Published in final edited form as:

Proc SPIE Int Soc Opt Eng. 2020 February ; 11316: . doi:10.1117/12.2549075.

Multi-reader multi-case analysis of variance software for diagnostic performance comparison of imaging modalities

Brian J. Smith^a, Stephen L. Hillis^b

^aDepartment of Biostatistics, University of Iowa, Iowa City, USA

^bDepartments of Radiology and Biostatistics, University of Iowa, Iowa City, USA

Abstract

A common study design for comparing the diagnostic performance of imaging modalities is to obtain modality-specific ratings from multiple readers of multiple cases whose true statuses are known. Typically, there is overlap between the modalities, readers, and/or cases for which special analytical methods are needed to perform statistical comparisons. We describe our new R software package MRMCaov, which is designed for multi-reader multi-case comparisons of two or more imaging modalities. The software allows for the comparison of reader performance metrics, such as area under the receiver operating characteristic curve (ROC AUC), with analysis of variance methods originally proposed by Obuchowski and Rockette (1995) and later unified and improved by Hillis and colleagues (2005, 2007, 2008, 2018). MRMCaov is an open-source package with an integrated command-line interface for performing multi-reader multi-case statistical analysis, plotting, and presenting results. Features of the package include (1) ROC curves estimated parametrically or non-parametrically; (2) reader-specific ROC curves and performance metrics; (3) user-definable performance metrics; (4) modality-specific estimates of mean performance along with confidence intervals and p-values for statistical comparisons; (5) support for factorial, nested, or partially paired study designs; (6) inference for random readers and cases, random readers and fixed cases, or fixed readers and random cases; (7) DeLong, jackknife, or unbiased covariance estimation; and (8) compatibility with Microsoft Windows, Mac OS, and Linux.

Keywords

multi-reader multi-case; ANOVA; ROC analysis; diagnostic radiology; software

1. INTRODUCTION

A common study design for comparing the diagnostic performance of imaging modalities, or diagnostic tests, is to obtain modality-specific ratings from multiple readers of multiple cases (MRMC) whose true statuses are known. In such a design, receiver operating characteristic (ROC) indices, such as area under the ROC curve (ROC AUC), can be used to quantify correspondence between reader ratings and case status. Indices can then be compared statistically to determine if there are differences between modalities. However,

special statistical methods are needed when readers or cases represent a random sample from a larger population of interest and there is overlap between modalities, readers, and/or cases. An ANOVA model designed for these characteristics of MRMC studies was initially proposed by Dorfman et al.¹ and Obuchowski and Rockette² and later unified and improved by Hillis and colleagues.^{3–6} Their models are implemented in the MRMCaov R package.⁷

MRMCaov performs multi-reader multi-case analysis of variance for the performance comparison of imaging modalities. This software is the first R implementation of Dr. Hillis unified methodology and builds upon his OR-DBM MRMC SAS software.⁸ It is designed to be user friendly, integrate with the R statistical computing and graphics environment, and offer new features and methodologies. Current features of the package are summarized below. Usage of the software is illustrated with a medical imaging example in the next section.

-
- MRMCaov Features
- ROC curves estimated parametrically or non-parametrically.
 - Reader-specific ROC curves and performance metrics.
 - Performance metric functions to compute area under the ROC curve, sensitivity for a specified specificity, and specificity for a specified sensitivity.
 - User-definable performance metrics.
 - Modality-specific estimates of mean performance along with confidence intervals and p-values for statistical comparisons.
 - Comparison of two or more modalities.
 - Support for factorial, nested, or partially paired study designs.
 - Inference for random readers and cases, random readers and fixed cases, or fixed readers and random cases.
 - DeLong, jackknife, or unbiased covariance estimation.
 - Compatibility with Microsoft Windows, Mac OS, and Linux.

2. USAGE

MRMCaov provides both graphical and tabular analysis results, including reader-specific ROC curves and AUC estimates, modality-specific estimates, confidence intervals, and p-values for statistical comparisons. It was released in July 2019 as a completely new R package, which includes a new method for unbiased covariance estimation as well as other features not collectively available in any other existing R package.

