

# Tutorial on Reader Study Designs and MRMC Analysis

Brandon D. Gallas, PhD

FDA/CDRH/OSEL/DIDSR

# Outline

## Reader Studies

aka Clinical Performance Studies, human-in-the-loop

Compare performance of a new imaging system to a reference imaging system

- Modality (imaging system, viewing condition)
  - FFDM: Full-field digital mammography vs. SFM: Screen-film mammography
  - WSI: Whole slide images (digitized glass slides) vs. Microscope
  - Image with computer aid vs. Image without computer aid
  - CT with new reconstruction method vs. CT with old reconstruction method
- Task/Performance (Detect, Classify, Grade, Measure)
- Readers (Clinician, Radiologist, Pathologist)
- Cases (Patients)

www.fda.gov Gallas, DRH Seminar Reader Study Designs and Analysis, 8 April 2022 5

## MRMC Variance Components

$$\text{var}(\widehat{AUC}_1 - \widehat{AUC}_2) = \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1} + \frac{\sigma_{01}^2}{N_0 N_1} + \frac{\sigma_R^2}{N_R} + \frac{\sigma_{0R}^2}{N_0 N_R} + \frac{\sigma_{1R}^2}{N_1 N_R} + \frac{\sigma_{01R}^2}{N_0 N_1 N_R}$$

www.fda.gov Gallas, DRH Seminar Reader Study Designs and Analysis, 8 April 2022 16

## MRMC Simulation

readerID	caseID	modID	score	Truth
reader1	case1	mod1	0.12	0
reader1	case1	mod2	...	...
reader1	case2	mod1	...	...
reader1	case2	mod2	...	...
reader2	case1	mod1	...	...
reader2	case1	mod2	...	...
reader2	case2	mod1	...	...
reader2	case2	mod2	...	...

www.fda.gov Gallas, DRH Seminar Reader Study Designs and Analysis, 8 April 2022 29

## Study Designs

www.fda.gov Gallas, DRH Seminar Reader Study Designs and Analysis, 8 April 2022 36

## Study Designs: Efficiency

- 2-Groups
- 3-Groups
- 4-Groups
- Fully-Crossed A
- Fully-Crossed B
- Readers Unpaired Across Modalities

www.fda.gov Gallas, DRH Seminar Reader Study Designs and Analysis, 8 April 2022 42

## VIPER Study

Validation of Imaging Premarket Evaluation and Regulation

www.fda.gov Gallas, DRH Seminar Reader Study Designs and Analysis, 8 April 2022 54

## Summary and Future Work

www.fda.gov Gallas, DRH Seminar Reader Study Designs and Analysis, 8 April 2022 74



# Reader Studies

aka Clinical Performance Studies, human-in-the-loop

**Compare performance of a new imaging system to a reference imaging system**

- **Modality** (imaging system, viewing condition)
  - **FFDM**: Full-field digital mammography vs. **SFM**: Screen-film mammography
  - **WSI**: Whole slide images (digitized glass slides) vs. Microscope
  - Image with computer aid vs. Image without computer aid
  - CT with new reconstruction method vs. CT with old reconstruction method
- **Task/Performance** (Detect, Classify, Grade, Measure)
- **Readers** (Clinician, Radiologist, Pathologist)
- **Cases** (Patients)



# Reader Studies

## Task/Performance

- Sensitivity
  - Success rate on diseased cases
- Specificity
  - Success rate on non-diseased cases

Binary Task  
Binary Decisions

- 
- ROC and Area Under ROC curve
    - Separation between conditional distributions
      - Scores on diseased cases
      - Vs.
      - Scores on non-diseased cases
    - Tradeoff between Sensitivity and Specificity

Binary Task  
Ordinal Scores

- 
- Mean-squared error, Correlation
  - Limits of Agreement, Bland-Altman Plots

Measure  
Quantitative Values

# Reader Studies

## MRMC Analysis



MRMC: Multi-reader, Multi-case Analysis

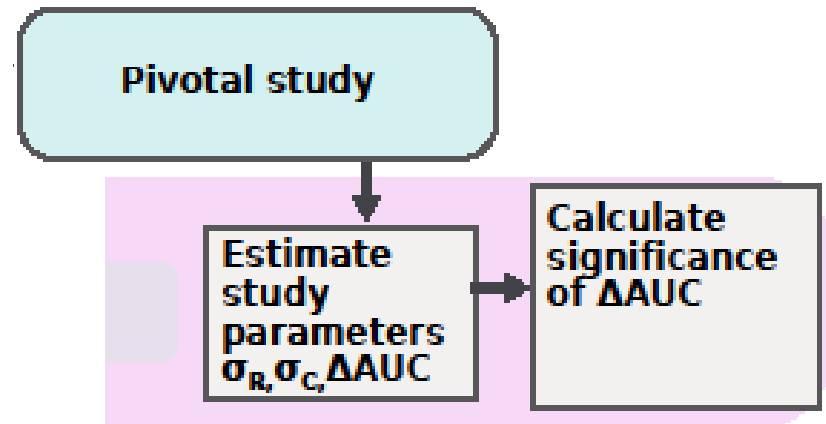
- Account for reader and case variability
- Account for reader and case correlations
- Analysis
  - Estimate variances, confidence intervals
  - Perform hypothesis tests
- Results Generalize to Population of Readers and Cases

# Reader Studies Study Design



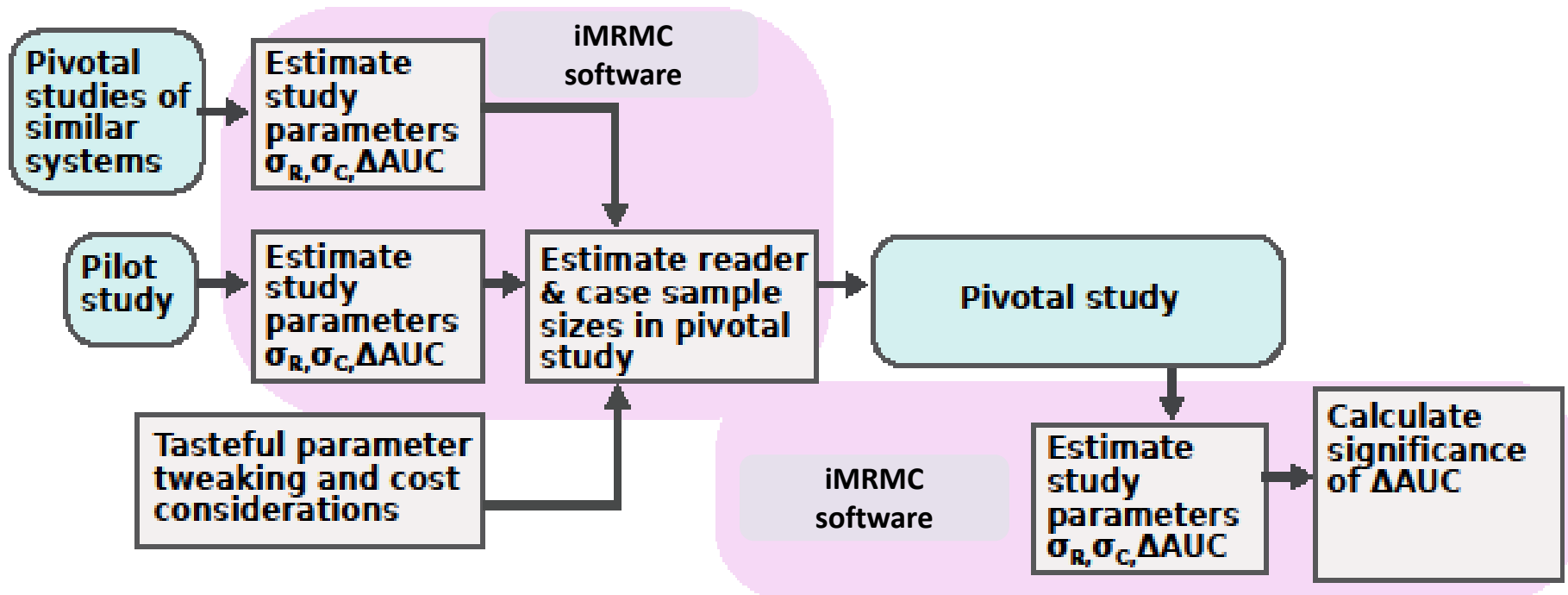
**Pivotal study**

# Reader Studies Study Design



Courtesy Weijie Chen

# Reader Studies Study Design



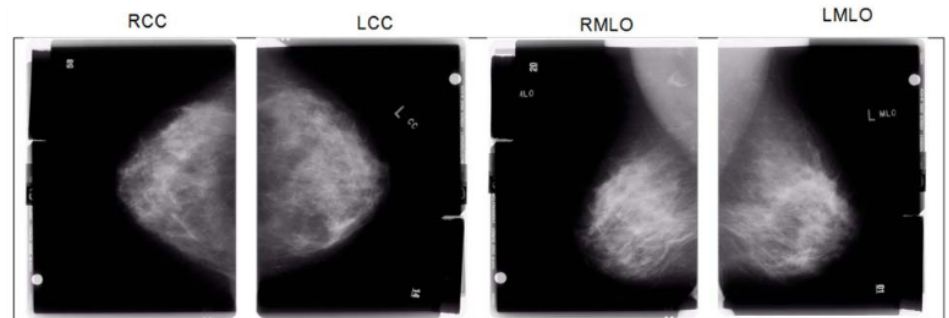
Courtesy Weijie Chen



# Reader Studies

## Data Collection

- Two steps
  - Binary patient management decision
  - More information (ROC scores)



Would you recall patient?

Yes  
 No

Being more quantitative in reporting your *Numeric Rating*:

- Are there no dense areas and no abnormal findings? If so, perhaps your *Numeric Rating* should be 1-25?
- Are there dense areas or benign findings, but not enough to prompt a decision to recall? If so, perhaps your *Numeric Rating* should be 75-100.
- Are the visual cues somewhere in the middle?

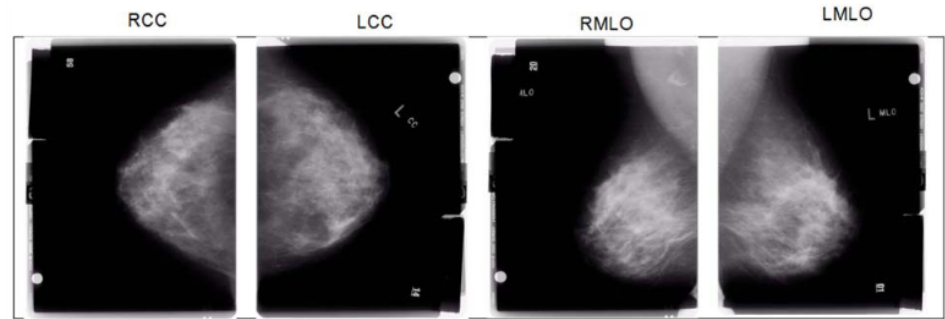


Numeric Score

--	--	--	--

# Reader Studies Data Collection

- Two steps
  - Binary patient management decision
  - More information (ROC scores)

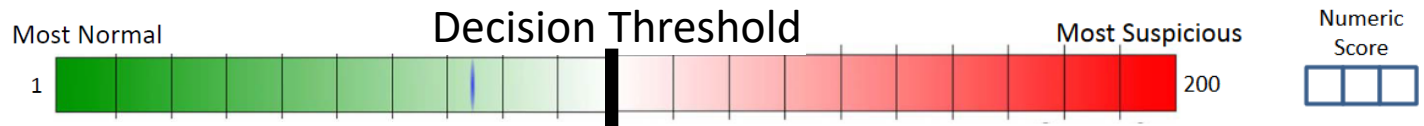


Would you recall patient?

- Yes
- No

Being more quantitative in reporting your *Numeric Rating*:

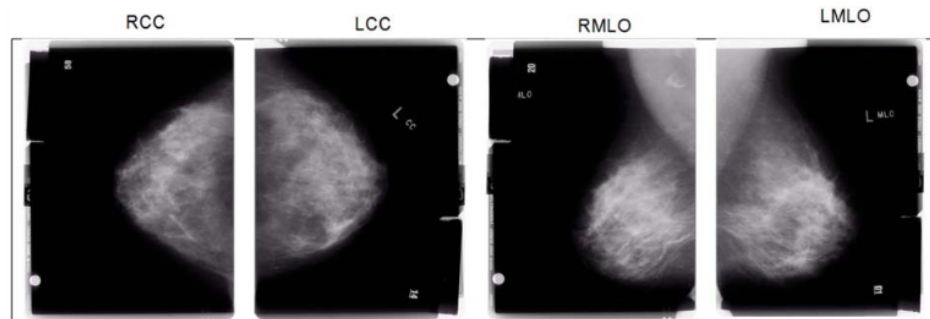
- Are there only a few inconclusive visual cues prompting your decision to recall? If so, perhaps your *Numeric Rating* should be 101-125?
- Are there many definitive visual cues prompting your decision to recall? If so, perhaps your *Numeric Rating* should be 175-200.
- Are the visual cues somewhere in the middle?



# Reader Studies

## Data Collection

- Two steps
  - Binary patient management decision
  - More information (ROC scores)
  
- Provide written instructions
  - Give clinician comfort
  - Not evaluating clinician
  
  - ROC scores foreign
  - Provide scoring rubric
  - Not asking for probabilities, too much baggage
  - Goal is to rank



Would you recall patient?

