RJafroc: an R implementation of JAFROC analysis for MRMC data

|  |  |
| --- | --- |
| Xuetong Zhai  University of Pittsburgh | Dev Chakraborty  University of Pittsburgh |
| Abstract  TBA  *Keywords*: JAFROC, ROC, FROC, software, R. |  |

# 1. Introduction

A common task in medical imaging is assessing whether a new imaging system is an improvement over an existing one. Observer performance measurements, widely used for this purpose, require data collection and analyses methods that fall under the rubric of what is loosely termed "ROC analysis", where ROC is an abbreviation for Receiver Operating Characteristic. The purpose of this work is to describe a package called RJafroc, which implements software for the analysis of data acquired using the ROC paradigm and its extensions. It is an enhanced implementation of a software suite called "JAFROC" (http://www.devchakraborty.com) that has been widely used in assessing medical imaging systems. In this section we introduce the terminology used in the "RJafroc-package" page of the documentation that accompanies this paper.

In an ROC study the radiologist is shown images from patients, the radiologist is “blinded,” of course, to the true disease states, and the task is to rate each patient for confidence in presence or absence of disease. The rating *r* is typically on a numeric scale, with higher values representing increasing confidence in presence of disease and lower values representing increasing confidence in absence of disease. Typically 5 or 6 integer ratings are used but the ratings could have higher precision. With a 6 rating scale a 1-rating would correspond to high confidence that patient is non-diseased and a 6-rating would correspond to high confidence that patient is diseased. The counts in the different ratings bins, cumulated separately for actually non-diseased and actually diseased patients, can be used to construct an operating point on the ROC plot contained within the unit square. For example, the cumulated counts in diseased ratings bins 3, 4 and 5, divided by the number of actually diseased images, yields the true positive fraction , and the corresponding cumulated counts for non-diseased images, divided by the total number of non-diseased images, yields the false positive fraction . It can be seen that as long as no bin has zero counts for both non-diseased and diseased images, an *R* rating ROC study will yield *R* - 1 non-trivial operating points . The origin (0,0) and the upper right corner (1,1) are trivial operating points that belong to any dataset. The empirical ROC is defined by connecting neighboring operating points (including the trivial ones) with straight lines. The trapezoidal area under the empirical ROC is frequently used as a non-parametric figure of merit (FOM) for quantifying observer performance. It can be shown to be equivalent to the Mann-Whitney-Wilcoxon 2-sample U-statistic. ROC studies are typically conducted with about 50/50 or more non-diseased/diseased patients. The patients are imaged in two or more imaging systems (also known as modalities or treatments) and the images are rated by a number of radiologists (typically about 5 to 10). This type of fully crossed study design is termed multiple reader multiple case (MRMC).

A limitation of the ROC paradigm is that it requires a single rating per image, which applies to the image as a whole, not to any specific region(s) in the image. Typically, disease manifests itself by the presence of localized diseased regions (malignant lesions). For example, cancer of the chest is often diagnosed by finding localized malignant lesions. There are two data collection paradigms that allow for localization information to be collected to different extents (a third paradigm, termed location ROC (LROC) is not included in this description, as it is not currently implemented in easily accessible software).

In the free-response paradigm the radiologist marks and rates regions that are suspicious for disease. The mark is classified as lesion localization (LL) if it successfully locates an actual lesion to within clinically relevant accuracy, or non-lesion localization (NL) otherwise. Unmarked lesions are assigned the –infinity rating. By treating the rating of the highest rated mark on a non-diseased image (or –infinity if the image has no marks) as its inferred FP rating, it is possible to define an inferred FPF that is analogous to true FPF obtained in an actual ROC study. By cumulating LL events and dividing by the total number of lesions it is possible to define a lesion localization fraction (LLF) that is analogous to TPF, but takes location into account. A plot of LLF along the ordinate vs. FPF is defined as the alternative FROC, or AFROC. The non-lesion localization fraction (NLF) is defined as the cumulated number of NLs divided by the total number of cases. The FROC plot is defined as that of LLF along the ordinate vs. NLF. By treating the rating of the highest rated mark on a diseased image (or –infinity if the image has no marks) as its inferred TP rating, it is possible to define an inferred TPF. The plot of inferred TPF vs. inferred FPF is the inferred ROC curve. Regarding the highest rated NL mark on any image as an inferred FP1 rating (the 1 denotes that NL marks on diseased cases could be contributing to this FP-like rating) and the corresponding AFROC1 plot is that of LLF vs. FPF1. By assigning clinically useful weights (relative importance of finding the lesions) to different lesions on the same diseased image, where the weights on any image add up to one, it is possible to define weighted LLF, weighted AFROC and weighted AFROC1 plots. With the exception of the FROC, the trapezoidal areas under all of these curves qualify as valid figures of merit (a valid figure of merit is one that rewards good decisions and penalizes bad decisions, where good and bad are defined with respect to patient outcome).

In the region of interest (ROI) paradigm each image is divided into Q regions of interest (typically Q is 4 or 5) and the reader gives a ROC-like rating to each region. Regarding each of the regions as a mini-image, it is possible to define ROC-like quantities TPF' and FPF', where the primes distinguish them from true FPF and TPF. For example, FPF' and TPF' can be defined for a dataset containing only diseased images, for which it would be impossible to define FPF. The data collection paradigms are summarized in Table 1.

Table 1: Data collection paradigms, operating characteristics, figure of merit and common terminology

|  |  |  |  |
| --- | --- | --- | --- |
| Data collection paradigm | Operating characteristic | FOM | Terminology |
| Receiver operating characteristic | ROC = TPF vs. FPF | Trapezoidal area under ROC | AUC |
| Free-response | AFROC = LLF vs. FPF | Trapezoidal area under AFROC | JAFROC, weighted JAFROC |
| AFROC' = LLF vs. FPF' | Trapezoidal area under AFROC' | JAFROC1, weighted JAFROC1 |
| FROC = LLF vs. NLF | Not recommended | |
| Inferred ROC | Trapezoidal area under inferred ROC | AUC |
| Region of interest | ROC'=TPF' vs. FPF' | Trapezoidal area under ROC' | AUC' |

Analysis of the data starts with estimation of the figure of merit for each treatment - reader combination. One object of the analysis is to determine the significance of the reader-averaged differences in FOMs between pairs of modalities. While several significance-testing methods have been proposed, see Table 2, we focus on two that are easily accessible and consequently in widespread use: the Dorfman-Berbaum-Metz (DBM) method and the Obuchowski-Rockette (OR) method, both of which have been significantly improved by contributions by Hillis, and are henceforth referred to as DBMH and ORH, respectively. A third method due to Gallas et al, often termed a mechanistic or first-principles approach to MRMC analysis, is also available online, that yields independent estimates of variability parameters used in DBMH and ORH analyses.

Table 2: Software availability of MRMC observer performance methods

|  |  |  |  |
| --- | --- | --- | --- |
| Significance testing methods | Online software name | Supported data collection paradigms | Supported FOMs |
| DBMH, ORH | OR-DBM MRMC | ROC | Wilcoxon and parametric fits |
| DBMH, Metz, Pesce et al | DBM MRMC | Wilcoxon and parametric fits |
| Gallas | iMRMC | Wilcoxon |
| Chakraborty | JAFROC | ROC, FROC, ROI | Trapezoidal areas under ROC, AFROC, AFROC1, weighted versions and ROC' |
| Toledano and Gatsonis | NA | ROC | NA |
| Song |
| Ishwaran and Gatsonis |
| Beiden-Wagner-Campbell |

Table 3 compares existing online software, and the new R package RJafroc, from several points of view: ease of data entry, whether or not they are open-source, and if so, the programming language expertise needed to modify them, whether or not they are cross platform applications, whether individual modules can be called from other languages, whether or not they include parametric fitting and integrated plotting routines, and the degree to which they accommodate location paradigms.

