incorrect decisions and AUC qualifies as such a metric (other FOMs are also used and will be introduced in context). When comparing two or more imaging systems (commonly referred to as modalities or treatments) one estimates AUC for each modality and reader and performs a statistical test to determine if the intermodality difference between the reader-averaged AUCs is sufficiently large compared to its statistical variability to be unlikely to be due to chance. The degree of "unlikelihood" is quantified by the p-value. The calculatiom the p-value and related quantities involves a process termed signficance testing which is very familiar to statisticians. Over the past 3 decades a specialized branch of statistics devoted to analysis of ROC ratings data, and loosely termed “ROC analysis”, has evolved that has developed better accounting of the available information and data strucuture and obtaining higher statistical power. The reason for the importance of ROC analysis is that critical decisions are made, based on MRMC ROC studies, on whether or not to accept new technological innovations in imaging technology and regulatory requirements often require adequately powered MRMC ROC studies before new imaging modalities, or variants of existing ones, can be marketed.

ROC analysis involves two steps: (1) selection of a FOM to quantify radiologist performance and (2) selection of an algorithm for significance testing. We have already mentioned two FOMs, the fitted area under the ROC curve and the trapezoidal area, denoted. The fitted area depends on the fitting model: the binormal model (Dorman ref. 1969) fitted area is denoted, the contaminated binormal model (ref) fitted area is denoted , the proper ROC (PROPROC) model fitted area is denoted . Some investigators prefer partial area measures, generically denoted; for example a partial area measure might be chosen as a suitably normalized area under the ROC curve that lies above a pre-selected FPF value (ref).   
Still other prefer the interpolated TPF at a specified FPF, denoted  .

There are several algorithms for significance testing. All algorithms report, at the very minimum, a p-value for rejecting the null hypothesis (NH) that the modalities are identical in performance. Additionally, they report confidence intervals of intermodality differences in figures of merit, and a number of other statistics. Current algorithms for ROC analysis are summarized in Table 1 along with the names of online downloadable implementations, where available, and the available choices of FOMs.

Table 1: Software availability of MRMC ROC signficance testing algorithms

|  |  |  |  |
| --- | --- | --- | --- |
| Significance testing algorithm | Software Names | FOMs | website |
| DBM-H(ref) | OR-DBM MRMC |  | XZ |
| JAFROC |  | XZ |
| DBM MRMC |  | XZ chicago |
| OR-H(ref) | OR-DBM MRMC |  | XZ iowa |
| BDG/BCK\* (2 ref) | iMRMC |  | XZ |
| OR (ref) | OBUMR2 | XZ cleaveland clinic |
| TG (ref) | NA | | |
| Song (ref) |
| IG (ref) |
| BWC (ref) |

Caption: \* BDG/BCK provides output that allows one to perform DBM-H or OR-H significance testing.

XZ: captions: spell out all abbreviations here ... also next version must have references; also, are you absolutely sure about the software names?

OR-DBM MRMC is maintained by the Medical Image Perception Laboratory of the University

of Iowa (website).

Table 2 compares the available algorithms, and RJafroc, the contribution of this paper, with respect to the following critera: (1) ease of data input (least for text files and most for Excel files); (2) whether or not the implementaion is open source, and if yes, the progrmming language expertise needed to customize it; (3) the portability of the software across different computing platforms; and (4) the ability to call the algorithm from other programming languages.

Table 2: Software comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Softwares | OR-DBM MRMC  DBM MRMC | JAFROC | iMRMC | JafrocR |
| Data Input | text | Excel | text | Excel |
| Open Source/Language | No | No | Yes/Java | Yes/R |
| Portability | No | No | Yes | Yes |
| Call from Other Languages | No | No | No | Yes |

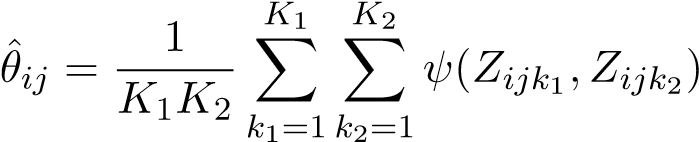
* JAFROC, an acronym for *j*ackknife *a*lternative *f* ree-response *r*eceiver *o*perating *c*haracteristic, is a free software applicable to the planning and analysis of FROC studies. The software is developed by Dr. Dev Chakraborty, Dr. Hong-Jun Yoon and Xuetong Zhai. For ROC data, JAFROC performs MRMC analysis using DBM method with AUC figure of method. Empirical estimation of AUC is computed using trapezoidal/Wilcoxon method, while no curve fitting method is included. JAFROC has the capability to analyze data accounting for localization and multiple lesions per image, which is described in section 4.
* iMRMC is a Java-based software. It is a open source project under the management of Dr. Brandon Gallas, Dr. Xin He and Rohan Pathare who are from US Food and Drug Administration. It estimates the variance components that are used in methods besides DBM and OR, including BDG, BCK and MS variance components. Hypothesis testing and confidence interval on the difference of two modalities are calculated using t-statistic with corresponding degrees of freedom. Similar with JAFROC, iMRMC also only selects AUC as the figure of merit, and estimates empirical AUC with trapezoidal/Wilcoxon method.

All of these three softwares are GUI (graphical user interface)-based, so it is difficult to call functions or grab the results from other program. But this is important for users with programming requirements. It is absolutely unreasonable to let users run softwares manually and copy the result to their own code every time. Considering R is widely used in statistic and data analysis, it is necessary to develop an R package with the capability to perform MRMC ROC analysis. Firstly, R users will be able to call functions in the package from their programs. Since all R packages are open source and R is a“higher”level language, users have the possibility to debug and modify the source code to satisfy their requirements without complicated programming skills. In addition, thanks to the amazing extensibility of R, packages will be available to be called from other languages, including C/C++, Python, Java and MATLAB, on all major platforms. Therefore, we developed the package JAFROCwR (JAFROC within R) to make the analysis easy and convenient, which implements all features in JAFROC. The package is available to download from the Comprehensive R Archive Network at http://CRAN.R-project.org/package=JAFROCwR. Table 2 shows a brief comparison of the three softwares and the package JAFROCwR.

This paper is organized as follows. The statistical model and method used in the package JAFROCwR are introduced in Section 2. Section 3 includes examples of the use of the package and analysis results of a same dataset from the package and other softwares. Localization issue and other additional features are shown in 4. Limitations and potential future updates is discussed in Section 5. Finally, the conclusions are described in Section 6.

# 2. Statistical Models and Methods

Let *Zijk* denote the rating given to the *k*th case by the *j*th reader using the *i*th modality with *i* = 1*,...,I*, *j* = 1*,...,J*, *k* = 1*,...,K*, where *I* is the number of modalities, *J* is the number of readers and *K* is the number of cases. Five or more integers are usually used with higher values meaning a higher confidence that the case contains lesions. The area under the ROC curve (AUC) can be used as a figure of merit, which is estimated by



where *k*1 and *k*2 denote the indices for non-diseased and diseased cases with *K*1 and *K*2 being the total numbers of them respectively (*K* = *K*1 + *K*2). The Wilcoxon function *ψ* is defined as,

 1 *X < Y*



*ψ*(*X,Y* ) = 0*.*5 *X* = *Y* (1)

 0 *X > Y*

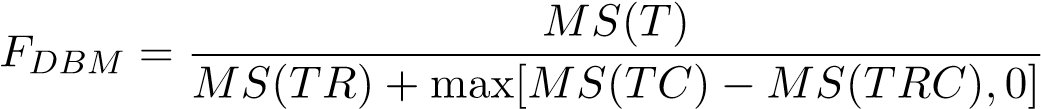
## **2.1. DBM Method**

For the DBM method, the pseudovalues of AUC are calculated using jackknife for combination of each modality, reader and case. Let *Yijk* indicate the AUC pseudovalue of modality *i*, reader *j* and case *k*. *Yijk* is defined as *Yijk* = *Kθ*ˆ*ij* − (*K* − 1)*θ*ˆ*ijk*, where *θ*ˆ*ijk* is the AUC estimate of modality *i*, reader *j* with the rating of case *k* removed. The statistical model is given by

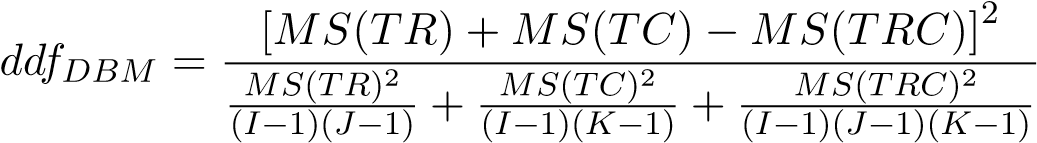
*Yijk* = *µ* + *τi* + *Rj* + *Ck* + (*τR*)*ij* + (*τC*)*ik* + (*RC*)*jk* + (*τRC*)*ijk* + *εijk*

where *τi* denotes the fixed effect of the *i*th modality, *Rj* denotes the random effect of the *j*th reader, *Ck* denotes the random effect of the *k*th case, and *εijk* is the error term. The interaction terms denote the mixed random effects of corresponding factors. All of the random terms are assumed to be independent and follow normal distributions with zero means and variances *σR*2 , *σC*2 , *στR*2 , *στC*2 , *σRC*2 , *στRC*2 and *σε*2.