Yes  
 No

Being more quantitative in reporting your *Numeric Rating*:

- Are there no dense areas and no abnormal findings? If so, perhaps your *Numeric Rating* should be 1-25?
- Are there dense areas or benign findings, but not enough to prompt a decision to recall? If so, perhaps your *Numeric Rating* should be 75-100.
- Are the visual cues somewhere in the middle?

Most Normal



Numeric Score

--	--	--	--

VIPER case report form and ROC scoring instructions

<https://didsr.github.io/viperData/>

# MRMC Variance Components

$$\begin{aligned}
 \text{var}(\widehat{AUC}_1 - \widehat{AUC}_2) &= \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1} + \frac{\sigma_{01}^2}{N_0 N_1} \\
 &\quad + \frac{\sigma_R^2}{N_R} \\
 &\quad + \frac{\sigma_{0R}^2}{N_0 N_R} + \frac{\sigma_{1R}^2}{N_1 N_R} + \frac{\sigma_{01R}^2}{N_0 N_1 N_R}
 \end{aligned}$$

# MRMC Variance Components

- Main Random Effects
  - case variability  
*difficulty*
  - reader variability  
*skill*
  - reader/case interaction  
*training, experience, cases encountered*

# MRMC Variance Components

- Main Random Effects
  - case variability  
*Non-disease + Disease + Interaction*
  - reader variability
  - reader/case interaction  
*Non-disease + Disease + Interaction*

# MRMC Variance Components

## U-statistic result

- Single Modality

- Gallas et al. (2009)

$$\begin{aligned}
 \text{var}(\widehat{AUC}_1) = & \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1} + \frac{\sigma_{01}^2}{N_0 N_1} && \leftarrow \text{Case Variability} \\
 & + \frac{\sigma_R^2}{N_R} && \leftarrow \text{Reader Variability} \\
 & + \frac{\sigma_{0R}^2}{N_0 N_R} + \frac{\sigma_{1R}^2}{N_1 N_R} + \frac{\sigma_{01R}^2}{N_0 N_1 N_R} && \leftarrow \text{Reader-Case Interaction}
 \end{aligned}$$

Non-diseased cases  
Diseased cases  
Interaction

Given U-statistic estimator of reader-averaged AUC

7 variance components  
7 coefficients

No modeling

# MRMC Variance Components

## U-statistic result

- Two Modalities

- Gallas et al. (2009)

$$\begin{aligned}
 \text{var}(\widehat{AUC}_1 - \widehat{AUC}_2) &= \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1} + \frac{\sigma_{01}^2}{N_0 N_1} \\
 &+ \frac{\sigma_R^2}{N_R} \\
 &+ \frac{\sigma_{0R}^2}{N_0 N_R} + \frac{\sigma_{1R}^2}{N_1 N_R} + \frac{\sigma_{01R}^2}{N_0 N_1 N_R}
 \end{aligned}$$

← Case Variability
← Reader Variability

← Reader-Case Interaction

Non-diseased cases

Diseased cases

Interaction

Different interpretation for these components

- AUC difference



# MRMC Variance Components

## U-statistic result

- Two Modalities

- Gallas et al. (2009)

$$\begin{aligned}
 \text{var}(\widehat{AUC}_1 - \widehat{AUC}_2) &= \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1} + \frac{\sigma_{01}^2}{N_0 N_1} \\
 &+ \frac{\sigma_R^2}{N_R} \\
 &+ \frac{\sigma_{0R}^2}{N_0 N_R} + \frac{\sigma_{1R}^2}{N_1 N_R} + \frac{\sigma_{01R}^2}{N_0 N_1 N_R}
 \end{aligned}$$

← Case Variability  
← Reader Variability

Reader-Case Interaction

Non-diseased cases

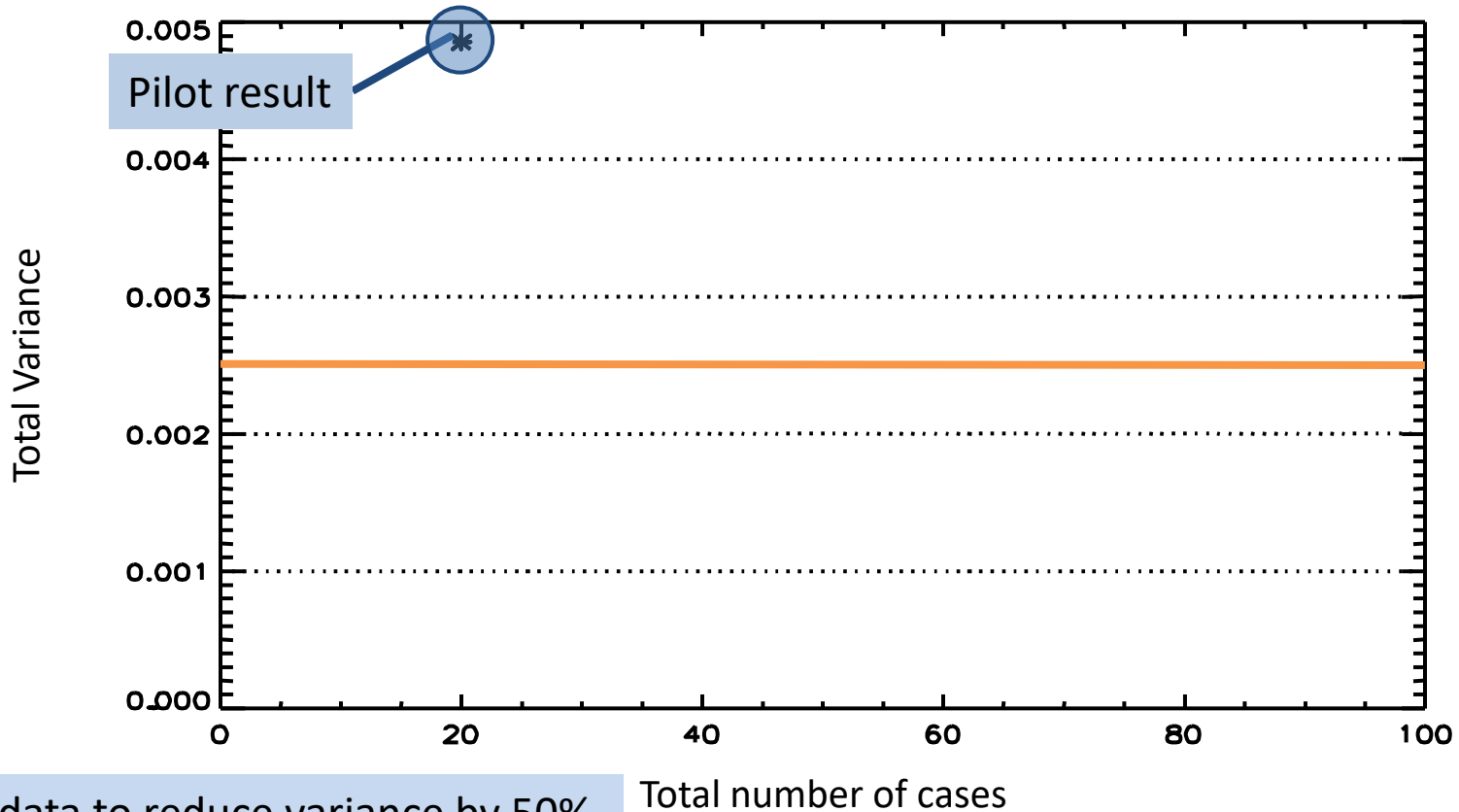
Diseased cases

Interaction

Sizing  
 Estimate components  
 Explore N0, N1, NR

# MRMC Variance Components

## Size a Trial



Add data to reduce variance by 50%.

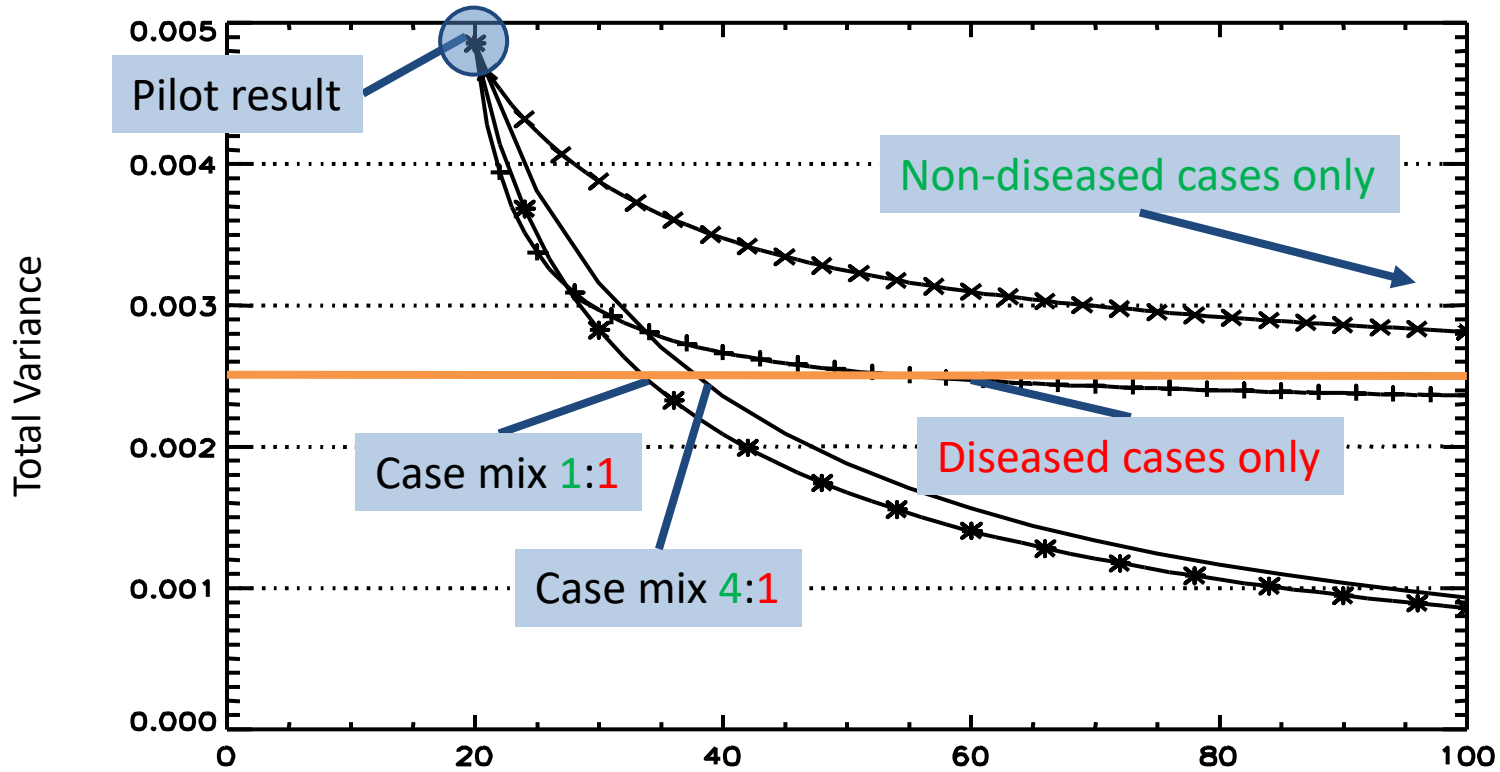
- 1:1 sampling
- 4:1 sampling

- Only Non-Diseased
- Only Diseased

Colposcopy Study  
Plot courtesy of Hsu, NCI.

# MRMC Variance Components

## Size a Trial



Add data to reduce variance by 50%. Total number of cases

1:1 sampling  
4:1 sampling

Only Non-Diseased  
Only Diseased

Colposcopy Study  
Plot courtesy of Hsu, NCI.

