Published in final edited form as:
Acad Radiol. 2012 December ; 19(12): 1508–1517. doi:10.1016/j.acra.2012.09.012.

Multi-Reader ROC studies with Split-Plot Designs: A Comparison of Statistical Methods

Nancy A. Obuchowski, PhD[*],
Department of Quantitative Health Sciences, JJN-3, Cleveland Clinic Foundation

Brandon D. Gallas, PhD, and
Division of Imaging and Applied Mathematics, Center for Devices and Radiological Health, US Food and Drug Administration

Stephen L. Hillis, PhD
Departments of Radiology and Biostatistics, The University of Iowa, Comprehensive Access and Delivery Research and Evaluation (CADRE) Center, Iowa City VA Health Care System

Abstract

Rationale and Objectives—Multi-reader imaging trials often use a factorial design, where study patients undergo testing with all imaging modalities and readers interpret the results of all tests for all patients. A drawback of the design is the large number of interpretations required of each reader. Split-plot designs have been proposed as an alternative, in which one or a subset of readers interprets all images of a sample of patients, while other readers interpret the images of other samples of patients. In this paper we compare three methods of analysis for the split-plot design.

Materials and Methods—Three statistical methods are presented: Obuchowski-Rockette method modified for the split-plot design, a newly proposed marginal-mean ANOVA approach, and an extension of the three-sample U-statistic method. A simulation study using the Roe-Metz model was performed to compare the type I error rate, power and confidence interval coverage of the three test statistics.

Results—The type I error rates for all three methods are close to the nominal level but tend to be slightly conservative. The statistical power is nearly identical for the three methods. The coverage of 95% CIs fall close to the nominal coverage for small and large sample sizes.

Conclusions—The split-plot MRMC study design can be statistically efficient compared with the factorial design, reducing the number of interpretations required per reader. Three methods of analysis, shown to have nominal type I error rate, similar power, and nominal CI coverage, are available for this study design.

Keywords

Multi-reader imaging study; MRMC study; ROC analysis; split-plot design

[*]Department of Quantitative Health Sciences, JJN-3, Cleveland Clinic Foundation, 9500 Euclid Ave, Cleveland, Ohio 44195, (216) 445-9549, obuchon@ccf.org.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Introduction

In imaging clinical trials investigators often compare the accuracy of clinicians' diagnostic interpretations of different imaging modalities, assessing the sensitivity, specificity, and/or receiver operating characteristic indices of the modalities [1–4]. In estimating the accuracy of mammography for detecting breast cancer, for example, mammograms are interpreted by trained radiologists who read the images to determine if suspicious lesions are present. It is well-known that there is variability between readers in their visual, cognitive, and perceptual abilities [4–7]; similarly, there is variability between patients in their anatomy, co-morbidities, and manifestation of disease. Thus, samples of both readers and patients are integral components in characterizing diagnostic test accuracy. The average accuracy of the readers is typically used as the measure of the test's accuracy. There has been a great deal of methodology development for the estimation and comparison of diagnostic tests' accuracy from multiple-reader studies [4,8–16].

Multi-reader imaging trials often use a factorial, or fully-crossed, design, where study patients undergo testing with all imaging modalities being compared, and study readers interpret the results of all of the tests for all patients. The rationale is that since both patients and readers introduce variability to the measurement of diagnostic accuracy, for comparing modalities variability from these sources can be reduced if study patients undergo all modalities and if study readers interpret all of the test results.

While the fully-crossed design is efficient in terms of the number of patients and readers required for the study, one drawback of the design is the number of interpretations required of each reader [17]. For a typical-sized study with 200 study patients and 2 modalities, each reader must interpret 400 images. If each test requires an average of 5 minutes to interpret, each study reader needs nearly a week to participate in the study. Some tests, such as CT colonoscopy, can take closer to 30 minutes to interpret (5 weeks of reading time).

When the number of interpretations per reader is a limiting factor in the execution of a trial, other study designs, such as the "hybrid" [17] or "mixed" [18] design, have been proposed, which are two different split-plot designs. In these designs one or a subset of readers interprets all of the images of a sample of patients, while other readers interpret all of the test results of other samples of patients. Compared to the factorial design, these designs can be more efficient for testing for a difference in modalities, with respect to the total number of reader interpretations, because they retain the reader and patient pairing across modalities while eliminating some of the positive correlations between readers. By reducing some of the between-reader correlations, these alternative study designs can reduce the overall numbers of interpretations needed per reader.

There are multiple statistical methods available for analyzing data from a fully-crossed multiple-reader multiple-case (MRMC) design [4, 8–16] but only a few papers discussing analysis of MRMC studies using other study designs [19–21]. In this paper we present three methods of analysis for the MRMC split-plot design: the OR method [9] modified for the split-plot design, a newly proposed marginal-mean ANOVA approach [21], and an extension of the three-sample U-statistic method [15, 20]. We perform a simulation study to compare the type I error rate, power and confidence interval coverage of three test statistics. The motivation for this study was an imaging trial assessing the accuracy of 36 readers who interpreted the mammograms of 200 patients without and with a computer aided detection (CAD) device. A four-block split-plot design was used such that each reader interpreted images from 50 patients. We present the trial in detail and analyze the results with the three test statistics. A discussion follows.

Methods

MRMC Split-Plot Study Design

Multi-reader multi-case (MRMC) trials are conducted to compare the diagnostic accuracy of two or more tests where the tests being studied require interpretation of cases by a trained reader. The primary goal of these studies is the comparison of the average accuracy of readers between the diagnostic tests. Common measures of test accuracy are sensitivity, specificity, and measures of accuracy from Receiver Operating Characteristic (ROC) curves [1–4]. In this paper we focus on the area under the ROC curve, AUC, but the methods are applicable to these other measures of test accuracy, as well.

Several MRMC study designs have been proposed [17], but the most common design to date is the fully-crossed design. In this design there are N_T total patients who have undergone each diagnostic test. J readers interpret all of the test results of the N_T patients (See Table 1). For a study comparing two diagnostic tests, there are $2 \times J \times N_T$ reader interpretations.

The split-plot MRMC study design [18] was proposed to reduce the number of interpretations required from each reader. In this design readers interpret all of the test results from a patient, but each reader interprets just a subset of the total study patients. Thus, the pairing across modalities is present (each patient imaged with all modalities and each reader interpreting results from all modalities), but some of the positive correlations between readers that would be present in a fully-crossed design are eliminated, typically improving the statistical efficiency relative to the fully-crossed design with respect to total number of reader interpretations.