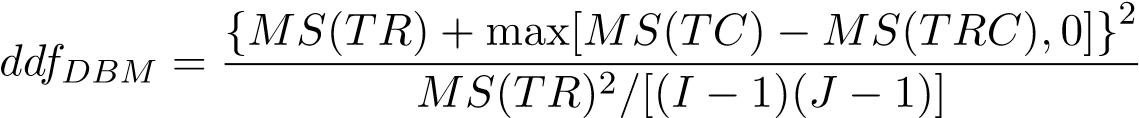
The *F* statistic for testing the null hypothesis of equal modality effects is defined,



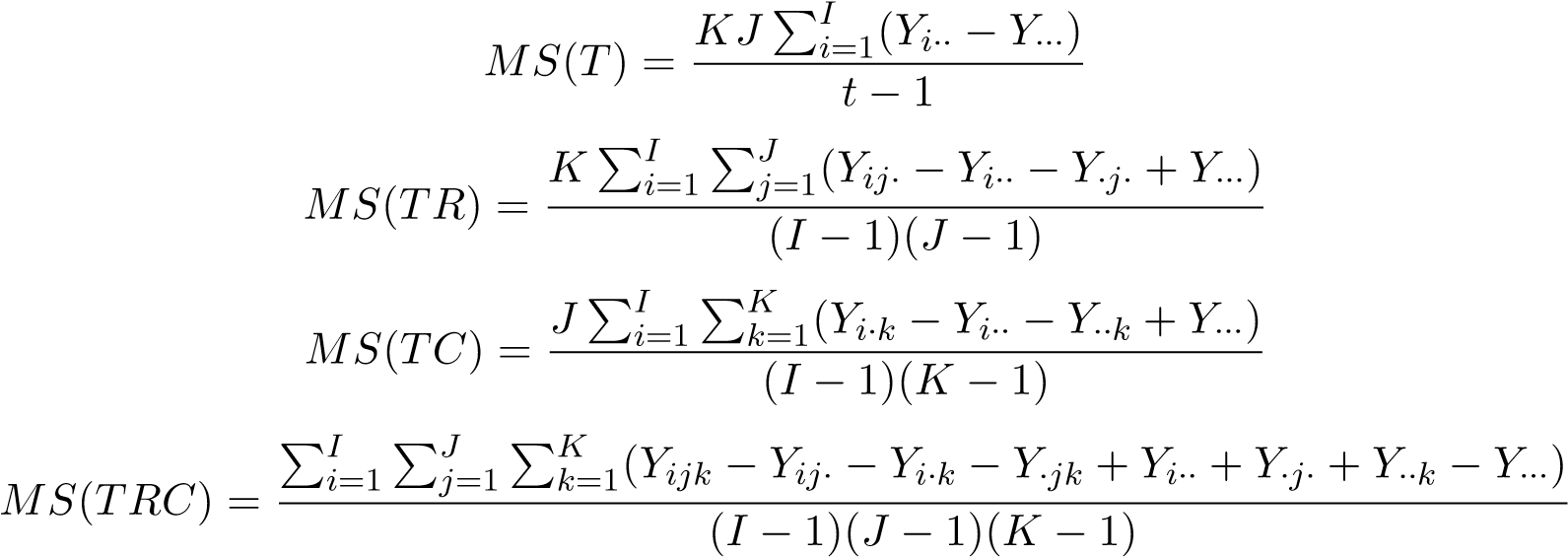
with the numerator degrees of freedom *t* − 1 and denominator degrees of freedom



It is corrected by Hillis,



where *MS*(*T*), *MS*(*TR*), *MS*(*TC*) and *MS*(*TRC*) are the pseudovalue means squares of modality, modality by reader, modality by case, modality by reader by case correspondingly, by definition,



where *Y* with a “·” subscript means the averaged pseudovalue over the replaced subscript.

## **2.2. OR Method**

The statistical model of OR method is defined by

*θ*ˆ*ij* = *µ* + *τi* + *Rj* + (*τR*)*ij* + *εij* (2)

where *θ*ˆ*ij* denotes the AUC estimate for modality *i* and reader *j*. Other terms except for *εij* are defined in the same way as DBM method. *εij* is not assumed to be independent here, since each case is read by each reader with each modality. The covariances of error terms are defined by

|  |  |  |
| --- | --- | --- |
|  Cov1    Cov(*εij,εi*0*j*0) = Cov2   Cov3 | *i* 6= *i*0*,j* = *j*0 (different modality, same reader) *i* = *i*0*,j* 6= *j*0 (same modality, different reader) *i* 6= *i*0*,j* 6= *j*0 (different modality, different reader) | (3) |

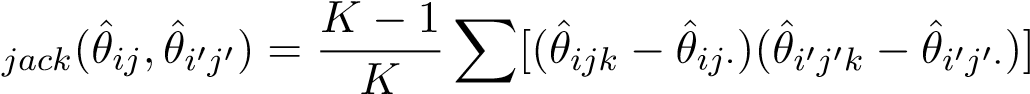
It is suggested that the covariances should be ordered as following sequence

Cov1 ≥ Cov2 ≥ Cov3 ≥ 0

In the condition of the model 2 (readers are regarded as fixed), covariances between error terms are also the corresponding covariances of the AUC estimates.

The OR variance and covariance components can be estimated using Jackknife, Bootstrap or DeLong’s method.

* Jackknife method calculates AUCs for each modality-reader combination with one case deleted from the dataset. Let *θ*ˆ*ij*· denote the the mean of *θ*ˆ*ijk* that is defined as same as in DBM model. The jackknife covariance estimate of the covariances between AUCs for readers *j*, *j*0 and modality *i*, *i*0 is given by

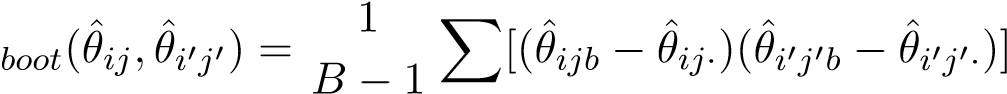
*K*

Cov

*k*=1

* Let *K*1 and *K*2 denote the number of non-diseased and diseased cases respectively. Sample *B* groups of *K* cases indices with replacement from *K*1 non-diseased cases and *K*2 diseased cases. Let *θ*ˆ*ijb* be the AUC estimate for modality *i* and reader *j* calculated using the *b*th bootstrap of sample. The bootstrap covariance estimate of the covariances between AUCs for readers *j*, *j*0 and modality *i*, *i*0 is given by

*B*

Cov

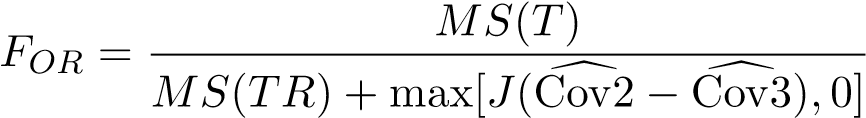
*b*=1

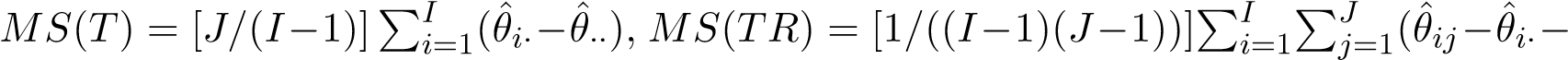
where *θ*ˆ*ij*· and *θ*ˆ*i*0*j*0· denote the means of AUCs using all bootstrap samples for corresponding reader and modality.

* DeLong derived the variance and covariance for trapezoidal AUC estimates. This method can be only used to trapezoidal AUC.