# MRMC Variance Components

## One-Shot Estimate of MRMC Variance: AUC<sup>1</sup>

Brandon D. Gallas

Academic Radiology, 2006

<https://doi.org/10.1016/j.acra.2005.11.030>

## A Framework for Random-Effects ROC Analysis: Biases with the Bootstrap and Other Variance Estimators

BRANDON D. GALLAS<sup>1</sup>, ANDRIY BANDOS<sup>2</sup>,  
FRANK W. SAMUELSON<sup>1</sup>, AND ROBERT F. WAGNER<sup>1</sup>

<sup>1</sup>NIBIB/CDRH Laboratory for the Assessment of Medical  
Imaging Systems, Silver Spring, Maryland, USA

<sup>2</sup>Department of Biostatistics, University of Pittsburgh, Pittsburgh,  
Pennsylvania, USA

Communications in Statistics - Theory and Methods, 2009

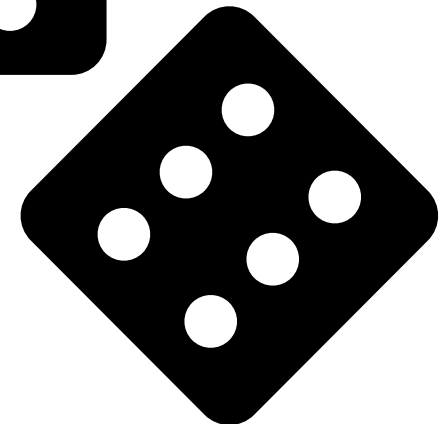
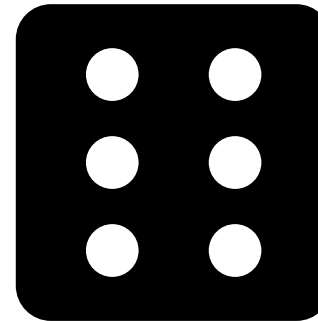
<https://doi.org/10.1080/03610920802610084>

## Published iMRMC Software

- 2013: Java Application - Google Code
  - Retired
- 2015: Java Application – GitHub
  - <https://github.com/DIDSR/iMRMC>
- 2017: R Package – CRAN
  - <https://cran.r-project.org/web/packages/iMRMC/index.html>

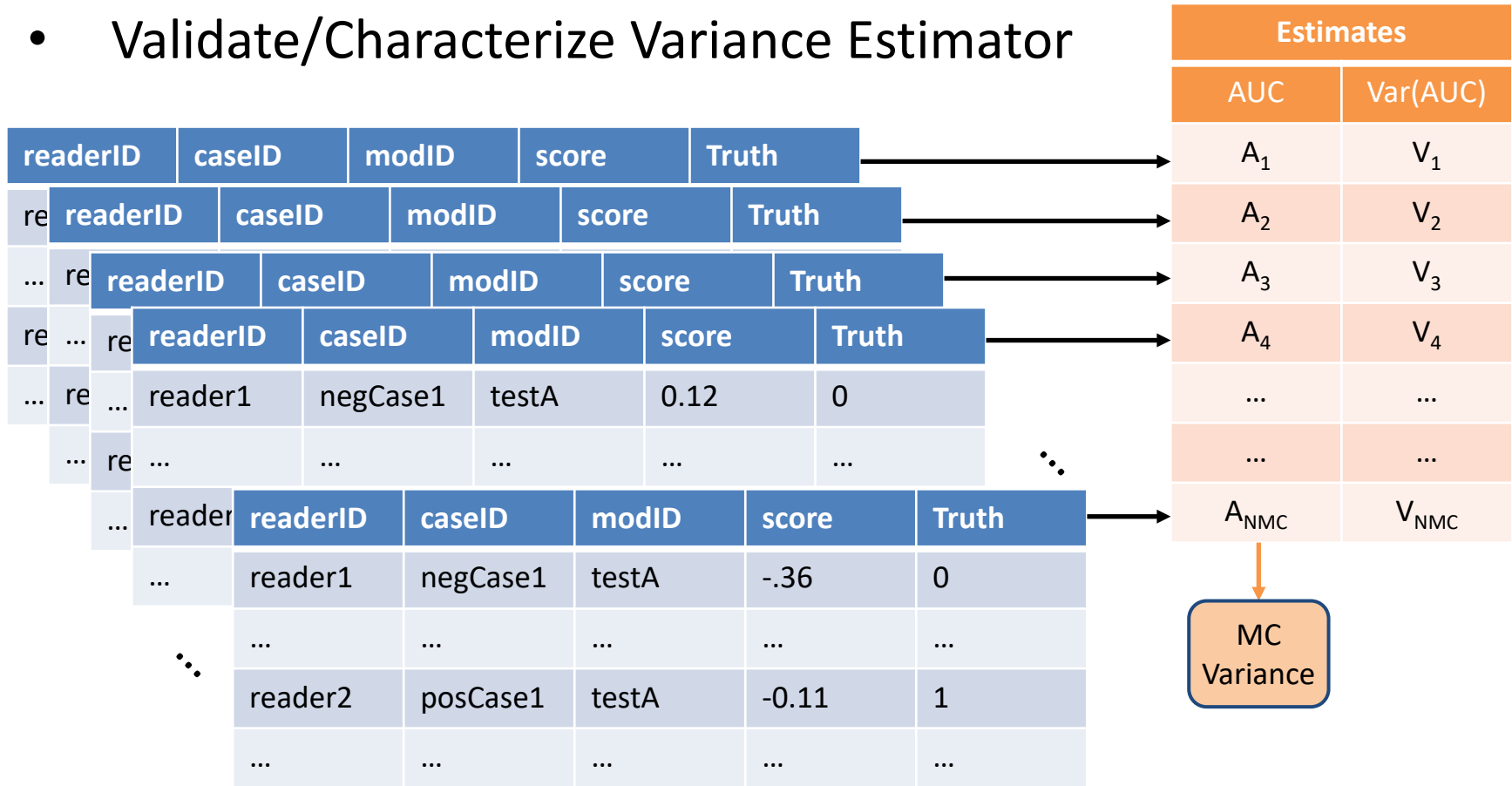
# MRMC Simulation

readerID	caseID	modID	score	Truth			
re	readerID	caseID	modID	score	Truth		
...	re	readerID	caseID	modID	score	Truth	
re	...	re	readerID	caseID	modID	score	Truth
...	re	...	reader1	negCase1	testA	0.12	0
...	re	...	...	...	...	...	...
...	re	...	readerID	caseID	modID	score	Truth
...	...	...	reader1	negCase1	testA	-0.36	0
...	...	...	...	...	...	...	...
...	...	...	reader2	posCase1	testA	-0.11	1
...	...	...	...	...	...	...	...



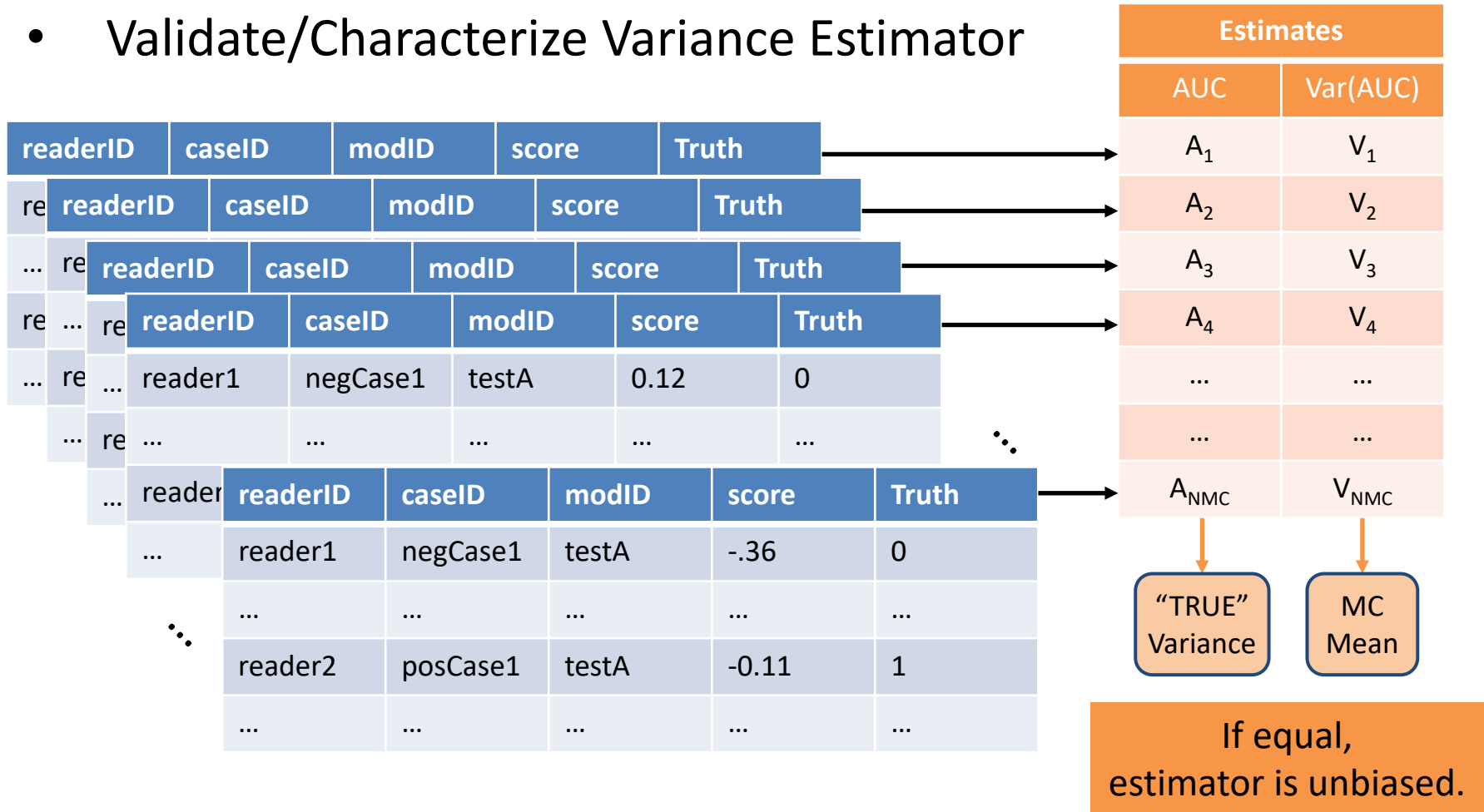
# MRMC Simulation

- Validate/Characterize Variance Estimator



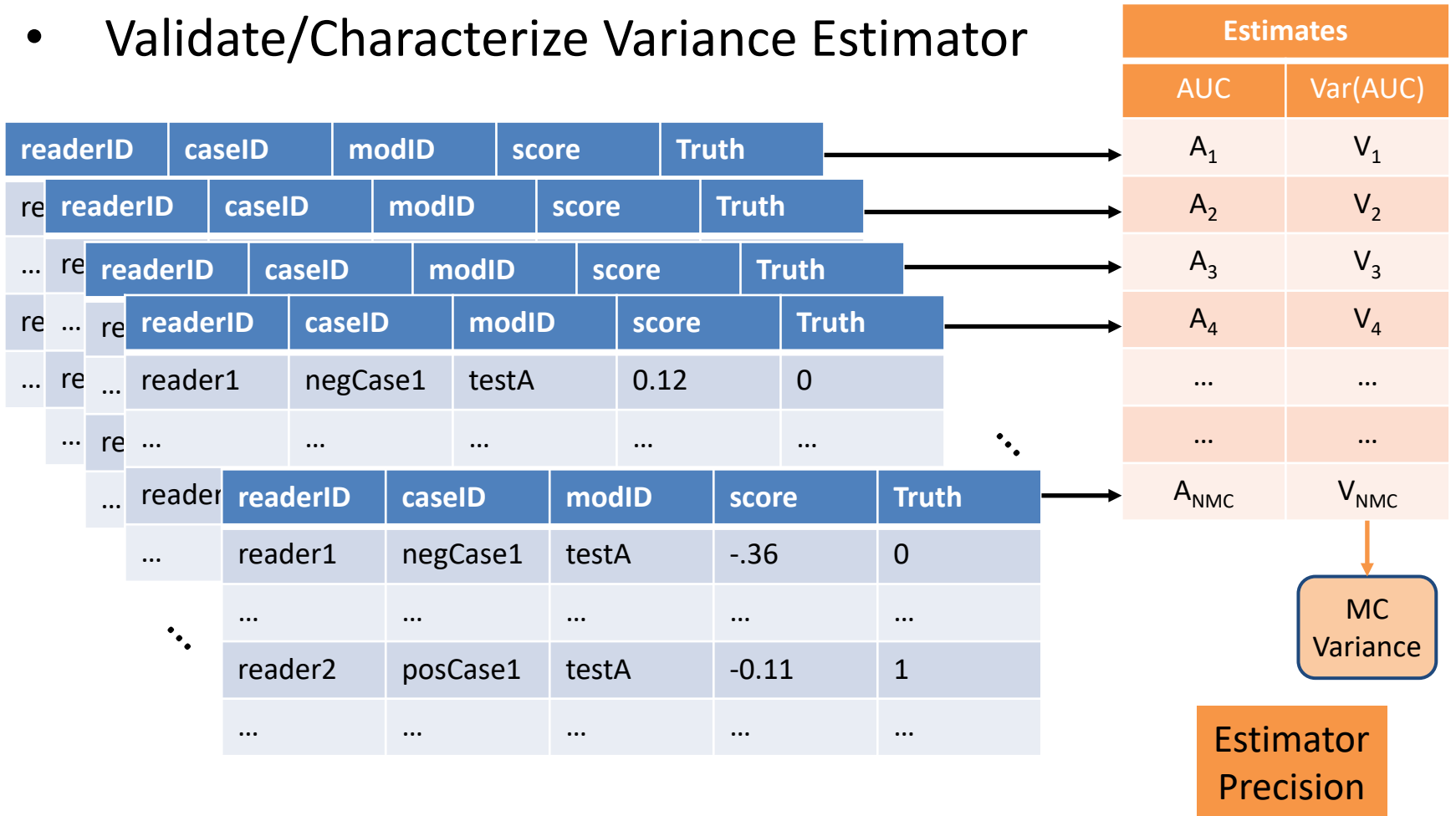
# MRMC Simulation

- Validate/Characterize Variance Estimator



# MRMC Simulation

- Validate/Characterize Variance Estimator





# MRMC Simulation

## Roe and Metz Model (1997)

- Simulation model for ROC scores
  - Multiple modalities (fixed effect)
  - Multiple readers
  - Multiple cases

### Signal-absent scores

$$X_{ijk0} = \tau_{i0}$$

$$+ C_{k0} \quad + [\tau C]_{ik0}$$

$$+ R_{j0} \quad + [\tau R]_{ij0}$$

$$+ [RC]_{jk0} \quad + [\tau RC]_{ijk0}$$

Fixed effect: Modality ( $i$ )

Random effects: (Independent Normal)

- Case ( $k$ )
- Reader ( $j$ )
- Interaction

# MRMC Simulation

## Roe and Metz Model (1997)

- Simulation model for ROC scores
  - Multiple modalities (fixed effect)
  - Multiple readers
  - Multiple cases

