iMRMC: Applet for Analyzing and Sizing MRMC Reader Studies

The primary objective of this java applet (iMRMC) is to assist investigators with analyzing and sizing multi-reader multi-case (MRMC) reader studies that compare the difference in the area under Receiver Operating Characteristic curves (AUCs) from two modalities. The core elements of this java applet include the ability to do MRMC variance analysis, the ability to size an MRMC trial, and a database containing components of variance from past MRMC studies.

Introduction to ROC Reader Studies

ROC reader studies are designed to evaluate and compare imaging devices and acquisition protocols, or generally evaluate image quality according to an objective task. The ROC task is a classification task, e.g., classifying a patient as non-diseased or diseased. Image quality is then defined as the ability of a reader (e.g., a radiologist) to perform such a task.

In a typical ROC reader study the reader is presented with one of two mutually exclusive alternatives (e.g. a tumor-present image or a tumor-absent image). The observer is then asked to rate his or her confidence level of which alternative is presented (e.g., the confidence level of tumor presence on an image). Any number of responses may be used to rate the confidence level. For example, in a traditional clinical reader study, a set of five confidence level responses is used with 1 representing “absolutely sure there is no tumor” and 5 representing “absolutely sure there is tumor present”. Alternatively, reader studies may ask the observer to use a “continuous” rating scale. Such scales are not really continuous but allow the reader to rate each case with a whole number ranging from 1 to 100. The rating values are collected for both non-diseased and diseased cases.

Given ratings for non-diseased and diseased cases, an ROC curve can be traced out by calculating the sensitivity/specificity (TPF/TNF) pair for each confidence level, or threshold, possible [1]. An ROC curve illustrates the tradeoff between sensitivity and specificity of the reader across all thresholds. This tradeoff is realized by a change in the reader’s threshold. In the case of breast cancer screening via mammography, when the threshold is made more aggressive the reader recalls more patients for additional imaging, increasing his or her sensitivity at the price of lower specificity. If the reader’s threshold is moved in the opposite direction, the reader will recall fewer patients; the reader is less aggressive, decreasing his or her sensitivity with the concomitant result of increased specificity. The area under this ROC curve (AUC) is a summary figure-of-merit for describing how well a reader is able to separate the population of diseased patients from non-diseased patients. One interpretation of AUC is that it is the reader’s average sensitivity over all possible specificities. As such, it is a global summary of task performance that avoids thresholds entirely.

To account for the variability in readers, an ROC study is often conducted in a multi-reader paradigm. The endpoint of such an ROC reader study is the reader-averaged AUC value. It is mathematically equivalent to the probability that a random reader will correctly rank a random pair of signal present/absent images in a 2-alternative forced choice (2AFC) setting. The uncertainty in the reader-averaged AUC suffers from two sources of variability: the readers and the cases. To account for both sources of variability, reader studies often involve several trained readers in addition to a dataset of diseased and non-diseased cases. One popular study design for estimating AUC is the fully crossed study design in which every reader reads every case. Statistical methods have been proposed in the literature to analyze fully-crossed MRMC data [1-9]. The origin of each method differs, and consequently, the estimation process of each method differs. Additionally, each method has at its foundation a different decomposition, or representation, of the total variance.

Methods

Variance Estimation

In this software, we utilize two methods to estimate the MRMC variance components of four widely used variance decompositions of the reader-averaged AUC. Then we use these variance decompositions to size a trial. The software can analyze data from a reader-study of two modalities that is fully crossed with readers and cases paired across modalities (every reader reads every case in two modalities).

The first variance estimation method uses U-statistics to provide unbiased estimates of the variance components [4]. This method lacks a positivity constraint and can lead to negative estimates of variance components and total variance. The second variance estimation method uses the non-parametric maximum likelihood estimate (MLE) of the distribution of readers and cases, which is the empirical distribution of readers and cases [2]. Efron and Tibshirani also refer to this estimation method as the "ideal" bootstrap. The MLE estimate of variance components and total variance cannot go negative. The tradeoff for positive variance estimates is a positive bias.

Components of Variance

Regardless of their original developments, the four variance decompositions that are included in this software can be derived from first principles for the reader average of *empirical* AUCs. As such they are related to one another with simple mappings [3]. The decompositions are

