Journal of
Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Paired split-plot designs of multireader multicase studies

Weijie Chen
Qi Gong
Brandon D. Gallas

SPIE.

Weijie Chen, Qi Gong, Brandon D. Gallas, "Paired split-plot designs of multireader multicase studies," *J. Med. Imag.* **5**(3), 031410 (2018), doi: 10.1117/1.JMI.5.3.031410.

Paired split-plot designs of multireader multicase studies

Weijie Chen,* **Qi Gong, and Brandon D. Gallas**
Food and Drug Administration, Center for Devices and Radiological Health, Office of Science and Engineering Laboratories, Division of Imaging, Diagnostics, and Software Reliability, Silver Spring, Maryland, United States

Abstract. The widely used multireader multicase ROC study design for comparing imaging modalities is the fully crossed (FC) design: every reader reads every case of both modalities. We investigate paired split-plot (PSP) designs that may allow for reduced cost and increased flexibility compared with the FC design. In the PSP design, case images from two modalities are read by the same readers, thereby the readings are paired across modalities. However, within each modality, not every reader reads every case. Instead, both the readers and the cases are partitioned into a fixed number of groups and each group of readers reads its own group of cases—a split-plot design. Using a *U*-statistic based variance analysis for AUC (i.e., area under the ROC curve), we show analytically that precision can be gained by the PSP design as compared with the FC design with the same number of readers and readings. Equivalently, we show that the PSP design can achieve the same statistical power as the FC design with a reduced number of readings. The trade-off for the increased precision in the PSP design is the cost of collecting a larger number of truth-verified patient cases than the FC design. This means that one can trade-off between different sources of cost and choose a least burdensome design. We provide a validation study to show the iMRMC software can be reliably used for analyzing data from both FC and PSP designs. Finally, we demonstrate the advantages of the PSP design with a reader study comparing full-field digital mammography with screen-film mammography. [DOI: 10.1117/1.JMI.5.3.031410]

Keywords: multireader multicase; split-plot design; reader studies; least burdensome approach; iMRMC software.

Paper 17328SSPRR received Nov. 2, 2017; accepted for publication Apr. 30, 2018; published online May 17, 2018.

1 Introduction

In multireader multicase (MRMC) studies, a number of readers (e.g., radiologists) read medical images of a number of patient cases for a specified clinical task (e.g., cancer detection) and the diagnostic performance is evaluated. In the most general case, both readers and cases are treated as random representative samples from their respective populations. By accounting for both reader and case variabilities, the reader-averaged diagnostic performance can be generalized to both the reader and case populations, thereby providing direct evidence of device efficacy. Because both sources of variability are often substantial, sufficiently large numbers of readers and cases are needed to achieve a desired precision such that a difference in performance between two imaging modalities can be found statistically significant. The theme of this paper is to investigate study design strategies for MRMC studies that may allow for reduced cost and increased flexibility.

The most widely used MRMC study design for comparing two modalities is the fully crossed (FC) design: every reader reads every case of both modalities.[1] By pairing both readers and cases across modalities, the FC design builds a positive correlation between the performances of two modalities and reduces the variability of the performance difference.[1,2] Formally, the variance of the performance difference is

$$\mathrm{Var}[\hat{A}^{(1)}-\hat{A}^{(2)}]=\mathrm{Var}[\hat{A}^{(1)}]+\mathrm{Var}[\hat{A}^{(2)}]-2\rho\sqrt{\mathrm{Var}\hat{A}^{(1)}\mathrm{Var}\hat{A}^{(2)}},$$

where $\hat{A}^{(i)}$ is the performance estimate for modality i and ρ is the correlation between $\hat{A}^{(1)}$ and $\hat{A}^{(2)}$. Typically, $\rho > 0$ and the correlation reduces the variability of the performance difference.

Throughout the paper, we assume the diagnostic performance metric is the area under the receiver operating characteristic curve (AUC) (unless noted otherwise) as this is a meaningful and widely used metric.[1] Now, by having every reader read every case, the FC design is cost-effective in terms of the total number of readers and the total number of cases.[3] However, as Obuchowski[3] pointed out for the FC design, "study length (i.e., the number of total readings) and the time commitment of individual readers (i.e., the number of readings per reader) can be great."

Alternative study designs have been investigated in the literature. Obuchowski[3] compared several designs including the traditional FC design, unpaired designs (either the cases or the readers or both are unpaired across modalities), and a hybrid design in which both readers and cases were paired across two modalities, but, within each modality, each reader read his/her own group of cases of both modalities. Obuchowski found that, not surprisingly, the unpaired designs had significant power disadvantages due to lack of correlation as explained in the previous paragraph. The FC design is powerful but requires a long study duration and a heavy workload for each reader. The "hybrid design" was shown to be very competitive in terms of study length and the time commitment of individual readers, but it requires a very large number of patient cases to be collected and truth-verified.

*Address all correspondence to: Weijie Chen, E-mail: weijie.chen@fda.hhs.gov

2329-4302/2018/$25.00 © 2018 SPIE

In an effort to reduce the number of cases, Obuchowski[4] proposed the "mixed" MRMC design in which both readers and cases were paired across two modalities and, within each modality, they were divided into a number of independent reader or case groups and each group of readers read their own group of cases. Obuchowski showed that the "mixed" design was a promising alternative borrowing strengths from both the FC design and the hybrid design. Later, this design was called the split-plot design and different analysis methods were compared[5] and refined.[6] In this paper, we call this design the "paired split-plot" (PSP) design to reflect the fact that both readers and cases are paired across modalities, and the split-plot design is used within each modality.

Despite the aforementioned publications showing potential advantages of the PSP design, we have rarely seen its use in real-world studies, either in the peer-reviewed literature or in the premarket applications submitted to the Food and Drug Administration for regulatory review. This is likely due to the deeply rooted notion that the FC design is the "most" powerful and that many validation studies for freely available software tools assume a FC design. However, the notion that the FC design is most powerful only means that, by having all the readers read all the cases, one can obtain the most information possible from the given numbers of readers and cases. This notion essentially considers the sample sizes (readers and cases) as the only resource needed to achieve certain statistical power. However, the workload of the readers (i.e., the number of readings) is another cost of a study, and frequently a major one. In this work, we show that, when both sample sizes and reader's workload are considered, the PSP design can be more cost-effective than the FC design, which is similar to the notion that Obuchowski[4] demonstrated by showing how the estimates of certain Obuchowski–Rockette model parameters are affected by different designs. In our work, we provide a mathematical proof with easy-to-understand formulas showing statistical efficiency/power gain of the PSP design compared with the FC design when the number of readings for each reader is the same. We then put the theoretical analysis into practical perspectives by comparing different designs in terms of power and cost trade-off under a variety of simulation conditions. Furthermore, we present a real MRMC study, the VIPER study (validation of imaging in premarket evaluation and regulation),[7] which used the PSP design. We use the parameters estimated from this real study to further compare different designs. Finally, in the appendix, we present a simulation study validating the freely available iMRMC software[8] and show that it works equally well for for both FC and PSP designs.

2 Efficiency Gain of Paired Split-Plot Designs

In this section, we present theoretical analyses to demonstrate that, with fixed number of readers and fixed workload per reader, the PSP design is more efficient than the FC design for both measuring the performance of a single modality and for comparing two modalities.

2.1 *Theoretical Analysis: Single Modality*

In a FC design with N_R readers each reading N_0 nondiseased cases and N_1 diseased cases, the reader-averaged empirical estimate of AUC for a single modality is

$$\hat{A} = \frac{\Sigma_{r=1}^{N_R}\Sigma_{i=1}^{N_0}\Sigma_{j=1}^{N_1} s(x_{ir}, y_{jr})}{N_R N_0 N_1}, \tag{1}$$

where x_{ir} and y_{jr} are the rating scores (e.g., level of confidence that cancer is present) of reader r on the nondiseased case i and diseased case j, respectively, and $s(x, y)$ is the kernel function

$$s(x,y) = \begin{cases} 1 & \text{if } x < y \\ 0.5 & \text{if } x = y \\ 0 & \text{if } x > y \end{cases}.$$

Note that we use the "hat" notation for "an estimate" of a population parameter and we will use upper-case X and Y to denote random variables corresponding to the observations x_{ir} and y_{jr}, respectively.