2.1 Installation

The R package is currently available for download from <https://github.com/brian-j-smith/MRMCaov>. Installation instructions are provided at the download site. Once installed, the package may be loaded for use in an R software⁹ session with the command `library(MRMCaov)`.

2.2 Data

Input data for MRMCaov analysis should be given as an R data frame with columns for reader, treatment, and case identifiers as well as event status truth and ratings and with their values in the rows. An example data frame named `VanDyke` is provided with the package. The data come from a study in which the relative performance of cinematic presentation of MRI (CINE MRI) was compared to single spin-echo magnetic resonance imaging (SE MRI) for the detection of thoracic aortic dissection.¹⁰ Forty five patients with aortic dissection and 69 without dissection were imaged with both modalities. Based on the images, five radiologists rated patient disease status as 1 = definitely no aortic dissection, 2 = probably no

aortic dissection, 3 = unsure about aortic dissection, 4 = probably aortic dissection, or 5 = definitely aortic dissection. Interest lies in estimating ROC curves for each combination of reader and modality and in comparing modalities with respect to summary statistics from the curves.

R Example: First 10 observations in VanDyke R data frame

	reader	treatment	case	truth	rating	case2	case3
1	1	1	1	0	1	1.1	1.1
2	1	2	1	0	3	1.1	2.1
3	2	1	1	0	2	2.1	1.1
4	2	2	1	0	3	2.1	2.1
5	3	1	1	0	2	3.1	1.1
6	3	2	1	0	2	3.1	2.1
7	4	1	1	0	1	4.1	1.1
8	4	2	1	0	2	4.1	2.1
9	5	1	1	0	3	5.1	1.1
10	5	2	1	0	2	5.1	2.1

Descriptions of the data frame variables are as follows.

reader: unique identifiers for the five radiologists.

treatment: identifiers for the imaging modality (1 = CINE MRI, 2 = SE MRI).

case: identifiers for the 114 cases.

truth: indicator for thoracic aortic dissection (1 = done, 0 = not done).

rating: five-point ratings given to case images by the readers.

case2: example identifiers representing nesting of cases within readers.

case3: example identifiers representing nesting of cases within treatments.

2.3 MRMC Analysis Specification

The first step in conducting a multi-reader, multi-case analysis with the package is an R command-line call to the function `mrmc` to specify the data inputs, performance metric of interest, and covariance estimation method. Below is a summary of the function arguments.

Multi-reader multi-case analysis function and arguments.

Usage

```
mrmc(response, test, reader, case, data, method = jackknife, design = NULL)
```

Arguments

response: response metric expressed in terms of a package-supported performance metric.

test: test identifiers variable.

reader: reader identifiers variable.

case: case identifiers variable.

data: data frame containing the response, tests, readers, and cases variables.

method: function, function call, or character string naming the method to use in calculating performance metric covariances.

design: one of the following study designs: 1 = factorial, 2 = cases nested within readers, 3 = cases nested within tests, or NULL to automatically set the design based on variable codings in data.

Study design is automatically inferred from the case identifiers. If the identifiers are different across readers and tests, a factorial design (`design = 1`) is assumed. Cases are assumed to be nested within readers if their identifiers are the same across readers and are assumed to be nested within tests if their identifiers are the same across tests. Examples of the latter two designs are given with the `case2` and `case3` variables in the VanDyke data frame. The case variable is coded for the full factorial design originally employed in the study. For the specification of response in `mrmc`, several performance metric functions are supplied by the package. Metrics include area under ROC curves, specificity, and sensitivity estimated either empirically or parametrically. Other, user-defined functions are supported.

```

Package-supplied performance metric functions.

Usage:
  empirical_auc(truth, rating)
  empirical_sens(truth, rating, spec)
  empirical_spec(truth, rating, sens)
  proproc_auc(truth, rating)
  proproc_sens(truth, rating, spec)
  proproc_spec(truth, rating, sens)

Arguments:
  truth: vector of true binary statuses.
  rating: vector of numeric ratings.
  sens/spec: numeric sensitivity/specificity at which to calculate specificity/sensitivity.