There are many possible configurations of the split-plot design, but in this paper we focus on balanced designs where the J readers and N_T patients are randomized to one of G blocks. In each block, the readers interpret all of the test results from the patients in that block. Table 2 illustrates the split-plot study design with G=3 blocks and two readers in each block (i.e. J=6). (Note that it can be shown that the design in Table 2 is a split-plot design with the reader and case combinations as the whole plots, test as the split-plot factor, and block as the between-whole-plots factor.)

To illustrate the efficiency gains possible with the split-plot design, we compared the efficiency of the 2-block split-plot design (J=6, N_T=120) to five alternatives: a 3-block split-plot design, a 4-block split-plot design, two fully-crossed designs (compared with the 2-block design, the Full-A study design has the same total number of readers, J, but half the total number of cases, $N_T/2$; and the Full-B study design has the same total number of readers and cases, J and N_T, respectively), and an unpaired reader study design (i.e. the readers are unpaired across modalities). In Table 3 we summarize the resource needs of the six study designs. From the table we can consider our ability to recruit the total number of readers, the cost of collecting the cases, the total number of interpretations required, and the total number of interpretations required per reader. The last two variables are likely to be proportional with total study time and time required per reader. The table also includes the statistical efficiency of each of the study designs relative to the 2-block study.

We see that the 2-block split-plot study takes only a moderate hit in efficiency relative to the fully-crossed study design with the same number of readers and cases (Full-B). This moderate loss in efficiency comes at a savings of half the reading time of each reader. We also see that the 2-block split-plot study is more efficient than the fully-crossed study design with the same number of readings per reader and total readings (Full-A). Splitting the study into even more blocks (each with three readers) saves even more time per reader and the additional readers increase the efficiency. It is important to realize, however, that these results are only for one particular data structure; for other structures, gains can be less or greater with respect to total readings.

OR Test Statistic Modified for Split-Plot Design

Obuchowski and Rockette (OR) [9] developed a general linear model of the estimate of the ROC area for the i-th modality by reader j for a fully-crossed (i.e. factorial) design:

$$\widehat{\theta}_{ij}=\tau_i+\mathrm{R}_j+(\tau \mathrm{R})_{ij}+\varepsilon_{ij}, \quad [1]$$

where τ_i is the fixed effect of the *i*th modality, R_j is the random reader effect, and $(\tau R)_{ij}$ is the random effect due to the interaction of modality and reader. The error term in Equation 1 is assumed to have a multivariate normal distribution with mean zero and covariance matrix defined as follows:

$$E(\varepsilon_{\mathrm{ij}},\varepsilon_{\mathrm{i}'\mathrm{j}'})=\begin{array}{ll} {\sigma_{\mathrm{c}}}^2 & \text{if } \mathrm{i}=\mathrm{i}' \text{ and } \mathrm{j}=\mathrm{j}' \\ \mathrm{cov}_1={\sigma_{\mathrm{c}}}^2\rho_1 & \text{if } \mathrm{i}\neq \mathrm{i}' \text{ and } \mathrm{j}=\mathrm{j}' \\ \mathrm{cov}_2={\sigma_{\mathrm{c}}}^2\rho_2 & \text{if } \mathrm{i}=\mathrm{i}' \text{ and } \mathrm{j}\neq \mathrm{j}' \\ \mathrm{cov}_3={\sigma_{\mathrm{c}}}^2\rho_3 & \text{if } \mathrm{i}\neq \mathrm{i}' \text{ and } \mathrm{j}\neq \mathrm{j}'. \end{array}$$

ρ_{1} denotes the correlation betweens errors corresponding to a reader reading the results of the same patients from different tests, ρ_2 denotes the correlation between different readers interpreting the same test, and ρ_3 denotes the correlation between different readers interpreting different tests.

The null and alternative hypotheses are

$$\begin{array}{c} \mathrm{H}_{\mathrm{o}}:\tau_1=\tau_2, \\ \mathrm{H}_{\mathrm{A}}:\tau_1\neq\tau_2. \end{array} \quad [2]$$

Obuchowski and Rockette [9] proposed the following test statistic, which approximately follows a central F-distribution under the null hypothesis for the factorial design:

$$\mathrm{F}^*=\mathrm{MS(T)}/\{\mathrm{MS(T}\times \mathrm{R})+max[\mathrm{J}\times\widehat{\varphi},0]\}. \quad [3]$$

MS is shorthand for mean square. Details of the calculation of the MS terms are given in the Appendix.

The last term of Equation 3 can be written as max[$J\times(\hat{\mathrm{cov}}_2-\hat{\mathrm{cov}}_3)$,0], where $\hat{\mathrm{cov}}_2$ and $\hat{\mathrm{cov}}_3$ are estimates of cov_2 and cov_3. This term is a correction factor proposed by Bhat [22] for the situation where the data are correlated. The estimates $\hat{\mathrm{cov}}_2$ and $\hat{\mathrm{cov}}_3$ are typically computed by averaging corresponding pairwise covariance estimates, which can be estimated by various methods such as the nonparametric method of DeLong et al [23], the jackknife, the bootstrap, or parametric methods.

For the factorial design Obuchowski and Rockette proposed that F* be compared to a central F distribution with (I−1) and (I−1)(J−1) degrees of freedom (dfs). (Note that dfs are values associated with the test statistic and used in hypothesis testing.) Hillis [14] showed that the denominator df for F* should be:

$$\text{ddf}=\{\text{MS}(\text{T}\times\text{R})+\mathit{max}[\text{J}\times(\widehat{\text{cov}}_2-\widehat{\text{cov}}_3),0]\}^2/\{\text{MS}(\text{T}\times\text{R})^2/((\text{I}-1)(\text{J}-1))\}. \quad [4]$$

An MRMC study using data collected in a split-plot design can be analyzed using the fully-crossed study design formulae with computation of the error covariances modified to account for zero covariances between AUCs from different blocks. Thus the factorial notation, factorial model mean square definitions, test statistic, and degrees of freedom for the OR method for the split-plot design are identical in appearance to those of the fully-crossed design. In the split-plot design, however, there are some between-reader covariances that are non-zero because the readers are in the same block and other between-reader covariances that are zero because the readers are in different blocks. For the split-plot design, $\hat{\varphi}$ is computed just like $\hat{\text{cov}}_2 - \hat{\text{cov}}_3$ for the factorial design, but with between-block pairwise covariance estimates set to zero.