Table 3: Software comparison

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Software | OR-DBM MRMC &  DBM-MRMC | iMRMC | JAFROC | RJafroc |
| Data entry | Plain text | Plain text | Excel file | Both |
| Open Source/Language | No/Fortran/C++ | Yes/Java | No/C++ | Yes/R |
| Cross Platforms | No | Yes | No | Yes |
| Call from Other Languages | No | No | No | Yes |
| ROC curve fitting | Yes | No | No | No |
| FROC curve fitting | No | No | Yes | Yes |
| Integrated plotting capability | No (exports csv files) | Yes (trapezoidal ROC) | Yes | Yes (empirical and fitted) |
| Localization paradigms (ROI and FROC) | No | No | Yes | Yes |
| Search model predicted operating characteristics | No | No | Yes | Yes |

Several parametric models are available for fitting ROC rating data. Obviously these cannot fit operating characteristics generated by location specific paradigms. A search model for describing data obtained in the free-response paradigm has been developed that accommodates the random number of marks per image that is characteristic of such data. It is characterized by three parameters: , a Poisson parameter describing the sampling of numbers of noise sites per image, i.e., suspicious regions on an image that do not correspond to true lesions; , the success probability of a binomial distribution, describing the sampling of numbers of signal sites per diseased image, i.e., suspicious regions on a diseased image that correspond to true lesions; and , a separation parameter between two unit variance normal distributions, describing the confidence level distribution for noise and signal sites.

# 2. Statistical Models and Methods

## **2.1. Figure of Merit**

Images are indexed by , where is the truth state index (1 for disease-free cases and 2 for diseased cases) and  indexes the cases for truth state , specifically,  and where is the number of disease-free cases and is the number of diseased cases. Let  denote the rating given to case by the reader *j* using modality *i* with *i* = 1*,...,I* and *j* = 1*,...,J*, where *I* is the number of modalities and *J* is the number of readers. While other performance measures are possible, in this paper we focus on the trapezoidal area under the ROC curve, , estimated for reader *j* in modality *i* by:



The kernel function  is defined by



## **2.2. DBMH Method**

The DBM method models the pseudovalues of , denoted ** for modality *i*, reader *j* and case *k* (*k* = 1, 2, ..., *K*; where  is the total number of cases), and defined by:



Here **is the estimate of  for modality *i*, reader *j* with case *k* removed from the analysis. Hillis has defined a centering transformation



The effect of this transformation is that the average of the centered pseudovalues over the case index is identical to the estimate of the figure of merit:



This has the practical advantage that all confidence intervals are correctly centered. While this transformation is unnecessary if one uses the Wilcoxon as the figure-of-merit, for generality with other possible figures of merit, *it is understood that all calculations from now on will use the centered pseudovalues*.

The DBM pseudovalue model is:



The right hand side consists of 2 fixed effects, , and 6 random effects modeled as samples from zero-mean normal distributions with variances (in the same order of appearance in the above equation) , , ,, and.

Using the dot symbol to denote an average over the corresponding index, the first term  can be estimated by averaging the observed left hand side over all three indices:





To estimate the modality effect one takes the difference as shown below:



The reader and case averaged difference between two modalities *i* and *i'* (often termed the *observed effect size*) is given by



Estimating the strengths of the random terms involves analysis of variance (ANOVA) methods specially adapted to this problem by Dorfman, Berbaum, Metz, Hillis and others. Only the final results are summarized here. The starting point is calculation of the mean squares. In the following definitions the Y subscript emphasizes that the relevant mean-square quantities are calculated using pseudovalues, not figure-of-merit values.



Hillis proposes the following statistic for testing the null hypothesis of no modality effect:



Here  is the unit step function, defined as unity for positive argument and zero otherwise. Hillis (ref) has shown that  is distributed as an F-statistic with numerator degrees of freedom *ndf* = *I* - 1 and ** denominator degrees of freedom, i.e.,:



The denominator degrees of freedom is defined by:



The analysis described so far treats both readers and cases as random factors, so it is termed random-reader random-case (RRRC) analysis. Special cases of the analysis, which regards either readers or cases as fixed factors, is possible, and the results are given in the appendix.

## **2.3. ORH Method**

The statistical model underlying the OR method is:



The left hand side is the estimated figure-of-merit for modality *i* and case set index {*c*}, where *c* = 1, 2, …, *C* denote different case sets (i.e., different *collections* of cases, not individual cases) sampled from the underlying population). In practice the dataset is limited to *c* = 1, but resampling and other methods, see below, are available to tease out the case-sample dependence. The first two terms on the right hand side of Eqn. xx have their usual meanings. The remaining terms are mutually independent random samples:denotes a random contribution to the figure-of-merit of reader *j*, modeled as a sample from a zero-mean normal distribution with variance ;  denotes a treatment-dependent random contribution of reader *j* (*j* = 1,2, ..., *J*) in modality *i*, modeled as a sample from a zero-mean normal distribution with variance . The error term is defined by a covariance matrix  with 4 parameters, , defined as follows:



OR have suggested that the 4 elements of the covariance matrix should be ordered as follows:



Resampling methods are used to estimate the parameters of the covariance matrix. Using the bootstrap method, where {*b*} is the *b*th bootstrap replicate, *b* = 1, 2, ..., *B*,



The averages, indicated by the bracket symbols, over modalities and readers are necessary since the co-variances in the OR model are assumed to be independent of modality and reader. The jackknife estimate is given by



DeLong et al have described a covariance estimation method that is applicable as long as one restricts to the ROC paradigm and Wilcoxon FOM (the bootstrap and the jackknife are more generally applicable to any figure of merit).

Because of the correlated structure of the error term a customized ANOVA is needed. The null hypothesis is that the true figure-of-merit of all modalities are identical, i.e.,



A modified F-statistic is needed, denoted and defined by:



Eqn. xx incorporates Hillis’ modification, which ensures that the constraint is always obeyed. The mean square (MS) terms are defined by (note the lack of the Y subscript, as these are calculated directly using FOM values):



According to Hillis, the observed statistic  is distributed as an F-statistic with *ndf* = I-1 and degrees of freedom:



where



For the Wilcoxon statistic, the two definitions of are equivalent.

## **2.4. Sample Size Calculation**

If a non-significant result is obtained (i.e., *p > α*) in the original study (termed the *pilot* study) then the investigator may wish to plan a new study (termed the *pivotal* study) that is sufficiently powered to detect a clinically relevant difference between two modalities of interest. The pilot study is used to get estimates of the variability components entering the figure of merit model. We will illustrate the procedure for the ORH method, and equivalent procedures for DBMH have been published (ref). The *observed effect size* (absolute value of the difference in figures of merit between the two modalities) is . Under the alternative hypothesis  the test statistic is distributed as a *non-central* F-distribution with *ndf* = 1 and to-be-determined *ddf* and non-centrality parameter . The sample size estimate follows the following procedure (ref). To be specific this procedure assumes random-readers and random cases. Different formulae apply when either readers or cases is treated as a fixed effect.

1. Specify the effect size *d*: typically, when dealing with area under the ROC curve as the figure of merit, one might choose *d* = 0.05.
2. Estimate the OR modality-reader interaction variance component:this is given by (ref):  
     
   

If this yields a negative variance, Hillis suggests setting it to zero.

1. Estimate the non-centrality parameter and the *ddf* of the F-distribution. Let  denote the number of cases in the pilot dataset, and let *J*, *K* be the numbers of readers, cases in the pivotal study. The non-centrality parameter ∆ and the *ddf* are estimated by:  
     
      
     
   
2. The statistical power  at significance level  can be calculated using:  
     
   

denotes the non-central F-distribution with degrees of freedom 1, ddf, and non-centrality parameter  and is the critical value of F such that fraction  of the central F distribution with degrees of freedom 1, *ddf* is below the critical value.

1. If the power is below the desired or target power, typically chosen to be 0.8, one tries successively larger value of K until the target power is reached. The procedure could be repeated with different values of J (depending on cost and other practicality issues, it might be better to have more reader each reading fewer cases to achieve the same target power).

# 3. Examples and Results Comparisons

The package comes pre-loaded with two datasets: an ROC dataset object named vanDykeData, which data has been repeatedly used by Hillis and colleagues to illustrate advances in ROC methodology, and an FROC dataset object named frocData. Their structures are shown below:

> str(vanDykeData)

List of 9

$ NL : num [1:2, 1:5, 1:114, 1] 1 3 2 3 2 2 1 2 3 2 ...

$ LL : num [1:2, 1:5, 1:45, 1] 5 5 5 5 5 5 5 5 5 5 ...

$ lesionNum : num [1:45] 1 1 1 1 1 1 1 1 1 1 ...