Gallas[9] showed that the MRMC variance of $\hat{A}$ can be written as

$$\begin{aligned}\operatorname{Var}\hat{A} = {}& \frac{1}{N_R}(c_1M_1 + c_2M_2 + c_3M_3 + c_4M_4) \\ &+ \frac{N_R - 1}{N_R}(c_1M_5 + c_2M_6 + c_3M_7 + c_4M_8) - M_8,\end{aligned} \tag{2}$$

where $c_i (i = 1, \ldots, 4)$ are determined by N_0 and N_1

$$c_1 = 1/N_0N_1, \quad c_2 = (N_0 - 1)/N_0N_1,$$
$$c_3 = (N_1 - 1)/N_0N_1, \quad c_4 = (N_0 - 1)(N_1 - 1)/N_0N_1,$$

and the $M_l (l = 1, \ldots, 8)$ are the second-order moments of the kernel function s (E denotes "expectation")

- $M_1 = E[s(X_{ir}, Y_{jr})^2]$,
- $M_2 = E[s(X_{ir}, Y_{jr})s(X_{i'r}, Y_{jr})](i \neq i')$,
- $M_3 = E[s(X_{ir}, Y_{jr})s(X_{ir}, Y_{j'r})](j \neq j')$,
- $M_4 = E[s(X_{ir}, Y_{jr})s(X_{i'r}, Y_{j'r})](i \neq i', j \neq j')$,
- $M_5 = E[s(X_{ir}, Y_{jr})s(X_{ir'}, Y_{jr'})](r \neq r')$,
- $M_6 = E[s(X_{ir}, Y_{jr})s(X_{i'r'}, Y_{jr'})](i \neq i', r \neq r')$,
- $M_7 = E[s(X_{ir}, Y_{jr})s(X_{ir'}, Y_{j'r'})](j \neq j', r \neq r')$,
- $M_8 = E[s(X_{ir}, Y_{jr})s(X_{i'r'}, Y_{j'r'})](i \neq i', j \neq j', r \neq r')$.

The unbiased estimators of these moments are provided in Gallas.[9] This approach has been developed based on a probabilistic foundation of MRMC analysis[9–12] and was later found to be identical to the U-statistic approach.[13,14]

We define two variance-component parameters. The first decreases with the number of readers [Eq. (3)]. The second is independent of the number of readers [Eq. (4)]

$$\begin{aligned}V_R \equiv {}& c_1(M_1 - M_5) + c_2(M_2 - M_6) + c_3(M_3 - M_7) \\ &+ c_4(M_4 - M_8)\end{aligned} \tag{3}$$

and

$$V_C \equiv c_1M_5 + c_2M_6 + c_3M_7 - (1 - c_4)M_8. \tag{4}$$

Then, the MRMC variance in the FC design as expressed in Eq. (2) can be written as

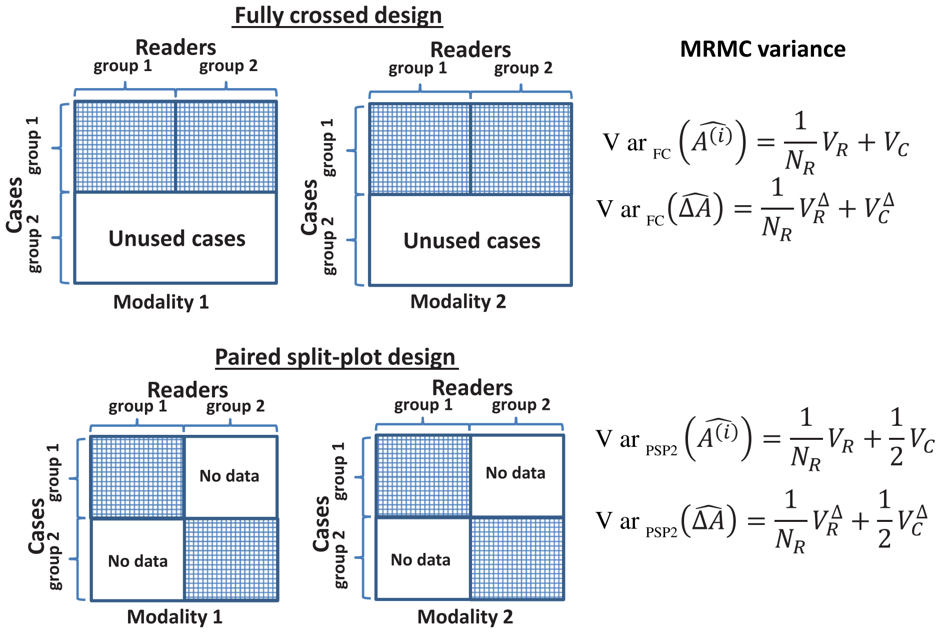

Fully crossed design
Readers
group 1
group 2
Cases
group 1
group 2
Unused cases
Modality 1
Modality 2
MRMC variance
$\mathrm{V\ ar}_{\mathrm{FC}}\left(\widehat{A^{(i)}}\right) = \frac{1}{N_R}V_R + V_C$
$\mathrm{V\ ar}_{\mathrm{FC}}\left(\widehat{\Delta A}\right) = \frac{1}{N_R}V_R^{\Delta} + V_C^{\Delta}$
Paired split-plot design
No data
$\mathrm{V\ ar}_{\mathrm{PSP2}}\left(\widehat{A^{(i)}}\right) = \frac{1}{N_R}V_R + \frac{1}{2}V_C$
$\mathrm{V\ ar}_{\mathrm{PSP2}}\left(\widehat{\Delta A}\right) = \frac{1}{N_R}V_R^{\Delta} + \frac{1}{2}V_C^{\Delta}$

Fig. 1 Illustration of the FC design and the PSP design (assuming the two groups have the same number of readers and each reader reads the same number of cases).

$$\mathrm{Var}_{\mathrm{FC}}\,\hat{A} = \frac{1}{N_R}V_R + V_C. \tag{5}$$

To understand the meaning of V_R and V_C, note that the MRMC variance in Eq. (2) can be decomposed into three components: reader variability purely due to the finite number of readers, case variability purely due to the finite number of cases, and variability due to reader by case interactions, which are expressed as

$$\mathrm{Var}_{\mathrm{reader}}\,\hat{A} = \frac{1}{N_R}(M_4 - M_8), \tag{6}$$

$$\mathrm{Var}_{\mathrm{case}}\,\hat{A} = c_1M_5 + c_2M_6 + c_3M_7 - (1-c_4)M_8, \tag{7}$$

and

$$\mathrm{Var}_{\mathrm{reader{\times}case}}\hat{A} = \frac{1}{N_R}[c_1M_1 + c_2M_2 + c_3M_3 - (1-c_4)M_4 \\ - c_1M_5 - c_2M_6 - c_3M_7 + (1-c_4)M_8], \tag{8}$$

respectively. It can be seen that V_C is identical to the case variability [Eqs. (4) and (7)] and V_R (when normalized by N_R) is the combination of reader variability and variability due to reader by case interactions.

