Contents

1 Standalone CAD vs. Radiologists

1.1 Abstract

Computer aided detection (CAD) research for screening mammography has so far focused on measuring performance of radiologists with and without CAD. Standalone performance of CAD algorithms is rarely measured. One reason is the lack of clear methodology for comparing CAD to radiologists interpreting the same cases. This work extends the method used in a recent study of standalone performance. The method is termed random-reader fixed case (1T-RRFC), since it only accounts for reader variability. The extension includes the effect of case-sampling variability. Since CAD is treated as an additional reader, the method is termed one-treatment random-reader random-case (1T-RRRC) analysis. The new method is based on existing methodology that allows comparing the average performance of readers in a single treatment to a constant value. The key modification is to regard the difference in performance between radiologists and CAD as a figure of merit, to which the existing work is directly applicable. The 1T-RRRC method was compared to 1T-RRFC and to an unorthodox usage of conventional ROC (receiver operating characteristic) analysis software, termed 2T-RRRC analysis, which involves replicating the CAD ratings as many times as there are radiologists, to simulate a second treatment, i.e., CAD is regarded as the second treatment. 1T-RRRC analysis has 3 random parameters as compared to 6 parameters in 2T-RRRC and one parameter in 1T-RRFC. As expected, since one is including an additional source of variability, both RRRC analyses (1T and 2T) yielded larger p-values and wider confidence intervals as compared to 1T-RRFC. For the F-statistic, degrees of freedom and p-value, both 1T-RRRC and 2T-RRRC analyses yielded exactly the same results. However, 2T-RRRC model parameter estimates were unrealistic; for example, it yielded zero for between-reader variance, whereas 1T-RRRC yielded the expected non-zero value, identical to that yielded by 1T-RRFC. The method is implemented in an open-source `R` package `RJafroc.`

1.2 Keywords

Technology assessment, computer-aided detection (CAD), screening mammography, standalone performance, single-treatment multi-reader ROC analysis.

1.3 Introduction

In the US the majority of screening mammograms are analyzed by computer aided detection (CAD) algorithms [@rao2010widely]. Almost all major imaging device manufacturers provide CAD as part of their imaging workstation display software. In the United States CAD is approved for use as a second reader [@fda2018guidance], i.e., the radiologist first interprets the images (typically 4 views, 2 views of each breast) without CAD and then CAD information (i.e., cued suspicious regions, possibly shown with associated probabilities of malignancies) is shown and the radiologist has the opportunity to revise the initial interpretation. In response to the second reader usage, the evolution of CAD algorithms has been guided mainly by comparing observer performance of radiologists with and without CAD.

Clinical CAD systems sometimes only report the locations of suspicious regions, i.e., it may not provide ratings. However, a (continuous variable) malignancy index for every CAD-found suspicious region is available to the algorithm designer [@edwards2002maximum]. Standalone performance, i.e., performance of designer-level CAD by itself, regarded as an algorithmic reader, vs. radiologists, is rarely measured. In breast cancer screening the authors are aware of only one study [@hupse2013standalone] where standalone performance was measured. [Standalone performance has been measured in CAD for computed tomography colonography, chest radiography and three dimensional ultrasound [@hein2010computeraided; @summers2008performance; @taylor2006computerassisted; @deBoo2011computeraided; @tan2012computeraided]].

One possible reason for not measuring standalone performance of CAD is the lack of an accepted assessment methodology for such measurements. The purpose of this work is to remove that impediment. It describes a method for comparing standalone performance of designer-level CAD to radiologists interpreting the same cases and compares the method to those described in two recent publications [@hupse2013standalone; @kooi2016comparison].

1.4 Methods

Summarized are two recent studies of CAD vs. radiologists in mammography. This is followed by comments on the methodologies used in the two studies. The second study used multi-treatment multi-reader receiver operating characteristic (ROC) software in an unorthodox or unconventional way. A statistical model and analysis method is described that avoids unorthodox, and perhaps unjustified, use of ROC software and has fewer model parameters.

1.4.1 Studies assessing performance of CAD vs. radiologists

The first study [@hupse2013standalone] measured performance in finding and localizing lesions in mammograms, i.e., visual search was involved, while the second study [@kooi2016comparison] measured lesion classification performance between non-diseased and diseased regions of interest (ROIs) previously found on mammograms by an independent algorithmic reader, i.e., visual search was not involved.

1.4.1.1 Study - 1 The first study [@hupse2013standalone] compared standalone performance of a CAD device to that of 9 radiologists interpreting the same cases (120 non-diseased and 80 with a single malignant mass per case). It used the LROC (localization ROC) paradigm [@starr1975visual; @metz1976observer; @swensson1996unified], in which the observer gives an overall rating for presence of disease (an integer 0 to 100 scale was used) and indicates the location of the most suspicious region. On a non-diseased case the rating is classified as a false positive (FP) but on a diseased case it is classified as a *correct localization* (CL) if the location is sufficiently close to the lesion, and otherwise it is classified as an *incorrect localization*. For a given reporting threshold, the number of correct localizations divided by the number of diseased cases estimates the probability of correct localization (PCL) at that threshold. On non-diseased cases the number of false positives (FPs) divided by the number of non-diseased cases estimates the probability of a false positive, or false positive fraction (FPF), at that threshold. The plot of PCL (ordinate) vs. FPF defines the LROC curve. Study - 1 used as figures of merit (FOMs) the interpolated PCL at two values of FPF, specifically FPF = 0.05 and FPF = 0.2, denoted $\text{PCL}_{0.05}$ and $\text{PCL}_{0.2}$, respectively. The t-test between the radiologist PCL_{FPF} values and that of CAD was used to compute the two-sided p-value for rejecting the NH of equal performance. Study - 1 reported p-value = 0.17 for $\text{PCL}_{0.05}$ and p-value ≤ 0.001, with CAD being inferior, for $\text{PCL}_{0.2}$.