$ lesionID : num [1:45, 1] 1 1 1 1 1 1 1 1 1 1 ...

$ lesionWeight: num [1:45, 1] 1 1 1 1 1 1 1 1 1 1 ...

$ maxNL : num 1

$ dataType : chr "ROC"

$ modalityID : chr [1:2] "1" "2"

$ readerID : chr [1:5] "1" "2" "3" "4" ...

> str(frocData)

List of 9

$ NL : num [1:5, 1:4, 1:200, 1:7] -Inf -Inf 1 -Inf -Inf ...

$ LL : num [1:5, 1:4, 1:100, 1:3] 4 5 4 5 4 3 5 4 4 3 ...

$ lesionNum : int [1:100] 1 1 1 1 1 1 1 1 1 1 ...

$ lesionID : num [1:100, 1:3] 1 1 1 1 1 1 1 1 1 1 ...

$ lesionWeight: num [1:100, 1:3] 1 1 1 1 1 1 1 1 1 1 ...

$ maxNL : num 7

$ dataType : chr "FROC"

$ modalityID : chr [1:5] "1" "2" "3" "4" ...

$ readerID : chr [1:4] "1" "3" "4" "5"

Since ROC data is a special case of free-response data, the same object structure is used to accommodate both. For ROC data the FP ratings are contained in the first  elements of the NL array. The ROC dataset has two modalities, five readers, 69 (= 114-45) non-diseased and 45 diseased cases. The FROC data set has five modalities, 4 readers, 100 non-diseased and 100 diseased cases.

## **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.

> resultDBM <- DBMAnalysis(data = dataVanDyke, analysisFOM = "ROC", alpha = 0.05)

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

+ covEstMethod = "Jackknife")

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, which is plotted using following code.

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

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

+ legendPosition = "bottom")

> rocCurveM1$ROCPlot

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)

All the results of these example codes are summrized in subsection 3.2. For more information and example, such as the meaning of each parameters and return varables, please read the help documentation at the page of JAFROCwR [http://CRAN.R-project.org/package= JAFROCwR.](http://CRAN.R-project.org/package=JAFROCwR)

## **3.2. Results Comparison**

Analysis results of the same dataset used in subsection 3.1 by other ROC softwares are compared in this subsection. AUC is selected as the figure of merit. DBM method or OR

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

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. JAFROC is omitted since the package JAFROCwR produces identical results. Table 3 shows the AUC calculation result of the three softwares. From the values in the table, we can see that AUC calculations of the three different wares are exactly identical.

Table 3: AUC calculation comparison for the dataset “Van Dyke” (“M” and “R” in column and row names indicate “Modality” and “Reader” respectively. Hereinafter the same.)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 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 |

Variance and covariance components estimation are given in table 4. The results of iMRMC are slightly different from the other two softwares. The reason is DBM and OR components are obtained through the linear combination of the BDG components instead of the original proposed estimation methods in iMRMC (Brandon D. Gallas 2014).

Table 5 shows the result of the null hypothesis that there is no modality effect. Confidence interval of the difference between the AUC of modality 0 and modality 1. iMRMC did not Table 4: Variance Components Comparison (Letters between the brackets indicate the source of variance. Covariances are defined in 2.3)

|  |  |  |  |
| --- | --- | --- | --- |
|  | OR-DBM MRMC | iMRMC | JAFROCwR |
| DBM Components  Var(R) | 1.5350E-03 | 1.5365E-03 | 1.5350E-03 |
| Var(C) | 2.7249E-02 | 2.6860E-02 | 2.7249E-02 |
| Var(T\*R) | 2.0040E-04 | 1.2090E-02 | 2.0040E-04 |
| Var(T\*C) | 1.1975E-02 | 2.0776E-04 | 1.1975E-02 |
| Var(R\*C) | 1.2265E-02 | 1.1793E-02 | 1.2265E-02 |
| Var(T\*R\*C) + Var(Error) | 3.9972E-02 | 3.9133E-02 | 3.9972E-02 |
| OR Components  Var(R) | 1.5350E-03 | 1.5365E-03 | 1.5350E-03 |
| Var(T\*R) | 2.0040E-04 | 2.0776E-04 | 2.0040E-04 |
| COV1 | 3.4661E-04 | 3.4167E-04 | 3.4661E-04 |
| COV2 | 3.4407E-04 | 3.3906E-04 | 3.4407E-04 |
| COV3 | 2.3903E-04 | 2.3561E-04 | 2.3903E-04 |
| Var(Error) | 8.0229E-04 | 7.8839E-04 | 8.0229E-04 |

test the null hypothesis. so the F value for iMRMC is not available. It gives the confidence interval and the statistics that are used to estimated the confidence interval. The results of iMRMC are still different from the others’ due to the same reason.

Table 5: Main Null Hypothesis Testing and Confidence Interval

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | F value | ddf | p-vlaue | 95% CI of modality 0 - 1 |
| Random Readers Random Cases  OR-DBM MRMC | 4.46 | 15.26 | 0.0517 | (-0.08796 , 0.00036) |
| iMRMC | — | 15.03 | 0.0556 | (-0.0879 , 0.0003) |
| JAFROCwR | 4.46 | 15.26 | 0.0517 | (-0.08796 , 0.00036) |
| Fixed Readers Random Cases  OR-DBM MRMC | 5.48 | 113 | 0.021 | (-0.08088 , -0.00672) |
| JAFROCwR | 5.48 | 113 | 0.021 | (-0.08088 , -0.00672) |
| Random Readers Fixed Cases  OR-DBM MRMC | 8.7 | 4 | 0.042 | (-0.08502 , -0.00258) |
| JAFROCwR | 8.7 | 4 | 0.042 | (-0.08502 , -0.00258) |

Figure 2 is the empirical ROC curves plotted by iMRMC and OR-DBM MRMC. Figure 3 are the fitted ROC curves using CBM (Dorfman, Berbaum, and Brandser 2000; Dorfman and Berbaum 2000a,b), PROPROC (Metz and Pan 1999; Pan and Metz 1997; Pesce and Metz 2007) and RSCORE (Dorfman and Alf 1969; Dorfman and Berbaum 1995) methods. Among these three softwares, only OR-DBM MRMC gives fitted ROC curves. Compare figure 1 and

figure 2, all of them provides identical ROC curves. Figure 2a and 3 are obtained by plotting the output points of OR-DBM using Excel, which is not convenent. In contrast, iMRMC only gives a plot of ROC curves. Re-plotting the curves using other tools or any changes to the plot (i.e. color, line type) are not allowed. Considering these two conditions, JAFROCwR plotting function returns both of the plot of ROC curves and the corresponding points data.