### Signal-present scores

$$\begin{aligned}
 Y_{ijk1} = & \tau_{i1} \\
 & + C_{k1} \quad + [\tau C]_{ik1} \\
 & + R_{j1} \quad + [\tau R]_{ij1} \\
 & + [RC]_{jk1} + [\tau RC]_{ijk1}
 \end{aligned}$$

Looks like  
3-way ANOVA

Warning  
Simulation for scores not AUC

# MRMC Simulation

## Build on Roe and Metz model

- Binary Data

- Parameters depend on truth and modality
- Analytical relationship
  - ROC scores
  - AUC components of variance

B70 J. Opt. Soc. Am. A/Vol. 24, No. 12/December 2007 Gallas *et al.*

**Multireader multcase variance analysis for binary data**

Brandon D. Gallas,\* Gene A. Pennello, and Kyle J. Myers

<https://doi.org/10.1364/JOSAA.24.000B70>

MedicalImaging.SPIEDigitalLibrary.org

<https://doi.org/10.1117/1.JMI.1.3.031011>

**Multireader multcase reader studies with binary agreement data: simulation, analysis, validation, and sizing**

2014

Weijie Chen  
Adam Wunderlich  
Nicholas Petrick  
Brandon D. Gallas

Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

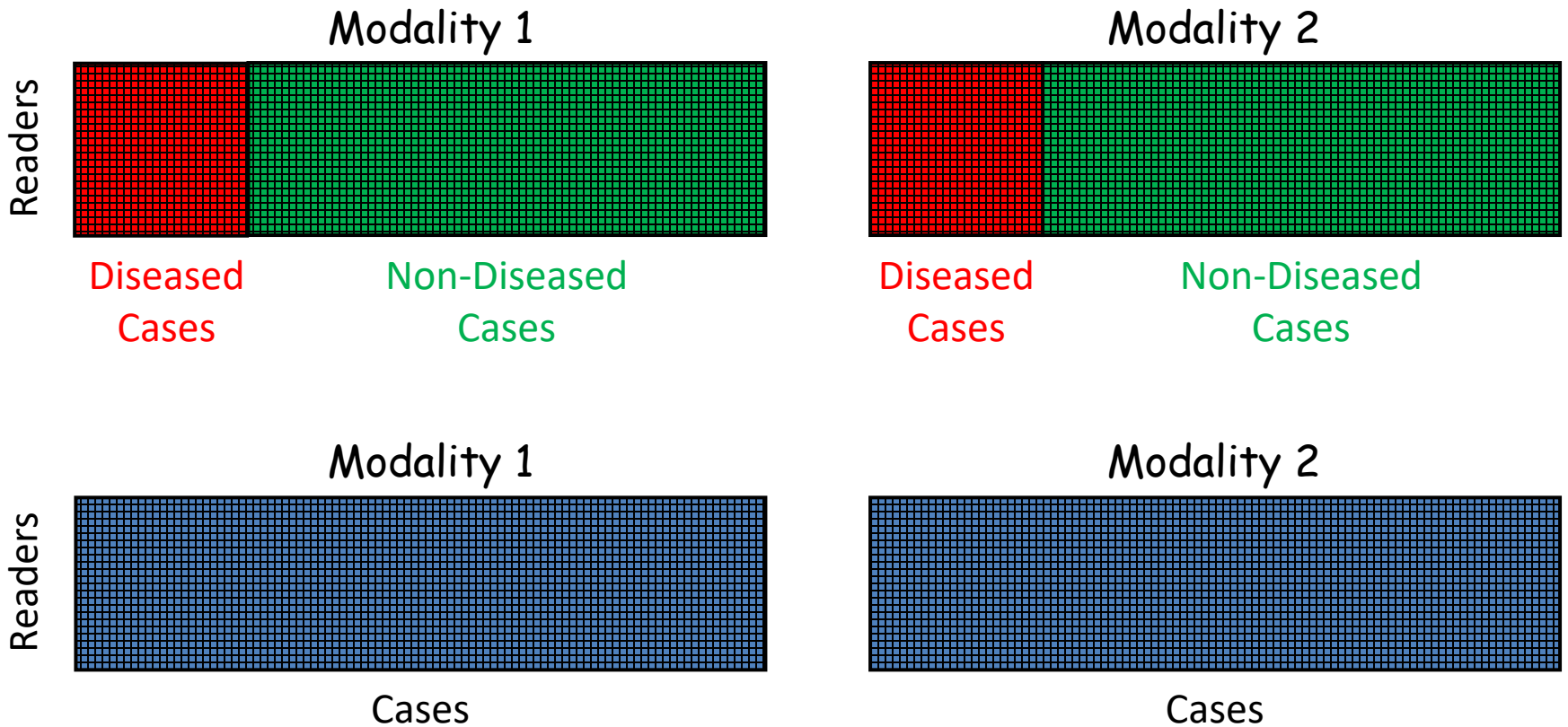
<https://doi.org/10.1117/1.JMI.1.3.031006>

**Generalized Roe and Metz receiver operating characteristic model: analytic link between simulated decision scores and empirical AUC variances and covariances**

Brandon D. Gallas  
Stephen L. Hillis

2014

# Study Designs

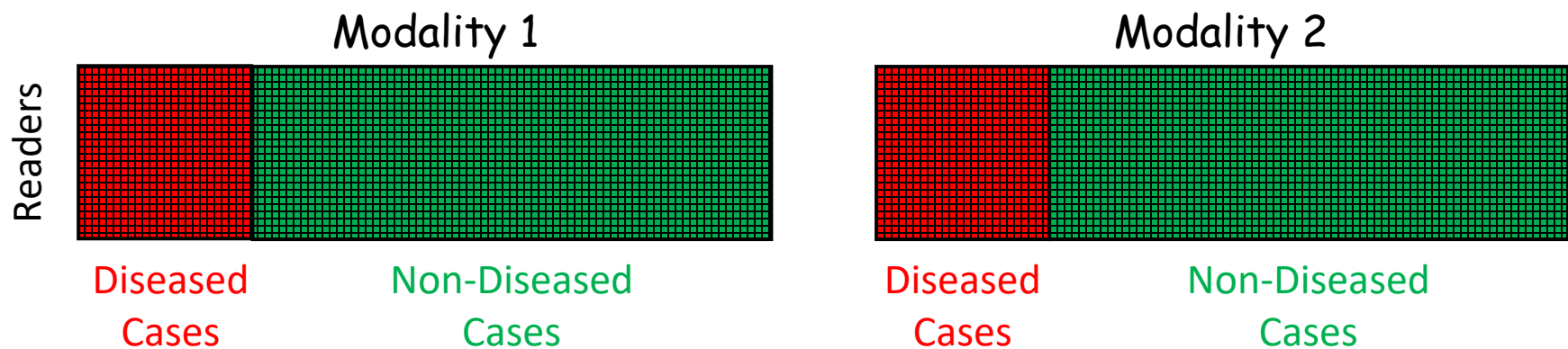


# Study Designs

## Fully-Crossed

- Fully-crossed study
  - All readers read all cases
  - Readers and cases are paired across modalities

Data Array  
Rows = readers  
Cols = cases



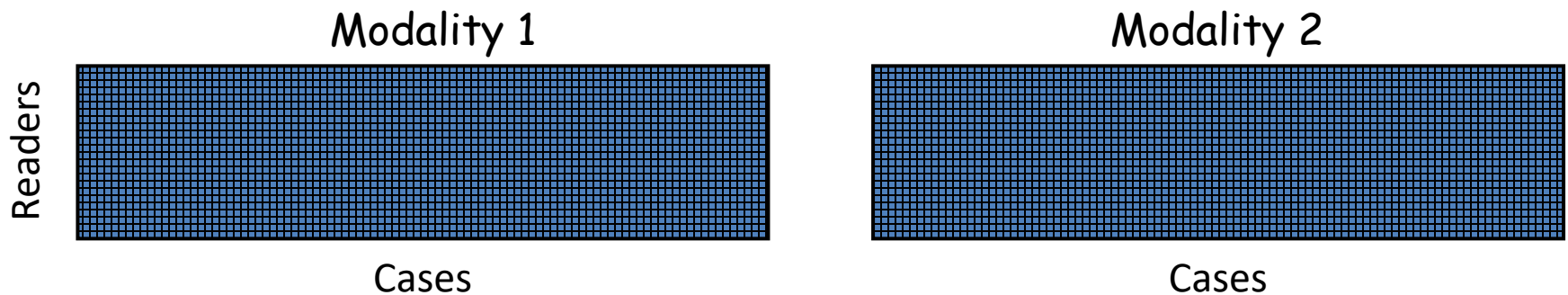
# Study Designs

## Fully-Crossed



- Fully-crossed study
  - All readers read all cases
  - Readers and cases are paired across modalities

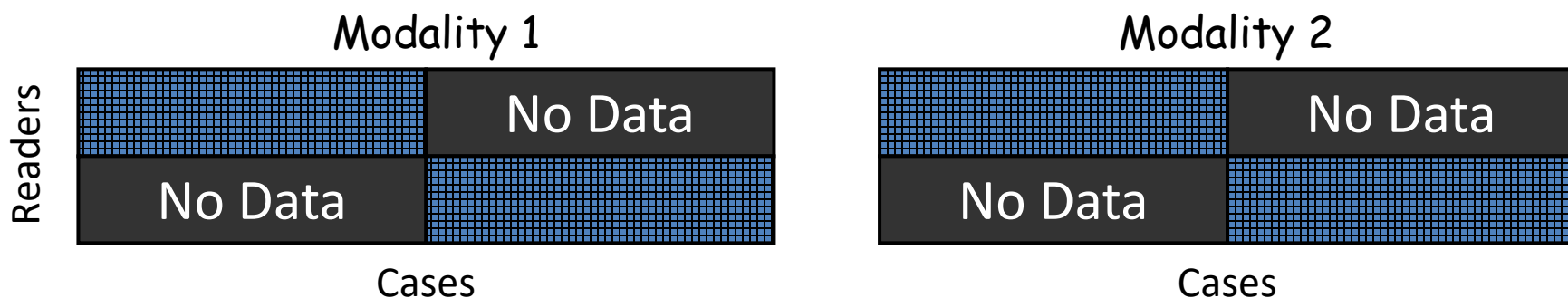
Remove truth labels to unclutter study design concepts.



# Study Designs

## Split-Plot

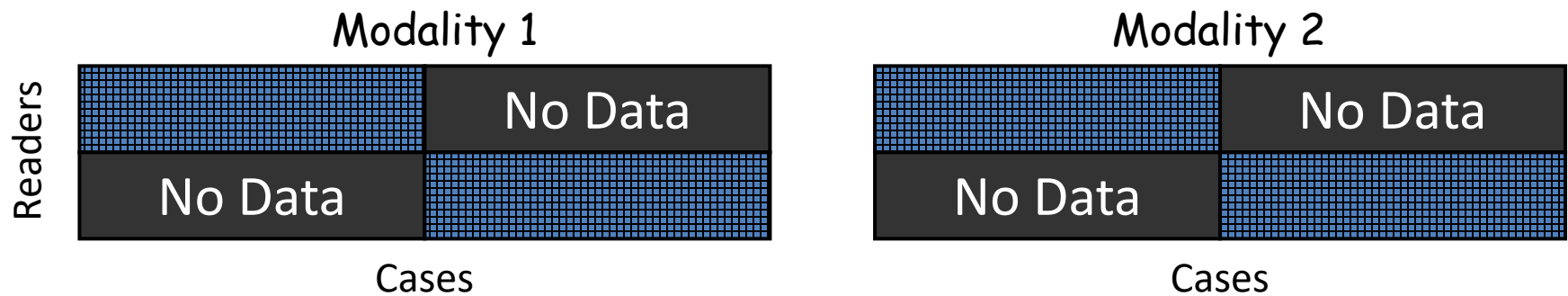
- Fully-crossed study is burdensome
  - All readers read all cases
  - Readers and cases are paired across modalities
- Split-plot study
  - Readers and cases split into 2 groups
  - Data is fully-crossed within a group



# Study Designs

## Split-Plot


- Fully-crossed is burdensome
  - A lot of reads per reader
  - A lot of reads total
- Split-plot studies can save time (and money)
  - Half the reads per reader
  - Half the reads total






# Study Designs

- Generalized analysis methods
  - Treat arbitrary study designs
  - Publications and Software



Available online at [www.sciencedirect.com](http://www.sciencedirect.com)

 **ScienceDirect**

Neural Networks 21 (2008) 387–397

Neural  
Networks

[www.elsevier.com/locate/neunet](http://www.elsevier.com/locate/neunet)

2008 Special Issue

Reader studies for validation of CAD systems<sup>\*</sup>

Brandon D. Gallas<sup>\*</sup>, David G. Brown

NIBIB/CDRH Laboratory for the Assessment of Medical Imaging Systems, FDA, Silver Spring, MD, 20993-0002, United States

Received 22 August 2007; received in revised form 7 December 2007; accepted 11 December 2007