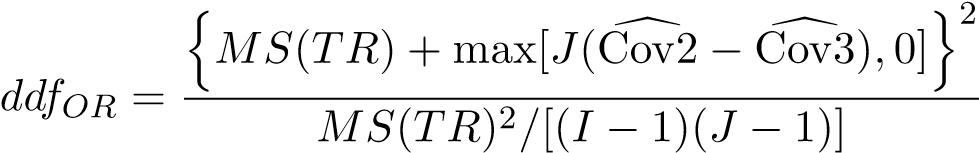
The OR covariances components Cov1, Cov2 and Cov3 can be calculated by averaging the covariances following the condition in equation 3

The *F* statistic for testing the null hypothesis of no modality effects is given by



where

*θ*ˆ·*j* +*θ*ˆ··), Cov2 andd Cov3 denote the estimates of Cov2 and Cov3.d *θ*ˆwith a“·”subscript means the averaged value over the replaced subscript. *FOR* approximately follows a *F* distribution with degrees of freedom *I* −1 and (*I* −1)(*J* −1). The denominator degrees of freedom *ddfOR* is corrected by Hillis, where

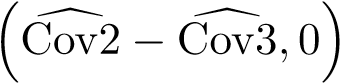


p-value can be calculated with the *F* value and degrees of freedom. Compare the p-value with the significance level *α*, we can determine if the null hypothesis should be rejected.

## **2.3. Sample Size Calculation**

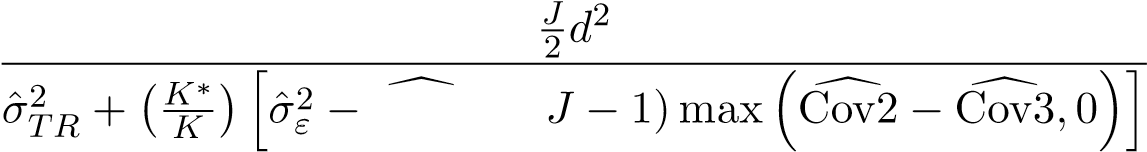
If a not significant result is obtained (i.e., *p >* 0*.*05) from a study, then the investigator may wish to estimate, from the observed variability of the data, how many cases and readers would be needed to achieve a given probability (typically chosen to be 80%) of detecting a true difference between the modalities in a subsequent study. This is termed sample size calculation. Method for statistical power calculation for ROC analysis is described in Ref. Briefly, we can calculate the statistical power with OR components via following steps. All OR components used are defined in the subsection 2.2.

1. *Specify the effect size.* Let *d* denote the effect size that is the absolute value of the difference of two AUCs. The effect size is defined by *d* = |*AUC*1 − *AUC*2|.
2. *Estimate OR parameter.* The variance of modality by reader interaction can be estimated by

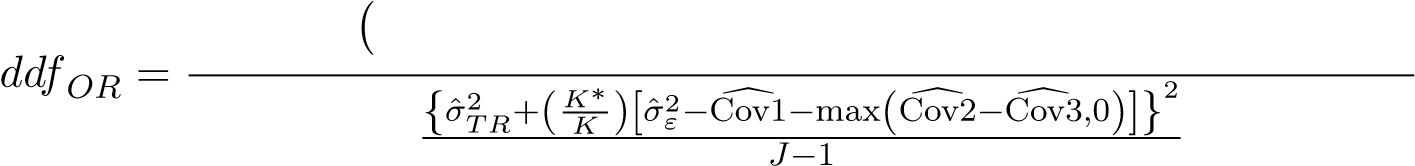
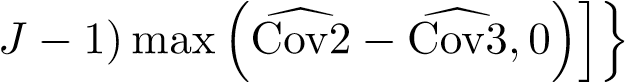
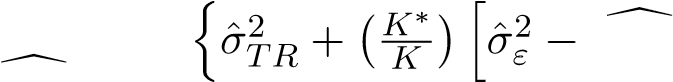
 Cov1 + maxd 

1. *Estimate the noncentrality parameter and the degrees of freedom*. Let *K*∗ denote the number of cases in current analysis, and let *J* and *K* be the numbers of readers and cases, which will be used to calculate the statistical power. The noncentrality parameter ∆ and the degrees of freedom ofˆ *FOR*’s denominatoris estimated by

∆ =ˆ

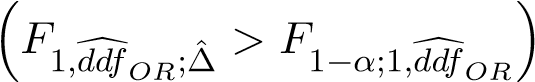
Cov1 + (

and

2

Cov1 + (

1. *Calculate the statistical power using the noncentrality parameter and degrees of freedom estimation.* The statistical power with significance level *α* can be calculated by

Power = Pr

To estimate the sample size for ROC analysis, we usually select a fixed number of readers(or cases) and calculate the statistical power by incrementing the number of cases(readers, vice versa) until the calculation result reaches the desired power. The number of readers and cases combination is a required sample size.

# 3. Examples and Results Comparisons

In this section, some examples for the usage of the package JAFROCwR are shown, and the analysis results of JAFROCwR and other softwares mentioned in Section 1 are provided and compared. The example dataset comes from Carolyn Van Dyke, MD. There are 45 diseased cases and 69 non-diseased cases in the dataset. Five radiologists (reader 0 to 4) read these images using two modalities (modality 0 and 1) and give their ratings for each case.

## **3.1. Analyze the Example Dataset Using** JAFROCwR

The first steps are loading the package and read the data file.

*> library(JAFROCwR)*

*> dataVanDyke <- readJAFROC("VanDyke.xlsx")*

Then we analyze the data using both DBM and OR method.

*> resultOR <- ORAnalysis(data = dataVanDyke, analysisFOM = "ROC", alpha = 0.05,*

*+ covEstMethod = "Jackknife")*

The result can be found in table

Following is an example of the plotting feature of this package. Figure 1 is the empirical ROC curves of each individual reader and their average performance using modality 0.

*> rocCurveM1 <- plotROC(data = dataVanDyke, plottingModalities = 1,*

*+ plottingReaders = list(1, 2, 3, 4, 5, c(1:5)),*

*+ legendPosition = "bottom")*

*> rocCurveM1$ROCPlot*

0.00

0.25

0.50

0.75

1.00

0.00

0.25

0.50

0.75

1.00

FPF

TPF

M−0

R−0

M−0

R−1

M−0

R−2

M−0

R−3

M−0

R−4

M−0

R−0 1 2 3 4

Figure 1: Plots of empirical ROC curves of each reader and their average performance using modality 0

The function for sample size calculation with effect size = 0.05 and desired power = 0.8 are as follows. Result can be found in table

*> calculateSampleSizeForData(data = dataVanDyke, alpha = 0.05,*

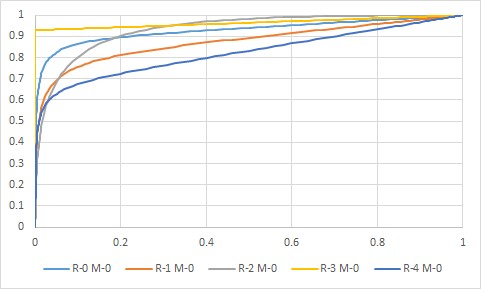
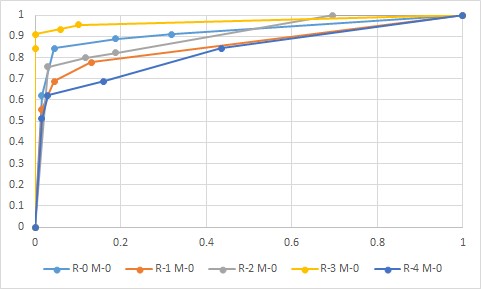
*+ Effect\_Size = 0.05, Desired\_Power = 0.8)*

## **3.2. Results of Other Softwares**

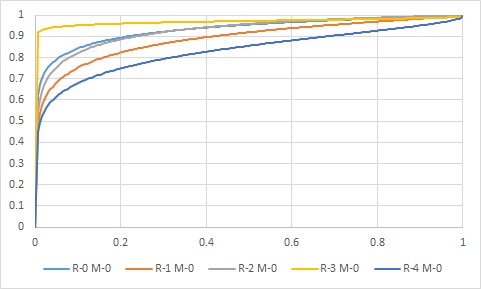
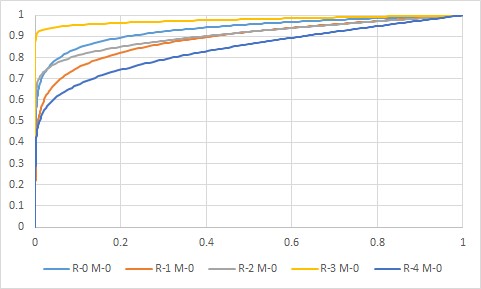
Analysis results of the same dataset used in subsection 3.1 by other common ROC softwares are shown in this subsection. AUC is selected as the figure of merit. DBM method or OR method using jackknife estimation is used for the analysis. JAFROCwR includes functions that save JAFROC data using formats of other softwares. Users can use these functions to convert data between different formats.

