Evaluating MRMC Studies with Binary Endpoints

Changhong Song* Dandan Xu † Xiaoqin Xiong ‡

Abstract

Multi-Reader Multi-Case (MRMC) study designs have been used to evaluate the diagnostic performance of medical devices or procedures in different areas such as radiological imaging and digital pathology. The primary statistical evaluations for MRMC studies have commonly been based on receiver operating characteristic (ROC) curve with observer performance evaluation paradigm using area under the curve (AUC). However, for many diagnostic clinical studies, readers only provide binary responses and therefore the primary performance metrics are sensitivity and specificity, instead of AUC of ROC. Most literature for MRMC studies focus on the AUC analysis. There are few systematic evaluations and comparisons of the various methods for MRMC studies with binary endpoints. In this study, we summarize and evaluate how some commonly used statistical methods for MRMC studies can be used to evaluate studies with binary study outputs.

1. Introduction

Multi-Reader Multi-Case (MRMC) study designs are commonly utilized to evaluate the clinical performance in a diagnostic evaluation study when reader interfaces with the device. It has been used for evaluating the diagnostic accuracy for devices in different areas such as radiological imaging, digital pathology, etc. The MRMC study design generally involves multiple readers who can represent intended reader population, multiple subjects (cases) from the target population, and multiple reading conditions or modalities. For example, for an MRMC study that evaluates whether a device can improve reader interpretation of the imaging results, a study design with multiple reading modalities may include one arm where readers interpret imaging results without the device and another arm where readers will interpret imaging results aided by the device. Fully-crossed study designs have been commonly used in MRMC studies, where all readers independently read all cases with all modalities. A fully crossed design has the greatest statistical power given a fixed number of cases and readers. Split plot designs have also been used in MRMC studies (Obuchowski et al. 2012), where readers read their own group of cases. The study endpoints for MRMC studies will vary depending on the clinical study design, device intended use, and clinical practice. For example, the endpoint may be an ordinal score that represent the reader's confidence in how likely a lesion in an image is malignant. The study endpoint can also be binary representing whether a suspected target condition is present or not.

The correlation structures in MRMC studies are complex, thus statistical analysis needs to account for the correlation structures. The correlation can arise due to different modalities read by same readers, the cases read by same reader under different modalities, the cases under same modality read by different readers. Many statistical methods have been proposed to address the correlations in MRMC studies (e.g. Dorfman et al. 1992, Obuchowski and Rockette 1995, Beiden et al. 2000, Gallas 2006), and most of them focus on the evaluation of receiver operating characteristic curve (ROC) and area under the curve

*Food and Drug Administration, Center for Devices and Radiological Health,10903 New Hampshire Avenue, Silver Spring, MD 20993 Email: Changhong.Song@fda.hhs.gov.

†Food and Drug Administration, Center for Devices and Radiological Health,10903 New Hampshire Avenue, Silver Spring, MD 20993 Email: Dandan.Xu@fda.hhs.gov

‡Food and Drug Administration, Center for Devices and Radiological Health,10903 New Hampshire Avenue, Silver Spring, MD 20993 Email: Xiaoqin.Xiong@fda.hhs.gov

(AUC). The statistical evaluation of the AUC generally requires a reader to report an ordinal score such as confidence score that a target condition is present. However, for diagnostic clinical studies in some areas such as digital pathology, an ordinal score may not be available and the reader may only report whether a target condition is present or not. Thus, the primary statistical analysis for these studies is based on analyses common to binary output/result such as sensitivity/specificity or positive/negative percent agreement (PPA/NPA) (Zhou et al., 2011). The methods proposed for the AUC analysis can also be extended to the binary data analysis. However, there are few systematic summary, evaluation, and comparison for the various methods that can evaluate MRMC studies with binary endpoints. For this paper, we summarize and evaluate some commonly used statistical methods for evaluating MRMC studies with binary study endpoints.

2. Statistical Analysis Method

2.1 Notations

Denote Y_{ijk} as the result for modality i, reader j, and case k with 1 being positive and 0 being negative. Denote μ as the population mean, α_i as the fixed effect of modality i with mean 0 and variance σ^2_α, R_j as the random effect for the jth reader with mean 0 and variance σ^2_R, C_k as the random effect for the kth case with mean 0 and variance σ^2_C. Denote αR_{ij}, αC_{ik}, RC_{jk}, αRC_{ijk} as the random interaction effect with mean 0 and variance $\sigma^2_{\alpha R}$, $\sigma^2_{\alpha C}$, σ^2_{RC}, and $\sigma^2_{\alpha RC}$ respectively. Denote Z_{ijk} as the random error with mean 0 and variance σ^2_Z for modality i, reader j, and case k. The variable D represents clinical truth with two possible values where 1 to indicate presence of disease and 0 to indicate absence of disease.

