iBinaryMRMC

User Manual

Version 1.0 Beta

Weijie Chen, PhD

Adam Wunderlich, PhD

US Food and Drug Administration
Center for Devices and Radiological Health
Office of Science and Engineering Labs
Division of Imaging and Applied Mathematics

March 7, 2014

Disclaimer

This software and documentation (the "Software") were developed at the Food and Drug Administration (FDA) by employees of the Federal Government in the course of their official duties. Pursuant to Title 17, Section 105 of the United States Code, this work is not subject to copyright protection and is in the public domain. Permission is hereby granted, free of charge, to any person obtaining a copy of the Software, to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, or sell copies of the Software or derivatives, and to permit persons to whom the Software is furnished to do so. FDA assumes no responsibility whatsoever for use by other parties of the Software, its source code, documentation or compiled executables, and makes no guarantees, expressed or implied, about its quality, reliability, or any other characteristic. Further, use of this code in no way implies endorsement by the FDA or confers any advantage in regulatory decisions. Although this software can be redistributed and/or modified freely, we ask that any derivative works bear some notice that they are derived from it, and any modified versions bear some notice that they have been modified.

This is a work in progress. To report bugs, problems, or any other questions related to this software or if you want to be notified when a new version is released, please contact

Weijie Chen, Ph.D.
Division of Imaging and Applied Mathematics
Office of Science and Engineering Laboratories
FDA/CDRH
10903 New Hampshire Avenue, WO62-4104
Silver Spring, MD 20993-0002
301-7962663 (phone) 301-796-9925 (fax)
weijie.chen@fda.hhs.gov

iBinaryMRMC: a software package for simulation, analysis, and sizing of MRMC reader studies with binary assessment

1. *Introduction*

Many medical diagnostic devices, especially those for medical imaging, are evaluated in the hands of physicians who use the device to make diagnostic assessment of patients. The effectiveness of a new device modality is typically established by showing that the performance of physicians using the new modality is non-inferior or superior to that using a conventional modality on a defined set of diagnostic tasks. Such studies are called multi-reader, multi-case(MRMC) studies, because they are based on a sample of representative physicians, or “readers” who read the output (e.g., images) from the device for a sample of representative patients, or “cases.”

In certain MRMC reader studies, the reader’s diagnostic assessment on a case is or can be converted in some meaningful way to a binary assessment, for example, whether the reader’s assessment is in agreement with a reference standard. This software package provides computer programs for simulating binary MRMC study data, validating statistical analysis methods, and sizing a reader study. In section 2 below, we describe how to use the software. The simulation modeling method and the relevant analysis methods in the literature are summarized in section 3.

2. *Using the software*

2.1 Platform

The software is written in Matlab (MathWorks, Inc., Natick, Massachusetts) and has been tested on Matlab Version 7.5.0.338 (R2007b). It should be runnable on Matlab of any version later than R2007b on either Windows or Linux systems.

2.2 Simulation of binary MRMC data

```
function [S1,S2] = iSimuBinaryMRMC(r,PC,Nr,Nc)
```

Generate binary MRMC data with specified parameters and sample size.
INPUTS:
r: a vector of length 7 representing 7 correlation coefficient parameters that characterize the correlations in the binary data.
r(1): Correlation between two cases from the modality 1 (conventional modality) read by the same reader.
r(2): Correlation between two cases from the modality 2 (new modality) read by the same reader.
r(3): Correlation between two readers reading the same case from modality 1.
r(4): Correlation between two readers reading the same case from modality 2.
r(5): Correlation between two modalities with the same reader reading the same case.

r(6): Correlation between two cases from different modalities read by the same reader.
r(7): Correlation between two readers reading the same case from the different modalities.

PC: a vector of length 2 representing the expected percentage correct (or agreement) for the two modalities.
PC(1): the expected percentage correct (or agreement) for modality 1.
PC(2): the expected percentage correct (or agreement) for modality 2.

Nr: the number of readers.

Nc: the number of cases.

Note:

The correlation parameters r and the expected PC parameters are ideally measured in a pilot study or adopted from past relevant studies. If they are specified arbitrarily (for example, to explore a range of possible parameters), the PC parameters are between 0 and 1 and all the correlations are between 0 and 1 and should satisfy the following constraints:
r(1) >= r(6), r(2) >= r(6), r(3) >= r(7), r(4) >= r(7), r(5) >= r(6)+r(7). For an explanation of these correlation parameters, see section 3.2.

OUTPUTS:
S1: binary assessment or success data for modality 1, Nc x Nr matrix, S1(k,j) is the binary assessment or success data for reader j assessing case k using modality 1, for example, 1 means agreement with the reference standard, 0 means disagreement with the reference standard.
S2: binary assessment or success data for modality 2, Nc x Nr matrix.