- **BDG components [3-5]**: The author Brandon D. Gallas (BDG) decomposed the total variance into eight moments from first principles in a fashion equivalent to U-statistics, which decomposes the total variance into seven conditional covariances. The "extra moment" is the mean squared, which is a part of each conditional covariance. There is a term for non-diseased cases, diseased cases, readers, and all combinations. The decomposition treats non-diseased cases separately from diseased cases such that the total variance can be easily generalized to new readers, new non-diseased cases, and new diseased cases.
- **BCK components [6]**: The authors Barrett, Clarkson, and Krupinski (BCK) decomposed the total variance into seven marginal variances from first principles. They are marginal in the sense that they average over the non-random effects rather than conditioning on them as above. The BCK components are thought of as pure variance terms. There is a term for non-diseased cases, diseased cases, readers, and all combinations. The BCK decomposition treats non-diseased cases separately from diseased cases such that the total variance can be easily generalized to new readers, new non-diseased cases, and new diseased cases.
- **DBM components [7]:** The authors Dorfman, Berbaum, and Metz (DBM) decomponsed the total variance into six components based on a mixed-model ANOVA. The components are: reader effect, case effect, reader-case effect, modality-reader effect, modality-case effect, and modality-reader-case effect. The DBM decomposition does not treat non-diseased cases separately from diseased cases. So, while the total variance can be easily generalized to new readers and to total cases (with a fixed disease prevalence), additional modeling is needed to separately generalize to new non-diseased cases and new diseased cases (i.e., population with arbitrary disease prevalence).
- **MS components [7]:** The MS decomposition is based on the same mixed-model ANOVA as the DBM components. MS stands for mean squares, which are estimated from the data first and then mapped to the DBM components. There are six MS components: reader effect, case effect, reader-case effect, modality-reader effect, modality-case effect, and modality-reader-case effect. The MS decomposition does not treat non-diseased cases separately from diseased cases. So, while the total variance can be easily generalized to new readers and to total cases (with a fixed disease

prevalence), additional modeling is needed to separately generalize to new non-diseased cases and new diseased cases (i.e., population with arbitrary disease prevalence).

- **OR components [8]:** The authors Obuchowski and Rockette (OR) decomposed the total variance into six components based on a two-factor ANOVA by modeling the accuracy of the jth reader using the ith diagnostic test. The components are: reader effect, modality-reader effect, same-reader-different-modality covariance, different-reader-same-modality covariance, different-reader-different-modality covariance, and residual error. The OR decomposition does not treat non-diseased cases separately from diseased cases. So, while the total variance can be easily generalized to new readers and to total cases (with a fixed disease prevalence), additional modeling is needed to separately generalize to new non-diseased cases and new diseased cases (i.e., population with arbitrary disease prevalence).

We emphasize that the software presents the four different variance decompositions of the reader average of *empirical* AUCs estimated by U-statistics and MLE [3]. The software does not estimate the DBM and OR components as originally proposed [7, 8]. They are obtained through linear combinations of the BDG components [3]. It is worth pointing out that the U-statistics and MLE estimation methods in this software are specific to the reader average of *empirical* AUCs, whereas the original DBM and OR *estimation* methods can be used for other performance measures.

Hypothesis Testing and Confidence Intervals

We use the methods of Hillis et al. 2008 [10] to determine the t-statistic, the corresponding degrees of freedom, and the p-value of the null hypothesis that the two modalities have equal AUCs (or that the single modality AUC = 0.5), and the confidence interval on the difference (or the confidence around the single modality AUC).

Sizing a Future Study

The variance components can be used to size future studies. In particular, the user specifies the significance level, expected effect size, and the number of readers, normal cases and abnormal cases. The software calculates the expected statistical power of the new study. The software implements two kinds of hypothesis testing. One is for a single modality; the null hypothesis is that the AUC is equal to a specified value. The second compares modalities, the null hypothesis is that the two modalities have the same AUC. The statistical power is computed in two ways. One uses the Z-test, and the other uses an F test from Hillis [9]. The Z-test assumes the variance is known even though it is actually estimated. The Z-test uses the ratio of the affect size over the square root of its estimated variance as the test statistic, and assumes the test statistic follows a standard normal distribution. The F-test does not assume the variance is known. It estimates the non-centrality parameter and the denominator degrees of freedom values using the components of variance (input or estimated) for the specified case and reader sample size and effect size. The statistical power is then computed using F distributions.

Database

The software includes a database that consists of results from simulated data and results from previous MRMC reader studies. We have accrued many reader study datasets and are working on the letters of permission to include them. The datasets are from a variety of sources including FDA sponsor data for imaging device approvals, as well as academic research studies. Each dataset contains

- **Basic information of the study:** this includes the source of the study, a summary of the study, related publications, and other information.
- **Key information of the study:** number of readers, number non-diseased cases, number of diseased cases, modalities used, and task.

- **Variance representations** as outlined above**.**

See Appendix A for details regarding the summary information.

In addition to the information outlined above, we are working on permission letters to share the ROC scores for each study. The ROC scores that we share are listed on the website. You can download a file with the study description that produced the ROC scores followed by the ROC scores themselves; the file is formatted for use with the iMRMC software. See the “Data Format for ROC Ratings” section for more information on the iMRMC file formatting and the “Sample Permission Letter” section below. We welcome more data. Please contact us.