<https://doi.org/10.1080/03610920802610084>

## Multi-reader ROC Studies with Split-plot Designs:

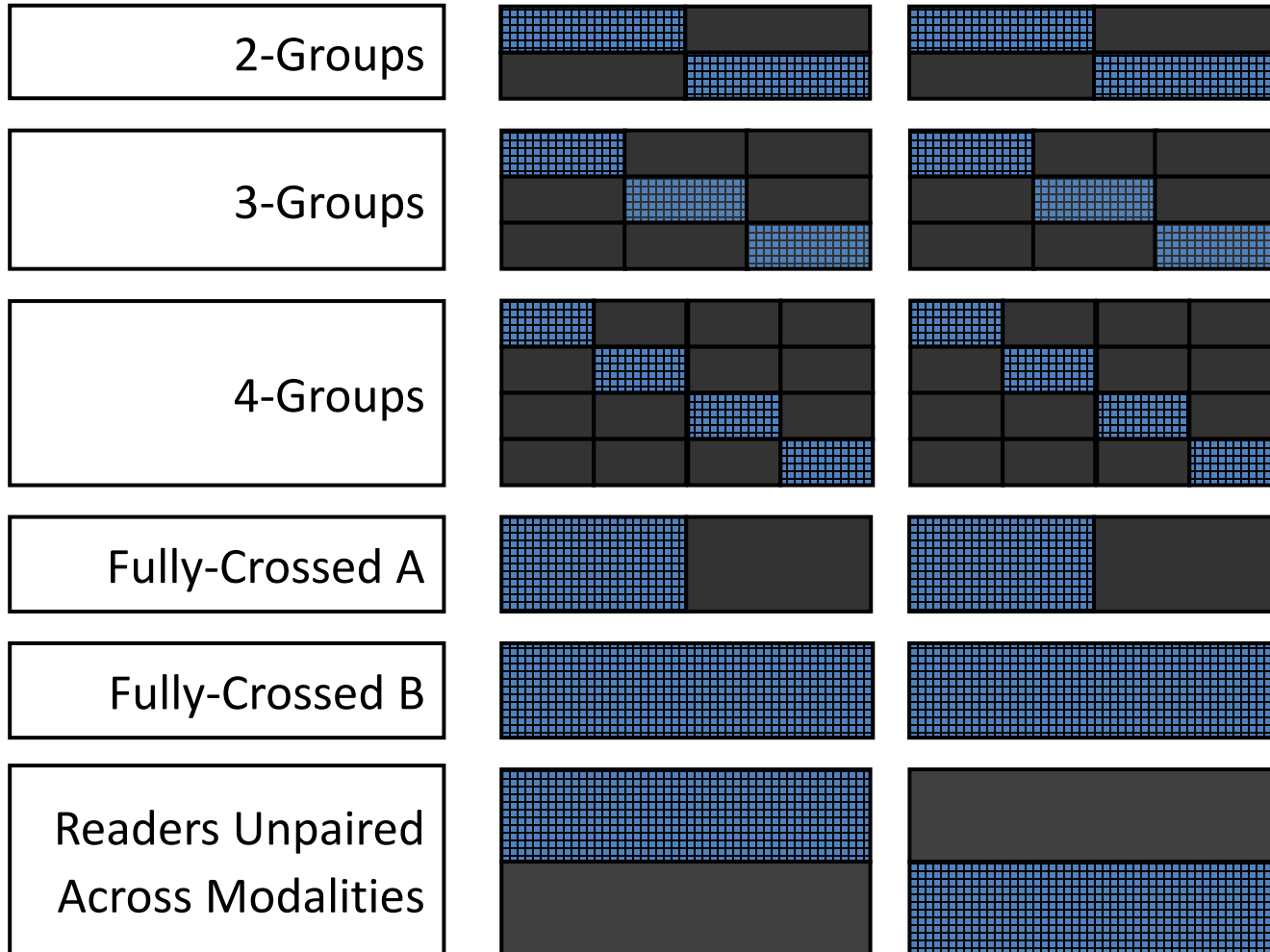
### A Comparison of Statistical Methods

Nancy A. Obuchowski, PhD, Brandon D. Gallas, PhD, Stephen L. Hillis, PhD

Academic Radiology, 2012

<https://doi.org/10.1016/j.acra.2012.09.012>

# Study Designs: Efficiency



# Study Designs: Efficiency

- MRMC framework that decouples variance components from study design
- Roe and Metz simulation
  - given description of scores, know the components of variance (numerical integration)
- Model parameters ( $\Delta\mu = 1.53$ )

<u>Var r</u>	<u>Var c</u>	<u>Var rc</u>	<u>Var tr</u>	<u>Var tc</u>	<u>Var trc</u>
0.011	0.100	0.200	0.030	0.100	0.200

# Study Designs: Efficiency

TABLE 3. Resources Needed for Different Study Designs

Study Design	Number of Readers (J)	Number of Patients*	Total Number of Image Interpretations	Number of Image Interpretations per Reader	Statistical Efficiency <sup>1</sup>
Two-block split-plot	6 (3/block)	120 (30 + 30)	720	120	1.0
Three-block split-plot	9 (3/block)	120 (20 + 20)	720	80	1.2
Four-block split-plot	12 (3/block)	120 (15 + 15)	720	60	1.33
Fully paired A	6	60 (30 + 30)	720	120	0.83
Fully paired B	6	120 (60 + 60)	1440	240	1.16
Unpaired reader	12	120 (60 + 60)	1440	120	0.90

## Examine trade off between

### Resources

- Number of Readers
- Number of cases
- Number of observations

### Statistical efficiency

$$\frac{\text{var}(\hat{A} \mid \text{Two-block split-plot})}{\text{var}(\hat{A} \mid \text{Study design X})}$$

# Study Designs: Efficiency



**TABLE 3. Resources Needed for Different Study Designs**

Study Design	Number of Readers ( <i>J</i> )	Number of Patients*	Total Number of Image Interpretations	Number of Image Interpretations per Reader	Statistical Efficiency <sup>†</sup>
Two-block split-plot	6 (3/block)	120 (30 + 30)	720	120	1.0
Three-block split-plot	9 (3/block)	120 (20 + 20)	720	80	1.2
Four-block split-plot	12 (3/block)	120 (15 + 15)	720	60	1.33
Fully paired A	6	60 (30 + 30)	720	120	0.83
Fully paired B	6	120 (60 + 60)	1440	240	1.16
Unpaired reader	12	120 (60 + 60)	1440	120	0.90

Take-away 1. It is possible (fairly easy) to compare study designs.

- Simulation
- Modeling

# Study Designs: Efficiency



TABLE 3. Resources Needed for Different Study Designs

Study Design	Number of Readers ( <i>J</i> )	Number of Patients*	Total Number of Image Interpretations	Number of Image Interpretations per Reader	Statistical Efficiency <sup>1</sup>
Two-block split-plot	6 (3/block)	120 (30 + 30)	720	120	1.0
Three-block split-plot	9 (3/block)	120 (20 + 20)	720	80	1.2
Four-block split-plot	12 (3/block)	120 (15 + 15)	720	60	1.33
Fully paired A	6	60 (30 + 30)	720	120	0.83
Fully paired B	6	120 (60 + 60)	1440	240	1.16
Unpaired reader	12	120 (60 + 60)	1440	120	0.90

Take-away 2. You pay a price when you don't pair readers across modalities

- More readers, more cases, more observations
- More variability – lower efficiency

# Study Designs: Efficiency



TABLE 3. Resources Needed for Different Study Designs

Study Design	Number of Readers (J)	Number of Patients*	Total Number of Image Interpretations	Number of Image Interpretations per Reader	Statistical Efficiency <sup>1</sup>
Two-block split-plot	6 (3/block)	120 (30 + 30)	720	120	1.0
Three-block split-plot	9 (3/block)	120 (20 + 20)	720	80	1.2
Four-block split-plot	12 (3/block)	120 (15 + 15)	720	60	1.33
Fully paired A	6	60 (30 + 30)	720	120	0.83
Fully paired B	6	120 (60 + 60)	1440	240	1.16
Unpaired reader	12	120 (60 + 60)	1440	120	0.90

Take-away 3. For the same number of observations, a split-plot study is more efficient.

- Need more cases.

# Study Designs: Efficiency



TABLE 3. Resources Needed for Different Study Designs

Study Design	Number of Readers (J)	Number of Patients*	Total Number of Image Interpretations	Number of Image Interpretations per Reader	Statistical Efficiency <sup>1</sup>
Two-block split-plot	6 (3/block)	120 (30 + 30)	720	120	1.0
Three-block split-plot	9 (3/block)	120 (20 + 20)	720	80	1.2
Four-block split-plot	12 (3/block)	120 (15 + 15)	720	60	1.33
Fully paired A	6	60 (30 + 30)	720	120	0.83
Fully paired B	6	120 (60 + 60)	1440	240	1.16
Unpaired reader	12	120 (60 + 60)	1440	120	0.90

Take-away 4. You can be more efficient by splitting more.

- Need more readers



# Study Designs: Efficiency



**TABLE 3. Resources Needed for Different Study Designs**

Study Design	Number of Readers ( <i>J</i> )	Number of Patients*	Total Number of Image Interpretations	Number of Image Interpretations per Reader	Statistical Efficiency <sup>†</sup>
Two-block split-plot	6 (3/block)	120 (30 + 30)	720	120	1.0
Three-block split-plot	9 (3/block)	120 (20 + 20)	720	80	1.2
Four-block split-plot	12 (3/block)	120 (15 + 15)	720	60	1.33
Fully paired A	6	60 (30 + 30)	720	120	0.83
Fully paired B	6	120 (60 + 60)	1440	240	1.16
Unpaired reader	12	120 (60 + 60)	1440	120	0.90

- Why are split-plot studies efficient?
  - Avoid diminishing returns
  - Observations on a case are correlated

# Study Designs: Efficiency



TABLE 3. Resources Needed for Different Study Designs

Study Design	Number of Readers (J)	Number of Patients*	Total Number of Image Interpretations	Number of Image Interpretations per Reader	Statistical Efficiency <sup>1</sup>
Two-block split-plot	6 (3/block)	120 (30 + 30)	720	120	1.0
Three-block split-plot	9 (3/block)	120 (20 + 20)	720	80	1.2
Four-block split-plot	12 (3/block)	120 (15 + 15)	720	60	1.33
Fully paired A	6	60 (30 + 30)	720	120	0.83
Fully paired B	6	120 (60 + 60)	1440	240	1.16
Unpaired reader	12	120 (60 + 60)	1440	120	0.90

- My rules of thumb:
  - Need 20 cases per class per reader  
-> Need to estimate individual reader performance.
  - Need at least 3 readers per case  
-> Need to estimate reader variability.

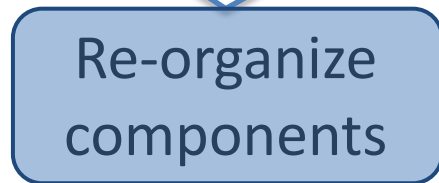
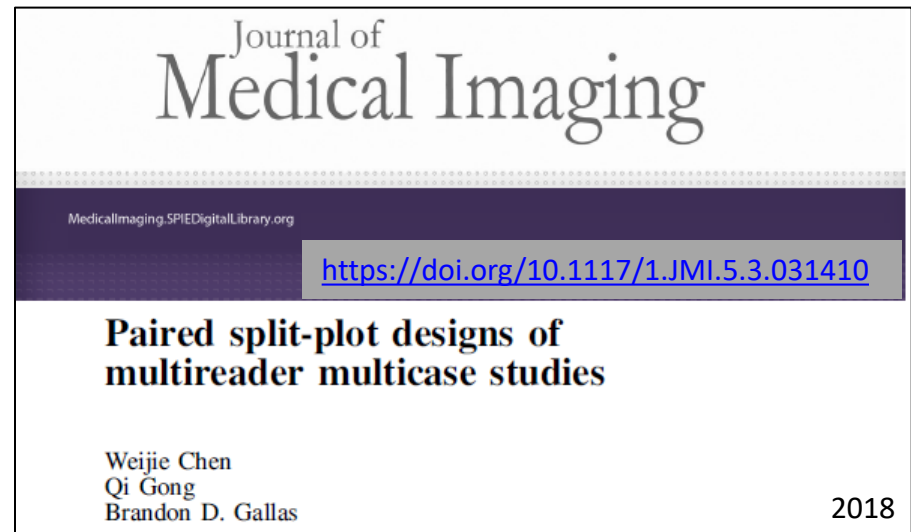
# Study Designs: Efficiency