2.3 Monte Carlo validation of an analysis method

```
function prob = iValidateBinaryMRMC(anaMethod,Nr, Nc, r, PC, nexp)
```

Validation of an analysis method using Monte Carlo simulation in terms of the empirical coverage probability of the 95% confidence interval estimated by the analysis method.

INPUTS:
anaMethod: a character string specifying the .m file name that implements an analysis method. The file anaMethod.m must be in the Matlab path (where ‘anaMethod’ is replaced by the actual file name). It is required that anaMethod.m takes as input two binary MRMC data matrices S1 and S2 as defined in section 2.2. The output of anaMethod.m should be a structure with a field “CI95”, which is the 95% confidence interval estimated by anaMethod.
Nr, Nc, r, PC: the same as those in function iSimuBinaryMRMC defined in section 2.2.
nexp: integer, number of Monte Carlo trials to calculate the empirical coverage probability of the estimated 95% confidence interval. Default value = 10,000.

OUPUT:
prob: empirical coverage probability of the estimated 95% confidence interval.

2.4 Monte Carlo method for power calculation

```
function pow = iPowerBinaryMRMC(anaMethod, Nr, Nc, r, PC, nim, nexp)
```

Calculation of empirical power of an analysis method (anaMethod) in a non-inferiority study using Monte Carlo simulations given a set of parameters and sample sizes.

INPUTS:
anaMethod, Nr, Nc, r, PC, nexp: the same as those function `iValidateBinaryMRMC` defined in section 2.3.
nim: non-inferiority margin.

OUTPUT:
pow: empirical power in nexp Monte Carlo trials.

2.5 Example

A Matlab script (test.m) is included in the software package for a demonstration of the use of the key functions.

```
% r = [r_c1,r_c2,r_r1,r_r2,r_t,r_tc,r_tr]
disp('arbitrary parameters for testing the program...');
r = [.007 .007 .25 .25 .50 .005 .20] %arbitrary parameters for testing the
program
PC = [.85 .85]
Nr = 10
Nc = 300

%Generate one dataset and analyze it using the OR method
[S1,S2] = iSimuBinaryMRMC(r,PC,Nr,Nc);
ret = iAnalyzeBinaryMRMC_OR(S1,S2);
disp('the estimated 95% confidence interval is:');
ret.CI95

anaMethod = 'iAnalyzeBinaryMRMC_OR';
nexp = 20000;
%validation of the OR method in terms of the empirical coverage probability
%of the estimated 95% confidence intervals
disp('the empirical coverage probability in 20000 Monte Carlo trials is:');
prob = iValidateBinaryMRMC(anaMethod,Nr, Nc, r, PC, nexp)

%Power calculation for a hypothetical non-inferiority test
nim = .04;
disp('the statistical power for the hypothetical parameters is:');
pow = iPowerBinaryMRMC('iAnalyzeBinaryMRMC_OR',Nr, Nc, r, PC, nim, nexp)
```

3. Simulation model and analysis/sizing methods: an overview

3.1. Setup

Let $i = 1, 2$ index the conventional and new modalities.

Let $j = 1, 2, \ldots, N_r$ index N_r readers.

Let $k = 1, 2, \ldots, N_c$ index patients (cases).

Let a_{ijk} denote the reading result for the i^{th} modality, j^{th} reader, and k^{th} case, $a_{ijk} = 1$ if the reading agrees with the reference standard (or "truth") and $\mathrm{a_{ijk}} = 0$ otherwise.

For a fully-crossed study, in which all readers read all cases, an unbiased point estimate of the percent agreement between the two modalities is:

$$\bar{d} = \frac{1}{N_r N_c} \sum_{j=1}^{N_r} \sum_{k=1}^{N_c} \left(a_{1jk} - a_{2jk}\right).$$

3.2. Simulation model

We use $\mathrm{X_{ijk}}$ to denote the underlying latent continuous random variable for the binary variable a_{ijk}. X_{ijk} can be modeled with a mixed-effect model:

$$X_{ijk} = \mu + \tau_i + R_j + C_k + (\tau R)_{ij} + (\tau C)_{ik} + (RC)_{jk} + e_{ijk}$$

where μ is the overall mean, τ_i is the fixed effect of modality i, R_j is the random effect for reader j , C_k is the random effect for case k, $(\tau R)_{ij}$, $(\tau C)_{ik}$, and $(RC)_{jk}$ are the corresponding two-way interactions (random), and e_{ijk} is a random error term (which includes the three-way interaction term that is indistinguishable from the random error with only one reading). The six random terms are assumed to be independent Gaussian random variables with zero mean and variances $\sigma_R^2, \sigma_C^2, \sigma_{\tau R}^2, \sigma_{\tau C}^2, \sigma_{RC}^2, \sigma_e^2$ respectively. The total variance of X_{ijk} is $\sigma_T^2 = \sigma_R^2 + \sigma_C^2 + \sigma_{\tau R}^2 + \sigma_{\tau C}^2 + \sigma_{RC}^2 + \sigma_e^2$.