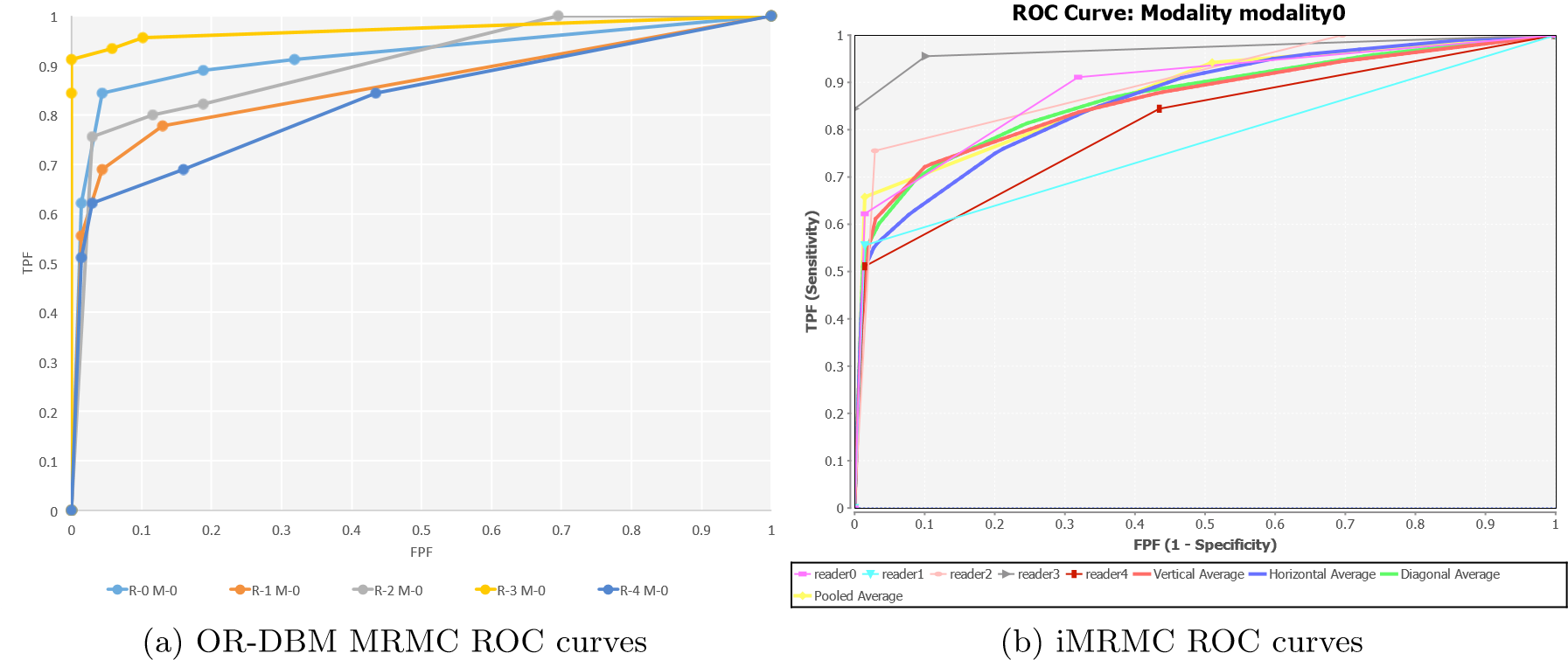


Figure 2: Empirical ROC curves output of OR-DBM MRMC and iMRMC

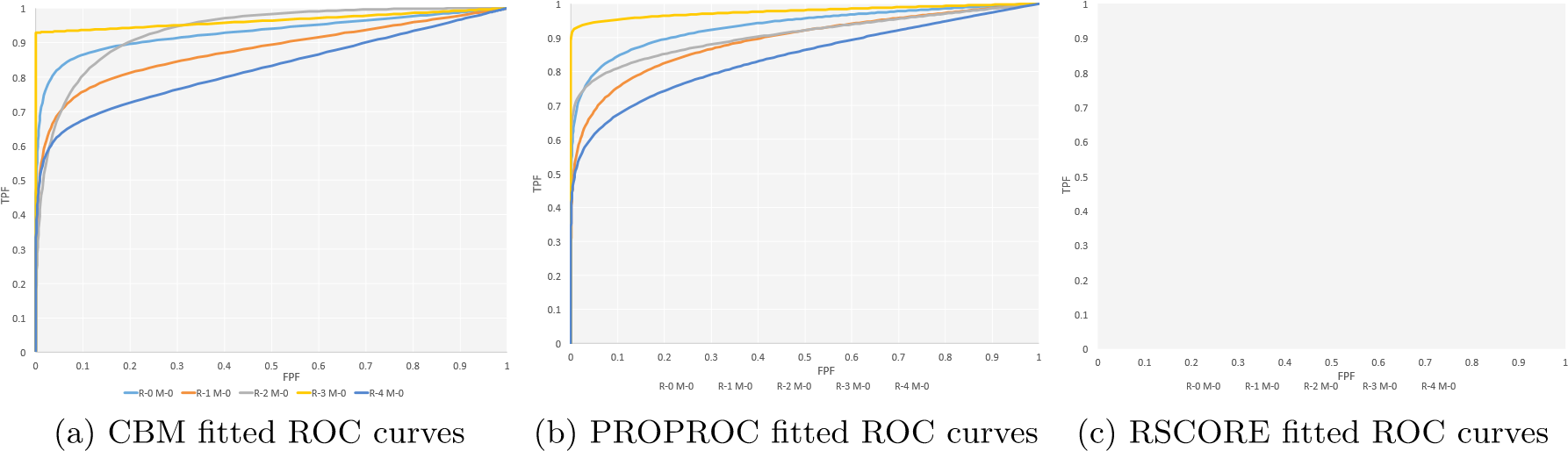


Figure 3: Fitted ROC Curves Plots

Table 6 contains the results of sample size calculation. Sample size (number of readers and cases) are estimated using JAFROCwR in the condition that significance level equals 0.05, effect size equals 0.05 and desired power equals 0.8. Then iMRMC is used to calculate the statistical power using corresponding sample size. The results are still slightly different due to the different estimation of variance components.

# 4. Additional Features of JAFROCwR

Table 6: Sample Size Calculation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| JAFROCwR | |  | iMRMC | |  |
| # of Readers | # of Cases | Power | # of Readers | # of Cases | Power |
| 5 | 213 | 0.80 | 5 | 213 | 0.79 |
| 6 | 170 | 0.80 | 6 | 170 | 0.80 |
| 7 | 148 | 0.80 | 7 | 148 | 0.80 |
| 8 | 134 | 0.80 | 8 | 134 | 0.80 |
| 9 | 125 | 0.80 | 9 | 125 | 0.82 |
| 10 | 119 | 0.80 | 10 | 119 | 0.82 |

However, the radiologists’ task is actually more than simply giving a single abnormality. In clinical situations, the lesion localizations (where it is) are required sometimes. It is important to indicate the radiologists’ capability both to detect and locate the lesions. In this case, some statistical approaches are proposed to analyze the performance on detecting and locating the lesions including localization ROC (LROC) (Starr, Metz, Lusted, and Goodenough 1975; Swensson 1996), region of interest (ROI) Obuchowski, Lieber, and Powell (2000); Rutter (2000) and free-response ROC (FROC) (Miller 1969; Bunch, Hamilton, Sanderson, and Simmons 1977; Chakraborty, Breatnach, Yester, Soto, Barnes, and Fraser 1986; Chakraborty 1989). These localization approaches are also implemented in JAFROCwR. Some figures of merit has been defined so DBM and OR method can be used for lesion localizations analysis. In LROC approach, the reader gives a single rating to each case and marks the suspicious region. The rating is regarded as a true positive if and only if the corresponding suspicious region is marked correctly. Otherwise, the rating is treated as a false positive. ROI and FROC approaches are introduced in following subsections.

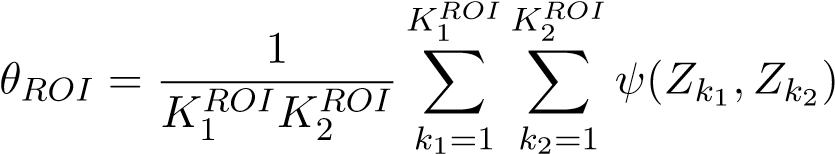
4.1. ROI and FROC Figures of Merit Used in JAFROCwR

## 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 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 (Obuchowski *et al.* 2000)

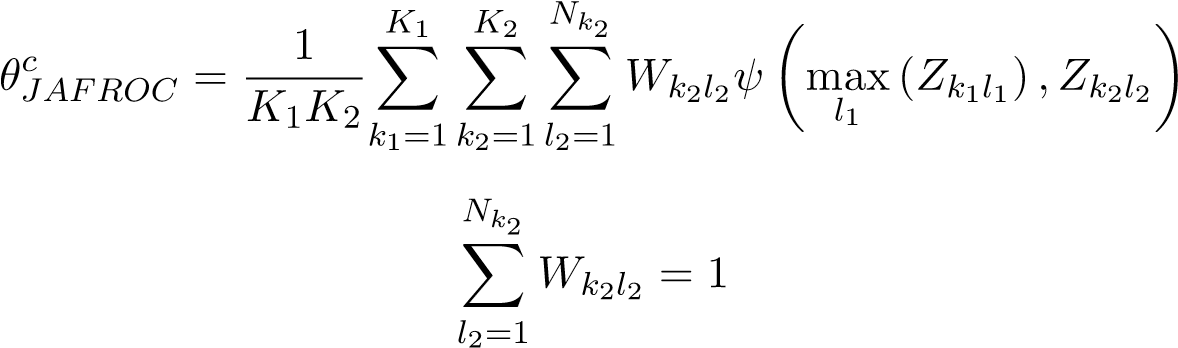


where *ψ* is the function defined by equation  is the number of non-diseased ROIs and  is the number of diseased ROIs, ) and) are the ratings for corresponding ROIs.