Marginal-mean ANOVA Test Statistic for Split-plot Design

The rationale for the marginal-mean ANOVA approach is summarized in the Appendix. Briefly, the method use a split-plot notation; i.e., Y_{ijg} denotes the AUC corresponding to the *j*th reader in block *g* reading the cases in block *g* under modality *i*. (In terms of the model in Equation 1 we have $Y_{111} = \hat{\theta}_{11}, Y_{121} = \hat{\theta}_{12}, Y_{112} = \hat{\theta}_{13}$, etc., if there are 2 readers in each block.) Let $r = J/G$ denote the number of readers in each block. Cov$_2$ and cov$_3$ are again defined as for the factorial model, but with the restriction that these covariances are only for pairs of outcomes from the same block. In work presently under review, Hillis [21] has proposed the following test statistic for testing the hypotheses in Equation [2] with the split-plot MRMC design:

$$F=\frac{\text{MS}(T)}{\text{MS}(T\times R(G))+\max\left[r(\widehat{\text{cov}}_2-\widehat{\text{cov}}_3),0\right]} \quad [5]$$

where

$$\text{MS}(\text{T}\times\text{R}(\text{G}))=\sum_{g=1}^{G}\sum_{i=1}^{I}\sum_{j=1}^{r}(\text{Y}_{\text{ijg}}-\text{Y}_{\text{i.g}}-\text{Y}_{\cdot\text{jg}}+\text{Y}_{\cdot\cdot\text{g}})^2/[\text{G}(\text{I}-1)(\text{r}-1)], \quad [6]$$

and $\hat{\text{cov}}_2$ and $\hat{\text{cov}}_3$ are computed as the averages of the corresponding estimated covariances within reader blocks; these covariances can be estimated using the same methods discussed previously for the OR statistic.

It is easy to show that the treatment mean square MS(T) in Equation 5 is equal to the treatment mean square in Equation A1. In contrast, MS(T*R(G)) in Equation 6 is the treatment-by-reader-nested-within-block mean square, which differs from the treatment-by-reader mean square in Equation A1 because each squared term in MS(T×R(G)) is a function only of AUCs from a particular block, whereas a squared term in MS(T×R) can be a function of AUCs from several blocks. Finally, it is straightforward to show that $(\hat{\text{cov}}_2 - \hat{\text{cov}}_3) = (J-1)\hat{\varphi}/(r-1)$, when the same covariance estimation method is used to compute $\hat{\text{cov}}_2 - \hat{\text{cov}}_3$ and $\hat{\varphi}$. Thus Equation 5 can be written as

$$F=\frac{\text{MS}(T)}{\text{MS}(T*R(G))+\max\left[\frac{J-1}{J-G}J\widehat{\varphi},0\right]}$$

The above expression shows the close relationship of F and F*: the numerators are the same, the first term in the denominators differ according to the treatment of blocks (nested or not), and the second term in the denominators differ by the factor (J−1)/(J−G).

The denominator degrees of freedom, derived in a manner similar to the degrees of freedom proposed by Hillis [14] for the factorial OR model, is

$$\mathrm{ddf}=\frac{\{\mathrm{MS}(T*R(G))+\max[r(\widehat{\mathrm{cov}}_2-\widehat{\mathrm{cov}}_3),0]\}^2}{\frac{[\mathrm{MS}(T*R(G))]^2}{G(I-1)(r-1)}}\qquad[7]$$

F is compared to a central F distribution with degrees of freedom (I−1) and ddf.

Three-Sample U-Statistic Test for Split-plot Design

Gallas [15] derived the variance of the reader-averaged non-parametric (trapezoidal) AUC and showed that it was equivalent to a three-sample U-statistics result; the three samples correspond to readers, non-diseased cases, and diseased cases [24]. Given that work, the covariance between $\hat{\theta}_{i\cdot}$ and $\hat{\theta}_{i'\cdot}$, the non-parametric reader-averaged AUCs from modalities i and i′, can be written as

$$\mathrm{Cov}\left(\widehat{\theta}_{i\cdot},\widehat{\theta}_{i'\cdot}\right)=\left(\frac{\alpha_{1ii'}}{N_0}+\frac{\alpha_{2ii'}}{N_1}+\frac{\alpha_{3ii'}}{N_0N_1}\right)+\left(\frac{\alpha_{4ii'}}{J}\right)+\left(\frac{\alpha_{4ii'}}{N_0J}+\frac{\alpha_{5ii'}}{N_1J}+\frac{\alpha_{6ii'}}{N_0N_1J}\right),\qquad[8]$$

where J is the number of readers, N_0 is the number of non-diseased cases, N_1 is the number of diseased cases, and each $\alpha_{ii'}$ is a variance when i=i′ and a covariance when i≠i′ [25, 26]. Gallas and Brown [20] generalized the variance derivation and estimation in Equation 8 to treat study designs that are not fully crossed by using scaling factors for the α's. This generalization is exactly what is used for the split-plot study designs examined here. Further details of Gallas' prior work are summarized in the Appendix.

The test statistic for the split-plot design is

$$T=\frac{\widehat{\theta}_1-\widehat{\theta}_2}{\sqrt{\widehat{V}_{UA}}}\qquad[9]$$

where $\hat{V}_{U\Delta}$ is the generalization of the three-sample U-statistic estimate of the variance of the difference in the reader-averaged empirical AUCs as described by Gallas and Brown [20]. When the sample size is large, we can assume that this test statistic is normal. For smaller sample sizes, we assume that T follows a Student's t-distribution. Here we consider an estimate of the degrees of freedom motivated by the approximate degrees of freedom derived by Hillis [14] and a general 4-way ANOVA with one fixed factor (modality) and three random factors (readers, non-diseased cases, and diseased cases). Specifically, we estimate the degrees of freedom with

$$d.f.=\frac{\left(\widehat{V}_{U\Delta}\right)^2}{\left(s_{BR}^2\right)^2/(N_R-1)^3+\left(s_{B0}^2\right)^2/(N_0-1)^3+\left(s_{B1}^2\right)^2/(N_1-1)^3}\qquad[10]$$

where

$$\begin{aligned} s_{BR}^2 &= \widehat{\alpha}_{B4ii} + \widehat{\alpha}_{B4i'i'} - 2\widehat{\alpha}_{B4ii'}, \\ s_{B0}^2 &= \widehat{\alpha}_{B1ii} + \widehat{\alpha}_{B1i'i'} - 2\widehat{\alpha}_{B1ii'}, \\ s_{B1}^2 &= \widehat{\alpha}_{B2ii} + \widehat{\alpha}_{B2i'i'} - 2\widehat{\alpha}_{B2ii'} \end{aligned}$$

are ideal bootstrap (method of moments) estimates of the components of variance for the three random effects: readers, non-diseased cases, and diseased cases. We can also refer to these as the non-parametric ML (maximum likelihood) estimates, and relate them to the mean squares of a 4-way ANOVA.

Note that if readers, non-diseased cases, or diseased cases are not paired across the modalities, the corresponding covariance term ($\alpha_{4ii'}$, $\alpha_{1ii'}$, or $\alpha_{2ii'}$) is zero by definition.