2.2 Empirical Evaluation of Diagnostic Accuracy/Agreement

2.2.1 Sensitivity and Specificity

Sensitivity is defined as the probability that the test result is positive given the target condition of interest is present, i.e., $P(Y = 1|D = 1)$. Specificity is defined as the probability that the test result is negative given the target condition of interest is not present, i.e., $P(Y = 0|D = 0)$. For an MRMC study, Table 1 summarizes the results based on the number of true positive $n_{11,ij}$, false positive $n_{01,ij}$, false negative $n_{10,ij}$, and true negative $n_{00,ij}$. For modality i and reader j, the empirical estimates of sensitivity and specificity are

$$\text{Sensitivity}_{ij} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} = \frac{n_{11,ij}}{n_{11,ij} + n_{10,ij}}$$

and

$$\text{Specificity}_{ij} = \frac{\text{True Negative}}{\text{True Negative} + \text{False Positive}} = \frac{n_{00,ij}}{n_{01,ij} + n_{00,ij}}.$$

When we assign equal weight for all readers, the average reader sensitivity and specificity for all readers for modality i can be calculated as

$$\text{Sensitivity}_i = \frac{1}{J}\sum_{j=1}^{J} \text{Sensitivity}_{ij}$$

and

$$\text{Specificity}_i = \frac{1}{J}\sum_{j=1}^{J} \text{Specificity}_{ij}.$$

Table 1: Test Versus Clinical Reference Standard for Modality i and Reader j

		Clinical Reference Standard		
		Positive	Negative	Total
Test Results	Positive	$n_{11,ij}$	$n_{01,ij}$	$n_{.1,ij}$
	Negative	$n_{10,ij}$	$n_{00,ij}$	$n_{.0,ij}$
	Total	$n_{1.,ij}$	$n_{0.,ij}$	$n_{..,ij}$

Table 2: Test Versus Comparator Method for Modality i and Reader j

		Comparator Method		
		Positive	Negative	Total
Test Results	Positive	$m_{11,ij}$	$m_{01,ij}$	$m_{.1,ij}$
	Negative	$m_{10,ij}$	$m_{00,ij}$	$m_{.0,ij}$
	Total	$m_{1.,ij}$	$m_{0.,ij}$	$m_{..,ij}$

2.2.2 PPA/NPA

In lieu of sensitivity and specificity, performance metrics are reported as positive percent agreement (PPA) and negative percent agreement (NPA) when the comparator method is not a clinical reference standard. PPA is the probability that test results is positive given that the comparator result is positive. NPA is the probability that the test results is negative given that the comparator result is negative. For an MRMC study, Table 2 summarizes the agreement of results for a test under evaluation and a comparator test. For modality i and reader j, the empirical way to calculate PPA and NPA is

$$\text{PPA}_{ij} = \text{Prob}(\text{Test} = +|\text{Comparator} = +) = \frac{m_{11,ij}}{m_{11,ij} + m_{10,ij}}$$

and

$$\text{NPA}_{ij} = \text{Prob}(\text{Test} = -|\text{Comparator} = -) = \frac{m_{00,ij}}{m_{01,ij} + m_{00,ij}}.$$

When we assign equal weight for all readers, the average PPA and NPA for all readers for modality i can be calculated as

$$\text{PPA}_i = \frac{1}{J}\sum_{j=1}^{J} \text{PPA}_{ij}$$

and

$$\text{NPA}_i = \frac{1}{J}\sum_{j=1}^{J} \text{NPA}_{ij}.$$

2.3 Statistical Methods to Account for Correlations

The empirical method in section 2.2.1 and 2.2.2 can provide point estimates for estimating sensitivity/specificity and PPA/NPA. While point estimates are straightforward, the standard errors needs to take the correlation into account in MRMC studies. In this paper, we

Table 3: Some Commonly Used Methods for Analyzing MRMC Studies with Binary Endpoints

Category	Methods	Methods to Account for Correlations	Strength and Limitation
Sampling Based	Bootstrap Method	preserving correlation through resampling	no assumption needed on model, but the method may fail if the number of cases or readers in the MRMC study are not adequate.
Non-Parametric	U Statistics	decomposing the variance of average accuracy measure	no assumption needed on model, may have larger variability on the estimates comparing to parametric methods.
Model Based	DBM Method	random effect linear model	Model is simple to run, but sensitivity and specificity estimates may be beyond 0 to 1 and there are model assumptions.
	OR Method	random effect linear model and correlated error term	Similar to DBM Method
	ORH Method	same as OR Method	OR Method with a revised degree of freedom
	GLMM Method	random effect generalized linear model	Can avoid an estimate of the sensitivity and specificity beyond 0 to 1, but it is not straightforward to integrate out random effects for sensitivity and specificity calculation. Model convergence may be an issue.

review 6 commonly used methods that can evaluate MRMC studies with binary endpoints and address the correlations. Table 3 summarizes the 6 methods and group them into different categories: re-sampling based approach, non-parametric approach, and model based approach. We will introduce each method briefly for applying them for binary data analysis in MRMC studies.