- Simulation informed theory
  - More groups = less variance

$$\text{var}(\widehat{AUC}_1 - \widehat{AUC}_2)$$

$$= \frac{1}{N_R} V_R + \frac{1}{N_G} V_C$$

Re-organize components

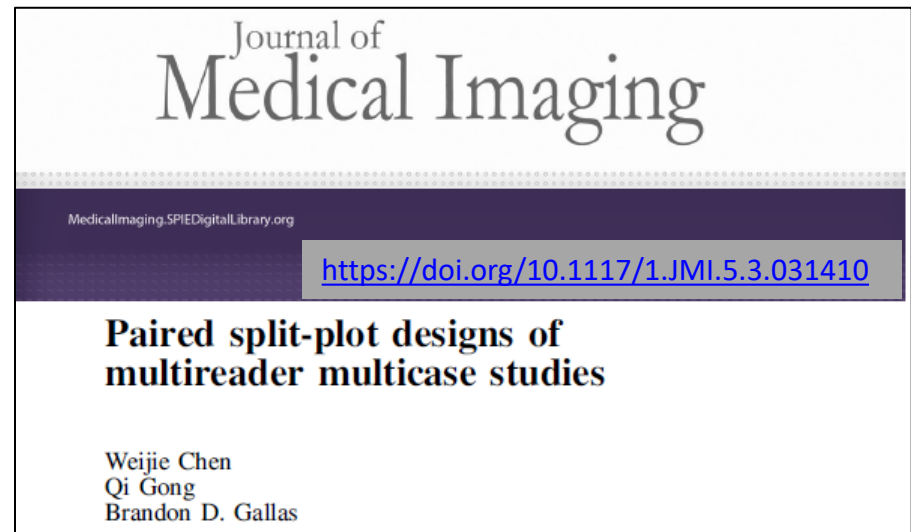
# Study Designs: Efficiency

- Simulation informed theory
  - More groups = less variance

$$\text{var}(\widehat{AUC}_1 - \widehat{AUC}_2)$$

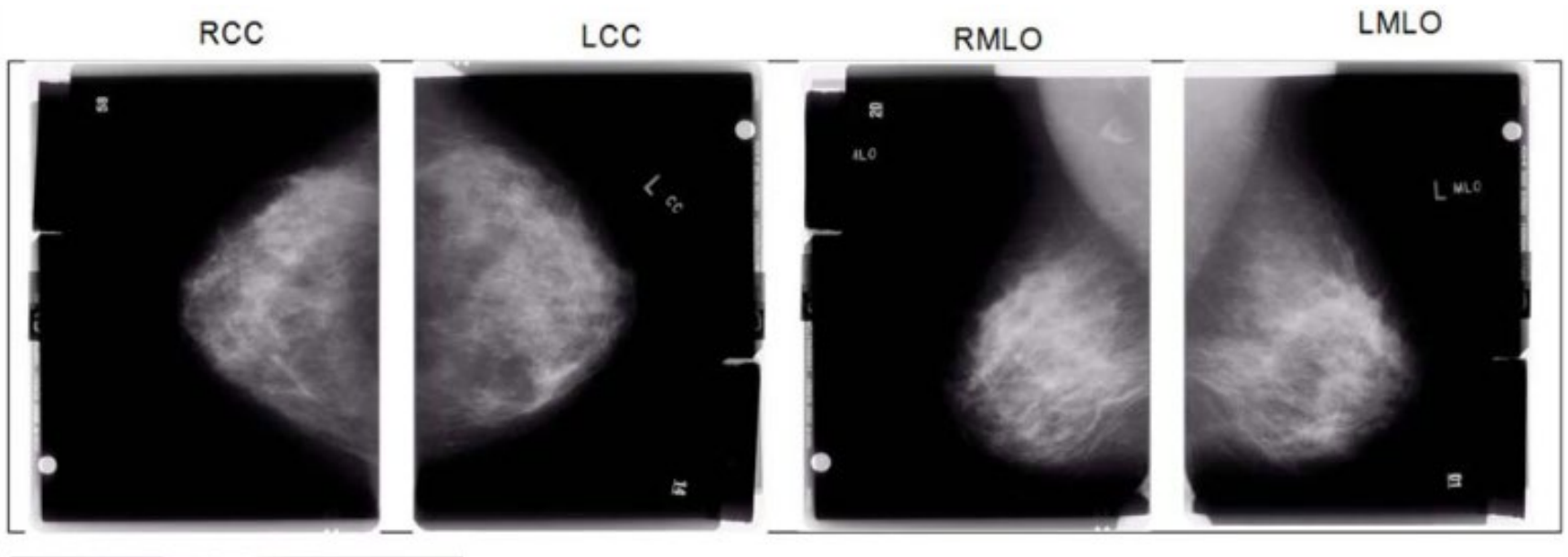
$$= \frac{1}{N_R} V_R + \frac{1}{N_G} V_C$$

More groups  
= less variance



# VIPER Study

Validation of Imaging Premarket Evaluation and Regulation



# VIPER Study

## Purpose and Setting

- Compare large prospective clinical trial to small controlled lab study
  - **DMIST**: Digital Mammography Imaging Screening Trial
    - 42,760 women
    - 1 reader case per FFDM and SFM
    - 85,520 observations
  - **VIPER**: Validation of Imaging Premarket Evaluation and Regulation
    - 716 women (images from DMIST)
    - 20 readers per case per FFDM and SFM
    - 20,382 observations
- Impact of Different Study Populations on Reader Behavior and Performance Metrics
  - Different levels of enrichment (range of prevalence)
  - Screening population vs. Challenge population

# VIPER Reader Study

## Purpose and Setting

aka Clinical  
Performance Study,  
human-in-the-loop

**Compare performance: new imaging system vs reference imaging system**

- **Modality**
  - **FFDM**: Full-field digital mammography vs. **SFM**: Screen-film mammography
- **Task/Performance**
  - Cancer detection: AUC, Sensitivity, Specificity
  - Truth by biopsy and follow up
- **Readers**
  - Mammography Quality Standards Act certified readers
- **Cases**
  - Women with dense breasts →
  - Challenging subgroup!

Why? DMIST found impressive  
performance improvement with FFDM  
AUC(FFDM) = 0.78  
AUC(SFM) = 0.68

# VIPER Study Design

- Original Plan
  1. **Screening study** (prevalence  $p=10\%$ )
    - 270 Non-cancer BIRADS 1-3
    - 30 Non-cancer BIRADS 0
    - 30 Cancers
  2. **Challenge study** (prevalence  $p=50\%$ )
    - 120 Non-cancer BIRADS 0
    - 120 Cancers
  
- Split-Plot research
  
- NEW PLAN
  - **5 sub-studies instead of 2!**

**BIRADS:** Breast Imaging-Reporting and Data System

- BIRADS 1-3 == Do not recall
- BIRADS 0 == Recall

Challenging non-cancers



# VIPER Study

## 5 sub-studies



### Screening Studies

1. 11% “Low” prevalence
2. 29% “Moderate” prevalence
3. 50% “High” prevalence

### Challenge Studies

- ~~1. 11% “Low” prevalence~~
4. 29% “Moderate” prevalence
5. 50% “High” prevalence

# VIPER Study Design

- 20 readers (rows)
- 716 cases (columns)

128 Cancer	156 Non-cancer  BIRADS 0 determined by FFDM	156 Non-cancer  BIRADS 0 determined by SFM	132 Non-cancer  BIRADS 1-3 determined by FFDM	144 Non-cancer  BIRADS 1-3 determined by SFM
---------------	--	---	--	---

Challenging Non-cancer Cases

# VIPER Study

## Split-Plot Study Design

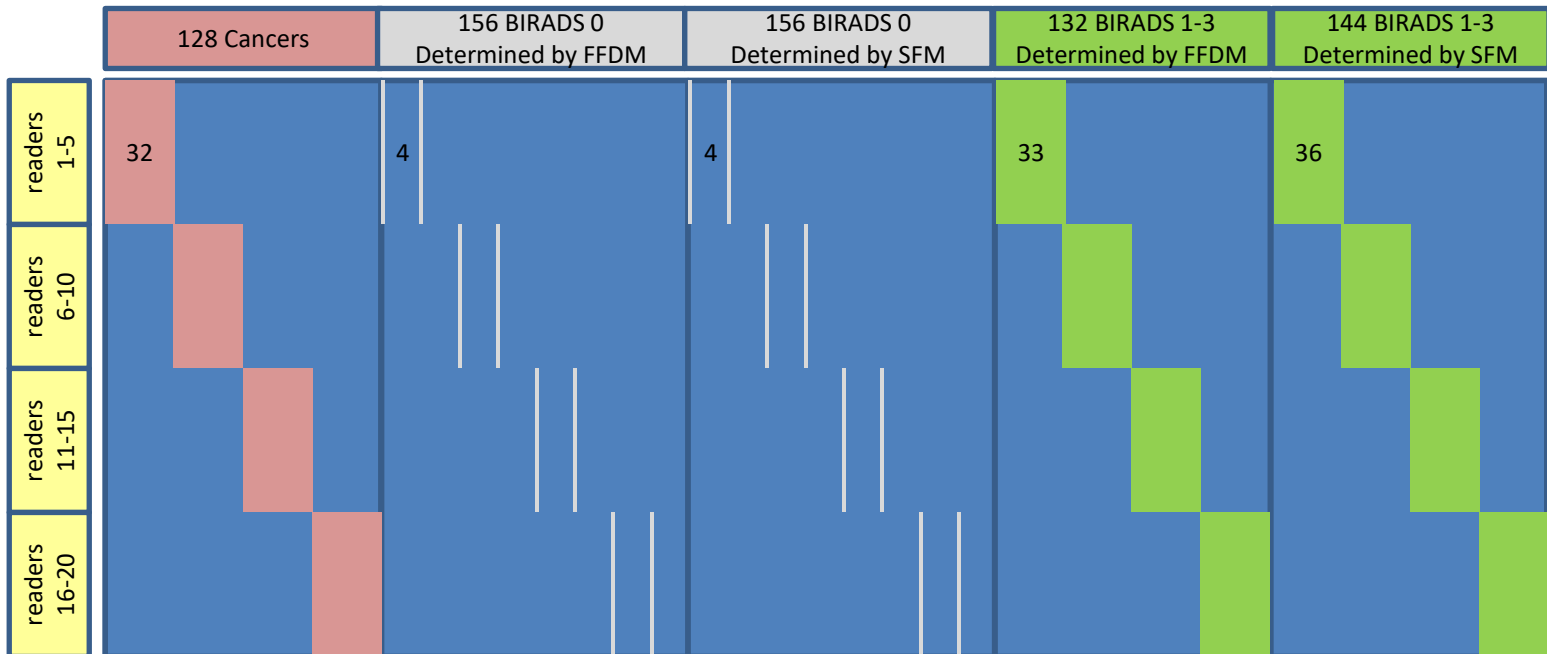
- 20 readers
- 4 split-plot groups

readers 1-5	readers 1-5
readers 6-10	readers 6-10
readers 11-15	readers 11-15
readers 16-20	readers 16-20

# VIPER Study

## Screening Study: 29% “Moderate” Prevalence

- Per modality we have
  - 109 cases per reader
- Total per modality we have
  - 436 cases total, 2180 obs.

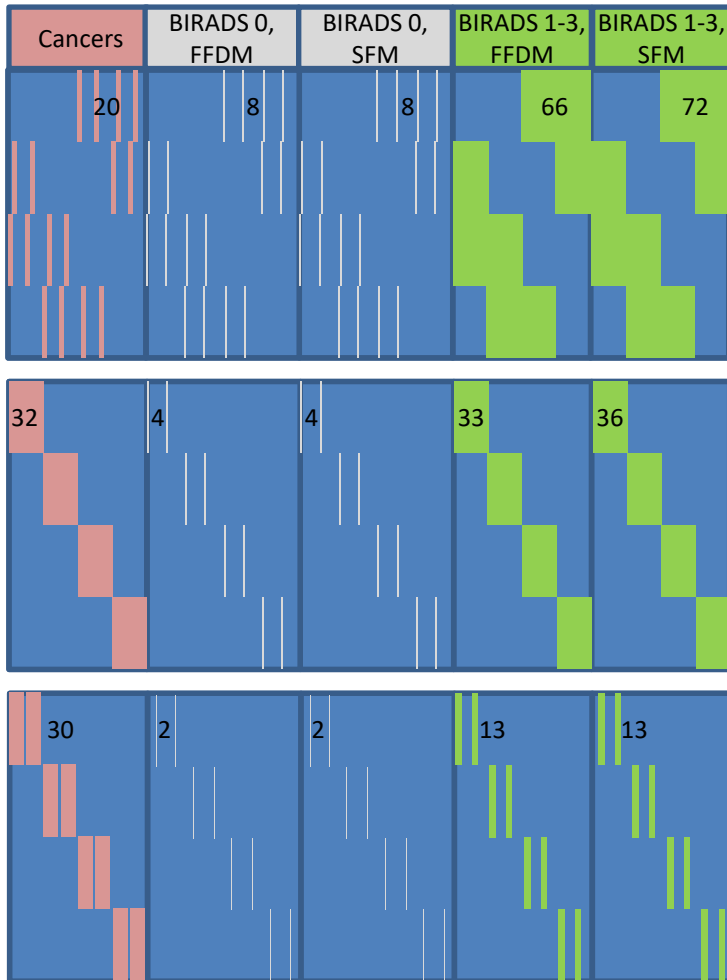


# VIPER Study 5 sub-studies



20,382 observations!  
20 readers per sub-study

Screening Studies

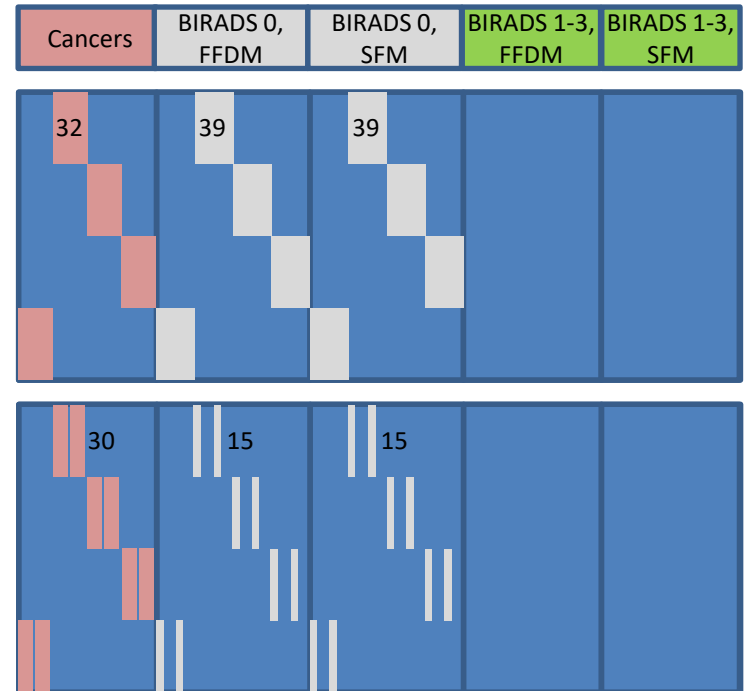


Low Prevalence  
P = 0.11

Moderate Prevalence  
P = 0.29

High Prevalence  
P = 0.50

Challenge Studies



# VIPER Study Design

- Readers could participate in more than one sub-study
  - As long as assigned to groups reading different cases
  - Many did (42 total readers vs. 100)

Recruiting lots of readers was challenging.