Under this model, it is straightforward to show that there are five non-trivial correlations in the continuous data:

- Correlation between two modalities with the same reader reading the same case ρ_τ

  $$\rho_\tau \equiv Corr\left(X_{ijk}, X_{i'jk}\right) = (\sigma_R^2 + \sigma_C^2 + \sigma_{RC}^2)/\sigma_T^2$$

  The model assumes that this correlation is the same for any reader and any case.

- Correlation between two readers reading the same case from the same modality ρ_R

  $$\rho_R \equiv Corr\left(X_{ijk}, X_{ij'k}\right) = (\sigma_C^2 + \sigma_{\tau C}^2)/\sigma_T^2$$

  The model assumes that this correlation is the same for any pair of readers, for any case, and any modality.

- Correlation between two readers reading the same case from the different modalities $\rho_{\tau R}$
  $$\rho_{\tau R} \equiv Corr\left(X_{ijk}, X_{i'j'k}\right) = \sigma_C^2/\sigma_T^2$$
  The model assumes that this correlation is the same for any pair of readers and any case.

- Correlation between two cases from the same modality read by the same reader ρ_C
  $$\rho_C \equiv Corr\left(X_{ijk}, X_{ijk'}\right) = (\sigma_R^2 + \sigma_{\tau R}^2)/\sigma_T^2$$
  The model assumes that this correlation is the same for any pair of cases, for any reader and any modality.

- Correlation between two cases from different modalities read by the same reader
  $$\rho_{\tau C} \equiv Corr\left(X_{ijk}, X_{i'jk'}\right) = \sigma_R^2/\sigma_T^2$$
  The model assumes that this correlation is the same for any pair of cases and for any reader.

To obtain binary MRMC data, a threshold is applied to dichotomize the continuous-valued data, where the threshold is chosen to achieve a specified reader-averaged percent agreement for each modality. Formally,

$$a_{ijk} = \mathrm{I}(X_{ijk} > TH_i),$$

where I is an indicator function, $\mathrm{I}(\text{true}) = 1, \mathrm{I}(\text{false}) = 0$, and $TH_i = \mu + \tau_i + \sigma_T \Phi^{-1}(1 - p_i)$ with p_i being the expected percentage agreement for modality i and Φ being the cumulative distribution function of the standard normal distribution.

There are five non-trivial correlations in the binary data, which are conceptually similar to those in the continuous domain. The correlation in the binary domain is a function of the threshold and the variance components in the continuous domain. The correlations in the two domains can be mapped to each other using numerical computation (that involves numerical integration).

Note:

A more general model allows $\sigma_{\tau R}^2$, $\sigma_{\tau C}^2$, and σ_e^2 to depend on the modality. This leads to 7 (instead of 5) correlation parameters with two versions of ρ_C and two versions of ρ_R. This more general model is implemented in the *iBinaryMRMC* software package.

3.3. Summary of methods for data analysis and sizing in the literature

The research on methodologies for data analysis and sizing of multi-reader multi-case reader studies has a history of over 20 years since the landmark paper by Dorfman, Berbaum, and Metz (DBM) [1]. The applications have been primarily in radiological imaging and the performance metric has been primarily the area under the receiver operating characteristic (ROC) curve (AUC). However, these methods deal with reader variability and case variability of a *performance metric* that is not necessarily AUC. In fact, many of these methods can be directly applied or adjusted in a straightforward manner so that they can be applied to the percentage agreement metric. Below we will briefly review part of this literature and point out necessary adjustments that are needed for binary data.

Obuchowski-Rockette (OR) method

The Obuchowski-Rockette (OR) method was first proposed in 1995 [2] [3]. In essence, this method uses a modality x reader ANOVA to model the performance of readers $\hat{\theta}_{ij}$ $(i = 1,2; j = 1,2, \ldots, N_r)$, where $\hat{\theta}_{ij}$ is the performance of the j^{th} reader on the i^{th} modality (the performance can be either AUC or percentage agreement). But unlike conventional ANOVA models they allow the errors to be correlated to account for the correlation structure in the MRMC problem. The method ends up with an adjusted F test statistic, which, with two modalities, is equivalent to an adjusted *t* statistic, where the word "adjusted" (rather than conventional) indicates accounting of correlations in readers' performance.