```

Performance metrics measure the degree to which higher case ratings are associated with positive case statuses, where positive status is taken to be the highest level of the truth variable. AUC is derived from ROC curves estimated empirically by `empirical_auc` and `trapezoidal_auc` and parametrically with a proper binormal model by `proproc_auc`. Details of empirical and parametric ROC AUC estimation and derived performance metrics can be found in the book of Pepe.¹¹ Methods for calculating covariances include DeLong,¹² jackknife,¹³ and unbiased.⁶ The methods that can be used in an analysis depends on the performance metric specified. Jackknife will work with any metric. DeLong and unbiased can be used with empirically estimated AUC. The `mrnc` function automatically checks for and informs users if called with an incompatible combination of covariance method and performance metric. In the next code example, `mrnc` is called to specify an analysis in which the VanDyke imaging modalities are compared with respect to area under the empirically estimated ROC curve and with unbiased covariance estimation. Results of the function call are saved to a new variable `est`. As can be seen from printing the results, the function calculates and returns several quantities needed for statistical comparison of the imaging modalities, including ROC estimates of false and true positive rates, areas under the ROC curves, analysis of variance sums of squares, and covariance estimates.

```

R Example: mrnc function call

> library(MRNCdev)
> est <- mrnc(empirical_auc(truth, rating), treatment, reader, case,
+             data = VanDyke, method = "unbiased")
> print(est, n = 10)

Positive truth status: 1

Response metric estimates:

  treatment reader  N empirical_auc
1          1      1 114      0.9196
2          2      2 114      0.9478
3          1      2 114      0.8588
4          2      2 114      0.9053
5          1      3 114      0.9039
6          2      3 114      0.9217
7          1      4 114      0.9731
8          2      4 114      0.9984
9          1      5 114      0.8298
10         2      5 114      0.9300

ANOVA Table:

      Df Sum Sq Mean Sq
treatment 1 0.0048 0.00480
reader     4 0.0153 0.00384
treatment:reader 4 0.0022 0.00055

Ubochowski-Rockette error variance and covariance estimates:

      Estimate Correlation
Error 0.0005943      NA
Cov1  0.0002327  0.3933
Cov2  0.0002944  0.3945
Cov3  0.0001608  0.2706

First 10 ROC performance metrics:

  treatment reader  FPR  TPR
1          1      1 0.0000 0.0000
2          1      1 0.0149 0.6222
3          1      1 0.0438 0.8444
4          1      1 0.1884 0.8889
5          1      1 0.3184 0.9111
6          1      1 0.0000 1.0000
7          2      1 0.0000 0.0000
8          2      1 0.0000 0.7333
9          2      1 0.1045 0.8444
10         2      1 0.3184 0.9778

```

By default, readers and cases are assumed to be random factors in the MRMC analysis. Either one of the factors may be designated as fixed in calls to `mrnc` with the syntax `fixed(<variable name>)`, where `<variable name>` is the name of the corresponding reader or case variable.

```
R Example: Designation of fixed readers or cases

## Fixed readers
mrmc(empirical_auc(truth, rating), treatment, fixed(reader), case,
data = VanDyke, method = unbiased)

## Fixed cases
mrmc(empirical_auc(truth, rating), treatment, reader, fixed(case),
data = VanDyke, method = unbiased)
```