## OR-DBM MRMC

OR-DBM MRMC performs similar analysis as JAFROCwR. The example of how to use the software can be found in the user manual. Results is compared in table . Figure 2 show the empirical and fitted ROC curves for all readers using modalities 0.



(a) Empirical ROC curves (b) CBM fitted ROC curves



(c) PROPROC fitted ROC curves (d) RSCORE fitted ROC curves

Figure 2: OR-DBM MRMC ROC Curves Plots

## iMRMC

iMRMC gives the test result of the difference between two modalities. Since we have only two modalities in the example datasets, we just choose them as the two modalities to be tested. Testing results are reorganized and compared with other softwares in table . Variance components used in statistical models other than DBM and OR are provided and shown below.

Table 3: BDG variance components

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | M1 | M2 | M3 | M4 |
| comp M0 | 8.6671E-01 | 8.4536E-01 | 8.1617E-01 | 8.0714E-01 |
| coeff M0 | 6.4412E-05 | 4.3800E-03 | 2.8341E-03 | 1.9272E-01 |
| comp M1 | 9.2320E-01 | 9.0951E-01 | 8.8986E-01 | 8.8622E-01 |
| coeff M1 | 6.4412E-05 | 4.3800E-03 | 2.8341E-03 | 1.9272E-01 |
| comp product | 8.6164E-01 | 8.5978E-01 | 8.4613E-01 | 8.4528E-01 |
| - coeff product | 1.2882E-04 | 8.7601E-03 | 5.6683E-03 | 3.8544E-01 |
| total | 4.2921E-06 | 1.5470E-04 | 3.9001E-05 | 5.3964E-04 |
|  | M5 | M6 | M7 | M8 |
| comp M0 | 8.3454E-01 | 8.2855E-01 | 8.0738E-01 | 8.0468E-01 |
| coeff M0 | 2.5765E-04 | 1.7520E-02 | 1.1337E-02 | -2.2911E-01 |
| comp M1 | 9.0132E-01 | 8.9642E-01 | 8.8668E-01 | 8.8517E-01 |
| coeff M1 | 2.5765E-04 | 1.7520E-02 | 1.1337E-02 | -2.2911E-01 |
| comp product | 8.5626E-01 | 8.5488E-01 | 8.4459E-01 | 8.4397E-01 |
| - coeff product | 5.1530E-04 | 3.5040E-02 | 2.2673E-02 | -4.5823E-01 |
| total | 6.0143E-06 | 2.6665E-04 | 5.5381E-05 | -4.3955E-04 |

Table 4: BCK variance components

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | N | D | N ˜D | R |
| comp M0 | 2.7003E-03 | 2.3874E-02 | 3.2916E-03 | 2.4661E-03 |
| coeff M0 | 1.4493E-02 | 2.2222E-02 | 3.2206E-04 | 2.0000E-01 |
| comp M1 | 1.5063E-03 | 1.1249E-02 | 3.3939E-03 | 1.0437E-03 |
| coeff M1 | 1.4493E-02 | 2.2222E-02 | 3.2206E-04 | 2.0000E-01 |
| comp product | 6.1991E-04 | 1.0910E-02 | 7.6453E-04 | 1.3141E-03 |
| - coeff product | 2.8986E-02 | 4.4444E-02 | 6.4412E-04 | 4.0000E-01 |
| total | 4.2996E-05 | 2.9559E-04 | 1.6607E-06 | 1.7633E-04 |
|  | N ˜R | D ˜R | R ˜N ˜D |  |
| comp M0 | 6.3239E-03 | 1.4348E-02 | 9.0360E-03 |  |
| coeff M0 | 2.8986E-03 | 4.4444E-03 | 6.4412E-05 |  |
| comp M1 | 2.1399E-03 | 1.2039E-02 | 6.6583E-03 |  |
| coeff M1 | 2.8986E-03 | 4.4444E-03 | 6.4412E-05 |  |
| comp product | 2.3462E-04 | 3.5850E-03 | 2.4796E-04 |  |
| - coeff product | 5.7971E-03 | 8.8889E-03 | 1.2882E-04 |  |
| total | 2.3172E-05 | 8.5409E-05 | 9.7896E-07 |  |

Empirical ROC curve can be plotted using iMRMC. Figure 3 is the plot containing individual and averaged ROC curves for all readers using modality 0.

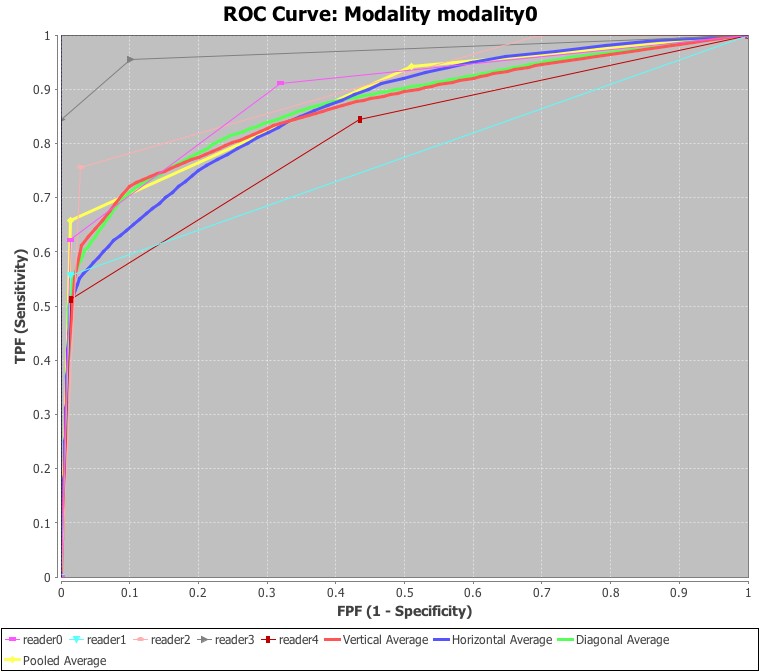


Figure 3: Individual and averaged ROC curves output by iMRMC

### 3.3. Results Comparison

In this section, we use some tables to show the results of the softwares. JAFROC is omitted since the package JAFROCwR produces identical results with it. Another dataset “franken” is also analyzed. Table 5

# 4. Additional Features of JAFROCwR

The main limitation of ROC analysis is that only one rating is used to describe each image. The radiologists may recognize a different region from where the true lesion is located, but a reasonable rating will still be given. The effect caused by this kind of errors is not revealed in the ROC analysis. The free-response receiver operating characteristic (FROC) analysis takes marks localization into account. In the free-response paradigm, the radiologist is free to mark and rate more than one suspicious regions that are considered. In this case, the number of Table 5: AUC calculation comparison for the dataset “Van Dyke”

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | OR-DBM MRMC | | | iMRMC | | | JAFROCwR | |
|  | M - 0 | M - 1 | M - 0 | | M - 1 | M - 0 | | M - 1 |
| R - 0 | 0.9196 | 0.9478 | 0.9196 | | 0.9478 | 0.9196 | | 0.9478 |
| R - 1 | 0.8588 | 0.9053 | 0.8588 | | 0.9053 | 0.8588 | | 0.9053 |
| R - 2 | 0.9039 | 0.9217 | 0.9039 | | 0.9217 | 0.9039 | | 0.9217 |
| R - 3 | 0.9731 | 0.9994 | 0.9731 | | 0.9994 | 0.9731 | | 0.9994 |
| R - 4 | 0.8298 | 0.9300 | 0.8298 | | 0.9300 | 0.8298 | | 0.9300 |

marks is also regarded as a random variable. This source of randomness is the main difficulty of FROC analysis. Some figures of merit has been defined so DBM and OR method can be used for FROC analysis.