- Each sub-study involved two sessions

Set A read in FFDM  
Set B read in SFM

X

Set B read in FFDM  
Set A read in SFM

- Cross-over design with washout (minimum 27 days, mean 68 days, median 50 days)

SPIE 2015  
[https://github.com/DIDSR/iMRMC/blob/gh-pages/000\\_resources/2015SPIE-MIworkshopBDG-4.pdf](https://github.com/DIDSR/iMRMC/blob/gh-pages/000_resources/2015SPIE-MIworkshopBDG-4.pdf)  
 Instructions and eCRF  
[https://github.com/DIDSR/iMRMC/tree/gh-pages/000\\_resources/VIPER](https://github.com/DIDSR/iMRMC/tree/gh-pages/000_resources/VIPER)

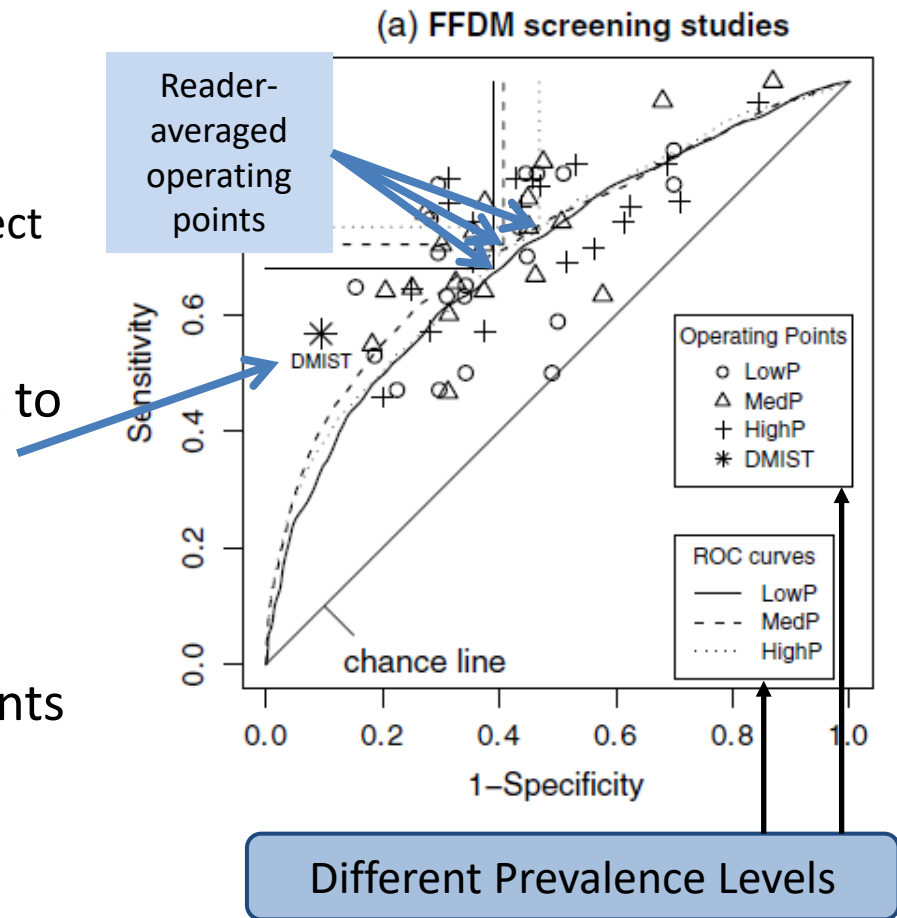
- Two-stage scoring system
  - Recall: yes/no
  - ROC score (202 point scale!)
  - Detailed scoring instructions + description of study population

# VIPER Study Results



## Screening Studies

- Nearly identical ROC curves
- Wide range of operating points
  - Appear overlapping with respect to prevalence
- DMIST operating point furthest to left
  - Lowest prevalence
  - Highest specificity (behavior)
- Reader-averaged operating points move up and to the right with increased prevalence.

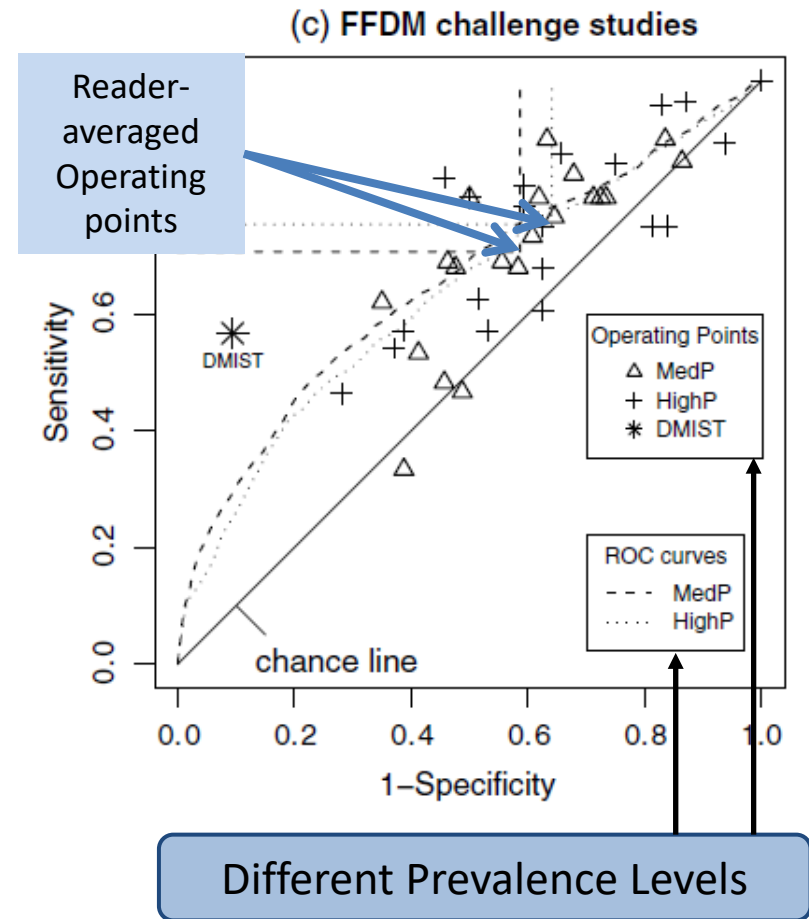


# VIPER Study Results



## Challenge Studies

- Nearly identical ROC curves
- Wide range of operating points
- Very hard task
  - Some points below the chance line!
- Reader-averaged operating points move up and to the right with increased prevalence.





# VIPER Study Results



- No difference in AUC from FFDM and SFM observed
  - Unable to reject null hypothesis
- Result robust to changes in prevalence

Table 3 MRMC performance differences for AUC, sensitivity, and specificity.

Reader study	Prevalence (%)	Number of observations	Difference FFDM-SFM	SE	95% confidence interval
Area under the ROC curve					
ScreeningLowP	10.6	6911	-0.029	0.024	(-0.078, 0.021)
ScreeningMedP	26.6	4325	-0.005	0.024	(-0.054, 0.043)
ScreeningHighP	45.6	2390	-0.025	0.025	(-0.075, 0.024)
ChallengeMedP	26.1	4377	-0.024	0.018	(-0.06, 0.013)
ChallengeHighP	45.2	2379	-0.047	0.023	(-0.093, -0.001)

# VIPER Study Efficiency

- Split-plot study design  
Versus
- Fully-crossed study
- **Less than half the reads**
- **Better precision**

	Viper Split-Plot, 4 groups Low Prevalence	Fully-crossed Low Prevalence
20 readers	SE (# of observations) # observations per reader # cases	SE (# of observations) # observations per reader # cases
Standard Error: AUC	<b>0.023</b> (3480 obs.)	<b>0.041</b> (8700 obs.)
Standard Error: Sensitivity	<b>0.038</b> (400 obs.) 20 cancers per reader <b>80 total</b>	<b>0.056</b> (2540 obs.) 30 cancers per reader <b>30 total</b>
Standard Error: Specificity	<b>0.040</b> (3080 obs.) 154 non-cancers per reader <b>308 total</b>	<b>0.039</b> (6160 obs.) 308 non-cancers per reader <b>308 total</b>

# VIPER Paper

## Journal of Medical Imaging

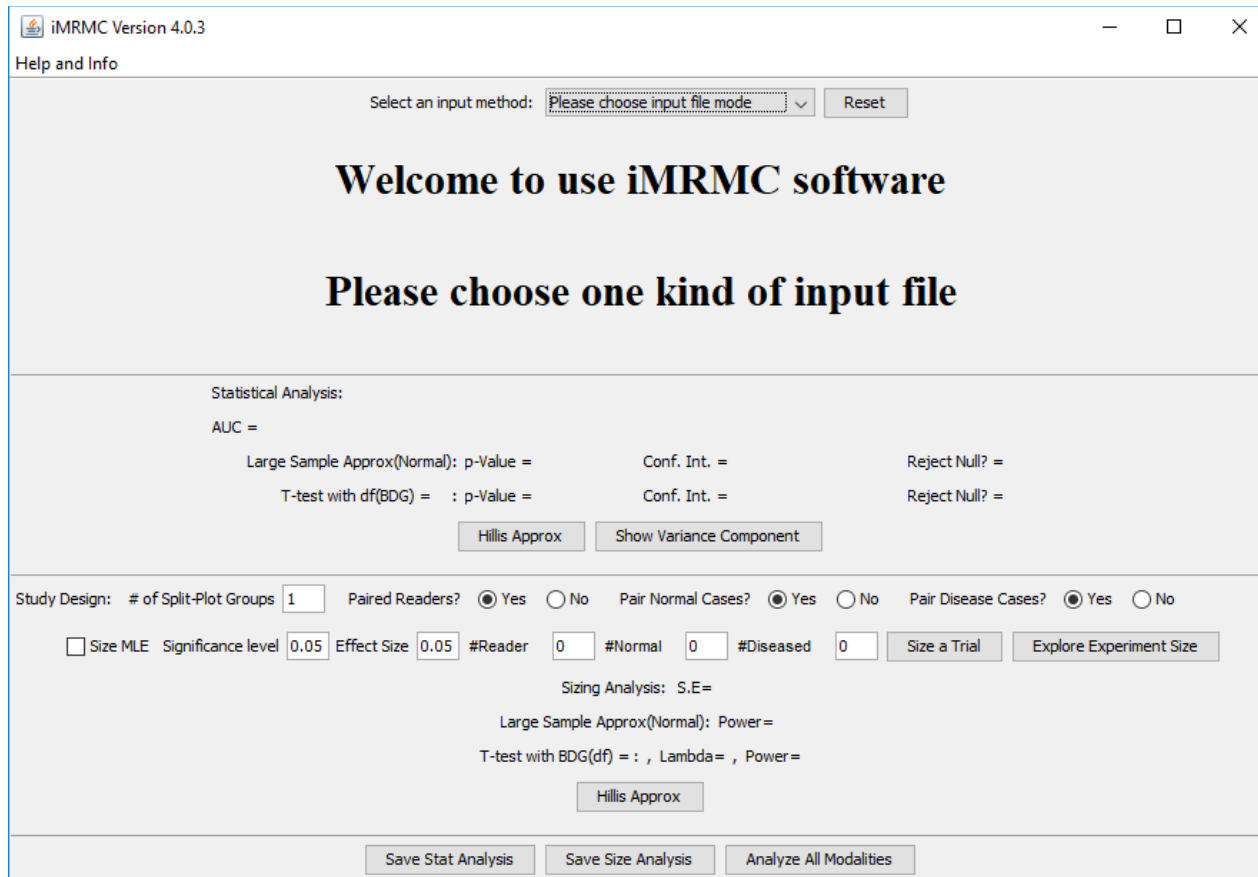
MedicalImaging.SPIEDigitalLibrary.org

doi: [10.1117/1.JMI.6.1.015501](https://doi.org/10.1117/1.JMI.6.1.015501)

### **Impact of prevalence and case distribution in lab-based diagnostic imaging studies**

Brandon D. Gallas  
Weijie Chen  
Elodia Cole  
Robert Ochs  
Nicholas Petrick  
Etta D. Pisano  
Berkman Sahiner  
Frank W. Samuelson  
Kyle J. Myers