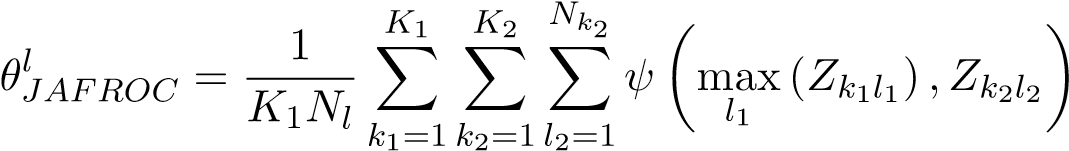
## FROC Figures of Merit

The free-response receiver operating characteristic 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 marks is also regarded as a random variable. This source of randomness is the main difficulty of FROC analysis. Some commonly used FROC figures of merit are defined as follows (Chakraborty and Berbaum 2004; Song, Bandos, Rockette, and Gur 2008).

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



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)



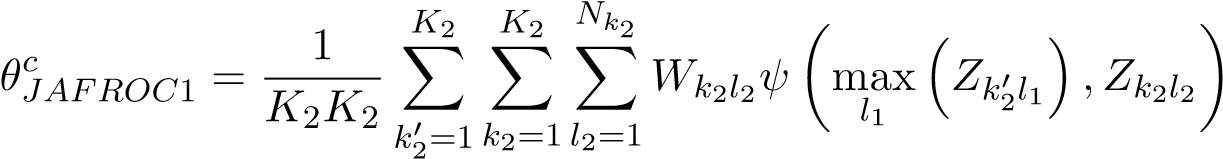
where *Nl* is the total number of lesions in all diseased cases by definition

*K*2

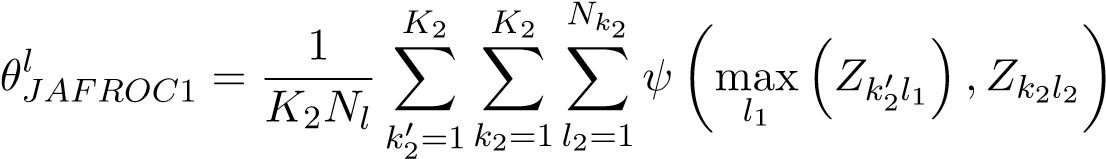
*Nl* = X *Nk*2

*k*2=1

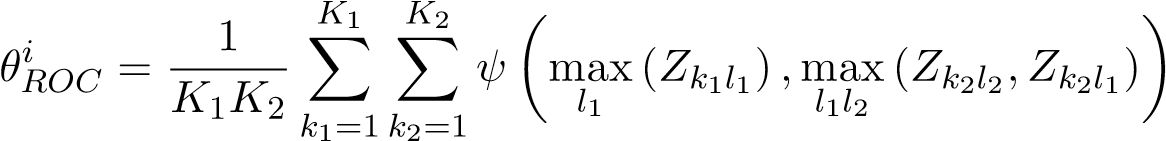
The weighted JAFROC1 figure of merit is defined by



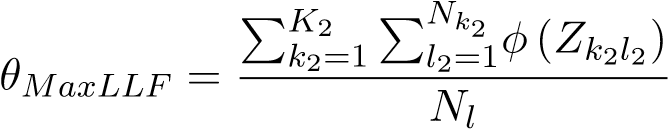
The un-weighted JAFROC1 figure of merit is defined by



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

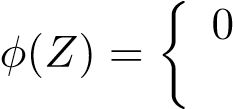


The maximum lesion localization fraction figure of merit is defined by

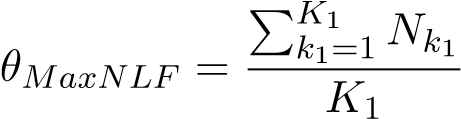


where *φ* is a function given by

*Z* is − ∞

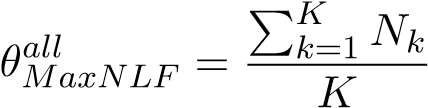
 1 otherwise

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

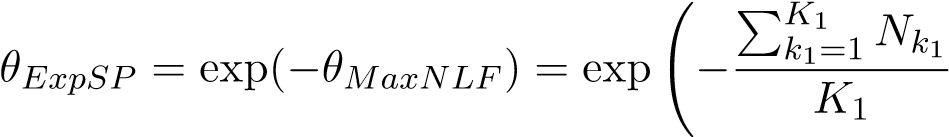


where *Nk*1 denotes the number of non-lesion localization marks in case *k*1.

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 Song *et al.* (2008).

### 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.

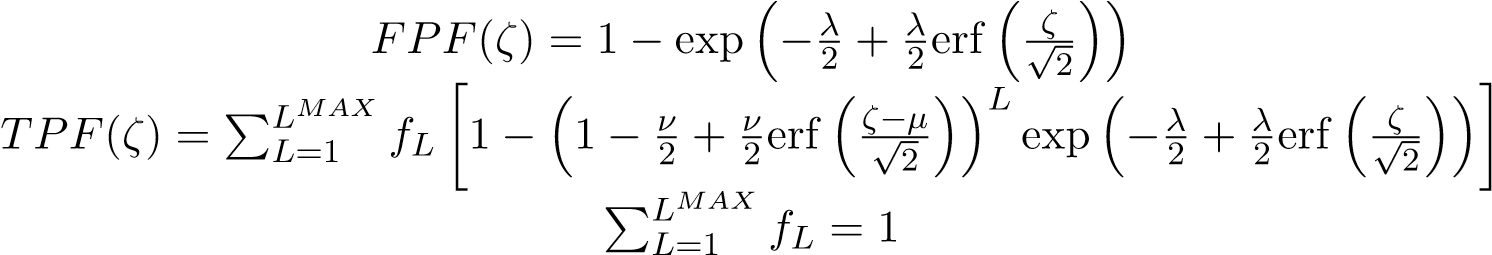
### 4.3. Search-Model

The search-model (Chakraborty 2006a,b) is used to estimate sample size and fit ROC and (A)FROC curves for FROC data in JAFROCwR. According to Kundel and Nodine (Kundel and Nodine 1983, 2004; Kundel, Nodine, Conant, and Weinstein 2007; Nodine and Kundel 1987), 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 (Chakraborty 2006a)

 (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 defined in Press (2007, p. 1235). *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 (Chakraborty and Yoon 2008)

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 (Chakraborty and Yoon 2008)

*NLF*(*ζ*) = *λ*(1 − Φ(*ζ*)) (6)

Likelihood functions are defined to fit the ROC/AFROC/FROC curves using search-model.

## 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

LROCb = [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

LROC = YLROCb

*b*=0

## 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

LAFROCb = [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

LFROCb = [NLF(ζb) − NLF(ζb+1)]Nb[LLF(ζb) − LLF(ζb+1)]Lb (9)

→−

To fit the curve, we need to estimate parameters *λ*, *ν*, *ζ* and *µ* that maximize the logarithm of the likelihood 7, 8 and 9. Following algorithm is used in JAFROCwR to estimate the search-model parameters of fitted ROC/AFROC/FROC curves.

→−

1. For given *λ*, *ν* and cutoffs *ζ* , determine *µ* by minimizing the Chi-square goodness of fit statistic
2. Calculate the logarithm of corresponding likelihood.

→−

1. Repeat preceding steps using varied *λ*, *ν* and *ζ* until reach the maximum.

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. Note that only ”ROI” figure of merit is available for ROI data. For FROC data, empirical and search model parametric ROC/AFROC/FROC curves are available. Figure 4, 5 and 6 are the empirical and parametric ROC/AFROC/FROC curves.