We now consider a PSP design with G groups (PSPG). For comparison, we assume that the total number of readers and the workload of each reader are the same as those in the FC design, i.e., N_R readers each reading N_0 nondiseased cases and N_1 diseased cases. The difference is that all the readers read the same $N_0 + N_1$ cases in the FC design whereas, in the PSP design, different groups of readers read different groups of cases. Within each group, however, all the readers read the same group of cases. These are schematically shown in Fig. 1 for the FC design and a PSP2 design. As each group is FC, the variance of the AUC averaged over the N_{R_g} readers in the g'th group is given as

$$\mathrm{Var}\,\hat{A}_g = \frac{1}{N_{R_g}}V_R + V_C.$$

As all the groups are independent, the performance estimates $\hat{A}_g (g = 1, \dots, G)$ in the G groups are independent (i.e., zero covariance across groups). Noting that $\Sigma_{g=1}^{G} N_{R_g} = N_R$, the variance of the average performance $\hat{\mathrm{A}} = \Sigma_{g=1}^{G} N_{R_g}\hat{A}_g/N_R$ is given as

$$\mathrm{Var}_{\mathrm{PSPG}}\,\hat{A} = \frac{1}{N_R^2}\Sigma_{g=1}^{G} N_{R_g}^2\,\mathrm{Var}\,\hat{A}_g \\ = \frac{1}{N_R}V_R + \frac{\Sigma_{g=1}^{G} N_{R_g}^2}{N_R^2}V_C, \tag{9}$$

$$= \frac{1}{N_R}V_R + \frac{1}{G}V_C\ (\text{if } N_{R_1} = \dots = N_{R_G}). \tag{10}$$

Comparing Eq. (5) with Eqs. (9) and (10), we see that, for study designs having the same number of readers with each reader reading the same number of cases, $\mathrm{Var}_{\mathrm{PSPG}}\,\hat{A} < \mathrm{Var}_{\mathrm{FC}}\,\hat{A}$, i.e., the PSP design is statistically more efficient in estimating the performance of a single modality. More specifically, we see that the precision gain of the PSP design with G groups is due to the shrinkage of the case variability component V_C by a factor of $1/G$. Of course the gain of efficiency is not free as when the number of cases read by each reader is held constant, the total number of cases in the PSPG design is G times that in the FC design. Basic statistics principles tell us that, when we have more case samples, we gain more information from cases and the uncertainty of the measured performance due to the finite case sample decreases. The theoretical analysis provided here shows how this happens exactly in the MRMC setting.

Table 1 Simulation parameters for the Roe and Metz model.

Structure	AUC$_1$	AUC$_2$	σ^2_R	$\sigma^2_{\tau R}$	σ^2_C	σ^2_{RC}	$\sigma^2_{\tau C}$	σ^2_E
HH	0.65	0.70	0.011	0.011	0.3	0.2	0.3	0.2
	0.80	0.85	0.030	0.030	0.3	0.2	0.3	0.2
	0.90	0.95	0.056	0.056	0.3	0.2	0.3	0.2
HL	0.65	0.70	0.0055	0.0055	0.3	0.2	0.3	0.2
	0.80	0.85	0.0055	0.0055	0.3	0.2	0.3	0.2
	0.90	0.95	0.0055	0.0055	0.3	0.2	0.3	0.2
LH	0.65	0.70	0.011	0.011	0.1	0.2	0.1	0.6
	0.80	0.85	0.030	0.030	0.1	0.2	0.1	0.6
	0.90	0.95	0.056	0.056	0.1	0.2	0.1	0.6
LL	0.65	0.70	0.0055	0.0055	0.1	0.2	0.1	0.6
	0.80	0.85	0.0055	0.0055	0.1	0.2	0.1	0.6
	0.90	0.95	0.0055	0.0055	0.1	0.2	0.1	0.6

Note: HH, high data correlation, high reader variance; HL, high data correlation, low reader variance; LH, low data correlation, high reader variance; LL, low data correlation, low reader variance.

On the one hand, when the total case sample size increases in the PSP design, the case variability decreases correspondingly although the number of cases read by each reader is the same. On the other hand, the other component in the total variance, namely V_R, is the same across designs because it is a function of the number of readers and the number of cases read per reader, which are set to be the same across designs.

2.2 *Theoretical Analysis: Comparing Two Modalities*

To compare two imaging modalities by testing the null hypothesis that the performances of the two modalities, $A^{(1)}$ and $A^{(2)}$, are equal, we need to estimate the performance difference $\widehat{\Delta A} = \widehat{A^{(1)}} - \widehat{A^{(2)}}$ and its variance $\widehat{\mathrm{Var}\widehat{\Delta A}}$ and construct a test statistic $z = \widehat{\Delta A}/\sqrt{\widehat{\mathrm{Var}\widehat{\Delta A}}}$. Note that we use superscript (i) to denote modality. For the FC design, we have

$$\widehat{\Delta A} = \frac{\Sigma_{r=1}^{N_R}\Sigma_{i=1}^{N_0}\Sigma_{j=1}^{N_1}[s^{(1)}(x_{ir}, y_{jr}) - s^{(2)}(x_{ir}, y_{jr})]}{N_R N_0 N_1}.$$

Comparing this equation with Eq. (1), we see that we just replace the kernel function s in the single-modality AUC formula with $s^{(1)} - s^{(2)}$ in the AUC difference formula here. It is straightforward to see that the variance formulas in Eqs. (2)–(10) for a single-modality AUC would hold for the variance of $\widehat{\Delta A}$ if we just replace s with $s^{(1)} - s^{(2)}$ in computing the moment parameters. For example, in the single-modality situation we have $M_1 = E[s(x_{ir}, y_{jr})^2]$. Then for the corresponding moment M_1^{Δ} for the variance of $\widehat{\Delta A}$, we have $M_1^{\Delta} = E\{[s^{(1)}(x_{ir}, y_{jr}) - s^{(2)}(x_{ir}, y_{jr})]^2\}$. Note that $M_1^{\Delta} = M_1^{(1)} + M_1^{(2)} - 2M_1^{(1\times 2)}$, where $M_1^{(i)} = E[s^{(i)}(x_{ir}, y_{jr})^2]$ is the moment parameter for the variance of AUC of modality i and $M_1^{(1\times 2)} = E[s^{(1)}(x_{ir}, y_{jr})s^{(2)}(x_{ir}, y_{jr})]$ is the moment parameter for the covariance between the two AUCs. The other seven moment parameters $M_k^{\Delta}(k = 2, \ldots, 8)$ are defined similarly with similar properties.

We define V_R^{Δ} and V_C^{Δ} in a fashion similar to V_R and V_C in Eqs. (3) and (4) by replacing the M_k parameters with $M_k^{\Delta}(k = 1, \ldots, 8)$. In the end, under the setting that the FC design and the PSPG design have the same number of readers each reading the same number of cases, we have the variance of AUC difference for these two designs in Eqs. (11) and (12), respectively. Again, we see that the V_C^{Δ} component of the variance would shrink by a factor of G in the PSPG design as compared with the FC design. Because statistical power is inversely related to the variance of the AUC difference, the PSPG design is more powerful in comparing two modalities given that the number of readers and the number of cases read per reader (workload) are the same

$$\mathrm{Var}_{\mathrm{FC}}\,\widehat{\Delta A} = \frac{1}{N_R}V_R^{\Delta} + V_C^{\Delta}, \tag{11}$$

$$\begin{aligned}\mathrm{Var}_{\mathrm{PSPG}}\,\widehat{A} &= \frac{1}{N_R}V_R^{\Delta} + \frac{\Sigma_{g=1}^{G}N_{Rg}^2}{N_R^2}V_C^{\Delta}\\ &= \frac{1}{N_R}V_R^{\Delta} + \frac{1}{G}V_C^{\Delta}\ (\text{if } N_{R1} = \cdots = N_{Rg}).\end{aligned} \tag{12}$$

3 Comparison of Paired Split-Plot with Fully Crossed Using Analytical Computations under the Roe and Metz Simulation Model

The purpose of this section is twofold. First, we put the theoretical variance analysis in the previous section into a more practical perspective by comparing the statistical power between the PSP design and the FC design under a variety of simulation conditions. This is not a simulation study, but we use a simulation model to explore different levels of reader and case variability (variance-component structures). Second, we show that a trade-off can be made between the total number of cases and the workload per reader in choosing the most cost-effective design for achieving the same power. In practice, the computational procedure presented here can be used to choose a cost-effective design based on real parameters (e.g., measured in a pilot study) rather than the simulation parameters as we do here.