Using the iMRMC Applet

Technical Specifications

The software is written in Java. It requires web browsers with Java Runtime Environment(JRE) 1.6.0_21-b07 Java HotSpot(TM) Client VM.

Please let us know what system and browser you have tried to run the applet, especially if your success is different from ours. We are still learning about portability.

First Time Use

Test that Java is working on your computer by going to
http://java.com/en/download/testjava.jsp

Enable Java in your web browser
http://java.com/en/download/help/enable_browser.xml

Enable clipboard access for java
https://www.member-data.com/rdc/help.aspx?topic=JavaClipboard#ptool

Run the applet
http://js.cx/~xin/mrmc.html

Input data

Select an input data option from the pull-down menu at the top titled “Input from database ...”. There are three options:

1. The user may choose a study from our database. After selecting a dataset, go to the “Data Analysis” section.
2. The user may choose to “Input raw data ...”, where “raw data” refers to ROC ratings from an MRMC reader study. See below “Data Format for ROC Ratings” to create a properly formatted file of ROC ratings. Click on the button “Input Pilot Study Data”. This opens an input window for the ROC ratings. “Select All” and “Copy” the contents of the ROC ratings file and “Paste” them into the input window. Click “OK” to continue on to the “Data Analysis” section. If the ROC ratings data is not properly formatted, the “OK” button will not close the input window.
3. The user may choose to input the variance components from a existing study manually. The user may choose to use other software to compute the components of variance and input those components manually into our applet to size a trial. The applet currently can size trials based on BDG components, DBM components, and OR components. After selecting a dataset, skip to the “Data Analysis” section.

Data Format for ROC Ratings

Here we describe a properly formatted file of the ROC ratings. These instructions and a sample input file can be accessed via buttons that appear after selecting “Input raw data ...”. You can “Select All” and “Copy” the contents of the sample input file.

The file format has two parts: the study description at the top followed by a list of the ROC ratings. The study description can include any information as free text. It must include three lines corresponding to the size of the experiment and then conclude with a line stating “BEGIN DATA”. We demonstrate the formatting of these lines in an example. If the study has 9 readers, 55 diseased patients, and 75 non-diseased patients, then a legitimate study description can be nothing more than the following lines.

```
N0: 75
N1: 55
NR: 9
BEGIN DATA
```

The list of ROC ratings has a row for each case and reader in the study (in any order). The current version of iMRMC expects fully-crossed data from two modalities: every reader reads every case in both modalities. Consequently, there should be NR*(N0+N1) rows. Each row has five fields: the reader id (integer), case id (integer), truth index (0=non-diseased or 1=diseased), and the ratings corresponding to modality 1 and 2 (integer or float). High ratings should indicate high likelihood or confidence of disease and low ratings should indicate less likelihood or confidence of disease. For example, the first six rows could look like the following.

```
1   1   0   1       3
1   2   0   2       3
1   3   0   2       3
2   1   0   1.876   2
2   5   1   5       2
3   6   0   1       2.001
```

In the example, we see the ROC ratings from three readers. The first three lines show ratings for reader 1 reading cases 1, 2, and 3, which are all non-diseased cases. The fourth line shows the ratings from reader 2 reading non-diseased case 1. The fifth line shows the ratings from reader 2 reading diseased case 5. The sixth line shows the ratings from reader 3 reading non-diseased case 6.

Data Analysis

You can estimate the components of variance for modality 1, modality 2, or the difference in modalities. Click the corresponding radio button and then click “OK”. This will populate the “Data Analysis” table.

As discussed, the components of variance can be given in five different representations: these are available by clicking on the corresponding tab of the “Data Analysis” table and reading the row labeled “components”. Each component contributes to the total variance; specifically, the total variance is a linear combination of the components of variance. For each component of variance, the corresponding coefficients/weights are given in the row labeled “coeff”, and the corresponding contributions to the total variance are given in the row labeled “total”. The contributions are summed to produce the total variance, and the square root of that is the standard error, which is displayed to the right of the “Data Analysis” table. See the brief summary of the GUI for a little more details on what they are and how they are estimated.

Sizing a Future Study

The second panel is for sizing a trial using the components of variance shown. Input the number of readers, non-diseased cases, and diseased cases, as well as the significance level and the

effect size. Then click “Size a Trial”. The results in this section are updates to the coefficients, the total contributions to the variance, and the resulting standard error. Other summary statistics are also produced (work in progress): power, d.f, p-value, confidence intervals, t-stat.

Generate a Report

The following information is summarized and displayed by the GUI. A report of the results may also be generated:

1) size of the existing study and AUC values
2) components of variance calculated from the existing study in BDG, BCK, DBM, OR, MS representations
3) size of the future study and AUC values
4) effect size, significance level, and corresponding statistical power
5) For input from database or input of raw data, the report may also include a description of the existing study if available. For the manual input, the report may not include all four kinds of components of variance, depending on whether the input components of variance are convertible to other types of components of variances. For example, if the user chooses to input DBM components of variance, OR components can be derived while BDG components cannot.