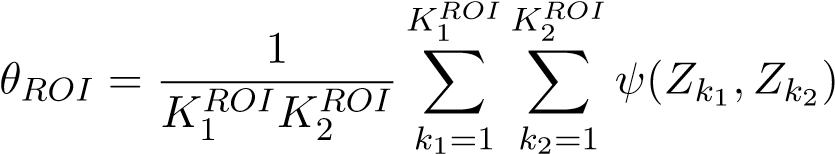
**4.1. ROI and FROC Figures of Merit Used in** JAFROCwR

## Notation

For ROI and FROC study, lesion localization is taken into consideration. Hence more than one marks are allowed for each case. Let *kt* denote the index of cases where t indicates the disease-status at the case (or patient) level, with *t* = 1 for non-diseased cases and *t* = 2 for diseased cases; *k*1 ranges from 1 to *K*1 for non-diseased cases and *k*2 ranges from 1 to *K*2 for diseased cases. Marks are indexed by *ls* where *s* indicates the truth at the location level, with *s* = 1 for a non-lesion localization and *s* = 2 for a lesion localization; *l*1 = 1*,*2*,...*, indexes marks of type *s* = 1 and *l*2 = 1*,*2*,...*, *nk*2, indexes marks of type *s* = 2, where *Nk*2 is the number of lesions visible in image *k*2. *Zktls* denotes the rating of mark *ls* on case *kt*. *Zk*2*l*2 is assigned −∞ if the corresponding lesion localization is unmarked.

## ROI Figure of Merit

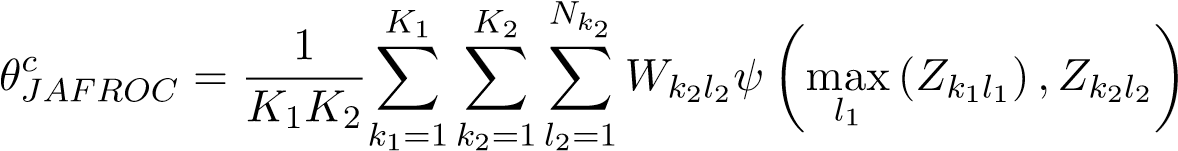
In the region of interest (ROI) paradigm the investigator segments the image into a number of regions of interest (ROIs) and the radiologist rates each ROI for presence of at least one lesion somewhere within the ROI. The data consisted of a rating for each ROI. The ROI figure of merit *θijROI* is defined by



where *ψ* is the function defined by equation  is the number of non-diseased ROIs and  is the number of diseased ROIs, ) and) are the ratings for corresponding ROIs.

## FROC Figures of Merit

The weighted JAFROC figure of merit is defined by (the superscript *c* denotes case-based)



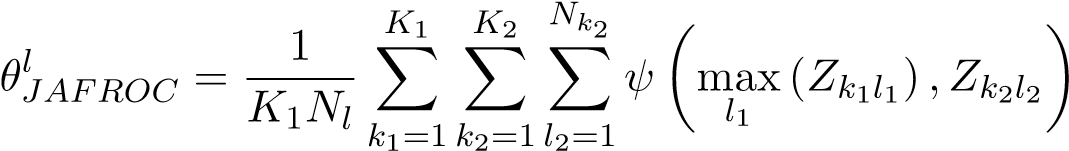
*Nk*2

X

*Wk*2*l*2 = 1

*l*2=1

where *ψ* is the function defined by equation 1, max*ls* (*Zktls*) is the maximum over the ratings of all *s* localizations on the case *kt* and *Wk*2*l*2 is the weight of lesion *l*2 in diseased case *k*2. The un-weighted JAFROC figure of merit is defined by (the superscript *l* denotes lesion-based)



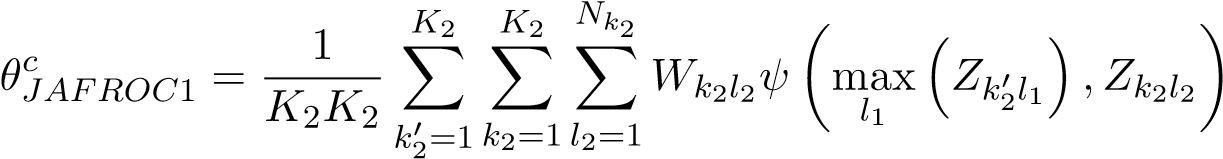
where *Nl* is the total number of lesions in all diseased cases by definition

*K*2

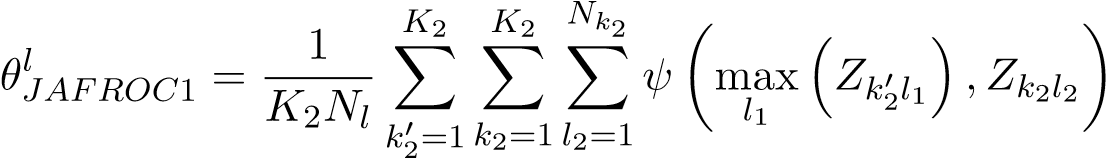
*Nl* = X *Nk*2

*k*2=1

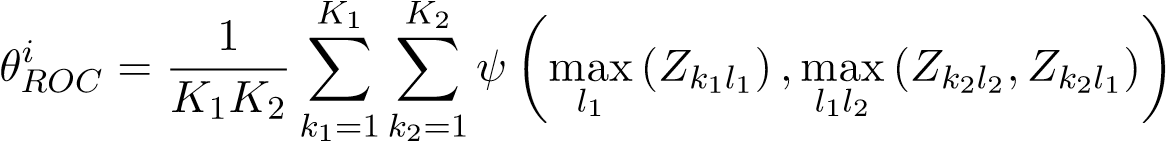
The weighted JAFROC1 figure of merit is defined by



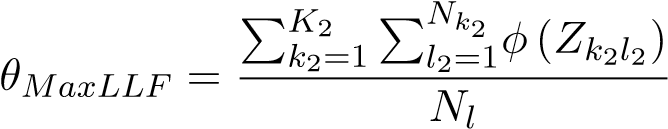
The un-weighted JAFROC1 figure of merit is defined by



The inferred-ROC figure of merit is defined by (the superscript i denotes inferred)

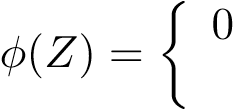


The maximum lesion localization fraction figure of merit is defined by

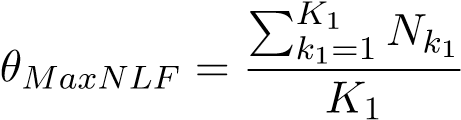


where *φ* is a function given by

*Z* is − ∞

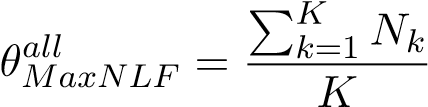
 1 otherwise

The maximum non-lesion localization fraction figure of merit is defined by

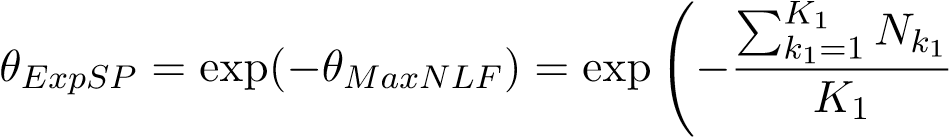


where *Nk*1 denotes the number of non-lesion localization marks in case *k*1.

The maximum non-lesion localization fraction for all cases figure of merit is defined by



where *Nk* denotes the number of non-lesion localization marks in case *k*. The exponential transformed specificity figure of merit is defined by

!

The Song figures of merit A0, A1 and A2 are also used in JAFROCwR. They are defined in

Ref.

### 4.2. ROC/AFROC/FROC Plots

Empirical and parametric ROC/AFROC/FROC curves can be plotted using functions in JAFROCwR. These curves are defined as follows.

## Empirical Plots

The ROC curve is a plot of true positive fraction (*TPF*) vs. false positive fraction (*FPF*), where *TPF* = number of positive decisions / number of diseased cases and *FPF* = number of positive decisions / number of non-diseased cases. For FROC data, *TPF* and *FPF* are calculated using the inferred ROC data by highest rating assumption. The AFROC curve is a plot of lesion localization fraction (*LLF*) vs. *FPF*, where *LLF* = number of lesions localizations / the total number of lesions. The FROC curve is a plot of *LLF* vs. non-lesion localization fraction (*NLF*), where *NLF* = number of non-lesion localizations / the number of cases.

## Parametric Plots

Parametric ROC/AFROC/FROC curves are parametric predicted (fitted) curves using search model, which are defined in subsection 4.3. The steps that fit the curves are also given in that part.