The simulation model and simulation parameters were initially developed by Roe and Metz[15] and have been frequently used in validating analysis methods. Roe and Metz[15] developed a linear mixed effect model to simulate reader study data

$$X_{ijkt} = \mu_{it} + R_{jt} + C_{kt} + [RC]_{jkt} + [\tau R]_{ijt} + [\tau C]_{ikt} + E_{ijkt}, \tag{13}$$

where X_{ijkt} denotes the rating by reader j using modality i for the likelihood of case k being diseased (e.g., a malignant lesion is present in the image), whereas the truth state is t ($t = 0$ for nondiseased and $t = 1$ for diseased). The Greek letter μ denotes a fixed modality effect and the remaining terms denote random effects, which are independent zero-mean Gaussian random variables with variance parameters denoted as σ_R^2, σ_C^2, σ_{RC}^2, $\sigma_{\tau R}^2$, $\sigma_{\tau C}^2$, σ_E^2, respectively. These variance parameters can vary with the modality and the truth state in general,[16] but, for simplicity, they were set in Roe and Metz[15] to be the same across modalities and across truth states. Roe and Metz[15] provided several sets of variance parameters, which we adopt as shown in Table 1. These parameters have different combinations of high or low data correlation (the first H or L letter in the "structure" column of Table 1) and high or low reader variability (the second H or L letter in the "structure" column of Table 1). To simulate three levels of AUC values (expectation over the population of readers and the population of cases), we change the separation of the scores from nondiseased and diseased cases by setting the μ_{it} parameter as $\mu_{i0} = 0$ and $\mu_{i1} = \Phi^{-1}[A^{(i)}] \times \sqrt{2(1 + \sigma_R^2 + \sigma_{\tau R}^2)}$, where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution and the "1" in this formula comes from the constraint that Roe and Metz[15] set: $\sigma_C^2 + \sigma_{RC}^2 + \sigma_{\tau C}^2 + \sigma_E^2 = 1$.

Given the Roe and Metz simulation parameters, we can analytically compute the moment parameters $M_l^{(i)} (i = 1,2)$ and $M_l^{(1\times2)} (l = 1, \dots, 8)$ using the method developed by Gallas and Hillis,[16] which has been implemented in the iRoeMetz software.[17] Using these moment parameters and specified sample sizes, we can compute the variance of the AUC difference using the methods described in Sec. 2. Under normal approximation, the statistical power for a two-sided test at the significant level α with critical values $\pm z_{\alpha/2}$ is

$$\text{Power} = \Phi\left[\frac{\Delta A}{\sqrt{\text{Var}(\Delta A)}} - z_{\alpha/2}\right] + \Phi\left[-\frac{\Delta A}{\sqrt{\text{Var}(\Delta A)}} - z_{\alpha/2}\right], \tag{14}$$

where $z_{\alpha/2} = 1.96$ for $\alpha = 0.05$.

Specifying the number of readers $N_R = 16$ with each reader reading $N_0 = 80$ nondiseased cases and $N_1 = 60$ diseased cases, we computed the variance and statistical power for the FC design and PSP designs (with 2, 4, and 8 groups, respectively) for each set of the Roe and Metz model parameters (Table 1). The results are shown in Table 2.

Table 2 Comparison between the FC design and PSP designs in terms of statistical efficiency/power by holding constant the number of readings per reader.

			$\text{Var}(\Delta A) \times 10^3$				Power (%)			
Structure	AUC_1	AUC_2	FC	PSP2	PSP4	PSP8	FC	PSP2	PSP4	PSP8
HH	0.65	0.70	1.42	0.83	0.53	0.38	26.4	41.2	58.2	72.3
	0.80	0.85	0.96	0.63	0.46	0.38	36.4	51.4	64.4	73.1
	0.90	0.95	0.42	0.30	0.24	0.21	68.0	82.0	89.5	93.0
HL	0.65	0.70	1.34	0.75	0.45	0.30	27.6	44.9	65.7	82.5
	0.80	0.85	0.77	0.43	0.26	0.17	43.7	67.6	87.7	96.9
	0.90	0.95	0.28	0.15	0.09	0.06	85.1	98.0	99.9	100
LH	0.65	0.70	0.71	0.52	0.43	0.38	46.6	58.9	67.5	72.5
	0.80	0.85	0.55	0.45	0.40	0.38	56.9	65.5	70.6	73.3
	0.90	0.95	0.27	0.23	0.22	0.21	86.5	90.5	92.4	93.3
LL	0.65	0.70	0.63	0.44	0.34	0.30	51.4	66.5	76.9	82.7
	0.80	0.85	0.35	0.25	0.20	0.17	75.8	88.5	94.4	96.8
	0.90	0.95	0.12	0.083	0.067	0.059	99.7	100	100	100

There are two kinds of cost that can be considered in comparing study designs. One is the cost associated with the radiologist's time, which can be represented by the total number of readings. The other is the cost for collecting and truth-verifying patient cases, which can be represented by the total number of cases. For the results shown in Table 2, we held the total number of readings the same across study designs and showed that the PSP designs gain efficiency/power with increased cost of collecting and truth-verifying more cases. We can also compare the number of readings and the number of cases needed to achieve the same statistical power, allowing a trade-off between the two kinds of cost being made such that the total cost is minimized.

We again assumed 16 readers in each of the four designs: FC, PSP2, PSP4, and PSP8. Utilizing Eq. (14), the fixed number of readers (i.e., 16), and a fixed ratio (3:4) of the number of diseased cases to the number of nondiseased cases, we iteratively solved the number of cases each reader needs to read such that a power of 80% is achieved. The results are shown as "cases per reader" in Table 3. Then, the total number of readings is simply the "case per reader" times 16 and therefore they are equivalent for comparison purpose. The total number of cases that need to be collected and truth-verified is the number of cases per reader times 1, 2, 4, and 8 for the FC, PSP2, PSP4, and PSP8 designs, respectively (shown as "total number of cases" in Table 3). From this table, we can see that one can make a trade-off between "cases per reader" and the "total number of cases" to choose the most cost-effective design. If patient cases are precious, one would certainly choose the FC design as it requires the least number of cases. If many cases are already available and reader's time is the major cost of the study, which is typical in many retrospective studies, one can choose a PSP design to reduce the workload of readers.

4 VIPER Study

From the results in Sec. 3, we have seen that the relative advantage of one design over the other would depend on the variance-component structure. It is known that some of the Roe and Metz simulation parameters may not be realistic.[18] Thus, it is useful to further compare different designs with realistic variance parameters estimated from real data. We used the iMRMC software[8] to analyze a real dataset, and, based on the estimated parameters, we compared the PSP design to the FC design. We have validated our software to analyze both FC and PSP data (see Appendix A, for a validation study using simulations).

The real dataset is a study on design methodologies surrounding the validation of imaging premarket evaluation and regulation called VIPER.[7] The VIPER study compared full-field digital mammography (FFDM) to screen-film mammography (SFM) for women with heterogeneously dense or extremely dense breasts. All cases and corresponding images were sampled from Digital Mammographic Imaging Screening Trial[19] archives. This is a retrospective reader study contracted to Medical University of South Carolina (MUSC). The institutional review board of MUSC approved the study.

Here, we analyze one of the VIPER reader studies and use the estimated variance parameters to compare a PSP design with a related FC design. In this VIPER reader study, we used the PSP4 design with 20 readers divided into four groups (i.e., each group had five readers). Using design, each group of readers were to read 60 cases with an enriched cancer prevalence of ∼50%. However, the split was not perfectly balanced and the numbers of diseased and nondiseased cases per group are slightly different from the original design. The case sample sizes for the four groups are shown in Table 4. In this table, we also show the moment parameters estimated from this dataset using our iMRMC software along with the AUC values, their

Table 3 Comparison between the FC design and PSP designs in terms of the number of cases read by each reader and the total number of cases to achieve 80% power (16 readers are assumed for all the designs and so the total number readings is 16 times the "case per reader").

			Cases per reader				Total number of cases			
Structure	AUC_1	AUC_2	FC	PSP2	PSP4	PSP8	FC	PSP2	PSP4	PSP8
HH	0.65	0.70	1197	630	352	198	1197	1260	1408	1584
	0.80	0.85	1348	709	410	237	1348	1418	1640	1896
	0.90	0.95	249	137	81	51	249	274	324	408
HL	0.65	0.70	774	398	217	128	774	796	868	1024
	0.80	0.85	385	202	112	67	385	404	448	536
	0.90	0.95	144	77	46	28	144	154	184	224
LH	0.65	0.70	513	333	249	198	513	666	996	1584
	0.80	0.85	578	375	305	237	578	750	1220	1896
	0.90	0.95	104	74	60	51	104	148	240	408
LL	0.65	0.70	333	217	153	125	333	434	612	1000
	0.80	0.85	158	104	79	65	158	208	316	520
	0.90	0.95	56	39	30	27	56	78	120	216

Table 4 VIPER's high-prevalence reader study results and sizing new studies based on these results.