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

> rocCurve <- plotROC(data = frocData, plottingModalities = 1,

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

+ legendPosition = "bottom")

> fittedRocCurve <- plotFittedROCCurve(data = frocData, plottingModalities = 1,

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

+ legendPosition = "bottom")

> multiplot(rocCurve$ROCPlot, fittedRocCurve$ROCPlot, cols = 2)

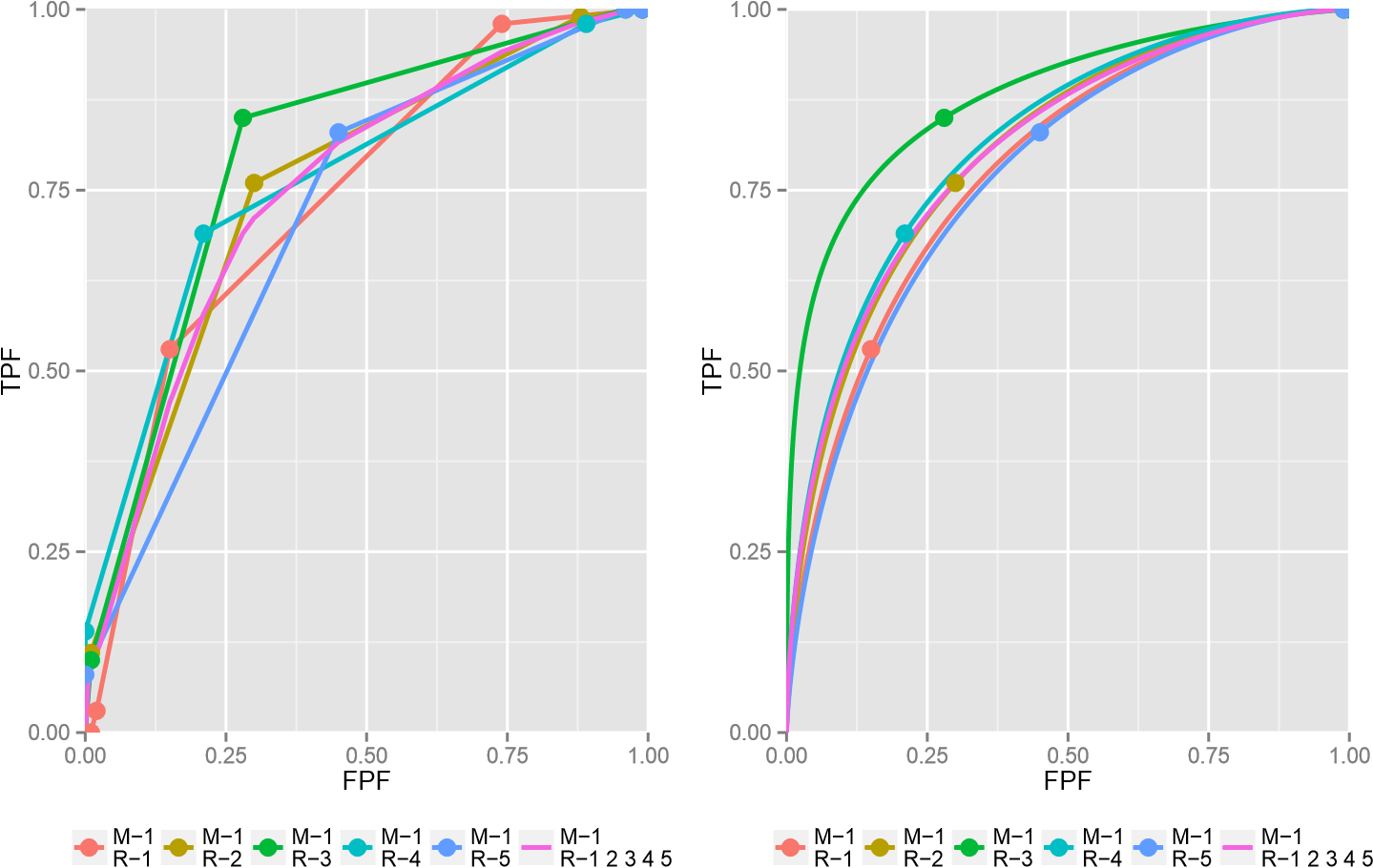


Figure 4: Plots of empirical (left) and fitted (right) ROC curves of each reader and their average performance using modality 0

> 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")

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

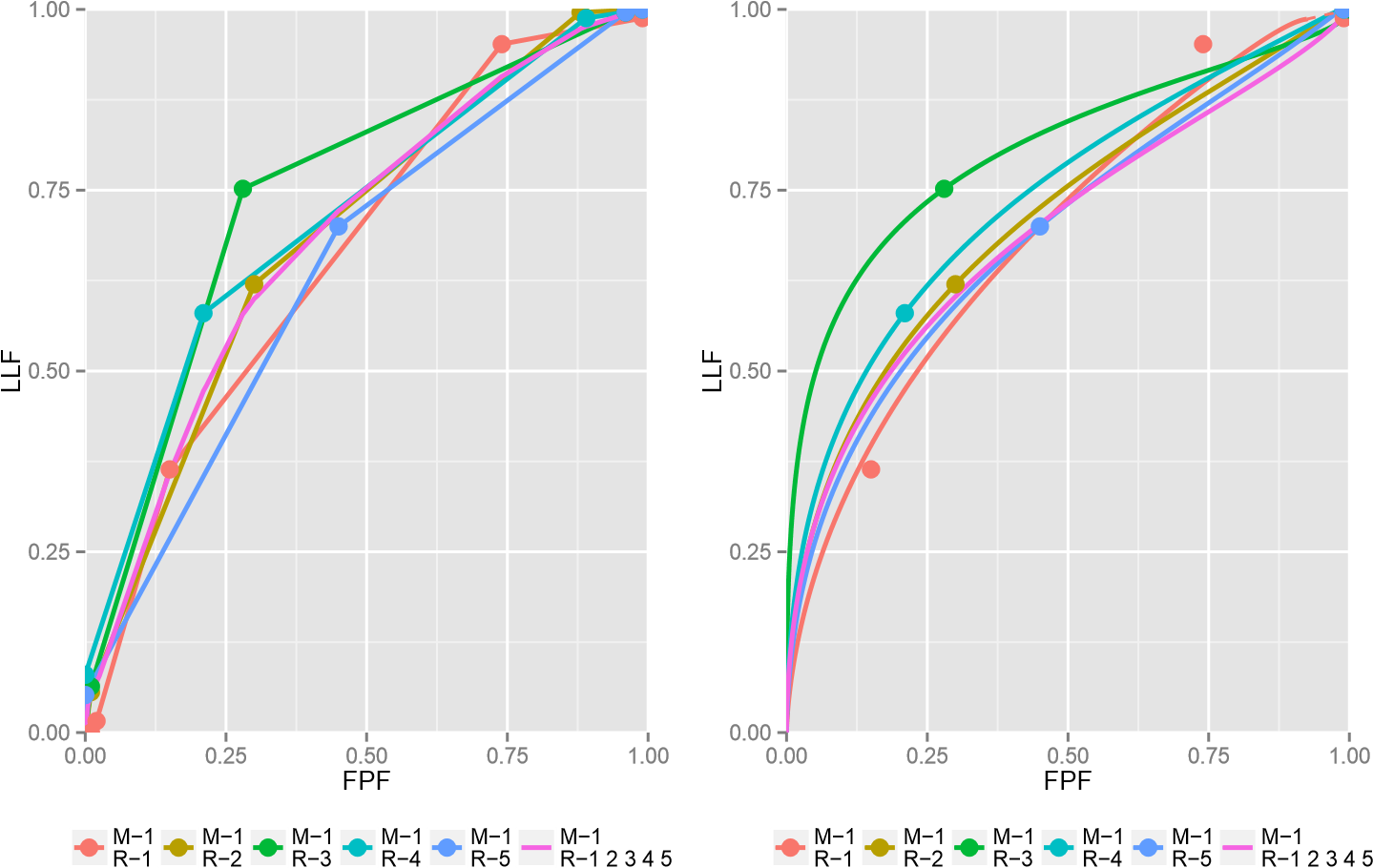


Figure 5: Plots of empirical (left) and fitted (right) AFROC curves of each reader and their average performance using modality 0

> 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 Chakraborty (2011). 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, 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.