Simulation Study

Roe and Metz [27] and Dorfman et al [28] described a method for simulating data for the MRMC fully-crossed study design. We have summarized their model in the Appendix. We utilized their general approach, modifying the model in Equation A2 for the split-plot study design under investigation here to reflect the nesting of reader and case within block. We examined designs where the study readers were divided evenly into 2 blocks. We considered reader sample sizes per block of 3, 5, or 7, and patient sample sizes per block of 60 (30 with disease and 30 without) and 120 (60 with disease and 60 without). We utilized an intermediate value for the mean ROC area under the null hypothesis around 0.90. We investigated scenarios where the readers' average ROC area with the two modalities was the same (null hypothesis), and scenarios where the readers' average ROC area with the two modalities increased by a small amount (0.030–0.032) (alternative hypothesis).

The values for the variance components were selected from values used by Roe and Metz [27] and Dorfman et al [28]. We generated test scores from an equal-variance binormal distribution (i.e., binormal parameter b = 1); the same variance components were used for diseased and non-diseased patients. There were a total of 36 different scenarios tested: 3 (reader set sizes) x 2 (case set sizes) x 3 (simulation configurations) x 2 (null & alternate hypothesis experiments). For each scenario, we simulated 2000 datasets, so that we would have at least 80% power to detect a type one error that differed by 0.015 or more from the nominal level of 0.05. The values of the variance components and fixed effects for the simulated data are summarized in Table 4.

The results of the simulation study (averaged over the 3 simulation configurations) are illustrated in Figures 1 and 2. The nonparametric estimate of the AUC was calculated for each reader and modality [23]. Both the OR and marginal-mean test statistics were based on covariances estimated using the DeLong et al [23] method.

The type I error rates (Figure 1) for all three methods are close to the nominal level but tend to be slightly conservative (i.e. run very slightly below the nominal level) even for larger numbers of patients and readers. The power is nearly identical for the three methods (results not shown). The coverage of 95% CIs (Figure 2) fall close to the nominal coverage for small and large sample sizes.

Example: Computer-Aided Detection of Breast Cancer

In this split-plot study 36 board-certified mammographers were randomized to the four blocks, such that there were 9 readers in each block. Similarly, 100 patients with breast cancer (biopsy-confirmed) and 100 patients without breast cancer (confirmed by biopsy or one year follow-up) were randomized to the four blocks, such that there were 25 cancer patients and 25 non-cancer patients in each block.

Each reader first interpreted a patient's mammogram without computer-aided detection (CAD) and reported his or her result. Specifically, the reader was asked to mark the location of any and all suspicious findings and assign a confidence score (1–100) to each finding, where 1=the lowest probability of malignancy, and 100=the highest probability of malignancy. The compilation of findings by the reader constitutes the reader's unaided findings and could not be altered by the reader.

Next, the reader was shown the CAD marks. The CAD system scans the image for abnormal features associated with malignancy and places a circle around each suspicious area. Readers were asked to consider each CAD mark, choosing either to dismiss the mark as a false hit or add the finding to their previous unaided findings. Readers were also allowed to increase or decrease the confidence scores of their previous findings. The compilation of findings by the reader after being shown the CAD marks constitutes the reader's aided findings. Thus, in this four-block design, each reader provided 100 interpretations (i.e. 50 images interpreted both with and without CAD). With 36 total readers, there were 3600 total interpretations.

If a patient had cancer, and the reader correctly located it, then the reader's confidence score for that lesion became the reader's score for the patient. If a patient had a cancer, and the reader did not locate it, then a confidence score of zero was assigned to the patient for that reader's interpretation. For a patient without cancer, the highest confidence score assigned by the reader to any false lesions was used as the reader's interpretation for that patient. If no false lesions were reported for a patient without cancer, then a confidence score of zero was assigned to that patient.

The nonparametric area under the ROC area was calculated for each reader without CAD and with CAD [23]. The method of DeLong et al. [23] was used to compute covariances for the OR and marginal-mean methods. Fourteen readers showed improvement in accuracy with CAD, 16 showed no change, and 6 showed reductions in accuracy with CAD. The mean ROC area over the 36 readers was 0.7735 without CAD and 0.7812 with CAD.

Table 5 summarizes the results of the three statistical methods. The SE of the difference is smallest for the three-sample U-statistic method and largest for the marginal mean ANOVA method. All three methods yield a non-significant result with 95% CIs containing zero. The lower and upper bounds on the CIs are quite similar, with the three-sample U-statistic method giving a slightly more narrow interval. Further results from the analysis are given in the Appendix.

Conclusions

For typical multi-reader imaging studies, readers and verified cases are of limited quantity; the fully-crossed design has been used in order to achieve maximum power with these limited resources. Split-plot designs, however, have a number of advantages over the traditional fully-crossed design. First, there is considerable savings in the number of interpretations required of each reader. This can be a useful recruiting tool for MRMC studies, especially when each interpretation is cumbered by lengthy CRFs, multi-levels of imaging to review, and/or complicated diagnoses. The results in Table 3 suggest that a study design with more splits of readers and cases can increase efficiency with respect to total number of readings, as long as more readers can be recruited. We note, however, that investigators should be warned not to split the cases into too small-sized blocks because there will be fewer empirical operating points sampling the ROC space. There are even fewer operating points if there are ties in the data, which is common for studies with human readers. In general we recommend at least 20 diseased and 20 non-diseased cases per block.

A second advantage of the split-plot design is that it can be used to efficiently study multiple imaging tests. For example, in the breast CAD study illustrated here, we presented results for the comparison of readers' unaided accuracy versus a CAD system, but, in fact, the study compared four CAD systems against readers' unaided accuracy. Table 6 illustrates the full study design. Each reader interpreted all 200 cases: 50 with and without CAD 1, 50 cases with and without CAD 2, 50 cases with and without CAD 3, and 50 cases with and without CAD 4. Each image was interpreted by all 36 readers unaided and with one of the four CAD systems. Thus, the study took advantage of the increased power offered by the split-plot design with respect to total readings, compared with the fully-crossed design, to simultaneously evaluate four CAD systems.