Retrospective reader study comparing SFM and FFDM								
Sample sizes: 20 readers in four groups								
			Group 1	Group 2	Group 3	Group 4	Total	
		N_0	32	31	32	35	130	
		N_1	28	29	28	24	109	
Estimated moment parameters								
	M_1	M_2	M_3	M_4	M_5	M_6	M_7	M_8
FFDM	0.7091	0.5904	0.5531	0.5060	0.5686	0.5509	0.5202	0.5075
SFM	0.7345	0.61470	0.5922	0.5433	0.59756	0.5774	0.5570	0.5439
Cross	0.5718	0.5549	0.5379	0.5239	0.5625	0.5519	0.5330	0.5257
Estimated AUC (with standard error)								
FFDM	0.713 (0.024)		SFM	0.738 (0.022)				
	Difference	−0.025 (0.025)		~95% CI	(−0.07, 0.02)			
FC with the same workload for each reader ($N_r = 20$, $N_0 = 32$, $N_1 = 27$):								
SE = 0.0394, ~95% CI (−0.10, 0.05)								
Size a noninferiority study (for 80% power, noninferiority margin $\delta = 0.05$)								
Assume effect size $\Delta A = 0$, diseased to nondiseased case ratio = 0.75								
	Cases per reader				Total number of cases			
	FC	PSP2	PSP4		FC	PSP2	PSP4	
$N_r = 16$	408	237	153		408	474	612	
$N_r = 20$	364	207	132		364	414	528	

difference, and the associated standard errors. The results show that FFDM's AUC is slightly inferior to the SFM's by 0.025 with a standard error (SE) of 0.025 and an ~95% confidence interval (CI) of (−0.07 and 0.02).

Given the components of variance in Table 4, we can estimate the variance of an FC study where 20 readers read the same 27 diseased cases and the same 32 nondiseased cases in both modalities [using Eq. (2) or Eq. (5)]. The SE of the AUC difference from such an FC study would be 0.039 [~95% CI (−0.10 and 0.05)]. Compared with the PSP4 study, the FC study has 25% the cases (59/239) and the same number of reads per reader and in total. The trade-off for the reduced cost of collecting and truth-verifying 25% of the cases is more uncertainty: the SE of the AUC difference from the FC study would be 56% larger than the PSP4 study.

For demonstration, we also show how to size a noninferiority study to achieve a specific power. A reasonable hypothesis to establish is that the diagnostic performance of FFDM is noninferior to that of SFM. We assume an effect size $\Delta A = 0$ based on the fact that their technological characteristics are similar and the clinical performance reported in the literature shows that FFDM is generally noninferior to the SFM. We specify a noninferiority margin δ of 0.05 in AUC. In addition, we assume the ratio of the number of diseased cases to the number of nondiseased cases to be 0.75. Then for a fixed number of readers, we iteratively solve the number of cases needed for a target power of 80%. The computational procedure was similar to that in Sec. 3 except that the power for a noninferiority study is (using normal approximation)

$$\text{Power}_{\text{noninferiority}} = \Phi\left(\frac{\Delta A + \delta}{\sqrt{\text{Var}(\Delta A)}} - z_{\alpha/2}\right).$$

We set the number of readers as 16 or 20 for the following designs: FC, PSP2, and PSP4. The results are shown in the bottom section of Table 4. We demonstrate again that one can assess the following in choosing a cost-effective design: how many readers are available, the cost of collecting and truth-verifying patient cases, and the cost of reading the cases by the readers. If a large number of cases are already available and the cost of collecting and truth-verifying cases is thus minimal, then a PSP design is preferred because the workload of each reader and the total number of reads are substantially reduced. For example, when $N_r = 16$, using the PSP4 design each reader would need to read 153 cases as compared with 408 in the FC design.

5 Discussions and Conclusion

In this work, we investigated the PSP design for MRMC reader studies and compared it with the widely used FC design. From the theoretical perspective, we analytically showed how statistical efficiency can be gained by the PSP design as compared with the FC design when the number of readings is the same across designs. We then used analytical computations to compare the two designs under a broad range of model parameters in terms of statistical efficiency/power, the number of reads per reader, and the number of cases that have to be collected and truth-verified. The results in Tables 2 and 3 showed that a trade-off can be made between the reader's workload and the collection and truth-verification of patient cases. Moreover, such a trade-off would depend on the true variance-component structure. Therefore, we further compared the two designs using variance-components estimated in a real study. The VIPER study results indicated that substantial precision can be gained by the PSP design with the same reading workload and more patient cases.

This means that, when the same cases are read again and again by multiple readers, the benefit of adding readers is subjected to diminishing returns. On the other hand, by having the readers read different cases (PSP) rather than the same cases (FC), substantial precision can be gained with fewer number of reads. In the meantime, it should be noted that this gain of precision may be associated with the extra cost of collecting and truth-verifying more cases. We further note that, in addition to these cost considerations, the PSP design offers practical flexibilities. For example, the study duration may be shortened because each reader reads fewer cases. Furthermore, in a study that involves multiple institutions, the readers may read the cases from their own institution thereby avoiding the need for shipping the cases around the country, assuming the study conditions can be well controlled and readers and patient cases from each institution are representative of their respective population. These findings are consistent with those of Obuchowski[4] who compared the PSP design with the FC design by showing how the Obuchowski–Rockette (OR) model parameters vary across the two designs. In this work, we employed the U-statistic variance analysis of the AUC to explicitly show the variance difference between the two designs using simple analytical formulas (Fig. 1).

These investigations have important practical implications. When patient cases are precious, one may choose the FC design to take full advantage of the cases that are available. On the other hand, when many cases are available and the study is mainly limited by the cost of the radiologist's time, one may choose to have the radiologists read fewer but different cases in a PSP design. If variance-component parameters are available (e.g., from pilot studies or previously published similar studies), one can quantify various sources of cost and compare the overall cost of different designs to choose the most cost-effective one.

We have assumed the same number of readers in comparing the FC design with the PSP design throughout this paper. This is mainly to make the analytical comparison easier. For example, in Fig. 1, we show that with the same number of readers each reading the same number of cases, the PSP design is more efficient (less variance) and it requires more cases. Graphically, the shaded area (that represents the number of readings) in this figure for the PSP design is the same as the FC design, but they are split in the vertical direction. Alternatively, we can split the shaded area in the horizontal direction and show that the PSP in that way is more efficient than the FC design. This is a setting that requires more readers. In reality, one may typically collect more cases, recruit more readers, or both, in the PSP design than in the FC design. The benefit is, as we have showed, fewer readings are needed to achieve the same precision (efficiency).