2.3.1 Three Way Bootstrap Method

Bootstrap method (Efron 1982) has been commonly used to address correlations in MRMC studies (e.g. Dorfman et al., 1995; Beiden et al., 2000; Kupinski et al. 2006; Bandos et al. 2007)). For MRMC studies, the bootstrap re-sampling should account for the correlations due to both cases and readers. The commonly used three way bootstrap method (Gallas et al. 2009) generally independently re-sample with replacement the same number of readers, normal cases, and diseased cases as observed in the clinical study. The variances and confidence intervals of the accuracy measure (e.g., sensitivity, specificity, PPA, NPA) is calculated based on the re-sampling distributions.

Bootstrap method has its limitations and can only be used when the underlying assumptions are satisfied. For example, one assumption for the bootstrap method is that the bootstrap re-samples should be able to represent the true underlying distribution of the population of interest. If the number of readers or cases in the MRMC studies are not sufficient,

the bootstrap samples may not be able to represent the true underlying distributions. For these cases, the statistical analysis based on bootstrap samples may under or over estimate the variations of relevant parameters. In addition, if the observed accuracy performance such as sensitivity is at or close to 0% or 100%, there may be no variation from the bootstrap re-samples and the statistical analysis for the variance or confidence intervals can also be incorrect.

2.3.2 U statistics

Gallas (2006) developed One-Shot Estimate of MRMC Variance for AUC from U statistic. Gallas, Pennello, Myers (2007) extended it to binary data analysis. This approach estimates the variance of the accuracy measure averaged over readers and cases by decomposing the variance into a system of equations and using the non-parametric variance derived in the literature on U statistics.

For a fully crossed study design, the percent correct rate PC_i (e.g. sensitivity, specificity, PPA, NPA) for modality i is calculated the same as the empirical method described in section 2.2 . The variance of the PC_i is calculated as

$$V(PC_i) = c_1 M_1 + c_4 M_4 + c_5 M_5 + c_8 M_8,$$

where coefficients c_1, c_4, c_5, c_8 and the moments M1, M4, M5, and M8 are notations derived in Gallas 2006 and 2007 paper.

2.3.3 DBM Method

Dorfman-Berbaum-Metz (DBM) method (Dorfman et al. 1992) has been commonly used to evaluate receiver operating characteristic curve (ROC) and AUC for MRMC studies. It can also be used to evaluate binary data evaluation. Let γ_{ijk} be the pseudovalue of accuracy measures (e.g. sensitivity, specificity) based on jackknife re-sampling for modality i, reader j, and case k, the statistical model for DBM method is

$$\gamma_{ijk} = \mu + \alpha_i + R_j + C_k + \alpha R_{ij} + \alpha C_{ik} + RC_{jk} + \alpha RC_{ij} + e_{ijk}.$$

Different from AUC analysis, which uses both positive and negative cases, sensitivity evaluation is based on all the positive cases by reference standard for analysis. Specificity evaluation is based on all the negative cases by reference standard for analysis. The average reader sensitivity/specificity and their confidence intervals can be derived based on the inference for $\mu + \alpha_i$ for modality i. Because we are modeling the sensitivity and specificity directly, a limitation is that it is possible for the confidence intervals of the sensitivity and specificity to be below 0 or above 1.

2.3.4 OR Method

Obuchowski Rockette (OR) method (Obuchowski and Rockette 1995) is another commonly used method to evaluate receiver operating characteristic curve (ROC) and AUC, which can be extended to evaluate binary endpoint in MRMC studies. Let PC_{ij} be the estimated reader-specific accuracy measure (i.e., sensitivity, specificity) for modality i and reader j. It can be modeled by a two-way, mixed effects ANOVA model:

$$PC_{ij} \quad = \mu + \alpha_i + R_j + \alpha R_{ij} + \epsilon_{ij}.$$

The random error term ϵ_{ij} is assumed to be normally distributed with mean zero and variance σ_ϵ^2, where σ_ϵ^2 represents both the variability due to the case sample and the within-reader variation. The error term ϵ_{ij} are not independent. Obuchowski and Rockette assumed equi-covariance between readers and modalities and therefore there are three possible covariances: $cov(\epsilon_{ij}, \epsilon_{ij\prime})$, $cov(\epsilon_{ij}, \epsilon_{i\prime j})$, and $cov(\epsilon_{ij}, \epsilon_{i\prime j\prime})$. Similar to DBM method, sensitivity is calculated based on all the positive cases by reference standard. Specificity is based on all the negative cases by reference standard. The average reader sensitivity/specificity and their confidence intervals can be derived based on the inference for $\mu+\alpha_i$ for modality i.