Table 1: The differences between the data structures in conventional DBM-MRMC analysis and the unorthodox application of the software used in Study - 2. There are four radiologists, labeled R1, R2, R3 and R4 interpreting 398 cases labeled 1, 2, . . . , 398, in two treatments, labeled 1 and 2. Sample ratings are shown only for the first and last radiologist and the first and last case. In the first four columns, labeled "Standard DBM-MRMC", each radiologist interprets each case twice. In the next four columns, labeled "Unorthodox DBM-MRMC", the radiologists interpret each case once. CAD ratings are replicated four times to effectively create the second "treatment". The quotations emphasize that there is, in fact, only one treatment. The replicated CAD observers are labeled C1, C2, C3 and C4.

Standard DBM-MRMC				Unorthodox DBM-MRMC			
Reader	Treatment	Case	Rating	Reader	Treatment	Case	Rating
R1	1	1	75	R1	1	1	75
...	...	...	...	...	...	...	...
R1	1	398	0	R1	1	398	0
...	...	...	...	...	...	...	...
R4	1	1	50	R4	1	1	50
...	...	...	...	...	...	...	...
R4	1	398	25	R4	1	398	25
R1	2	1	45	C1	2	1	55
...	...	...	...	...	...	...	...
R1	2	398	25	C1	2	398	5
...	...	...	...	...	...	...	...
R4	2	1	95	C4	2	1	55
...	...	...	...	...	...	...	...
R4	2	398	20	C4	2	398	5

1.4.1.2 Study - 2 The second study [@kooi2016comparison] used 199 diseased and 199 non-diseased ROIs extracted by an independent CAD algorithm. These were interpreted using the ROC paradigm (i.e., rating only, no localization required) by a different CAD algorithmic observer from that used to determine the ROIs, and by four expert radiologists. The figure of merit was the empirical area (AUC) under the respective ROC curves (one per radiologist and one for CAD). The p-value for the difference in AUCs between the average radiologist and CAD was determined using an unorthodox application of the Dorfman-Berbaum-Metz [@dorfman1992receiver] multiple-treatment multiple-reader multiple-case (DBM-MRMC) software with recent modifications [@hillis2008recent]. The unorthodox application was that in the input data file *radiologists and CAD were entered as two treatments.* In conventional (or orthodox) DBM-MRMC each reader provides two ratings per case and the data file would consist of paired ratings of a set of cases interpreted by 4 readers. To accommodate the paired data structure assumed by the software, the authors of Study - 2 *replicated the CAD ratings four times in the input data file*, as explained in the caption to Table 1. By this artifice they converted a single-treatment 5-reader (4 radiologists plus CAD) data file to a two-treatment 4-reader data file, in which the four readers in treatment 1 were the radiologists, and the four “readers” in treatment 2 were CAD replicated ratings. Note that for each case the four readers in the second treatment had identical ratings. In Table 1 the replicated CAD observers are labeled C1, C2, C3 and C4.

Study − 2 reported a not significant difference between CAD and the radiologists (p = 0.253).

1.4.1.3 Comments For the purpose of this work, which focuses on the respective analysis methods, the difference in observer performance paradigms between the two studies, namely a search paradigm in Study - 1 vs. an ROI classification paradigm in Study − 2, is inconsequential. The paired t-test used in Study - 1 treats the case-sample as fixed. In other words, the analysis is not accounting for case-sampling variability but it is accounting for reader variability. While not explicitly stated, the reason for the unorthodox analysis in Study – 2 was the desire to include case-sampling variability. [1]

In what follows, the analysis in Study – 1 is referred to as random-reader fixed-case (1T-RRFC) while that in Study – 2 is referred to as dual-treatment random-reader random-case (2T-RRRC).

1.4.2 The 2T-RRRC analysis model

This could be termed the conventional or the orthodox method. There are two treatments and the study design is fully crossed: each reader interprets each case in each treatment. The method is termed 2T-RRRC analysis. The following approach uses the Obuchowski and Rockette (OR) figure of merit model [@obuchowski1995hypothesis] for analyzing such studies, instead of the pseudovalue model used in the original DBM paper [@dorfman1992receiver]. Hillis has shown the two to be equivalent [@hillis2005comparison]. For fully crossed multiple-treatment multiple-reader interpretations (i.e., assuming the data structure in the left half of Table 1) the OR model is:

$$\theta_{ij\{c\}} = \mu + \tau_i + (\tau \mathrm{R})_{ij} + \epsilon_{ij\{c\}} \tag{1}$$