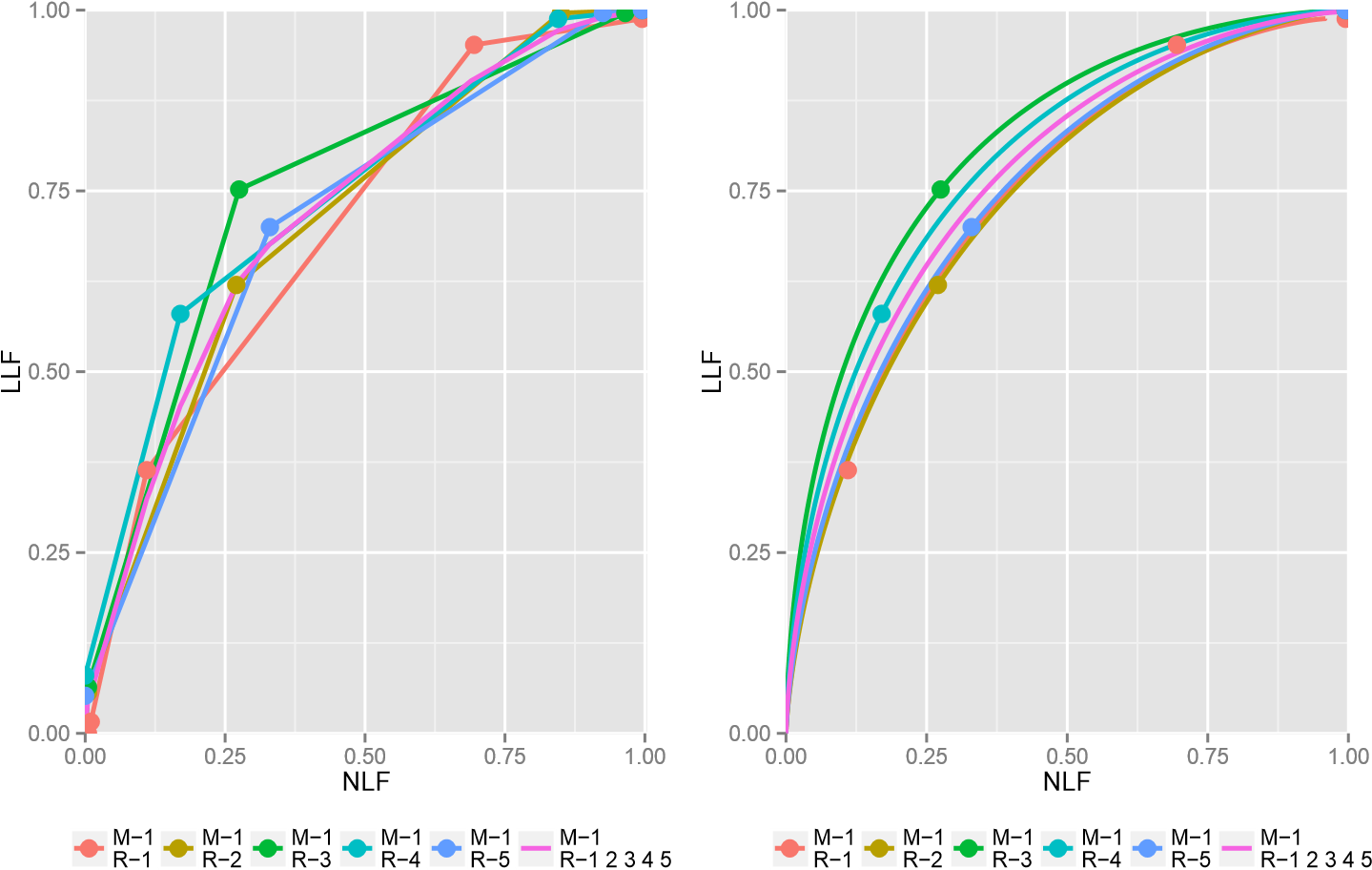


Figure 6: Plots of empirical (left) and fitted (right) FROC curves of each reader and their average performance using modality 0

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 7 shows example plots for *µ* = 2, *λ* = 1, *ν* = 0.6, 30 single-lesioncases and 35 two-lesion-cases in the NH modality and *µ* = 3 in AH modality. The four plots show how *µ* affects the observer’s performance. Other parameters’ effects can be studied by changing and compare the plots.

> lesionNumList <- list(rbind(c(1, 30), c(2, 35)),

+ rbind(c(1, 30), c(2, 35)))

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

+ nu = c(0.6, 0.6),

+ lesionNumTable = lesionNumList,

+ legendPosition = "bottom")

> oprtChrctResults$aucROC

[1] 0.8279472 0.8562302

> oprtChrctResults$aucJAFROC

[1] 0.6323901 0.6641375

> multiplot(oprtChrctResults$ROCPlot, oprtChrctResults$FROCPlot,

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

# 5. Discussion

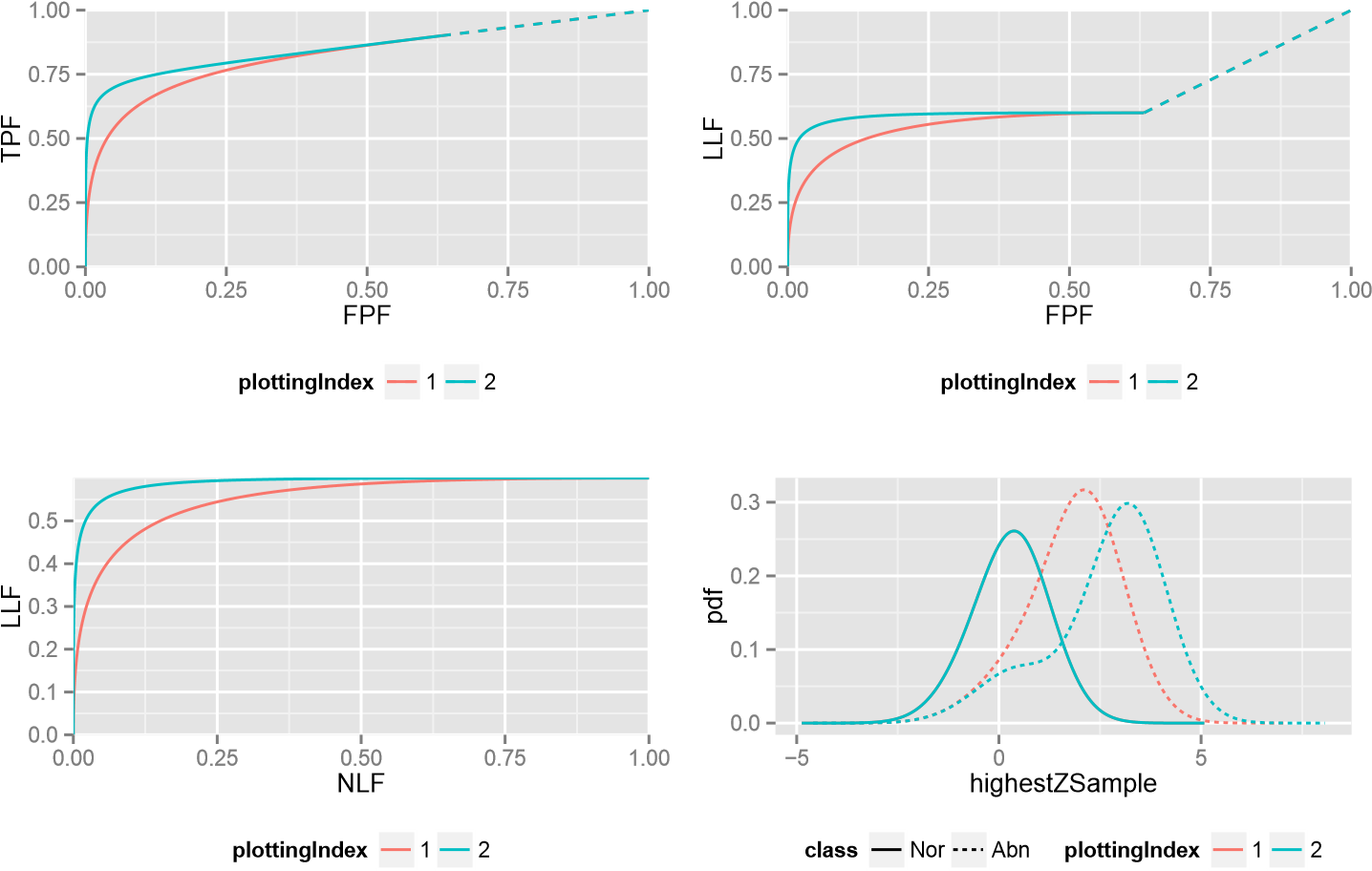


Figure 7: Operating Characteristics curves for *µ* = 2, *λ* = 1, *ν* = 0.6 (red) and *µ* = 3, *λ* = 1, *ν* = 0.6 (green)