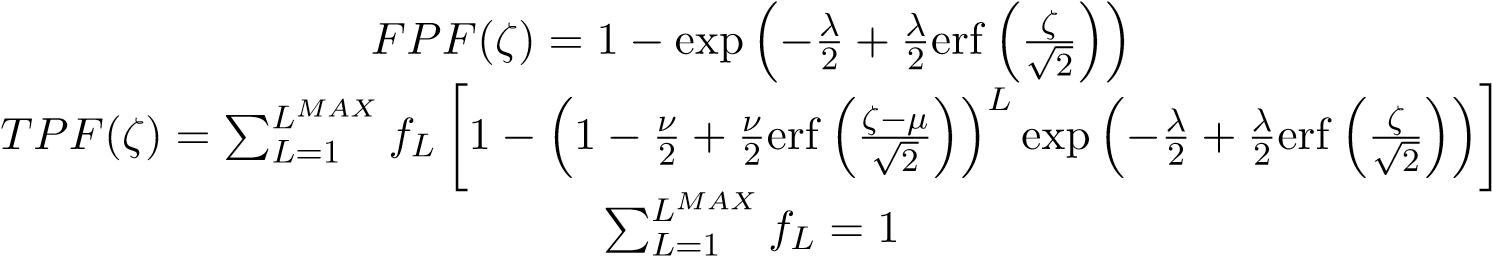
### 4.3. Search-Model

The search-model is used to estimate sample size and fit ROC and (A)FROC curves for FROC data in JAFROCwR. According to Kundel and Nodine, image interpretation involves a search stage and a decision-making stage. During the search stage the radiologist quickly identifies suspicious regions. The search-stage is characterized by a random number of suspicious regions. During the decision making stage the observer calculates a rating at each suspicious region, and if it exceeds the minimum reporting threshold the region is marked.

Suspicious regions are termed *noise-sites* or *signal-sites* for benign tumors or lesions, respectively. The number of noise-sites on an image is assumed to be sampled from a Poisson distribution with mean *λ*. The number of signal-sites on a diseased image is assumed to be sampled from a binomial distribution with success probability *ν* and the total number of lesions on the diseased case *k*2. The ratings form noise-site is assumed sampled from a normal distribution *N*(0*,*1) and that from signal-site is assumed sampled from *N*(*µ,*1).

## Search-model predicted ROC curves

The ROC curve of FROC data can be inferred using the highest rating assumption. The ROC curve predicted by the search-model is defined by

 (4)

where *ζ* is the cutoff parameter determining an operating point on the ROC curve, *L* is the number of lesions in a diseased case, *fL* is the fraction of diseased cases with *L* lesions, and erf(*x*) is the error function. *TPF*(*ζ*) is a weighted average of true positive fraction for cases with *L* = 1*,*2*,...,LMAX* lesions.

## Search-model predicted AFROC curves

The expression for the FPF predicted by the search-model has already been given in equation 4. That for the lesion localization fraction (LLF) is

*LMAX*

*LLF*(*ζ*) = X *fL* [*ν* (1 − Φ(*ζ* − *µ*))] = *ν* (1 − Φ(*ζ* − *µ*)) (5)

*L*=1

where Φ(*ζ*) is the cumulative distribution function (CDF) of standard normal distribution.

## Search-model predicted FROC curves

The expression for the LLF predicted by the search-model has already been given in equation 5. That for the non-lesion localization fraction (NLF) is

*NLF*(*ζ*) = *λ*(1 − Φ(*ζ*)) (6)

## ROC Likelihood function

Let (*Fb,Tb*) denote the number of false positive and true positives, respectively, in ratings bin *b* defined by neighboring cutoffs (*ζb,ζb*+1), where *b* = 0*,*1*,...,R* where *R* is the number of bins, and *ζ*0 = −∞ and *ζR*+1 = +∞. For example, *F*0 and *T*0 represent the number of non-diseased and diseased cases with no marks respectively. The contribution of the ROC likelihood function from bin *b* is

L*ROCb* = [*FPF*(*ζb*) − *FPF*(*ζb*+1)]*Fb*[*TPF*(*ζb*) − *TPF*(*ζb*+1)]*Tb* (7)

The net likelihood L*ROC* is the product of the ROC likelihood function from all *R* + 1 bins,

*R*

L*ROC* = YL*ROCb*

*b*=0

→−

To fit the ROC curve, we need to estimate parameters *λ*, *ν*, *ζ* and *µ* that maximize the logarithm of the likelihood 7. Following algorithm is used in JAFROCwR

→−

1. For given *λ*, *ν* and cutoffs *ζ* , determine *µ* by minimizing the Chi-square goodness of fit statistic
2. Calculate the log-likelihood of L*ROC*
3. Repeat preceding steps using varied *λ*, *ν* and →−*ζ* until reach the maximum of L*ROC*.

Search-model fitted ROC curve for (F)ROC data can be plotted using model 4 with optimized parameters.

## AFROC Likelihood function

Let (*Fb,Lb*) denote the number of false positive and lesion localizations, respectively, in bin *b* between neighboring cutoffs (*ζb,ζb*+1). The contribution of the AFROC likelihood function from bin *b* is

L*AFROCb* = [*FPF*(*ζb*) − *FPF*(*ζb*+1)]*Fb*[*LLF*(*ζb*) − *LLF*(*ζb*+1)]*Lb* (8)

## FROC Likelihood function

Let (*Nb,Lb*) denote the number of non-lesion localizations and lesion localizations, respectively, in bin *b* between neighboring cutoffs (*ζb,ζb*+1). The contribution of the FROC likelihood function from bin *b* is

L*FROCb* = [*NLF*(*ζb*) − *NLF*(*ζb*+1)]*Nb*[*LLF*(*ζb*) − *LLF*(*ζb*+1)]*Lb* (9)

Search-model fitted AFROC and FROC curves can be plotted using model 8 and 9 with optimized parameters.

**4.4. FROC Example**

## ROI and FROC Analysis

ROI and FROC data file are able to be analyzed by JAFROCwR using figures of merit that were mentioned in subsection 4.1. The usage of functions for ROI/FROC data analysis are same as that for ROC analysis. For FROC data, empirical and search model parametric AFROC/FROC curves are available. Figure 4 and 5 are the empirical and parametric AFROC/FROC curves.

*> frocData <- readJAFROC("frocDataFile.xls")*

*> afrocCurve <- plotAFROC(data = frocData, plottingModalities = 1,*

*+ plottingReaders = list(1, 2, 3, 4, 5, c(1:5)),*

*+ legendPosition = "bottom")*

*> fittedAfrocCurve <- plotFittedAFROCCurve(data = frocData, plottingModalities = 1,*

*+ plottingReaders = list(1, 2, 3, 4, 5, c(1:5)),*

*+ legendPosition = "bottom")*

*> frocCurve <- plotFROC(data = frocData, plottingModalities = 1,*

*+ plottingReaders = list(1, 2, 3, 4, 5, c(1:5)),*

*+ legendPosition = "bottom")*

*> fittedFrocCurve <- plotFittedFROCCurve(data = frocData, plottingModalities = 1,*