Assuming two treatments, i ($i = 1, 2$) is the treatment index, j ($j = 1, ..., J$) is the reader index, and k ($k = 1, ..., K$) is the case index, and $\theta_{ij\{c\}}$ is a figure of merit for reader j in treatment i and case-sample $\{c\}$. A case-sample is a set or ensemble of cases, diseased and non-diseased, and different integer values of c correspond to different case-samples. The first two terms on the right hand side of Eqn. (1) are fixed effects (average performance and treatment effect, respectively). The next two terms are random effect variables that, by assumption, are sampled as follows:

$$\left.\begin{aligned} R_j &\sim N\left(0, \sigma_R^2\right) \\ (\tau R)_{ij} &\sim N\left(0, \sigma_{\tau R}^2\right) \end{aligned}\right\} \tag{2}$$

The terms R_j represents the random treatment-independent contribution of reader j, modeled as a sample from a zero-mean normal distribution with variance σ_R^2, $(\tau R)_{ij}$ represents the random treatment-dependent contribution of reader j in treatment i, modeled as a sample from a zero-mean normal distribution with variance $\sigma_{\tau R}^2$. The sampling of the last (error) term is described by:

$$\epsilon_{ij\{c\}} \sim N_{I \times J}\left(\vec{0}, \Sigma\right) \tag{3}$$

Here $N_{I \times J}$ is the $I \times J$ variate normal distribution and $\vec{0}$, a $I \times J$ length zero-vector, represents the mean of the distribution. The $\{I \times J\}\ times \{I \times J\}$ dimensional covariance matrix Σ is defined by 4 parameters, Var, Cov_1, Cov_2, Cov_3, defined as follows:

$$\mathrm{Cov}\left(\epsilon_{ij\{c\}}, \epsilon_{i'j'\{c\}}\right) = \begin{Bmatrix} \mathrm{Var}\ (i = i', j = j') \\ \mathrm{Cov1}\ (i \neq i', j = j') \\ \mathrm{Cov2}\ (i = i', j \neq j') \\ \mathrm{Cov3}\ (i \neq i', j \neq j') \end{Bmatrix} \tag{4}$$

Software {U of Iowa and `RJafroc`} yields estimates of all terms appearing on the right hand side of Eqn. (4). Excluding fixed effects, the model represented by Eqn. (1) contains six parameters:

$$\sigma_R^2, \sigma_{\tau R}^2, \mathrm{Var}, \mathrm{Cov}_1, \mathrm{Cov}_2, \mathrm{Cov}_3 \tag{5}$$

The meanings the last four terms are described in [@hillis2007comparison; @obuchowski1995hypothesis; @hillis2005comparison; @chakraborty2017observer]. Briefly, Var is the variance of a reader's FOMs, in

[1] Prof. Karssemeijer (private communication, 10/27/2017) had consulted with a few ROC experts to determine if the procedure used in Study – 2 was valid, and while the experts thought it was probably valid they were not sure.

a given treatment, over interpretations of different case-samples, averaged over readers and treatments; Cov_1/Var is the correlation of a reader's FOMs, over interpretations of different case-samples in different treatments, averaged over all different-treatment same-reader pairings; Cov_2/Var is the correlation of different reader's FOMs, over interpretations of different case-samples in the same treatment, averaged over all same- treatment different-reader pairings and finally, Cov_3/Var is the correlation of different reader's FOMs, over interpretations of different case-samples in different treatments, averaged over all different-treatment different-reader pairings. One expects the following inequalities to hold:

$$\text{Var} \geq \text{Cov}_1 \geq \text{Cov}_2 \geq \text{Cov}_3 \tag{6}$$

In practice, since one is usually limited to one case-sample, i.e., $c = 1$, resampling techniques [@efron1994introduction] – e.g., the jackknife – are used to estimate these terms.

1.4.3 The (1T-RRRC) analysis model

In this work the key difference from the approach in Study - 2 is to regard standalone CAD as a different reader, not as a different treatment. Therefore, needed is a single treatment method for analyzing readers and CAD, where the latter is regarded as an additional reader. Accordingly the proposed method is termed single-treatment RRRC (1T-RRRC) analysis.

The starting point is the [@obuchowski1995hypothesis] model for a single treatment, which for the radiologists (i.e., *excluding* CAD) interpreting in a single-treatment reduces to the following model:

$$\theta_{j\{c\}} = \mu + R_j + \epsilon_{j\{c\}} \tag{7}$$

$\theta_{j\{c\}}$ is the figure of merit for radiologist j ($j = 1, 2, ..., J$) interpreting case-sample $\{c\}$; R_j is the random effect of radiologist j and $\epsilon_{j\{c\}}$ is the error term. For single-treatment multiple-reader interpretations the error term is distributed as:

$$\epsilon_{j\{c\}} \sim N_J\left(\vec{0}, \Sigma\right) \tag{8}$$

The $J \times J$ covariance matrix Σ is defined by two parameters, Var and Cov_2, as follows:

$$\Sigma_{jj'} = \text{Cov}\left(\epsilon_{j\{c\}}, \epsilon_{j'\{c\}}\right) = \begin{cases} \text{Var} & j = j' \\ \text{Cov}_2 & j \neq j' \end{cases} \tag{9}$$