In this paper we presented three test statistics for the split-plot MRMC study design and compared their performance. These three methods have important differences. First, the variance *estimators* that the different methods use are not equivalent: the marginal-mean ANOVA method uses a correlated-error, three-way split-plot ANOVA with three covariances, the modified OR method uses a correlated error two-way factorial ANOVA with three covariances, and Gallas uses U-statistics. Second, the full three-sample U-statistic result differs from the OR and Hillis variance models in its level of detail. The full U-statistic result has 21 terms, 7 for the variance of modality 1, 7 for modality 2, and 7 for the covariance. The OR and Hillis models each have 7 variance-component parameters total (Note that without replications, only 6 of these parameters are estimable). The OR and Hillis variance models pool information across modalities. In contrast, the U-statistic variance includes all of the variance components related to the reader, case, modality, and disease status interactions, which can be particularly important for sizing future studies, though there is a cost. The U-statistic variance does not immediately generalize for measures of test accuracy that are not U-statistics, while the OR and Hillis methods can be applied to these other measures of test accuracy. Lastly, the marginal-mean ANOVA and OR methods require complete data, and the marginal-mean ANOVA model requires a balanced design (i.e. the same number of readers and cases in each block). In contrast, the U-statistic method can be used for incomplete and/or unbalanced designs.

Despite these differences, the three test statistics performed quite similarly. The type I error rates of the three methods tended to run slightly below the nominal level, the power of the three methods was nearly identical, and the confidence interval coverage was at the nominal level. Thus, all three test statistics performed well and similarly for the datasets in our simulation study.

For comparison, we also investigated a standard normal distribution for the pivotal statistic in Equation 9 (results not shown). As expected, the type one errors were inflated but tended to the nominal rate as the number of readers increased: 0.06–0.09 for J/G=3 and 5 and 0.05–0.065 for J/G=7.

Our simulation study does have several limitations. First, our study was limited to normally distributed decision scores. In many imaging trials, an ordinal scale (e.g. 1–5) is used to measure reader confidence. All three proposed methods can be applied to studies using an ordinal scale, although we did not evaluate the performance of the methods for ordinal data. Similarly, decision scores expressed on a semi-continuous scale (e.g. 0–100) often do not follow a normal distribution. For example, in the breast CAD study the distribution of confidence scores for non-cancer patients was skewed to the right with 65% of cases assigned a score of zero. The distribution of confidence scores for the cancer patients was skewed very slightly to the left, but the mode occurred at a confidence score of zero (27% of cancer cases had a confidence score of zero). We evaluated the type I error rate of the three test statistics when low scores are binned at a confidence score of zero similar to the data in the breast CAD study, and found that the type I error rates of all three methods remain close to the nominal level: 0.053 for all three methods. Second, we only considered two-block designs; future work should include an expanded range of study designs. Lastly, we only considered the simple case where the variance and the number of patients in the experiment are the same for diseased and non-diseased patients and the same for both modalities. More research is needed under other conditions to determine if any of the methods has clear advantages.

Acknowledgments

Stephen Hillis is supported by National Institute of Biomedical Imaging and Bioengineering (NIBIB), grants R01EB000863 and R01EB013667

References

1. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978; 8:283–298. [PubMed: 112681]
2. Zweig MH, Campbell G. Receiver operating characteristic plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993; 39:561–577. [PubMed: 8472349]
3. Pepe, MS. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press; 2004.
4. Zhou, XH.; Obuchowski, NA.; McClish, DL. Statistical Methods in Diagnostic Medicine. 2. Wiley and Sons, Inc; New York: 2011.
5. Wagner RF, Metz CE, Campbell G. Assessment of medical imaging systems and computer aids: a tutorial review. Acad Radiol. 2007:723–748. [PubMed: 17502262]
6. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: Findings from a national sample. Arch Intern Med. 1996; 156:209–213. [PubMed: 8546556]
7. Chakraborty DP. Analysis of location specific observer performance data: validated extensions of the jackknife free-response (JAFROC) method. Academic Radiology. 2006; 13:1187–1193. [PubMed: 16979067]
8. Dorfman DD, Berbaum KS, Metz CE. ROC rating analysis: generalization to the population of readers and cases with the jackknife method. Invest Radiol. 1992; 27:723. [PubMed: 1399456]
9. Obuchowski NA, Rockette HE. Hypothesis Testing of the Diagnostic Accuracy for Multiple Diagnostic Tests: An ANOVA Approach with Dependent Observations. Communications in Statistics: Simulation and Computation. 1995; 24:285–308.
10. Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: and alternative method for random-effects, receiver operating characteristic analysis. Academic Radiol. 2000; 7:341.
11. Beiden SV, Wagner RF, Campbell G, Metz CE, Jiang Y. Components-of-variance models for random-effects ROC analysis: The case of unequal variance structures across modalities. Academic Radiol. 2001; 8:605.
12. Beiden SV, Wagner RF, Campbell G, Chan H-P. Analysis of uncertainties in estimates of components of variance in multivariate ROC analysis. Academic Radiol. 2001; 8:616.
13. Obuchowski NA, Beiden SV, Berbaum KS, et al. Multireader, multicase receiver operating characteristic analysis: An empirical comparsion of five methods. Acad Radiol. 2004; 11:980–995. [PubMed: 15350579]
14. Hillis SL. A comparison of denominator degrees of freedom methods for multiple observer ROC analysis. Stat Med. 2007; 26:596–619. [PubMed: 16538699]
15. Gallas BD. One-shot estimate of MRMC variance: AUC. Acad Radiol. 2006; 13:353–362. [PubMed: 16488848]

16. Bandos AI, Rockette HE, Gur D. A permutation test for comparing ROC curves in multireader studies. Acad Radiol 2006. 2006; 13:414–420.

17. Obuchowski NA. Multi-reader ROC Studies: A Comparison of Study Designs. Academic Radiology. 1995; 2:709–716. [PubMed: 9419629]

18. Obuchowski NA. Reducing the number of reader interpretations in MRMC studies. Acad Radiol. 2009; 16:209–217. [PubMed: 19124107]

19. Gallas BD, Pennello GA, Myers KJ. Multireader multicase variance analysis for binary data. J Opt Soc Am A Opt Image Sci Vis. 2007; 24:B70–80. [PubMed: 18059916]

20. Gallas BD, Brown DG. Reader studies for validation of CAD systems. Neural Networks. 2008; 21:387–397. [PubMed: 18215501]

21. Hillis SL. A marginal-mean ANOVA approach for analyzing multi-reader radiological imaging data. Jul 9.2012 Submitted to Statistics in Medicine.

22. Bhat BR. On the distribution of certain quadratic forms in normal variates. J Royal Statistical Society B. 1962; 24:148–151.

23. DeLong E, DeLong D, Clarke-Pearson D. Comparing areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. Biometrics. 1988; 44:837–845. [PubMed: 3203132]

24. Gallas BD, Bandos A, Samuelson F, Wagner RF. A framework for random-effects ROC analysis: Biases with the bootstrap and other variance estimators. Commun Stat A-Theory. 2009; 38:2586–2603.

25. Barrett; Kupinski, MA.; Clarkson, E. Probabilistic foundations of the MRMC method. In: Eckstein, MP.; Jiang, Y., editors. Medical Imaging 2005: Image Perception Observer Performance, and Technology Assessment; Proc SPIE; p. 21-31.