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 Localization 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 (Wickham 2009) is used to plot all these curves. It gives fancy layouts, colors and legends, and ggplot2 objects are easy

to be modified. Both the plotting data points and ggplot2 objects are returned. Users can output the data points then plot the curves using other tools, or modify the ggplot2 object as their own 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. In the data import part, xlsx is used to read Excel file. The package provides the capibility to handle the operations to Excel fiel, but the running speed is slow. We are looking for a replacement of xlsx (Dragulescu 2014) to reduce the data reading time.
2. *Other Figure of Merit and Methods*. Only AUC is used as the ROC analysis figure of merit and only search model is used to fit the curves. Other figure of merit and model should also be implemented and included in the future.
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. Sample size calculation helps to estimate sample size for given effect size and statistical power in future study. Through the cross validation between the analysis of same dataset by JAFROCwR and other softwares, the analysis results of JAFROCwR is reliable.

Empirical and parametric ROC/AFROC/FROC curves are also given by the package. Search model is used to fit the 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 as well as provide more features.

# References

Brandon D Gallas Xin He RP (2014). *iMRMC User Manual Version 2.2*. URL [https:](https://code.google.com/p/imrmc/)

[//code.google.com/p/imrmc/.](https://code.google.com/p/imrmc/)

Bunch PC, Hamilton JF, Sanderson GK, Simmons AH (1977). “A free response approach to the measurement and characterization of radiographic observer performance.” In *Application of Optical Instrumentation in Medicine VI*, pp. 124–135. International Society for Optics and Photonics.

Chakraborty D (2006a). “ROC curves predicted by a model of visual search.” *Physics in medicine and biology*, **51**(14), 3463.

Chakraborty D (2006b). “A search model and figure of merit for observer data acquired according to the free-response paradigm.” *Physics in medicine and biology*, **51**(14), 3449.

Chakraborty D, Yoon HJ (2008).“Operating characteristics predicted by models for diagnostic tasks involving lesion localization.” *Medical physics*, **35**(2), 435–445.

Chakraborty DP (1989). “Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data.” *Medical physics*, **16**(4), 561–568.

Chakraborty DP (2011).“New developments in observer performance methodology in medical imaging.” In *Seminars in nuclear medicine*, volume 41, pp. 401–418. Elsevier.

Chakraborty DP, Berbaum KS (2004). “Observer studies involving detection and localization: modeling, analysis, and validation.” *Medical physics*, **31**(8), 2313–2330.

Chakraborty DP, Breatnach ES, Yester MV, Soto B, Barnes G, Fraser R (1986). “Digital and conventional chest imaging: a modified ROC study of observer performance using simulated nodules.” *Radiology*, **158**(1), 35–39.

DeLong ER, DeLong DM, Clarke-Pearson DL (1988).“Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach.” *Biometrics*, pp. 837–845.

Dorfman DD, Alf E (1969).“Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals—rating-method data.” *Journal of mathematical psychology*, **6**(3), 487–496.

Dorfman DD, Berbaum KS (1995). “Degeneracy and discrete receiver operating characteristic rating data.” *Academic Radiology*, **2**(10), 907–915.

Dorfman DD, Berbaum KS (2000a). “A contaminated binormal model for ROC data: Part II. A formal model.” *Academic radiology*, **7**(6), 427–437.

Dorfman DD, Berbaum KS (2000b). “A contaminated binormal model for ROC data: Part III. Initial evaluation with detection ROC data.” *Academic radiology*, **7**(6), 438–447.

Dorfman DD, Berbaum KS, Brandser EA (2000). “A contaminated binormal model for ROC data: Part I. Some interesting examples of binormal degeneracy.” *Academic radiology*, **7**(6), 420–426.

DORFMAN DD, BERBAUM KS, METZ CE (1992).“Receiver Operating Characteristic Rating Analysis: Generalization to the Population of Readers and Patients with the Jackknife Method.” *Investigative radiology*, **27**(9), 723–731.

Dragulescu AA (2014). xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.5.7.

Efron B, Stein C (1981). “The jackknife estimate of variance.” *The Annals of Statistics*, pp. 586–596.

Hillis SL (2007). “A comparison of denominator degrees of freedom methods for multiple observer ROC analysis.” *Statistics in medicine*, **26**(3), 596–619.

Hillis SL, Obuchowski NA, Berbaum KS (2011). “Power estimation for multireader ROC methods: an updated and unified approach.” *Academic radiology*, **18**(2), 129–142.

Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS (2005). “A comparison of the Dorfman– Berbaum–Metz and Obuchowski–Rockette methods for receiver operating characteristic (ROC) data.” *Statistics in medicine*, **24**(10), 1579–1607.

Kundel H, Nodine C (1983). “A visual concept shapes image perception.” *Radiology*, **146**(2), 363–368.

Kundel HL, Nodine CF (2004). “Modeling visual search during mammogram viewing.” In *Medical Imaging 2004*, pp. 110–115. International Society for Optics and Photonics.

Kundel HL, Nodine CF, Conant EF, Weinstein SP (2007). “Holistic component of image perception in mammogram interpretation: gaze-tracking study 1.” *Radiology*, **242**(2), 396– 402.

Metz CE, Pan X (1999). ““Proper” binormal ROC curves: theory and maximum-likelihood estimation.” *Journal of mathematical psychology*, **43**(1), 1–33.

Miller H (1969). “The FROC curve: a representation of the observer’s performance for the method of free response.” *The Journal of the Acoustical Society of America*, **46**(6B), 1473– 1476.

Nodine CF, Kundel HL (1987). “Using eye movements to study visual search and to improve tumor detection.” *Radiographics*, **7**(6), 1241–1250.

Obuchowski NA, Lieber ML, Powell KA (2000). “Data analysis for detection and localization of multiple abnormalities with application to mammography.” *Academic radiology*, **7**(7), 516–525.

Obuchowski Jr NA, Rockette Jr HE (1995). “Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations.” *Communications in Statistics-Simulation and Computation*, **24**(2), 285–308.

Pan X, Metz CE (1997). “The “proper” binormal model: parametric receiver operating characteristic curve estimation with degenerate data.” *Academic radiology*, **4**(5), 380–389.

Pesce LL, Metz CE (2007). “Reliable and computationally efficient maximum-likelihood estimation of “proper” binormal ROC curves.” *Academic radiology*, **14**(7), 814–829.

Press WH (2007). Numerical recipes 3rd edition: The art of scientific computing. Cambridge university press.

Rutter CM (2000). “Bootstrap estimation of diagnostic accuracy with patient-clustered data.” *Academic radiology*, **7**(6), 413–419.

Song T, Bandos AI, Rockette HE, Gur D (2008). “On comparing methods for discriminating between actually negative and actually positive subjects with FROC type data.” *Medical physics*, **35**(4), 1547–1558.

Starr SJ, Metz CE, Lusted LB, Goodenough DJ (1975). “Visual Detection and Localization of Radiographic Images 1.” *Radiology*, **116**(3), 533–538.

Stephen L Hillis Kevin M Schartz KSB (2014). *OR-DBM MRMC 2.5 User Guide*. URL [http://perception.radiology.uiowa.edu/Software/](http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/MRMCAnalysis/tabid/116/Default.aspx)

[ReceiverOperatingCharacteristicROC/MRMCAnalysis/tabid/116/Default.aspx.](http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/MRMCAnalysis/tabid/116/Default.aspx)

Swensson RG (1996). “Unified measurement of observer performance in detecting and localizing target objects on images.” *Medical physics*, **23**(10), 1709–1725.

Van Dyke C, White R, Obuchowski N, Geisinger M, Lorig R, Meziane M (1993). “Cine MRI in the diagnosis of thoracic aortic dissection. 79th RSNA Meetings.” *Chicago, IL*, **28**.

Wickham H (2009). *ggplot2: elegant graphics for data analysis*. Springer Science &amp; Business Media.

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 |