Database

At the bottom of the applet is the means by which we are sharing the whole database (basic information of each study, size of the study, and variance decompositions). The “database” is populated by simulated datasets right now. The buttons at the bottom are related to downloading a spread sheet of that data (work in progress).

References

[1] Metz, C. E., “Basic principles of ROC analysis.”, Semin Nucl Med. 1978 Oct;8(4):283-98.

[2] Efron, B. & Tibshirani, R. J, “An Introduction to the Bootstrap: Monographs on Statistics and Applied Probability”, Chapman & Hall, New York, N.Y, 1993.

[3] Gallas, B. D.; Bandos, A.; Samuelson, F. & Wagner, R. F. (2009), 'A Framework for Random-Effects ROC Analysis: Biases with the Bootstrap and Other Variance Estimators', Commun Stat A-Theory 38(15), 2586-2603

[4] Gallas, B. D. (2006), 'One-Shot Estimate of MRMC Variance: AUC', *Acad Radiol* **13**(3), 353-362.

[5] Gallas, B. D. & Brown, D. G. (2008), 'Reader Studies for Validation of CAD Systems', *Neural Networks* **21**(2-3), 387-397.

[6] Clarkson, E.; Kupinski, M. A. & Barrett, H. H. (2006), 'A Probabilistic Model for the MRMC Method. Part 1.Theoretical Development', *Acad Radiol* **13**(11), 1410-1421.

[7] Dorfman, D. D.; Berbaum, K. S. & Metz, C. E. (1992), 'Receiver Operating Characteristic Rating Analysis: Generalization to the Population of Readers and Patients with the Jackknife Method', *Invest Radiol* **27**(9), 723-731.

[8] Obuchowski, N. A. (1995), 'Multireader, Multimodality Receiver Operating Characteristic Curve Studies: Hypothesis Testing and Sample Size Estimation Using an Analysis of Variance Approach with Dependent Observations', *Acad Radiol* **2**(Suppl 1), S22-S29.

[9] Hillis S.L., Berbaum K. S. “Power estimation for the Dorfman-Berbaum-Metz method.” Acad Radiol. 2004 Nov;11(11):1260-73, 2004.

[10] Hillis, S. L.; Berbaum, K. S. & Metz, C. E. (2008), 'Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis.', Acad Radiol 15(5), 647--661.

Appendix A

In this appendix we describe the basic and key information we intend to provide for each dataset.

Basic Information

The following fields summarize the data source, setting, experimental conditions, and reference material related to the reader ROC ratings.

1. **Source: (Owner of data, free text)** Data owner or representative, affiliation, mailing address, phone, and email.
2. **Summary: (free text)** A succinct description of the reader study (readers, cases, reference method, rating method) with more details on the task and modalities (treatments) under investigation.
3. **Publication: (free text)** Full references to articles that describe or utilize the reader study or data. This field should also include other sources of information related to the study or data (links to webpages, protocol identifiers).
4. **Notes: (free text)** This field can alert the user to any peculiar issues related to the data. For example, is there any missing data or corrections to the data?
5. **Acknowledgements: (free text)**
6. **Nr: (Number of readers, integer)**
7. **N0: (Number of normal cases, integer)**
8. **N1: (Number of abnormal cases, integer)**
9. **AUC1: (AUC Modality 1, floating point)**
10. **AUC2: (AUC Modality 2, floating point)**
11. **DAUC: (Difference in AUCs, floating point)**

Key Information

The following items are critical for determining the application area of the components of (co)variance. They may be renamed and regrouped over time as deemed useful.

12. **Mod1: (Modality 1, limited choices)**
13. **Mod2: (Modality 2, limited choices)**
    a. simulation
    b. film radiography
    c. digital radiography
    d. film with CAD
    e. film mammography
    f. digital mammography
    g. SE MRI
    h. CINE MRI
    i. Multi-detector row CT
    j. Digital breast tomosythesis
    k. Mammography + ultrasound
    l. Mammography + ultrasound + CAD
    m. Ultrasound
    n. Ultraound + CAD
14. **Task: (limited choices)**
    a. simulation
    b. chest and abdominal abnormality detection
    c. breast abnormality detection
    d. breast cancer detection
    e. detection of thoracic aortic dissection
15. **TaskAdd: (limited choices)** This field indicates a modification or qualifier to the task. Multiple qualifiers may be indicated. More liberty will be given to data suppliers in defining these choices. This list is expected to grow.

a. pediatric
b. malignant mass
c. malignant calcification
d. nodule
e. pneumothorax
f. interstitial disease