26. Clarkson E, Kupinski MA, Barrett HH. A probabilistic model for the MRMC method. Part 1. Theoretical development. Acad Radio. 2006; 13:1410–1421.

27. Roe CA, Metz CE. Variance-component modeling in the analysis of receiver operating characteristic index estimates. Academic Radiol. 1997; 4:587.

28. Dorfman DD, Berbaum KS, Lenth RV, Chen YF, Donaghy BA. Monte Carlo validation of a multireader method for receiver operating characteristic discrete rating data: factorial experimental design. Acad Radiol. 1998; 5:9–19. [PubMed: 9442202]

29. Hillis SL, Obuchowski NA, Berbaum KS. Power estimation for multireader ROC methods an updated and unified approach. Acad Radiol. 2011 Feb; 18(2):129–42. [PubMed: 21232681]

Appendix

Calculation of Mean Square Terms for OR method

In equation 3, MS(T) is the mean square of the modality effect and MS(T×R) is the mean square of the interaction of reader and modality:

$$\begin{gathered}\mathrm{MS(T)}=\mathrm{J}\sum_{\mathrm{i}}(\widehat{\theta}_{i.}-\widehat{\theta}_{..})^2/(\mathrm{I}-1),\ \text{and}\\ \mathrm{MS(T\times R)}=\{1/(\mathrm{J}-1)(\mathrm{I}-1)\}\sum_{\mathrm{i}}\sum_{\mathrm{j}}(\widehat{\theta}_{ij}-\widehat{\theta}_{i.}-\widehat{\theta}_{.j}+\widehat{\theta}_{..})^2.\end{gathered} \qquad [A1]$$

Note that, in the expressions above, $\hat{\theta}_{i.}$ is the estimate of the ROC area for modality i, averaged over all J readers, $\hat{\theta}_{.j}$ is the estimate of the ROC area for reader j, averaged over all I modalities, and $\hat{\theta}_{..}$ is the estimate of the ROC area, averaged over all readers and modalities:

$$\begin{gathered}\widehat{\theta}_{i.}=\sum_{\mathrm{j}}\widehat{\theta}_{ij}/\mathrm{J}\\ \widehat{\theta}_{.j}=\sum_{\mathrm{i}}\widehat{\theta}_{ij}/\mathrm{I}\\ \widehat{\theta}_{..}=\sum_{\mathrm{i}}\sum_{\mathrm{j}}\widehat{\theta}_{ij}/\mathrm{I}\times\mathrm{J}\end{gathered}$$

where i=1, … I, and j=1,…J.

Rationale for Marginal-Mean Model Approach

The OR model for the usual fully-crossed modality-by-reader-by-case design can be shown to be the same as the model for the marginal means, across cases, of a conventional modality-by-reader-by-case ANOVA with reader and case as random factors and all possible interactions included. The OR F statistic can be derived by modifying the conventional F statistic by replacing mean squares involving case by error-covariance estimates for the marginal-means model. This produces a valid ANOVA F statistic for the marginal-means ANOVA model, and hence also for the OR model.

This general approach can easily be extended to the split-plot design as follows. For the conventional ANOVA model corresponding to the split-plot design with reader and case as random factors, with block and test as fixed factors and all possible interactions included, the resulting model for the marginal means across cases is given by

$$Y_{ijg}=\mu+\tau_i+\gamma_g+R_{j(g)}+(\tau\gamma)_{ig}+(\tau R)_{i\,j(g)}+\varepsilon_{ijg}$$

where $g = 1,\ldots,G$, $i = 1,\ldots, t$, $j = 1,\ldots, r$, where G is the number of blocks, t is the number of tests, r is the number of readers in each block, τ_i denotes the fixed effect of test, γ_g denotes the fixed effect of block, and $(\tau\gamma)_{ig}$ denotes the fixed test-by-block interaction. The $R_{j(g)}$ and $(\tau R)_{ij(g)}$ are random reader and test-by-reader effects, nested within block; they are mutually independent and normally distributed with zero means and respective variances $\sigma^2_{R(G)}$ and $\sigma^2_{\tau R(G)}$, where the subscript $R(G)$ is read "reader nested within group," etc. The ε_{ijg} are normally distributed with zero mean and variance σ^2_{ε}. The ε_{ijg} are independent of the $R_{j(g)}$ and $(\tau R)_{ij(g)}$. The covariances are defined by $\text{Cov}_1 \equiv \text{Cov}\ (\varepsilon_{ijg}, \varepsilon_{i'jg})$, $\text{Cov}_2 \equiv \text{Cov}\ (\varepsilon_{ijg}, \varepsilon_{ij'g})$, and $\text{Cov}_3 \equiv (\text{Cov}\ (\varepsilon_{ijg}, \varepsilon_{ijg})$ where $i \neq i'$, $j \neq j'$, and are subject to these constraints: $\text{Cov}_1 \geq \text{Cov}_3$, $\text{Cov}_2 \geq \text{Cov}_3$, $\text{Cov}_3 \geq 0$. Thus this is a 3-way split-plot ANOVA with correlated errors, with test and block crossed and reader nested within block. Thus readers are the whole plots, test is the split-plot factor, and block is the whole-plot factor.

The F statistic given by Equation 5 resulted from modifying the F statistic for the conventional split-plot ANOVA model by replacing mean squares involving case by error-covariance estimates for the corresponding marginal-means model.

Background Work for the Three-Sample U-sample approach

Gallas [15] derived the variance of the reader-averaged non-parametric (trapezoidal) AUC from a fully-crossed study design and expressed it as a linear combination of success moments, second-order moments of the AUC kernel. Gallas also provided unbiased estimates of the moments, which, consequently, yield unbiased estimates of the variance of the reader-averaged AUC itself. This estimate was referred to as the one-shot estimate, as it did not rely on any resampling (the jackknife or bootstrap). Gallas et al. [24] later recognized that the variance derived was equivalent to a three-sample U-statistics result; the three samples correspond to readers, non-diseased cases, and diseased cases, resulting in the estimate of the covariance between $\hat{\theta}_{i.}$ and $\hat{\theta}_{i'.}$ given in equation 8. This decomposition was introduced by Barrett, Clarkson, and Krupinski (BCK) [25, 26]. The first component $\alpha_{1ii'}$ is that due to the non-diseased population, $\alpha_{2ii'}$ is that due to the diseased population, $\alpha_{4ii'}$ is that due to the readers, and the remaining α's are due to interactions of these three populations.