2.3.5 ORH Method

Although the statistical models for DBM method and OR method appear quite different, Hillis (2005) showed that F Statistics for DBM and OR methods have the same form and will typically have similar values but there are differences in the denominator degree of freedom for the two methods. Hillis (2007) proposed a new degree of freedom estimate that can be used for both DBM and OR methods. Chen et al. (2014) adapted the Obuchowski Rockette Hillis (ORH) method for the simulation, analysis, validation, and sizing of MRMC studies with binary agreement data.

2.3.6 Generalized Linear Mixed Effects Model

Generalized linear mixed models have been frequently considered to evaluate MRMC studies with binary study endpoints. For example, with logit link function, we can model the probability of the correct results PC_{ijk} for results Y_{ijk} as a function of fixed modality effect, random readers, random cases, and random interactions between/among modalities, readers, and cases.

$$\text{Logit}(PC_{ijk}) = \mu + \alpha_i + R_j + C_k + \alpha R_{ij} + \alpha C_{ik} + RC_{jk} + \alpha RC_{ijk}.$$

The population averaged percent correct rate for modality i can be calculated by integrating out the random subject and reader effect. However, the calculation of the population averaged sensitivity can be quite complex. Some investigators simply calculate the percent correct rate for modality i as

$$\text{PC}_i = \frac{exp(\mu + \alpha_i)}{1 + exp(\mu + \alpha_i)}.$$

However, the estimated percent correct rate using this approach can be biased greatly compared to the population averaged estimates. Further research is needed for the evaluation of MRMC studies using the generalized linear mixed effects model.

3. Summary

Many MRMC clinical studies have binary endpoints and the primary statistical analysis will be based on sensitivity/specificity or PPA/NPA. It is important to address the correlations in these MRMC studies. There are multiple statistical methods available to address the correlations in the MRMC study with binary endpoints. However, most literature for MRMC studies are about the AUC analysis. There are few systematic evaluations/comparisons of the various methods for MRMC studies with binary endpoints. Further research is needed about the strength and weakness of the various methods that evaluate MRMC studies with binary endpoint.

4. Disclaimer

This study reflects the views of the authors and should not be construed to represent FDA's views or policies. The authors gratefully thank Dr. Bipasa Biswas for the help and support.

REFERENCES

Bandos, A. I., Rockette, H. E., Gur, D. . Exact bootstrap variances of the area under the ROC curve. Commun. Statist. A Theor. 2007. 36(13):2443 2461.

Beiden, S. V., Wagner, R. F., Campbell, G. Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects, receiver operating characteristic analysis. Acad. Radiol. 2000. 7(5):341 349.

Chen, W., Wunderlich, A., Petrick, N,, Gallas, B,D. Multireader multicase reader studies with binary agreement data: simulation, analysis, validation, and sizing. J Med Imaging (Bellingham). 2014;1(3):031011. doi:10.1117/1.JMI.1.3.031011.

Dorfman, D.D., Berbaum, K.S., Metz, C.E. Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. Investigative Radiology. 1992; 27:723-731.

Dorfman, D. D., Berbaum, K. S., Lenth, R. V. Multireader, multicase receiver operating characteristic methodology: a bootstrap analysis. Acad. Radiol. 1995; 2(7): 626 633.

Efron, B. The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia, PA: Society for Industrial and Applied Mathematics. 1982.

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

Gallas, B.D., Pennello, G.A., and Myers, K.L. Multireader multicase variance analysis for binary data. J Opt Soc Am A Opt Image Sci Vis. 2007; 24(12): B70-80.

Gallas, B.D., Bandos, A., Samuelson, F.W., Wagner, R.F. A framework for random-effects ROC analysis: Biases with the bootstrap and other variance estimators. Commun. Stat. Theory Methods. 2009; 38: 2586 2603

Hillis, S.L. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Statistics in Medicine. 2007; 26:596-619.

Hillis, S.L., Obuchowski, N.A., Schartz, K.M., Berbaum, K.S.. 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.

Kupinski, M. A., Clarkson, E., Barrett, H. H. A probabilistic model for the MRMC method. Part 2. Validation and applications. Acad. Radiol. 2006. 13(11):1422 1430.

Obuchowski, N., Gallas, B.D., Hillis, S.L. Multi-reader ROC studies with split-plot designs: A comparison of statistical methods. Acad Radiol. 2012; 19: 1508 1517.

Obuchowski, N.A., Rockette, H.E. Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: An ANOVA approach with dependent observations. Communications in Statistics, Part B: Simulation and Computation. 1995; 24:285-308.

Zhou, X.-H., Obuchowski, N. A., and McClish, D. K., 2011. Statistical Methods in Diagnostic Medicine. 2nd ed. s.l.:Wiley.