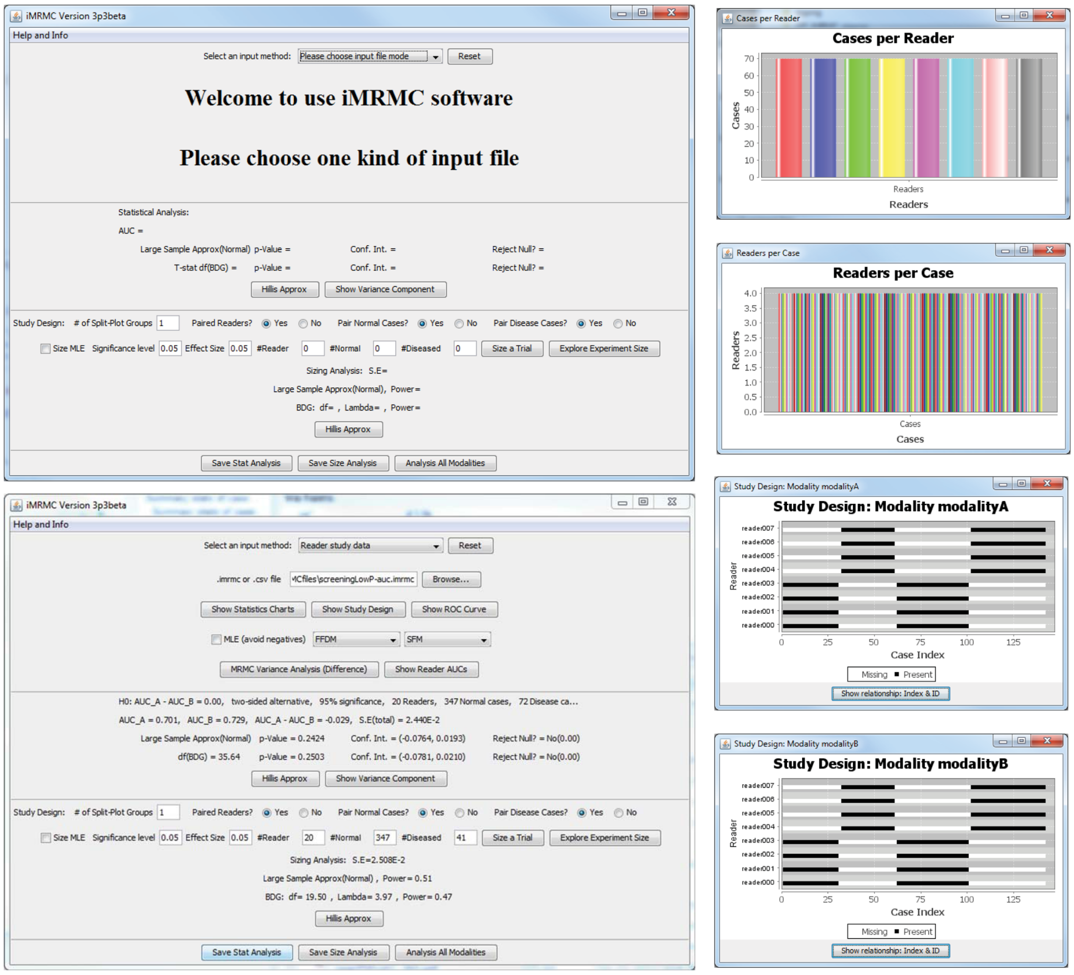
We note that the amount of PSP-versus-FC efficiency gain given the same reader workload depends on the variance components of the problem, as we have showed in Table 2. This is also evident from our analytical results in Eqs. (11) and (12). Because only one variance component, namely V_C, is reduced in the PSP design compared with the FC design, the efficiency gain can be small if this component is very small compared with the other variance components (i.e., $V_C \ll V_R$). Similar analytical results may be obtained in terms of the OR model parameters using the marginal-mean ANOVA approach developed by Hillis.[6] This is interesting future work because one can survey the published studies analyzed by either methods to compare study designs in a broad range of real-world applications. For readers who are familiar with the OR model, we showed a connection between the U-statistic parameters (V_C and V_R) and the OR model parameters in Appendix B.

The iMRMC[8] and iRoeMetz[17] software can be used to aid the design process as it can compute the variance and statistical power for different designs and sample sizes. We also validated the iMRMC software's statistical inference functionality using simulations and showed that it can be reliably used to analyze data from both the FC and PSP designs. However, we should point out that the current version of the iMRMC software only supports the binary performance endpoint (e.g., sensitivity and specificity) and the AUC endpoint estimated by the trapezoidal/Wilcoxon method, which is appropriate for ROC data collected on the multilevel ordinal or the (quasi-)continuous scale. Alternatively, one can use the software from the University of Iowa[20] that implements the Obuchowski–Rockette method for MRMC study sizing and analysis,[5,21,22] especially when partial AUC or semiparametric estimate of AUC is the preferred endpoint.

In conclusion, the PSP design is a useful alternative to the widely used FC design in MRMC studies. The PSP design may substantially reduce the cost of reader studies as compared with the FC design in many applications.

Appendix A: Validation of the iMRMC Software in Analyzing Paired Split-Plot Study Data

The freely available iMRMC software[8] can be used to aid the design (sizing) of an MRMC study and analyze data for both the FC and PSP designs. The software is platform independent with a graphical user interface (see Fig. 2 for screen shots). It is well documented and actively maintained. In its core, the software uses the trapezoidal/Wilcoxon method to estimate the AUC (a U statistic) and applies the U-statistic method to estimate the MRMC variance, which has been validated as an unbiased estimator.[9,13,23] For statistical inference, the test statistic



Select an input method: Please choose input file mode

Welcome to use iMRMC software

Please choose one kind of input file

Statistical Analysis:

AUC =

Large Sample Approx(Normal) p-Value = Conf. Int. = Reject Null? =

T-stat df(BDG) = p-Value = Conf. Int. = Reject Null? =

Study Design: # of Split-Plot Groups 1 Paired Readers? ● Yes ○ No Pair Normal Cases? ● Yes ○ No Pair Disease Cases? ● Yes ○ No

□ Size MLE Significance level 0.05 Effect Size 0.05 #Reader 0 #Normal 0 #Diseased 0

Sizing Analysis: S.E=

Large Sample Approx(Normal), Power=

BDG: df= , Lambda= , Power=

Select an input method: Reader study data

.imrmc or .csv file ᐱCfiles\screeningLowP-auc.imrmc

□ MLE (avoid negatives) FFDM SFM

H0: AUC_A - AUC_B = 0.00, two-sided alternative, 95% significance, 20 Readers, 347 Normal cases, 72 Disease ca...

AUC_A = 0.701, AUC_B = 0.729, AUC_A - AUC_B = -0.029, S.E(total) = 2.440E-2

Large Sample Approx(Normal) p-Value = 0.2424 Conf. Int. = (-0.0764, 0.0193) Reject Null? = No(0.00)

df(BDG) = 35.64 p-Value = 0.2503 Conf. Int. = (-0.0781, 0.0210) Reject Null? = No(0.00)

Study Design: # of Split-Plot Groups 1 Paired Readers? ● Yes ○ No Pair Normal Cases? ● Yes ○ No Pair Disease Cases? ● Yes ○ No

□ Size MLE Significance level 0.05 Effect Size 0.05 #Reader 20 #Normal 347 #Diseased 41

Sizing Analysis: S.E=2.508E-2

Large Sample Approx(Normal) , Power= 0.51

BDG: df= 19.50 , Lambda= 3.97 , Power= 0.47

Fig. 2 Graphical user interface of the iMRMC software.

$$t = \frac{|\widehat{A^{(1)}} - \widehat{A^{(2)}}|}{\sqrt{\widehat{\mathrm{Var}[A^{(1)} - A^{(2)}]}}},$$

is modeled as a Student's t statistic.[5]

We validated the statistical inference functionality of the software by simulating data under the null hypothesis (i.e., two modalities have equal AUC performance) and investigated the empirical type I error rate, which we expected to be close to the nominal level of 5%. Specifically, we simulated MRMC datasets using the Roe and Metz model [Eq. (13)] with specified parameters and analyzed it with our iMRMC software to see if the difference of performance between the two modalities was statistically significant at the significant level 0.05. We repeated the simulations 100,000 times and computed the proportion of experiments that showed a statistically significant AUC difference, which is the empirical type I error rate. We did such simulation validation for three study designs: FC, PSP2, and PSP3. For each design, we varied the following parameters in a fully factorial fashion:

- Variance component parameters with three AUC levels: we used the 12 sets of parameters in Table 1 except that we changed the values of AUC_2 such that $\mathrm{AUC}_2 = \mathrm{AUC}_1$ because we intended to simulate a null hypothesis,
- The total number of readers: $N_R = 6,12,18$, and
- The total number of cases: $N_C = 48,90$ cases per class.

The simulation results are plotted in Fig. 3. The results showed that the iMRMC software controlled the type I error rate reliably across a broad range of simulation parameters and study designs.

We do not recommend a reader study with too few readers because they may not represent the reader population. But just to test our software with extreme parameter values, we further performed simulation studies simulating only three readers reading $25 + 25$ cases in an FC fashion. The summary results (Fig. 4) show that the type I error is controlled below the nominal level of 0.05 across all the simulations (though it is a little conservative for some simulation conditions).