Gallas and Brown [20] generalized the variance derivation and estimation in Equation 8 to treat study designs that are not fully crossed. The only difference between the variance given in Equation 8 and that of a split-plot study design are the scaling factors for the α's. The estimation of the scaling factors and the α's is controlled by a design matrix that indicates whether or not a reader×case observation is included in the dataset. This generalization is used for the split-plot study designs.

Simulation Model

Roe and Metz [27] and Dorfman et al [28] described a method for simulating data for the MRMC fully-crossed study design. They assumed a linear effects model for the decision variables (i.e. test scores), X_{ijkt},

$$X_{ijkt}=\mu_t+\tau_{it}+R_{jt}+C_{kt}+(\tau R)_{ijt}+(\tau C)_{ikt}+(RC)_{jkt}+(\tau RC)_{ijkt}+E_{ijkt}, \quad [A2]$$

where X_{ijkt} is the test score assigned by the j-th reader to the k-th case with truth state t (t=0 for non-diseased patients and t=1 for diseased patients) imaged with modality i. Every effect on the right-hand side depends on the truth state t: μ_t is an intercept term, τ_{it} is the fixed effect due to the i-th modality, R_{jt} is the random effect due to the j-th reader, C_{kt} is the random effect due to the k-th case, $(\tau R)_{ijt}$ is the random effect due to the interaction between modality and reader, $(\tau C)_{ikt}$ is the random effect due to the interaction between modality and case, $(RC)_{jkt}$ is the random effect due to the interaction between reader and case, $(\tau RC)_{ijkt}$ is the random effect due to the three-way interaction between modality, reader, and patient, and E_{ijkt} is the pure random error term.

Parameter Estimates from Breast Cancer CAD study

The estimated variances and covariances from the CAD breast cancer example are presented in Tables A1 (from the three-sample U statistic method) and A2 (from the OR and marginal mean ANOVA methods). These estimates are often useful for planning the sample size of future MRMC studies [29]. Note that the average between-reader correlations were 0.463 (between readers interpreting the same cases with the same modality) and 0.457 (between readers interpreting the same cases with different modalities); it is the difference between these two correlations, 0.006, that affects the efficiency of the split-plot design. Although this difference is quite small, the 4-group-split-plot used in the breast CAD study is 5 times more efficient than the fully-crossed design *for the same total number of reader interpretations*, though the better efficiency does take four times as many cases.

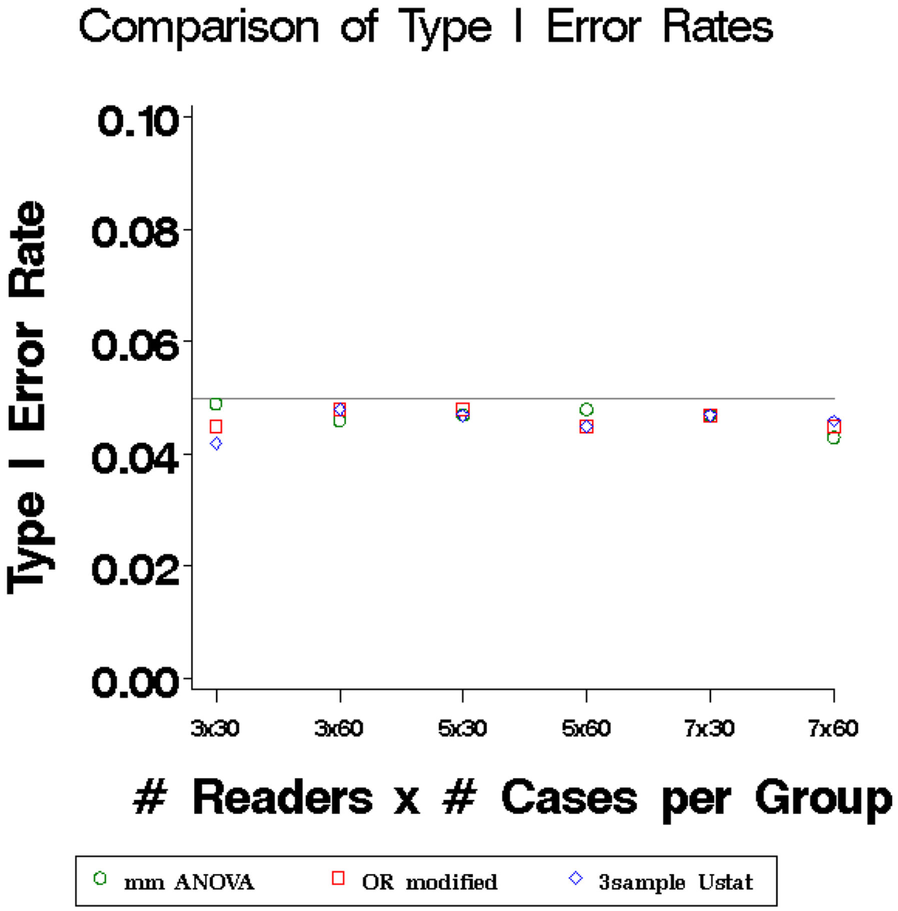

Figure 1.

Type I error rates of three methods: marginal mean ANOVA test statistic (mm ANOVA) (Equation 6) plotted with circles, modified OR test statistic (Equation 3) plotted with squares, and three-sample U-statistic (equation 10) plotted with diamonds. The nominal type I error rate was 0.05.

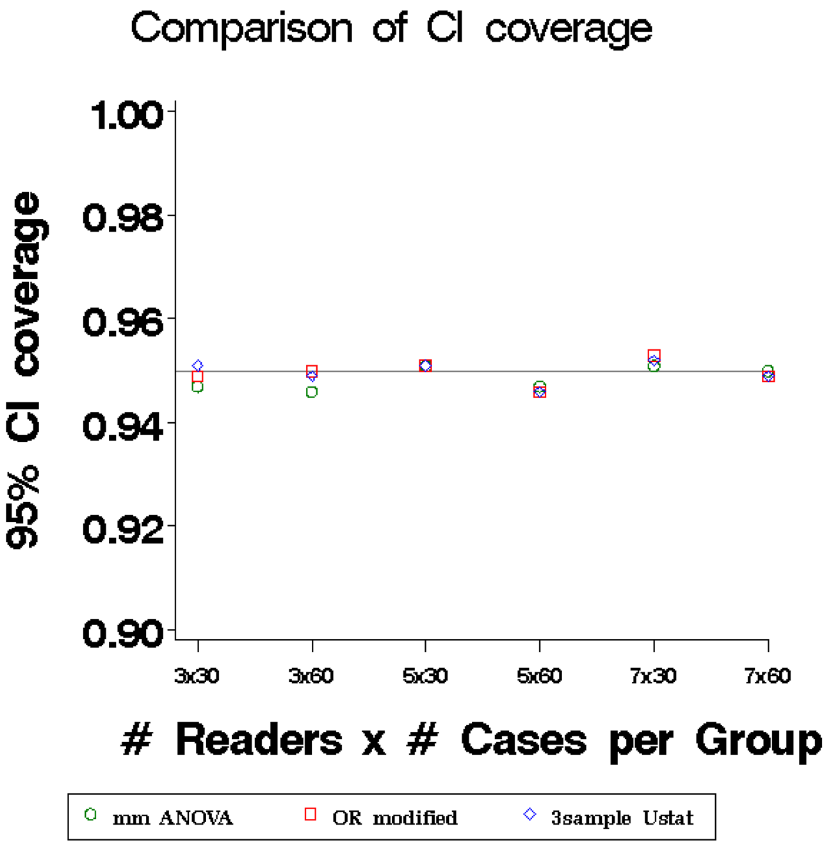

Figure 2.