The terms Var and Cov_2 are estimated using resampling methods. Using the jackknife, and denoting the difference FOM with case k removed by $\psi_{j(k)}$ (the index in parenthesis denotes deleted case k, and since one is dealing with a single case-sample, the case-sample index c is superfluous). The covariance matrix is estimated using (the dot symbol represents an average over the replaced index):

$$\Sigma_{jj'}|_{\text{jack}} = \frac{K-1}{K} \sum_{k=1}^{K} \left(\psi_{j(k)} - \psi_{j(\bullet)}\right) \left(\psi_{j'(k)} - \psi_{j'(\bullet)}\right) \tag{10}$$

The final estimates of Var and Cov_2 are averaged (indicated in the following equation by the angular brackets) over all pairings of radiologists satisfying the relevant equalities/inequalities shown just below the closing angular bracket:

$$\left.\begin{aligned} \text{Var} &= \left\langle \Sigma_{jj'}|_{\text{jack}} \right\rangle_{j=j'} \\ \text{Cov}_2 &= \left\langle \Sigma_{jj'}|_{\text{jack}} \right\rangle_{j \neq j'} \end{aligned}\right\} \tag{11}$$

Hillis' formulae [@hillis2005comparison; @hillis2007comparison] permit one to test the NH: $\mu = \mu_0$, where μ_0 is a pre-specified constant. One could set μ_0 equal to the performance of CAD, but that would not be accounting for the fact that the performance of CAD is itself a random variable, whose case-sampling variability needs to be accounted for.

Instead, the following model was used for the figure of merit of the radiologists and CAD ($j = 0$ is used to denote the CAD algorithmic reader):

$$\theta_{j\{c\}} = \theta_{0\{c\}} + \Delta\theta + R_j + \epsilon_{j\{c\}} j = 1, 2, ...J \tag{12}$$

$\theta_{0\{c\}}$ is the CAD figure of merit for case-sample $\{c\}$; $\Delta\theta$ is the average figure of merit increment of the radiologists over CAD. To reduce this model to one to which existing formulae are directly applicable, one subtracts the CAD figure of merit from each radiologist's figure of merit (for the same case-sample), and defines this as the difference figure of merit $\psi_{j\{c\}}$, i.e.,

$$\psi_{j\{c\}} = \theta_{j\{c\}} - \theta_{0\{c\}} \tag{13}$$

Then Eqn. (12) reduces to:

$$\psi_{j\{c\}} = \Delta\theta + R_j + \epsilon_{j\{c\}} j = 1, 2, ...J \tag{14}$$

Eqn. (14) is identical in form to Eqn. (7) with the difference that the figure of merit on the left hand side of Eqn. (14) is a difference FOM, that between the radiologist's and CAD. Eqn. (14) describes a model for J difference radiologists interpreting a common case set, each of whose performances is measured relative to that of CAD. Under the NH the expected difference is zero: NH:$\Delta\theta = 0$. The method [@hillis2005comparison; @hillis2007comparison] for single-treatment multiple-reader analysis is now directly applicable to the model described by Eqn. (14).

Apart from fixed effects, the model in Eqn. (14) contains three parameters:

$$\sigma_R^2, \mathrm{Var}, \mathrm{Cov}_2 \tag{15}$$

Setting $\mathrm{Var} = 0, \mathrm{Cov}_2 = 0$ yields the 1T-RRFC model, which contains only one random parameter, namely σ_R^2. [One expects identical estimates of σ_R^2 using 1T-RRFC, 2T-RRRC or 1T-RRRC analyses.]

1.5 Computational details

The three analyses, namely random-reader fixed-case (1T-RRFC), dual-treatment random-reader random-case (2T-RRRC) and single-treatment random-reader random-case (1T-RRRC), are implemented in `RJafroc`, an R-package [@packageRJafroc].

The following code shows usage of the three analyses. Note that `datasetCadLroc` is the LROC dataset and `dataset09` is the corresponding ROC dataset.

```
RRFC_1T_PCL_0_05 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.05, method = "1T-RRFC")
RRRC_2T_PCL_0_05 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.05, method = "2T-RRRC")
RRRC_1T_PCL_0_05 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.05, method = "1T-RRRC")
```

```
RRFC_1T_PCL_0_2 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.2, method = "1T-RRFC")
RRRC_2T_PCL_0_2 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.2, method = "2T-RRRC")
RRRC_1T_PCL_0_2 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 0.2, method = "1T-RRRC")

RRFC_1T_PCL_1 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 1, method = "1T-RRFC")
RRRC_2T_PCL_1 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 1, method = "2T-RRRC")
RRRC_1T_PCL_1 <- StSignificanceTestingCadVsRad (datasetCadLroc,
FOM = "PCL", FPFValue = 1, method = "1T-RRRC")

RRFC_1T_AUC <- StSignificanceTestingCadVsRad (dataset09,
FOM = "Wilcoxon", method = "1T-RRFC")
RRRC_2T_AUC <- StSignificanceTestingCadVsRad (dataset09,
FOM = "Wilcoxon", method = "2T-RRRC")
RRRC_1T_AUC <- StSignificanceTestingCadVsRad (dataset09,
FOM = "Wilcoxon", method = "1T-RRRC")
```

The results are organized as follows:

- `RRFC_1T_PCL_0_05` is a list containing the results of 1T-RRFC analysis for figure of merit $= PCL_{0.05}$.
- `RRRC_2T_PCL_0_05` is a list containing the results of 2T-RRFC analysis for figure of merit $= PCL_{0.05}$.
- `RRRC_1T_PCL_0_05` is a list containing the results of 1T-RRFC analysis for figure of merit $= PCL_{0.05}$.
- `RRFC_1T_PCL_0_2` is a list containing the results of 1T-RRFC analysis for figure of merit $= PCL_{0.2}$.
- `RRRC_2T_PCL_0_2` is a list containing the results of 2T-RRRC analysis for figure of merit $= PCL_{0.2}$.
- `RRRC_1T_PCL_0_2` is a list containing the results of 1T-RRRC analysis for figure of merit $= PCL_{0.2}$.
- `RRFC_1T_AUC` is a list containing the results of 1T-RRFC analysis for the Wilcoxon figure of merit.
- `RRRC_2T_AUC` is a list containing the results of 2T-RRRC analysis for the Wilcoxon figure of merit.
- `RRRC_1T_AUC` is a list containing the results of 1T-RRRC analysis for the Wilcoxon figure of merit.

The structures of these objects are illustrated next with three examples.

1.5.1 The first example shows the structure of `RRFC_1T_PCL_0_2`.

```
print(fom_individual_rad)
#>         rdr1 rdr2    rdr3  rdr4       rdr5       rdr6   rdr7  rdr8  rdr9
#> 1 0.69453125 0.65 0.80625 0.725 0.65982143 0.76845238 0.7375 0.675 0.675
print(stats)
#>       fomCAD  avgRadFom avgDiffFom        varR     Tstat df          pval
#> 1 0.59166667 0.71017278 0.11850612 0.002808612 6.7083568  8 0.00015139664
print(ConfidenceIntervals)
#>       CIAvgRadFom CIAvgDiffFom
#> Lower  0.66943619  0.077769525
#> Upper  0.75090938  0.159242710
```

The results are displayed as three data frames.

The first data frame :

- `fom_individual_rad` shows the figures of merit for the nine radiologists in the study.

The next data frame summarizes the statistics.

- `fomCAD` is the figure of merit for CAD.
- `avgRadFom` is the average figure of merit of the nine radiologists in the study.
- `avgDiffFom` is the average difference figure of merit, RAD - CAD.
- `varR` is the variance of the figures of merit for the nine radiologists in the study.
- `Tstat` is the t-statistic for testing the NH that the average difference FOM `avgDiffFom` is zero.
- `df` is the degrees of freedom of the t-statistic, whose square is the F-statistic.
- `pval` is the p-value for rejecting the NH. In the example shown below the value is highly signficant.

The last data frame summarizes the 95 percent confidence intervals.

- `CIAvgRadFom` is the 95 percent confidence interval, listed as pairs `Lower`, `Upper`, for `avgRadFom`.
- `CIAvgDiffFom` is the 95 percent confidence interval for `avgDiffFom`.
- If the pair `CIAvgDiffFom` excludes zero, the difference is statistically significant.
- In the example the interval excludes zero showing that the FOM difference is significant.

1.5.2 The next example shows the structure of `RRRC_2T_PCL_0_2`.

```
print(fom_individual_rad)
#>         rdr1 rdr2    rdr3  rdr4       rdr5       rdr6   rdr7  rdr8  rdr9
#> 1 0.69453125 0.65 0.80625 0.725 0.65982143 0.76845238 0.7375 0.675 0.675
print(stats1)
#>       fomCAD  avgRadFom avgDiffFom
#> 1 0.59166667 0.71017278 0.11850612
print(stats2)
#>             varR         varTR          cov1         cov2          cov3
#> 1 -7.5894152e-19 0.00026488983 0.00076136841 0.0022942211 0.00076136841
#>            Var     FStat        df        pval
#> 1 0.0034336373 4.1576797 937.24371 0.041726262
```

In addition to the quantities defined previously, the output contains the covariance matrix for the Obuchowski-Rockette model, summarized in Eqn. (1) – Eqn. (4).

- `varTR` is $\sigma_{\tau R}^2$.
- `cov1` is Cov_1.
- `cov2` is Cov_2.
- `cov3` is Cov_3.
- `Var` is Var.
- `FStat` is the F-statistic for testing the NH.
- `ndf` is the numerator degrees of freedom, equal to unity.
- `df` is denominator degrees of freedom of the F-statistic for testing the NH.
- `Tstat` is the t-statistic for testing the NH that the average difference FOM `avgDiffFom` is zero.
- `pval` is the p-value for rejecting the NH. In the example shown below the value is signficant.

Notice that including the variability of cases results in a higher p-value for 2T-RRRC as compared to 1T-RRFC.

Shown next are the confidence interval statistics `x$ciAvgRdrEachTrt` for the two treatments ("trt1" = CAD, "trt2" = RAD):

```
print(x$ciAvgRdrEachTrt)
#>        Estimate      StdErr        DF    CILower    CIUpper         Cov2
#> trt1 0.59166667 0.058028349       Inf 0.47793319 0.70540014 0.0033672893
#> trt2 0.71017278 0.039156365 193.10832 0.63294372 0.78740185 0.0012211529
```

- `Estimate` contains the difference FOM estimate.
- `StdErr` contains the standard estimate of the difference FOM estimate.
- `DF` contains the degrees of freedom of the t-statistic.
- `t` contains the value of the t-statistic.
- `PrGtt` contains the probability of exceeding the magnitude of the t-statistic.
- `CILower` is the lower confidence interval for the difference FOM.
- `CIUpper` is the upper confidence interval for the difference FOM.