Appendix B Relationship between the U-Statistic Variance Parameters and the Obuchowski–Rockette Model Parameters

In Sec. 2.2, we showed that the variance of the estimate of difference in AUC $[\widehat{\Delta A} = \hat{A}^{(1)} - \hat{A}^{(2)}]$ in a FC design can be expressed in terms of U-statistic parameters as

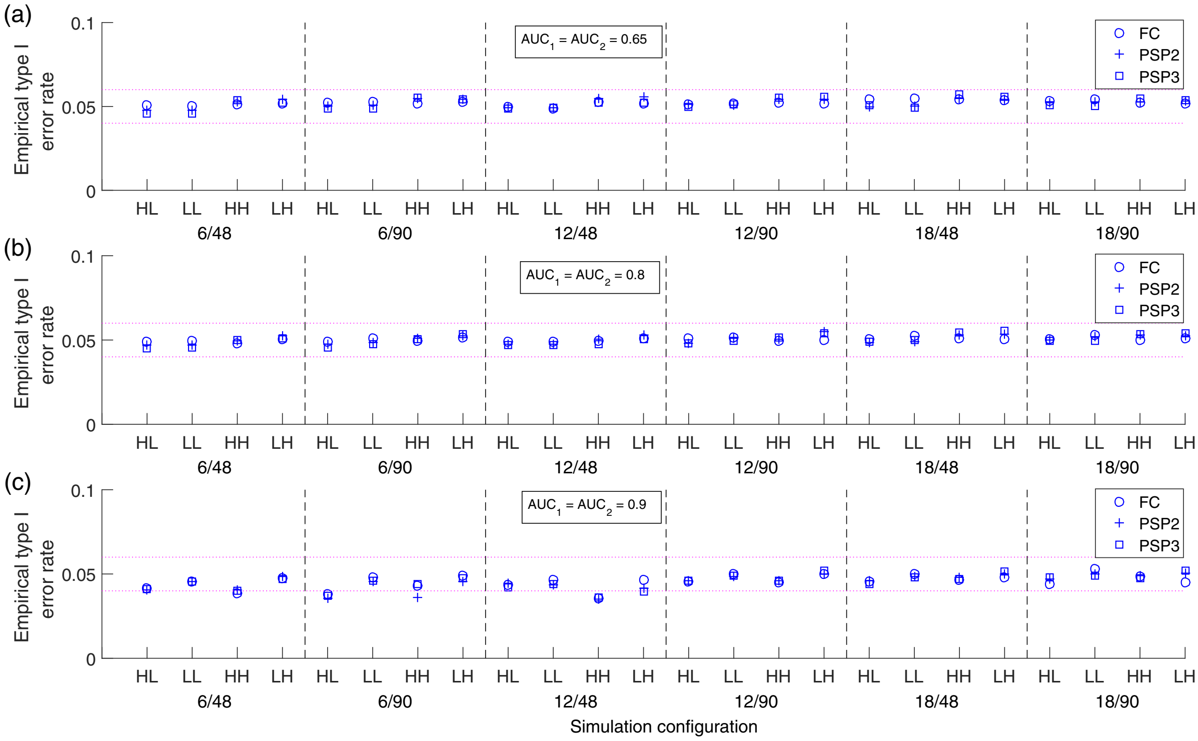


Fig. 3 Validation of the iMRMC software for analyzing MRMC data.

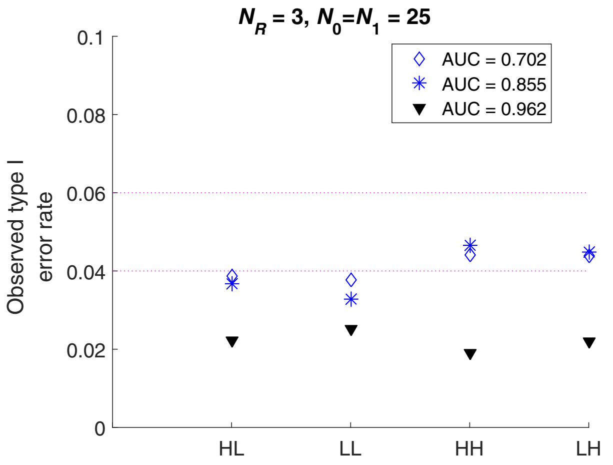


Fig. 4 Small sample size performance of the iMRMC software: FC design.

$$\mathrm{Var}_U \widehat{\Delta A} = \frac{1}{N_R} V_R^\Delta + V_C^\Delta, \tag{15}$$

where N_R is the number of readers $V_R^\Delta \equiv c_1(M_1^\Delta - M_5^\Delta) + c_2(M_2^\Delta - M_6^\Delta) + c_3(M_3^\Delta - M_7^\Delta) + c_4(M_4^\Delta - M_8^\Delta)$, $V_C^\Delta \equiv c_1 M_5^\Delta + c_2 M_6^\Delta + c_3 M_7^\Delta - (1 - c_4) M_8^\Delta$, and $M_k^\Delta = M_k^{(1)} + M_k^{(2)} - 2M_k^{(1\times 2)}$.

The estimate of variance of $\widehat{\Delta A}$ can be expressed in terms of OR parameters as[6]

$$\widehat{\mathrm{Var}_{\mathrm{OR}}(\widehat{\Delta A})} = \frac{2}{N_R}\mathrm{MS}(T*R) + 2\mathrm{Max}(\widehat{\mathrm{COV}}_2 - \widehat{\mathrm{COV}}_3, 0), \tag{16}$$

where $\mathrm{MS}(T*R)$ is modality × reader mean squares, COV_2 is the between-reader within-modality covariance of AUC, COV_3 is the between-reader between-modality covariance of AUC, and Max is a constraint of setting $\widehat{\mathrm{COV}}_2 - \widehat{\mathrm{COV}}_3$ to zero if it is negative.

Note that the U-statistic variance expression [Eq. (15)] is for population parameters, whereas the OR expression [Eq. (16)] is an estimator. The main purpose of this appendix is to show the expectation ("E") of the latter equals to the former, i.e., $E[\widehat{\mathrm{Var}_{\mathrm{OR}}(\widehat{\Delta A})}] = \mathrm{Var}_U \widehat{\Delta A}$. More specifically, we will show that

$$V_R^\Delta = 2E[\mathrm{MS}(T*R)] \tag{17}$$

and

$$V_C^\Delta = 2(\mathrm{COV}_2 - \mathrm{COV}_3). \tag{18}$$

Hillis[6] has shown that

$$E[\mathrm{MS}(T*R)] = \sigma_{TR}^2 + \sigma_\epsilon^2 - \mathrm{COV}_1 - \mathrm{COV}_2 + \mathrm{COV}_3, \tag{19}$$

where σ_{TR}^2 is the variance of the modality × reader interaction term in the OR model, σ_ϵ^2 is the expected variance of a fixed-reader AUC, and COV_1 is the within-reader between-modality covariance of AUC. Next, we express each of the parameters on the r.h.s of Eq. (19) in terms of U-statistic parameters.