*+ plottingReaders = list(1, 2, 3, 4, 5, c(1:5)),*

*+ legendPosition = "bottom")*

*> multiplot(afrocCurve$AFROCPlot, fittedAfrocCurve$AFROCPlot, cols = 2)*

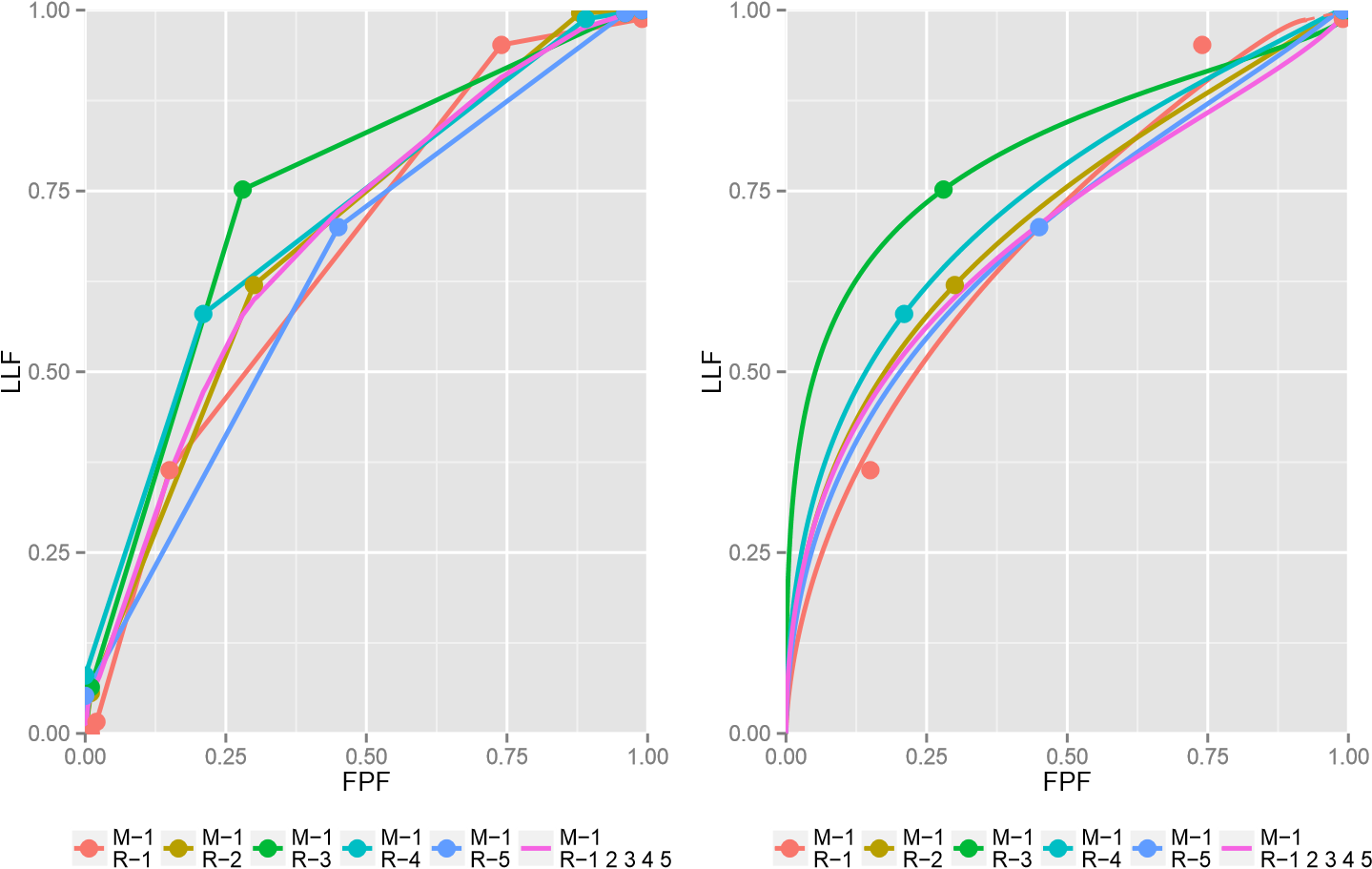


Figure 4: Plots of empirical (left) and fitted (right) AFROC curves of each reader and their average performance using modality 0

*> frocCurve <- plotFROC(data = frocData, plottingModalities = 1,*

*+ plottingReaders = list(1, 2, 3, 4, 5, c(1:5)),*

*+ legendPosition = "bottom")*

*> fittedFrocCurve <- plotFittedFROCCurve(data = frocData, plottingModalities = 1,*

*+ plottingReaders = list(1, 2, 3, 4, 5, c(1:5)),*

*+ legendPosition = "bottom")*

*> multiplot(frocCurve$FROCPlot, fittedFrocCurve$FROCPlot, cols = 2)*

## Sample Size Estimation for FROC Data

For FROC data, sample size estimation can be performed using similar methods. Details are described in Ref. In short, estimate the search-model parameters of pilot study, then use them to calculate the area under ROC curve by numerical integration of the predicted ROC curve (equation 4). Add the desired ROC effect size, and new values of parameters can be determined. Calculate effect size in JAFROC unit with these new parameters. Finally,

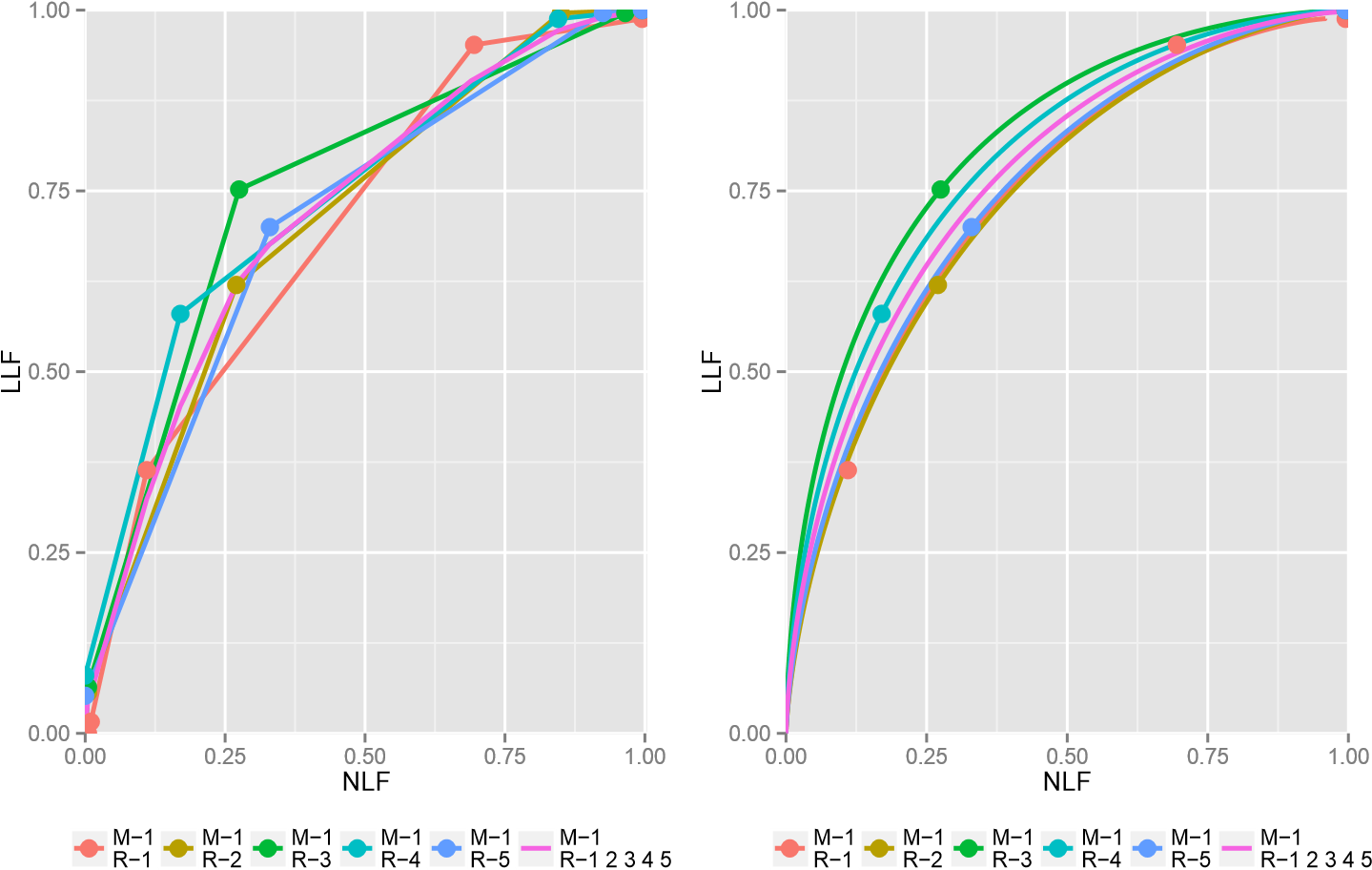


Figure 5: Plots of empirical (left) and fitted (right) FROC curves of each reader and their average performance using modality 0

required sample size to achieve the JAFROC unit effect size and statistical power can be estimated using same steps for ROC sample size estimation.

## Operating Characteristics

Operating characteristics function provides the ability to input specific values for the search model parameters to visualize the effects on the predicted ROC, AFROC and FROC curves. It is a convenient tool to understand how changes in reader characteristics (parameters are defined in subsection 4.3) affect reader performance. The function accommodates one or more modalities and varying number of lesions per case and calculates ROC and AFROC areas for all modalities. Figure 6 shows example plots for *µ* = 2, *λ* = 1, *ν* = 0.6, 15 single-lesion-cases and 35 two-lesion-cases in the NH modality and *µ* = 3, *λ* = 0.5, *ν* = 0.9, 20 single-lesion-cases and 30 two-lesion-cases in the AH modality.