Shown next are the confidence interval statistics `x$ciDiffFom` between the two treatments ("trt1-trt2" = CAD - RAD):

```
print(x$ciDiffFom)
#>             Estimate      StdErr        DF         t       PrGTt     CILower
#> trt2-trt1 0.11850612 0.058118615 937.24371 2.0390389 0.041726262 0.004448434
#>             CIUpper
#> trt2-trt1 0.2325638
```

The difference figure of merit statistics are contained in a dataframe `x$ciDiffFom` with elements:

- `Estimate` contains the difference FOM estimate.
- `StdErr` contains the standard estimate of the difference FOM estimate.
- `DF` contains the degrees of freedom of the t-statistic.
- `t` contains the value of the t-statistic.
- `PrGtt` contains the probability of exceeding the magnitude of the t-statistic.
- `CILower` is the lower confidence interval for the difference FOM.
- `CIUpper` is the upper confidence interval for the difference FOM.

The figures of merit statistic for the two treatments, 1 is CAD and 2 is RAD.

- `trt1`: statistics for CAD.
- `trt2`: statistics for RAD.
- `Cov2`: Cov$_2$ calculated over individual treatments.

1.5.3 The last example shows the structure of `RRRC_1T_PCL_0_2`.

```
RRRC_1T_PCL_0_2
#> $fomCAD
#> [1] 0.59166667
#>
```

```
#> $fomRAD
#> [1] 0.69453125 0.65000000 0.80625000 0.72500000 0.65982143 0.76845238 0.73750000
#> [8] 0.67500000 0.67500000
#> 
#> $avgRadFom
#> [1] 0.71017278
#> 
#> $CIAvgRad
#> [1] 0.59611510 0.82423047
#> 
#> $avgDiffFom
#> [1] 0.11850612
#> 
#> $CIAvgDiffFom
#> [1] 0.004448434 0.232563801
#> 
#> $varR
#> [1] 0.002808612
#> 
#> $varError
#> [1] 0.0053445377
#> 
#> $cov2
#> [1] 0.0030657054
#> 
#> $Tstat
#>      rdr2 
#> 2.0390389 
#> 
#> $df
#>      rdr2 
#> 937.24371 
#> 
#> $pval
#>        rdr2 
#> 0.041726262
```

The differences from `RRFC_1T_PCL_0_2` are listed next:

- `varR` is σ_R^2 of the single treatment model for comparing CAD to RAD, Eqn. (15).
- `cov2` is Cov_2 of the single treatment model for comparing CAD to RAD.
- `varError` is Var of the single treatment model for comparing CAD to RAD.

Notice that the `RRRC_1T_PCL_0_2` p value, i.e., 0.04172626, is identical to that of `RRRC_2T_PCL_0_2`, i.e., 0.04172626.

1.6 Results

The three methods, in historical order 1T-RRFC, 2T-RRRC and 1T-RRRC, were applied to an LROC dataset similar to that used in Study – 1 (I thank Prof. Karssemeijer for making this dataset available).

Shown next, Table 2, are the significance testing results corresponding to the three analyses.

Table 2: Significance testing results of the analyses for an LROC dataset. Three sets of results, namely RRRC, 2T-RRRC and 1T-RRRC, are shown for each figure of merit (FOM). Because it is accounting for an additional source of variability, each of the rows labeled RRRC yields a larger p-value and wider confidence intervals than the corresponding row labeled 1T-RRFC. [θ_0 = FOM CAD; $\theta_\bullet$ = average FOM of radiologists; $\psi_\bullet$ = average FOM of radiologists minus CAD; CI= 95 percent confidence interval of quantity indicated by the subscript, F = F-statistic; ddf = denominator degrees of freedom; p = p-value for rejecting the null hypothesis: $\psi_\bullet = 0$.]

FOM	Analysis	θ_0	CI_{θ_0}	$\theta_\bullet$	$CI_{\theta_\bullet}$	$\psi_\bullet$	$CI_{\psi_\bullet}$	F	ddf	p
PCL_0_05	1T-RRFC	4.5e-01	0	4.93e-01	(4.18e-01,5.68e-01)	4.33e-02	(-3.16e-02,1.18e-01)	1.77e+00	8e+00	2.2e-01
	2T-RRRC		(2.58e-01,6.42e-01)		(3.76e-01,6.11e-01)		(-1.57e-01,2.44e-01)	1.79e-01	7.84e+02	6.7e-01
	1T-RRRC		NA		(2.93e-01,6.94e-01)					
PCL_0_2	1T-RRFC	5.92e-01	0	7.1e-01	(6.69e-01,7.51e-01)	1.19e-01	(7.78e-02,1.59e-01)	4.5e+01	8e+00	1.51e-04
	2T-RRRC		(4.78e-01,7.05e-01)		(6.33e-01,7.87e-01)		(4.45e-03,2.33e-01)	4.16e+00	9.37e+02	4.2e-02
	1T-RRRC		NA		(5.96e-01,8.24e-01)					
PCL_1	1T-RRFC	6.75e-01	0	7.83e-01	(7.4e-01,8.27e-01)	1.08e-01	(6.48e-02,1.52e-01)	3.3e+01	8e+00	4.33e-04
	2T-RRRC		(5.71e-01,7.79e-01)		(7.12e-01,8.54e-01)		(4.5e-03,2.12e-01)	4.2e+00	4.93e+02	4.1e-02
	1T-RRRC		NA		(6.8e-01,8.87e-01)					
Wilcoxon	1T-RRFC	8.17e-01	0	8.49e-01	(8.26e-01,8.71e-01)	3.17e-02	(8.96e-03,5.45e-02)	1.03e+01	8e+00	1.24e-02
	2T-RRRC		(7.52e-01,8.82e-01)		(8.07e-01,8.9e-01)		(-3.1e-02,9.45e-02)	9.86e-01	8.78e+02	3.2e-01
	1T-RRRC		NA		(7.86e-01,9.11e-01)					