Coverage of 95% Confidence Intervals of three methods: marginal mean ANOVA test statistic (circles), modified OR test statistic (squares), and three-sample U-statistic (diamonds).

Table 1

Lay-Out of Fully-Crossed MRMC Study Design

	Reader 1		Reader 2 …		Reader J	
	Test 1	**Test 2**	**Test 1**	**Test 2**	**Test 1**	**Test 2**
Patient 1	X_{111}	X_{211}	X_{121}	X_{221}	X_{1J1}	X_{2J1}
Patient 2	X_{112}	X_{212}	X_{122}	X_{222}	X_{1J2}	X_{2J2}
Patient 3	X_{113}	X_{213}	X_{123}	X_{223}	X_{1J3}	X_{2J3}
…						
Patient N_T	X_{11N}	X_{21N}	X_{12N}	X_{22N}	X_{1JN}	X_{2JN}

X_{ijk} denotes the test score assigned by the j-th reader to the k-th patient imaged with modality i.

Table 2

Lay-out of Split-Plot MRMC Study Design with Three Blocks

Block 1		**Block 2**		**Block 3**	
Reader 1	*Reader 2*				
X_{1111}, X_{2111}	X_{1211}, X_{2211}				
X_{1121}, X_{2121}	X_{1221}, X_{2221}				
X_{1131}, X_{2131}	X_{1231}, X_{2231}				
		Reader 3	*Reader 4*		
		X_{1342}, X_{2342}	X_{1442}, X_{2442}		
		X_{1352}, X_{2352}	X_{1452}, X_{2452}		
		X_{1362}, X_{2362}	X_{1462}, X_{2462}		
				Reader 5	*Reader 6*
				X_{1573}, X_{2573}	X_{1673}, X_{2673}
				X_{1583}, X_{2583}	X_{1683}, X_{2683}
				X_{1593}, X_{2593}	X_{1693}, X_{2693}

In a split-plot reader design with G blocks, J readers are randomized to a block and N_T patients are randomized to a block such that in each block there are J/G readers and N_T/G patients. In the 3-block split-plot design illustrated here, J=6 and N_T=9; thus, two readers are randomized to each of the three reader blocks and three patients are randomized to each of the three blocks.

X_{ijkg} denotes the test score assigned by the j-th reader in block g to the k-th patient in block g imaged with modality i.

Table 3

Resources Needed for Different Study Designs

Study Design	# readers, J	# patients[*]	Total # of image interpretations	# image interpretations per reader	Statistical Efficiency[**]
2-block split-plot	6 (3/block)	120 (30+30)	720	120	1.0
3-block split-plot	9 (3/block)	120 (20+20)	720	80	1.2
4-block split-plot	12 (3/block)	120 (15+15)	720	60	1.33
Fully-paired A	6	60 (30+30)	720	120	0.83
Fully-paired B	6	120 (60+60)	1440	240	1.16
Unpaired-reader	12	120 (60+60)	1440	120	0.90

[*] Total sample size (number of non-diseased and diseased patients per block)

[**] Statistical efficiency is defined as the variance of the two-block split plot design divided by the variance of the specified alternative design [19]. An efficiency >1 indicates that the variance of the 2- block split design is larger than the variance of the specified alternative design; an efficiency <1 indicates that the variance of the 2-block split design is smaller than the variance of the specified alternative. The calculations are based on the first set of model parameters in our simulation study (see Section 3).

Table 4

Parameter Values for Simulated Test Scores

Parameter	**Test Values**
Intercept	For non-diseased patients, μ_0=0. For diseased patients, μ_1=1.53.
Fixed modality effect	Under the null hypothesis, τ_{it}=0 for i=1 and 2 and t=0 and 1. Under the alternative hypothesis, τ_{i0}=0 for i=1 and 2, and τ_{11}=0 and τ_{21}=0.25.
Random effect due to reader j	Two values of σ_r^2 were tested: 0.011 and 0.056 to represent small and large inter-reader variability.
Random effect due to case k	σ_C^2 was set to 0.1.
Random effect due to modality × reader	Two values of $\sigma_{\tau R}^2$ were tested: 0.03 and 0.06.
Random effect due to modality × case	$\sigma_{\tau C}^2$ was set to 0.1.
Random effect due to reader × case	σ_{RC}^2 was set to 0.2.
Random effect due to pure error	$\sigma_{\tau RC}^2$ was set to 0.2.

Table 5

Summary of Results of Three Methods for Breast Cancer Example

Method	Test statistic, p-value	Estimated difference (SE)	95% CI for difference
Marginal mean ANOVA	F=3.12, p=0.0786	0.0076 (0.00431)	[−0.0009, 0.0161]
Modified OR	F=3.37, p=0.0678	0.0076 (0.00415)	[−0.0005, 0.0159]
3-sample U-stats	F=3.55, p=0.0644	0.0076 (0.00404)	[−0.0005, 0.0157]

Table 6

Lay-out of 4-Block Split-Plot Design simultaneously evaluating 4 CAD systems

	Reader Block 1	Reader Block 2	Reader Block 3	Reader Block 4
Patient Block 1	Unaided vs. CAD 1	Unaided vs. CAD 2	Unaided vs. CAD 3	Unaided vs. CAD 4
Patient Block 2	Unaided vs. CAD 4	Unaided vs. CAD 1	Unaided vs. CAD 2	Unaided vs. CAD 3
Patient Block 3	Unaided vs. CAD 3	Unaided vs. CAD 4	Unaided vs. CAD 1	Unaided vs. CAD 2
Patient Block 4	Unaided vs. CAD 2	Unaided vs. CAD 3	Unaided vs. CAD 4	Unaided vs. CAD 1

• Each reader block contained 9 readers. Each patient block contained 25 patients with cancer and 25 patients without cancer.