The empirical fixed-reader AUC is

$$\hat{A}_r = \frac{\Sigma_{i=1}^{N_0} \Sigma_{j=1}^{N_1} s(x_{ir}, y_{jr})}{N_0 N_1},$$

and its U-statistic variance is given by Refs. 9 and 14 $\mathrm{Var}(\hat{A}_r|r) = c_1 M_{1|r} + c_2 M_{2|r} + c_3 M_{3|r} + (c_4 - 1) M_{4|r}$, where

$M_{k|r}$ is the M_k conditional on reader r. For example, $M_{1|r} = E[s(X_{ir}, Y_{jr})^2|r]$. Then M_k is the expectation of $M_{k|r}$ over the reader population, i.e., $M_k = E_r(M_{k|r})$. In the two-modality FC setting, $\sigma_\epsilon^2 = (E_r\{\text{Var}[\widehat{A}_r^{(1)}|r]\} + E_r\{\text{Var}[\widehat{A}_r^{(2)}|r]\})/2$. So we have

$$\sigma_\epsilon^2 = \frac{1}{2}\Sigma_{i=1}^2[c_1M_1^{(i)} + c_2M_2^{(i)} + c_3M_3^{(i)} + (c_4 - 1)M_4^{(i)}]. \tag{20}$$

Similarly, we have

$$\text{COV}_1 = c_1M_1^{(1\times2)} + c_2M_2^{(1\times2)} + c_3M_3^{(1\times2)} + (c_4 - 1)M_4^{(1\times2)}, \tag{21}$$

$$\text{COV}_2 = \frac{1}{2}\Sigma_{i=1}^2[c_1M_5^{(i)} + c_2M_6^{(i)} + c_3M_7^{(i)} + (c_4 - 1)M_8^{(i)}], \tag{22}$$

and

$$\text{COV}_3 = c_1M_5^{(1\times2)} + c_2M_6^{(1\times2)} + c_3M_7^{(1\times2)} + (c_4 - 1)M_8^{(1\times2)}. \tag{23}$$

To express σ_{TR}^2 in terms of U-statistic parameters, we first apply the conditional variance identity theorem[24]

$$\text{Var}[\hat{A}_r^{(1)} - \hat{A}_r^{(2)}] = E\{\text{Var}[\hat{A}_r^{(1)} - \hat{A}_r^{(2)}]|r\} + \text{Var}\{E[\hat{A}_r^{(1)} - \hat{A}_r^{(2)}|r]\},$$

to the OR model, then we have $\text{Var}\{E[\hat{A}_r^{(1)} - \hat{A}_r^{(2)}|r]\} = 2\sigma_{TR}^2$. On the other hand, based on U-statistics, $\text{Var}\{E[\hat{A}_r^{(1)} - \hat{A}_r^{(2)}|r]\} = M_4^\Delta - M_8^\Delta$. So we have

$$\sigma_{TR}^2 = \frac{1}{2}(M_4^\Delta - M_8^\Delta). \tag{24}$$

By inserting Eqs. (20)–(24) back to Eqs. (17)–(19), one can find Eqs. (17) and (18) hold.

Disclosures

No conflicts of interest, financial or otherwise, are declared by the authors.

Acknowledgments

Part of this work was presented at the SPIE Medical Imaging Symposium (Houston, Texas, February 12, 2018).[25]

References

1. B. D. Gallas et al., "Evaluating imaging and computer-aided detection and diagnosis devices at the FDA," *Acad. Radiol.* **19**, 463–477 (2012).
2. R. F. Wagner, C. E. Metz, and G. Campbell, "Assessment of medical imaging systems and computer aids: a tutorial review," *Acad. Radiol.* **14**(6), 723–748 (2007).
3. N. A. Obuchowski, "Multireader receiver operating characteristic studies: a comparison of study designs," *Acad. Radiol.* **2**(8), 709–716 (1995).
4. N. A. Obuchowski, "Reducing the number of reader interpretations in MRMC studies," *Acad. Radiol.* **16**, 209–217 (2009).
5. N. Obuchowski, B. D. Gallas, and S. L. Hillis, "Multi-reader ROC studies with split-plot designs: a comparison of statistical methods," *Acad. Radiol.* **19**, 1508–1517 (2012).
6. S. L. Hillis, "A marginal-mean ANOVA approach for analyzing multireader multicase radiological imaging data," *Stat. Med.* **33**, 330–360 (2014).
7. B. D. Gallas et al., "Impact of different study populations on reader behavior and performance metrics: initial results," *Proc. SPIE* **10136**, 101360A (2017).
8. B. D. Gallas, "iMRMC v4.0.0 application for analyzing and sizing MRMC reader studies," Division of Imaging, Diagnostics, and Software Reliability, OSEL/CDRH/FDA, Silver Spring, Maryland, 2017, https://github.com/DIDSR/iMRMC/releases.
9. B. D. Gallas, "One-shot estimate of MRMC variance: AUC," *Acad. Radiol.* **13**(3), 353–362 (2006).
10. H. H. Barrett, M. A. Kupinski, and E. Clarkson, "Probabilistic foundations of the MRMC method," *Proc. SPIE* **5749**, 21–31 (2005).
11. E. Clarkson, M. A. Kupinski, and H. H. Barrett, "A probabilistic model for the MRMC method. Part 1. Theoretical development," *Acad. Radiol.* **13**(11), 1410–1421 (2006).
12. M. A. Kupinski, E. Clarkson, and H. H. Barrett, "A probabilistic model for the MRMC method, part 2: validation and applications," *Acad. Radiol.* **13**, 1422–1430 (2006).
13. B. D. Gallas et al., "A framework for random-effects ROC analysis: biases with the bootstrap and other variance estimators," *Commun. Stat.—Theory Methods* **38**(15), 2586–2603 (2009).
14. W. Chen, B. D. Gallas, and W. A. Yousef, "Classifier variability: accounting for training and testing," *Pattern Recognit.* **45**, 2661–2671 (2012).
15. C. A. Roe and C. E. Metz, "Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic (ROC) data: validation with computer simulation," *Acad. Radiol.* **4**, 298–303 (1997).
16. B. D. Gallas and S. L. Hillis, "Generalized Roe and Metz receiver operating characteristic model: analytic link between simulated decision scores and empirical AUC variances and covariances," *J. Med. Imaging* **1**(3), 031006 (2014).
17. B. D. Gallas, "iRoeMetz v2.1: application for modeling and simulating MRMC reader studies," Division of Imaging, Diagnostics, and Software Reliability, OSEL/CDRH/FDA, Silver Spring, Maryland, 2013, https://github.com/DIDSR/iMRMC/releases (14 May 2018).
18. S. L. Hillis, "Relationship between Roe and Metz simulation model for multireader diagnostic data and Obuchowski–Rockette model parameters," *Stat. Med.* **37**(13), 2067–2093 (2018).
19. E. D. Pisano et al., "Diagnostic performance of digital versus film mammography for breast-cancer screening," *N. Engl. J. Med.* **353**(17), 1773–1783 (2005).
20. S. L. Hillis, K. M. Schartz, and K. S. Berbaum, "OR-DBM MRMC Software version 2.5.," 2014, http://perception.radiology.uiowa.edu/ (14 May 2018).
21. N. A. Obuchowski and H. E. Rockette, "Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: an ANOVA approach with dependent observations," *Commun. Stat. Simul. Comput.* **24**(2), 285–308 (1995).
22. S. L. Hillis, K. S. Berbaum, and C. E. Metz, "Recent developments in the Dorfman-Berbaum-Metz procedure for multireader ROC study analysis," *Acad. Radiol.* **15**, 647–661 (2008).
23. B. D. Gallas and D. G. Brown, "Reader studies for validation of CAD systems," *Neural Networks* **21**(2–3), 387–397 (2008).
24. G. Casella and R. L. Berger, *Statistical Inference*, Duxbury Advanced Series, 2nd ed., Duxbury/Thomson Learning, Pacific Grove, California (2002).
25. W. Chen, Q. Gong, and B. D. Gallas, "Efficiency gain of paired split-plot designs in MRMC ROC studies," *Proc. SPIE* **10577**, 105770F (2018).

Weijie Chen received his PhD in medical physics from the University of Chicago, Chicago, Illinois, in 2007. Since then, he has been a scientist at the Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland. His research

interests include statistical assessment methodology for diagnostic devices in general, and ROC methodology and reader studies for medical imaging and computer-aided diagnosis in particular.

Qi Gong is a research fellow at the Center for Devices and Radiological Health, U.S. Food and Drug Administration, Silver Spring, Maryland. His research areas include medical image processing, reader studies statistical analysis, and programming. He received his Master in science degree in electrical engineering from the George Washington University in 2014.

Brandon D. Gallas provides mathematical, statistical, and modeling expertise to the evaluation of medical imaging devices at the FDA. His main areas of contribution are in the design and statistical analysis of reader studies, image quality, computer-aided diagnosis (CAD), and imaging physics. Before working at the FDA, he was in Dr. Harrison Barrett's research group at the University of Arizona, earning his PhD from the Graduate Interdisciplinary Program in Applied Mathematics.