Results are shown for the following FOMs: $\text{PCL}_{0.05}$, $\text{PCL}_{0.2}$, PCL_1, and the empirical area (AUC) under the ROC curve estimated by the Wilcoxon statistic. The first two FOMs are identical to those used in Study – 1. Columns 3 and 4 list the CAD FOM θ_0, and its 95% confidence interval CI_{θ_0}, columns 5 and 6 list the average radiologist FOM $\theta_\bullet$ (the dot symbol represents an average over the radiologist index) and its 95% confidence interval $CI_{\theta_\bullet}$, columns 7 and 8 list the average difference FOM $\psi_\bullet$, i.e., radiologist minus CAD, and its 95% confidence interval $CI_{\psi_\bullet}$, and the last three columns list the F-statistic, the denominator degrees of freedom (ddf) and the p-value for rejecting the null hypothesis. The numerator degree of freedom of the F-statistic, not listed, is unity.

In Table 2 identical values in adjacent cells in vertical columns have been replaced by the common values. The last three columns show that 2T-RRRC and 1T-RRRC analyses yield *identical F-statistics, ddf and p-values*. So the intuition of the authors of Study – 2, that the unorthodox method of using DBM – MRMC software to account for both reader and case-sampling variability, turns out to be correct. If interest is solely in these statistics one is justified in using the unorthodox method.

Commented on next are other aspects of the results evident in Table 2.

1. Where a direct comparison is possible, namely 1T-RRFC analysis using and as FOMs, the p-values in Table 2 are similar to those reported in Study – 1.
2. All FOMs (i.e., θ_0, $\theta_\bullet$ and $\psi_\bullet$) in Table 2 are independent of the method of analysis. However, the corresponding confidence intervals (i.e., CI_{θ_0}, $CI_{\theta_\bullet}$ and $CI_{\psi_\bullet}$) depend on the analyses.
3. Since 1T-RRFC analysis ignores case sampling variability, the CAD figure of merit is a constant, with zero-width confidence interval. For compactness the CI is listed as 0, rather than two identical values in parentheses. The confidence interval listed for 2T-RRRC analyses is centered on the corresponding CAD value, as are all confidence intervals in Table 2.
4. The LROC FOMs increase as the value of FPF (the subscript) increases. This should be obvious, as PCL increases as FPF increases, a general feature of any partial curve based figure of merit.
5. The area (AUC) under the ROC is larger than the largest PCL value, i.e., $AUC \geq \text{PCL}_1$. This too should be obvious from the general features of the LROC [@swensson1996unified].
6. The p-value for either RRRC analyses (2T or 1T) is larger than the corresponding 1T-RRFC value. Accounting for case-sampling variability increases the p-value, leading to less possibility of finding a significant difference.

7. Partial curve-based FOMs, such as PCL$_{FPF}$, lead, depending on the choice of FPF, to different conclusions. The p-values generally decrease as FPF increases. Measuring performance on the steep part of the LROC curve (i.e., small FPF) needs to account for greater reader variability and risks lower statistical power.
8. Ignoring localization information (i.e., using the AUC FOM) led to a not-significant difference between CAD and the radiologists ($p = 0.3210$), while the corresponding FOM yielded a significant difference ($p = 0.0409$). Accounting for localization leads to a less "noisy" measurement. This has been demonstrated for the LROC paradigm [@swensson1996unified] and I have demonstrated this for the FROC paradigm [@chakraborty2008validation].
9. For 1T-RRRC analysis, is listed as NA, for not applicable, since is not a model parameter, see Eqn. (14).

Shown next, Table 3, are the model-parameters corresponding to the three analyses.

Table 3: Parameter estimates for the analyses; NA = not applicable.