*> lesionNumList <- list(rbind(c(1, 15), c(2, 35)),*

*+ rbind(c(1, 20), c(2, 30)))*

*> oprtChrctResults <- operatingCharacteristics(mu = c(2, 3), lambda = c(1, 0.5),*

*+ nu = c(0.6, 0.9),*

*+ lesionNumTable = lesionNumList,*

*+ legendPosition = "bottom")*

*> oprtChrctResults$aucROC*

[1] 0.8473118 0.9726777

*> oprtChrctResults$aucJAFROC*

[1] 0.6323901 0.9229876

*> multiplot(oprtChrctResults$ROCPlot, oprtChrctResults$FROCPlot,*

*+ oprtChrctResults$AFROCPlot, oprtChrctResults$PDFPlot, cols = 2)*

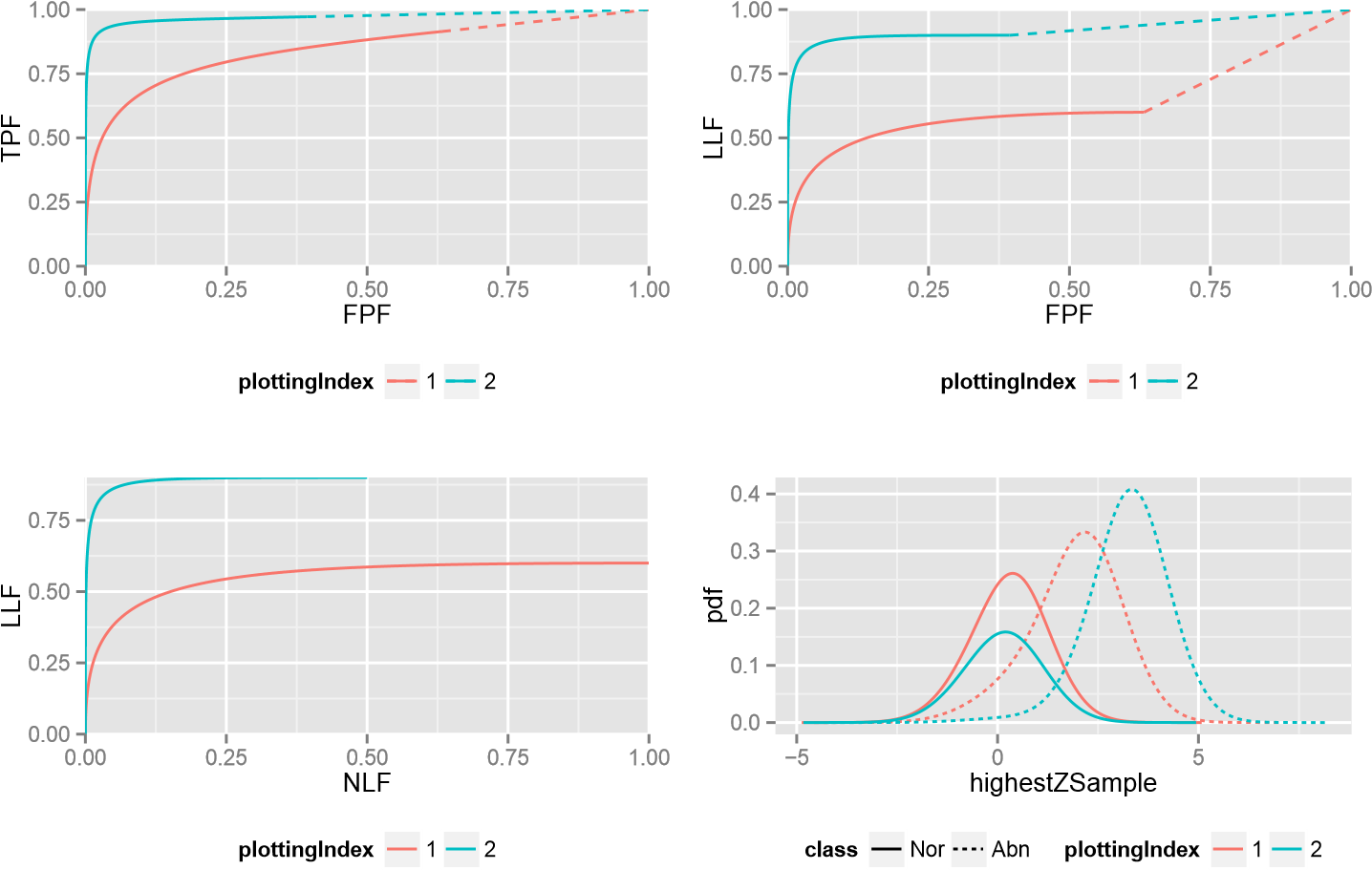


Figure 6: Operating Characteristics curves for *µ* = 2, *λ* = 1, *ν* = 0.6 (red) and *µ* = 3, *λ* =

0.5, *ν* = 0.9 (green)

# 5. Discussion

Compared with iMRMC and OR-DBM MRMC, advantages and limitations of JAFROCwR are summed up in following parts.

## **5.1. Advantages of** JAFROCwR

In summary, JAFROCwR performs better than the others in following aspects.

1. *Programmabilty*. Both of OR-DBM MRMC and iMRMC are GUI-based softwares, so it is almost impossible to call them and catch the results from users’ own program. JAFROCwR is distributed as a package of R. All functions in the package can be used in users’ own R code. Considering the good interactivity between R and other programming languages, JAFROCwR can also be easily called from programming environments other than R including but not limited to C/C++, Java, Python and MATLAB. In additional, JAFROCwR is an open source package. Users can find the source code of the package online, and it is convenient to debug and modify the source code if the original version cannot meet users’ specific requirements.
2. *Ability to Analyze FROC Data*. JAFROC and JAFROCwR are the only available software for free-response studies, and JAFROCwR implements OR method for FROC data for the first time. Operating Characteristics is a useful tool to help users understand the effect of FROC parameters.
3. *Plotting Features* JAFROCwR provides empirical and search model fitted parametric ROC/AFROC/FROC curves. The R package ggplot2 is used to plot all these curves. It gives fancy layouts, colors and legends, and ggplot2 objects are also easy to be modified as the users requirements.

## **5.2. Limitations and Potential Future Updates**

JAFROCwR has limitations need to be improved in the future. There are also some more features should be useful to be added in the package. General ideas are given as follows.

1. *Running Speed*. Since all the code of JAFROCwR are written using R and a lot of resampling operations (jackknife and bootstrap) are implemented by loops, the running time, especially for larger data file, is longer than that of the softwares written using “lower-level”languages such as C/C++. We are considering that convert some functions with many loop operations to C/C++ in the future. Rcpp will be used to combine them into the package. So the efficiency can be improved.
2. *Binormal Model*. Only search model is used to fit the curves in JAFROCwR. Binormal model estimation should also be included.
3. *Simulator*. Simulation plays an important role in ROC/FROC researches. It is always used to validate a new statistical method. Hence the simulator design is very significant. We are planning to provide a simulator in JAFROCwR to help users generate simulation data.
4. *Graphical User Interface*. GUI is necessary for some users, especially for the old users of JAFROC. An R based GUI package for JAFROCwR is in development. We will publish it when it is finished. JAFROC can be completely replaced at that time.

# 6. Conclusions

The JAFROCwR package is able to perform MRMC analysis of ROC/FROC/ROI data. The package provides OR and DBM methods for performing MRMC analysis, each of which can be used with different figures of merit. Empirical and parametric ROC/AFROC/FROC curves are also given by the package. Sample size calculation helps to estimate sample size for given effect size and statistical power in future study.

Search model is used to fit ROC/AFROC/FROC curves. Operating Characteristics shows the effect of search model parameters on the predicted ROC, AFROC and FROC curves. It is convenient to demonstrate how the parameters affect readers’ performance.

Through the provided examples, JAFROCwR will be definitely helpful if you are using R as a main programming language. Future version of this package will attempt to reduce the running time, and provide more features.

**Affiliation:**

Xuetong Zhai

Department of Radiology

University of Pittsburgh

Pittsburgh, PA 15213

United States of America

E-mail: xuz19@pitt.edu

|  |  |
| --- | --- |
| *Journal of Statistical Software* | <http://www.jstatsoft.org/> |
| published by the American Statistical Association | <http://www.amstat.org/> |
| Volume VV, Issue II | *Submitted:* yyyy-mm-dd |
| MMMMMM YYYY | *Accepted:* yyyy-mm-dd |