The OR method was later refined by Hillis [4] [5], whose contribution is a better method for calculating the degree of freedom of the test statistic. Based on the updated OR method, Hillis et al. provided a method for power estimation and sizing a MRMC reader study [6]. More recently, Hillis put the OR model in the marginal-mean ANOVA framework that allows an easier derivation and interpretation of the OR method [7].

Because the OR method models the observed reader performance, it can be applied directly to the problem of analyzing the binary data.

Nonparametric method for variance estimation

Gallas et al. investigated a nonparametric method for variance analysis in MRMC studies, which applies regardless of whether the endpoint is AUC [8] or an average of binary data [9] like percentage agreement. Note that the Gallas et al. only studied the variance of the endpoint in one modality, which was proven to be unbiased. It is straightforward to get the variance of the performance difference between two modalities $(\bar{d})$ as the Gallas et al. method only needs the addition of the covariance between the two modalities, which has a similar formula to the variance formula.

For hypothesis testing, one can use a normal approximation for a ballpark estimation. Empirical experience has shown that, when the number of cases or the number of readers is very small, the normal approximation may result in inflated type I error. A better inference is to formulate a *t* test statistic and use a sophisticated method (such as that of Hillis [5]) for calculating its degree of freedom.

Equivalence of DBM, OR, and nonparametric methods for fully-crossed designs

Although the DBM, OR, and nonparametric methods mentioned above are conceptually different and approach the MRMC problem from different angles, they have been generally shown to be consistent with each other in practice. Moreover, under certain conditions, they can be proved to be mathematically equivalent. Currently, the *iBinaryMRMC* package supplies the OR method for statistical analysis of fully-crossed study designs.

Flexible study designs

Most of the methods reviewed above were initially designed for a fully-paired study design, i.e., every reader reads every case for both modalities. These methods can be adjusted for analyzing alternative designs, such as the “reader-patient” design where readers are divided into subgroups and cases are divided into subgroups and every subgroup of readers read its own subgroup of cases, or even a hybrid design which is a mix of partially fully-paired design and partially “reader-patient” design [7],[10].

Simulation approach to sizing and validation

It is well recognized that simulation is a useful tool to validate the analysis methods and aid sizing. The model in Section 3.2 can be used to simulate binary MRMC data and a Monte Carlo approach can be used to help sizing the study or validate the performance of an analysis method. In this approach, one repeatedly performs the simulation experiments (i.e., drawing finite data from the simulation model). In each repetition, the analysis method is applied to estimate the variance of the agreement difference and the 95% confidence interval. One can then estimate the coverage probability of the confidence interval from repeated experiments. Likewise, one can validate the accuracy of a variance estimator by comparing the mean of the estimates with the Monte Carlo truth. Similarly, this approach can aid sizing by simulating the alternative hypothesis. With specified sample sizes (# of readers and # of cases), one repeatedly performs the simulation experiments and examine the statistical power (i.e., empirical probability that the analysis method correctly rejects the null hypothesis). Then the sample sizes can be varied iteratively to achieve the desired statistical power.

References

[1] 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.

[2] Obuchowski, N. A., Rockette, H. E. (1995a). Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations. Commun. Statist. Simul. 24(2):285–308.

[3] Obuchowski, N. A. (1995b). 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.

[4] Hillis SL, Obuchowski NA, Schartz KM, Berbaum KS. A comparison of the Dorfman–Berbaum–Metz and Obuchowski–Rockette Methods for receiver operating characteristic (ROC) data. Statistics in Medicine 2005; 24:1579–1607. DOI: 10.1002/sim.2024.

[5] Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Statistics in Medicine 2007; 26:596–619. DOI: 10.1002/sim.2532.

[6] Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for multireader ROC methods: an updated and unified approach. Acad Radiol 2011; 18:129–142.

[7] Hillis SL. A marginal-mean ANOVA approach for analyzing multireader multicase radiological imaging data. Statistics in Medicine 2014; 33:330–360.

[8] Gallas, B. D. (2006). One-shot estimate of MRMC variance: AUC. Acad. Radiol. 13(3):353–362.

[9] Gallas, B. D., Pennello, G. A., Myers, K. J. (2007). Multi-reader multi-case variance analysis for binary data. J. Opt. Soc. Amer. A 24(12):B70–B80.

[10] Obuchowski NA, Gallas BD, Hillis SL. (2012). Multi-reader ROC studies with split-plot designs: a comparison of statistical methods. Academic Radiology; 19:1508–1517. DOI: 10.1016/j.acra.2012.09.012.