FOM	Analysis	σ_R^2	$\sigma_{\tau R}^2$	Cov1	Cov2	Cov3	Var
PCL_0_05	1T-RRFC	9.5e-03	NA	NA	NA	NA	NA
	2T-RRRC	1.84e-18	-5.71e-03	1.31e-03	6.01e-03	1.31e-03	1.65e-02
	1T-RRRC	9.5e-03	NA	NA	9.4e-03	NA	3.03e-02
PCL_0_2	1T-RRFC	2.81e-03	NA	NA	NA	NA	NA
	2T-RRRC	-7.59e-19	2.65e-04	7.61e-04	2.29e-03	7.61e-04	3.43e-03
	1T-RRRC	2.81e-03	NA	NA	3.07e-03	NA	5.34e-03
PCL_1	1T-RRFC	3.2e-03	NA	NA	NA	NA	NA
	2T-RRRC	1.63e-18	1e-03	6.43e-04	1.86e-03	6.43e-04	2.46e-03
	1T-RRRC	3.2e-03	NA	NA	2.44e-03	NA	3.64e-03
Wilcoxon	1T-RRFC	8.78e-04	NA	NA	NA	NA	NA
	2T-RRRC	2.98e-19	2.01e-04	2.62e-04	7.24e-04	2.62e-04	9.62e-04
	1T-RRRC	8.78e-04	NA	NA	9.24e-04	NA	1.4e-03

The following characteristics are evident from Table 3.

1. For 2T-RRRC analyses $\sigma_R^2 = 0$. Actually, the analysis yielded very small values, of the order of 10^{-18} to 10^{-19}, which, being smaller than double precision accuracy, were replaced by zeroes in Table 2. $\sigma_R^2 = 0$ is clearly an incorrect result as the radiologists do not have identical performance. In contrast, 1T-RRRC analyses yielded more realistic values, identical to those obtained by 1T-RRFC analyses, and consistent with expectation – see comment following Eqn. (15).
2. Because 2T analysis found zero reader variability, it follows from the definitions of the covariances [@obuchowski1995hypothesis], that $Cov_1 = Cov_3 = 0$, as evident in the table.
3. When they can be compared (i.e., σ_R^2, Cov$_2$ and Var), all variance and covariance estimates were smaller for the 2T method than for the 1T method.
4. For the 2T method the expected inequalities, Eqn. (6), are not obeyed (specifically, $Cov_1 \geq Cov_2 \geq Cov_3$ is not obeyed).

For an analysis method to be considered statistically valid it needs to be tested with simulations to determine if it has the proper null hypothesis behavior. The design of a ratings simulator to statistically match a given dataset is addressed in Chapter 23 of reference [@chakraborty2017observer]. Using this simulator, the 1T-RRRC method had the expected null hypothesis behavior (Table 23.5, ibid).

1.7 Discussion

TBA TODOLAST The argument often made for not measuring standalone performance is that since CAD will be used only as a second reader, it is only necessary to measure performance of radiologists without and with CAD. It has been stated [@nishikawa2011fundamental]:

> High stand-alone performance is neither a necessary nor a sufficient condition for CAD to be truly useful clinically.

Assessing CAD utility this way, i.e, by measuring performance with and without CAD, may have inadvertently set a low bar for CAD to be considered useful. As examples, CAD is not penalized for missing cancers as long as the radiologist finds them and CAD is not penalized for excessive false positives (FPs) as long as the radiologist ignores them. Moreover, since both such measurements include the variability of radiologists, there is additional noise introduces that presumably makes it harder to determine if the CAD system is optimal.

Described is an extension of the analysis used in Study − 1 that accounts for case sampling variability. It extends [@hillis2005comparison] single-treatment analysis to a situation where one of the "readers" is a special reader, and the desire is to compare performance of this reader to the average of the remaining readers. The method, along with two other methods, was used to analyze an LROC data set using different figures of merit.

1T-RRRC analyses yielded identical overall results (specifically the F-statistic, degrees of freedom and p-value) to those yielded by the unorthodox application of DBM-MRMC software, termed 2T-RRRC analyses, where the CAD reader is regarded as a second treatment. However, the values of the model parameters of the dual-treatment analysis lacked clear physical meanings. In particular, the result $\sigma_R^2 = 0$ is clearly an artifact. One can only speculate as to what happens when software is used in a manner that it was not designed for: perhaps finding that all readers in the second treatment have identical FOMs led the software to yield $\sigma_R^2 = 0$. The single-treatment model has half as many parameters as the dual-treatment model and the parameters have clear physical meanings and the values are realistic.

The paradigm used to collect the observer performance data - e.g., receiver operating characteristic (ROC) [@metz1986rocmethodology], free-response ROC (FROC) [@Chakraborty1986DigitalVsConv], location ROC (LROC) [@starr1975visual] or region of interest (ROI) [@obuchowski2010data] - is irrelevant − all that is needed is a scalar performance measure for the actual paradigm used. In addition to PCL and AUC, RJafroc currently implements the partial area under the LROC, from FPF = 0 to a specified value as well other FROC-paradigm based FOMs.

While there is consensus that CAD works for microcalcifications, for masses its performance is controversial27,28. Two large clinical studies TBA 29,30 (222,135 and 684,956 women, respectively) showed that CAD actually had a detrimental effect on patient outcome. A more recent large clinical study has confirmed the negative view of CAD31 and there has been a call for ending Medicare reimbursement for CAD interpretations32.

In my opinion standalone performance is the most direct measure of CAD performance. Lack of clear-cut methodology to assess standalone CAD performance may have limited past CAD research. The current work hopoefully removes that impediment. Going forward, assessment of standalone performance of CAD vs. expert radiologists is strongly encouraged.

1.8 References