2.4 Statistical Analysis

Statistical comparisons of treatment modalities is performed by calling the summary function with output returned by mrmc. The summary call produces ANOVA results for a global test of equality of ROC AUC means across all treatment modalities and tests of pairwise differences, along with confidence intervals for the differences and intervals for the individual modalities.

```
R Example: Summary results of MRMC statistical analysis

> summary(est)
Factor types: Random Readers and Random Cases
Experimental design: factorial
Obuchowski-Rockette variance component and covariance estimates:

      Estimate Correlation
reader 0.0015697      NA
treatment:reader 0.0002641      NA
Error 0.0005943      NA
Cov1 0.0002337 0.3933
Cov2 0.0002344 0.3945
Cov3 0.0001608 0.2706

ANOVA global test of equal treatment empirical_auc:

      MS(T)  MS(T:R)  Cov2  Cov3 Denominator  F df1 df2 p-value
1 0.004796 0.000551 0.0002344 0.0001608 0.0009191 5.218 1 11.13 0.04295

95% CIs and tests for treatment empirical_auc pairwise differences:

      Comparison Estimate StdErr df CI.Lower CI.Upper t p-value
1 1 - 2 -0.0438 0.01917 11.13 -0.080943 -0.005658 -2.284 0.04295

95% treatment empirical_auc CIs (each analysis based only on data for the
specified treatment):

      Estimate MS(R) Cov2 StdErr df CI.Lower CI.Upper
1 0.8970 0.003083 0.0001255 0.03069 9.339 0.8280 0.9661
2 0.9408 0.001305 0.0001433 0.02011 9.602 0.8958 0.9859
```

The confidence interval level (default: 95%) can be changed with the conf.int argument to summary.

```
MRMC statistical analysis summary function.

Usage
summary(object, conf.level = 0.95, ...)

Arguments
x: result of a call to mrmc.
conf.int: confidence level for confidence intervals.
```

3. CONCLUSIONS

MRMCAov brings a new statistical package for MRMC analysis to the R software environment and its large community of users. The package enables comparison of imaging modalities with an interactive interface and flexible options for performance metrics, study designs, covariance estimation methods, and statistical estimation and testing of performance. A demonstration of features currently implemented in the package is provided in this paper. New features, such as partial ROC AUC performance metrics and methodological extensions, are under development and planned for inclusion in future versions. Proper statistical methods and readily available software are crucial for the evaluation and comparison of multi-reader multi-case studies of imaging modalities. The MRMCAov software is designed to help ensure the application of such methods.

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health grant R01 EB025174.

REFERENCES

- [1]. Dorfman DD, Berbaum KS, and Metz CE, "Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method.," *Investigative Radiology* 27, 723–731 (1992). [PubMed: 1399456]
- [2]. Obuchowski NA and Rockette HE, "Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations," *Communications in Statistics–Simulation and Computation* 24, 285–308 (1995).
- [3]. Hillis SL, Obuchowski NA, Schartz KM, and Berbaum KS, "A comparison of the Dorfman-Berbaum-Metz and Obuchowski-Rockette methods for receiver operating characteristic (ROC) data," *Statistics in Medicine* 24, 1579–1607 (2005). [PubMed: 15685718]
- [4]. Hillis SL, "A comparison of denominator degrees of freedom methods for multiple observer ROC analysis," *Statistics in Medicine* 26, 596–619 (2007). [PubMed: 16538699]
- [5]. Hillis SL, Berbaum KS, and Metz CE, "Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis," *Academic Radiology* 15, 647–661 (2008). [PubMed: 18423323]
- [6]. Hillis SL, "Relationship between Roe and Metz simulation model for multireader diagnostic data and Obuchowski-Rockette model parameters," *Statistics in Medicine* 37, 2067–2093 (2018). [PubMed: 29609206]
- [7]. Smith BJ and Hillis SL, *MRMCAov: Multi-Reader Multi-Case Analysis of Variance* (2020). R package version 0.1.3.
- [8]. Schartz KM, Hillis SL, Pesce LL, Berbaum KS, and Metz CE, *OR-DBM MRMC* (2019). version 2.52.
- [9]. R Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2019).
- [10]. VanDyke CW, White RD, Obuchowski NA, Geisinger MA, Lorig RJ, and Meziane MA, "Cine MRI in the diagnosis of thoracic aortic dissection," 79th Radiological Society of North America Meetings (1993).
- [11]. Pepe MS, [The Statistical Evaluation of Medical Tests for Classification and Prediction], Oxford University Press, New York (2003).
- [12]. DeLong ER, DeLong DM, and Clarke-Pearson DL, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," *Biometrics* 44, 837–845 (1988). [PubMed: 3203132]
- [13]. Efron B, [The Jackknife, the bootstrap and other resampling plans], SIAM, Philadelphia (1982).

Figure 1.
R Example: Plots of ROC estimates from mrmc function call.