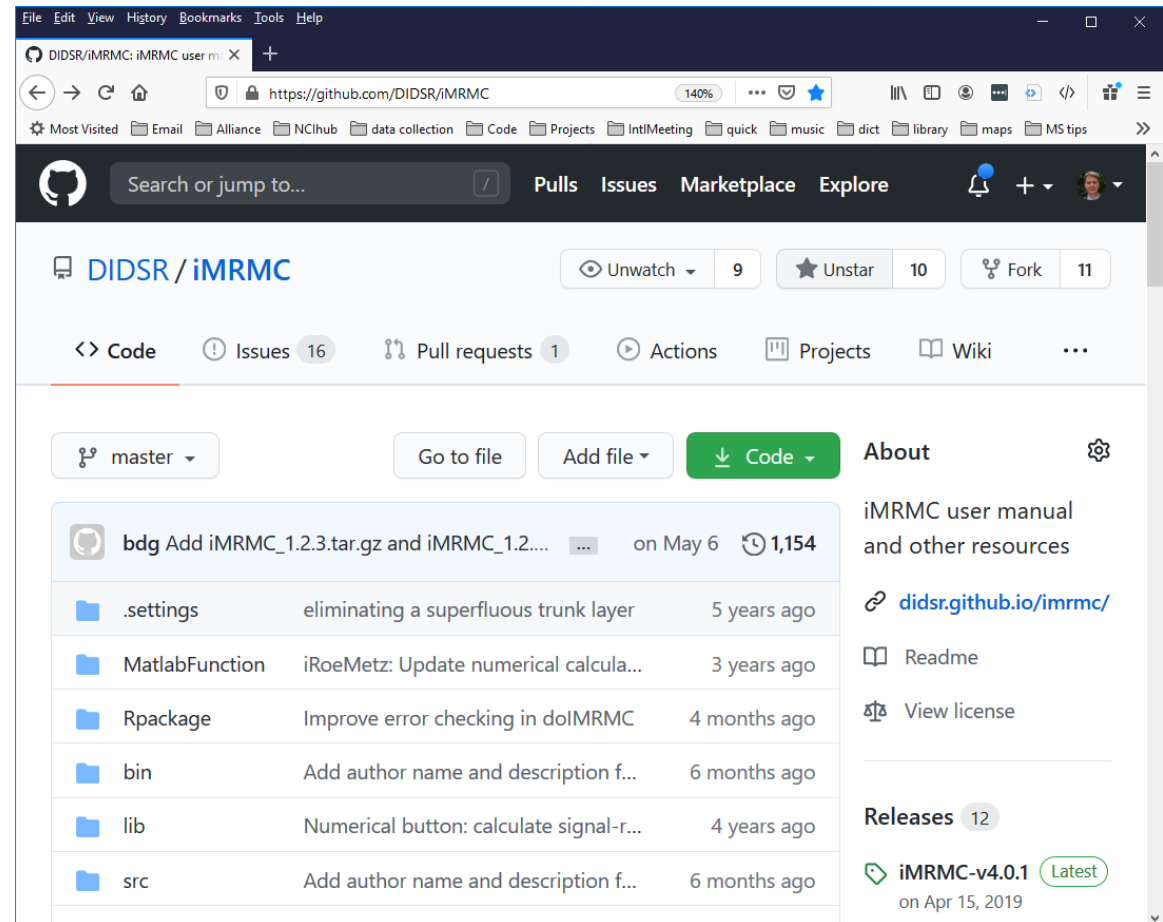
# MRMC Tools



# MRMC Tools

## iMRMC Software, GitHub Repository

- GitHub:
  - Version Control
  - Collaboration
  - Issue tracking
  - Dissemination
- Java Package
- R Package
  - Hosted at CRAN
- iMRMC features
  - Size MRMC study
  - Analyze MRMC study
  - Produce ROC curves
- Wiki
  - Adapt for binary data
  - Links to data packages



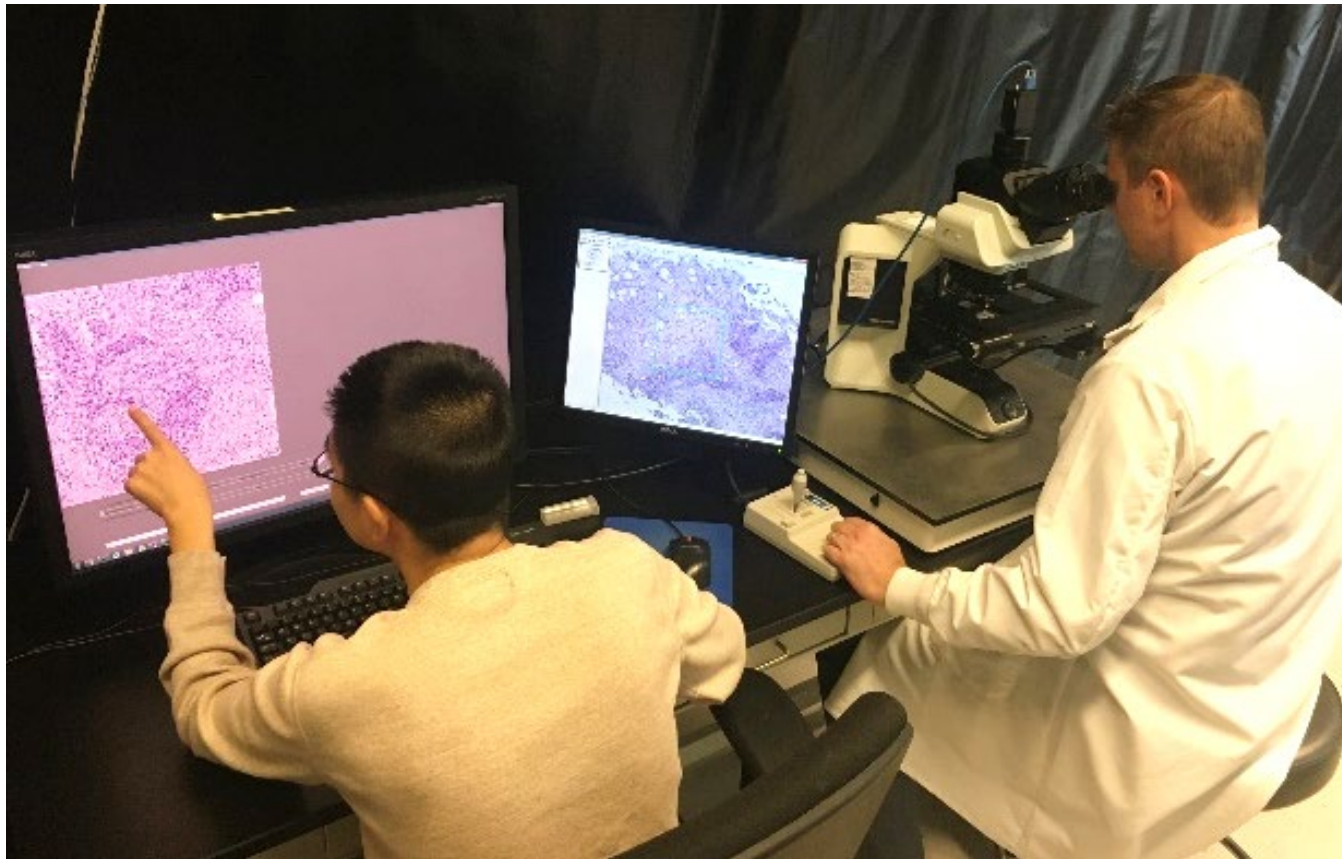
# MRMC Tools

## iMRMC Demo



- Download java app
  - <https://github.com/DIDSR/iMRMC>
  - Backup on desktop
- Check out Wiki
  - Explore VIPER data package
  - Backup in Zotero

# Summary and Future Work



# Summary

- Reader studies compare new imaging modalities to old imaging modalities (*clinical performance*)
  - with the clinician in the loop
  - performing objective tasks
  - on a specific population of cases
- Reader studies are a healthy portion of DIDSr's review responsibilities
- MRMC analyses are not trivial
  - Account for reader and case variability
  - Account for reader and case correlations



# Summary

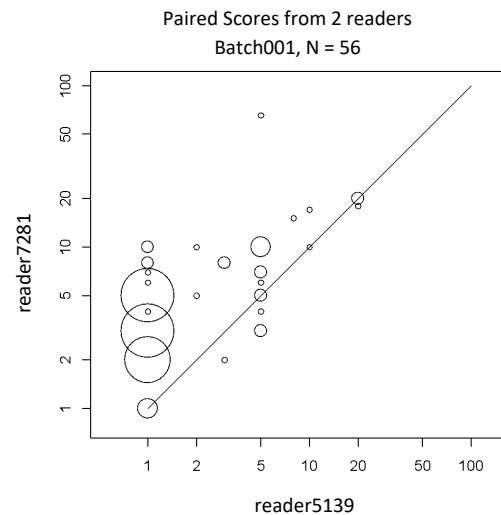
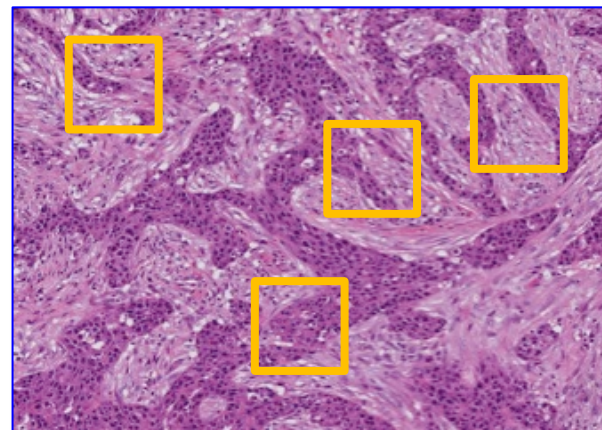
- MRMC variance of AUC framework allows study sizing
  - Variance components
  - Coefficients that correspond to experiment size
  
- Framework (and simulation) allow study of tradeoffs
  - Resources (Number of readers, cases, and observations)
  - Statistical efficiency
  
- Split-plot studies are less burdensome than fully-crossed studies
  - Avoid diminishing returns from collecting correlated data

# Summary

- VIPER study collected 20,382 observations
  - Real radiologists
  - Clinical images
  - Five sub-studies
    - Explore enrichment
    - Explore changes to study population
  - Demonstrated modeling and theory concepts
  - Found AUC to be
    - Robust to enrichment
    - Moderately robust to differences in study population
  - Demonstrated software
  - Reproducible (data and scripts on GitHub)

# Current Work

- Cluster / Nested Data
  - Multiple regions per case
  - Regions within a case are correlated
  - Manuscript submitted
  
- Quantitative Measurements
  - Between-reader agreement
  - Within-reader agreement
  - Algorithm-reader agreement
  
  - Generalizing MRMC methods and simulation
  - Correlation, Mean-squared error
  - Limits of Agreement, Bland-Altman Plots



# Bibliography

- Beiden, S. V., Wagner, R. F., & Campbell, G. (2000). Components-of-variance models and multiple-bootstrap experiments: An alternative method for Random-Effects, receiver operating characteristic analysis. *Acad Radiol*, 7(5), 341–349.
- Chen, W., Gong, Q., & Gallas, B. D. (2018). Paired split-plot designs of multireader multicase studies. *Journal of Medical Imaging*, 5, 031410. <https://doi.org/10.1117/1.JMI.5.3.031410>
- Chen, W., Wunderlich, A., Petrick, N. A., & Gallas, B. D. (2014). Multireader multicase reader studies with binary agreement data: Simulation, analysis, validation, and sizing. *J Med Img*, 1(3), 031011. <https://doi.org/10.1117/1.JMI.1.3.031011>
- Dorfman, D. D., Berbaum, K. S., & Metz, C. E. (1992). Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. *Invest Radiol*, 27(9), 723–731.
- Gallas, B. D. (2006). One-shot estimate of MRMC variance: AUC. *Acad Radiol*, 13(3), 353–362. <https://doi.org/10.1016/j.acra.2005.11.030>
- Gallas, B. D., Bandos, A., Samuelson, F., & Wagner, R. F. (2009). A framework for random-effects ROC analysis: Biases with the bootstrap and other variance estimators. *Commun Stat A-Theory*, 38(15), 2586–2603. <https://doi.org/10.1080/03610920802610084>
- Gallas, B. D., & Brown, D. G. (2008). Reader studies for validation of CAD systems. *Neural Networks Special Conference Issue*, 21(2), 387–397. <https://doi.org/10.1016/j.neunet.2007.12.013>
- Gallas, B. D., Chen, W., Cole, E., Ochs, R., Petrick, N., Pisano, E. D., Sahiner, B., Samuelson, F. W., & Myers, K. J. (2019). Impact of prevalence and case distribution in lab-based diagnostic imaging studies. *Journal of Medical Imaging*, 6(1), 015501. <https://doi.org/10.1117/1.JMI.6.1.015501>
- Gallas, B. D., & Hillis, S. L. (2014). Generalized Roe and Metz ROC model: Analytic link between simulated decision scores and empirical AUC variances and covariances. *J Med Img*, 1(3), 031006. <https://doi.org/doi:10.1117/1.JMI.1.3.031006>
- Gallas, B. D., Pennello, G. A., & Myers, K. J. (2007). Multireader multicase variance analysis for binary data. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, 24(12), B70-80. <https://doi.org/10.1364/josaa.24.000b70>

# Bibliography

- Hillis, S. L. (2014). A marginal-mean ANOVA approach for analyzing multireader multicase radiological imaging data. *Stat Med*, 33(2), 330–360. <https://doi.org/10.1002/sim.5926>
- Metz, C. E. (1978). Basic principles of ROC analysis. *Semin Nucl Med*, 8(4), 283–298. [https://doi.org/10.1016/S0001-2998\(78\)80014-2](https://doi.org/10.1016/S0001-2998(78)80014-2)
- Obuchowski, N. A., Gallas, B. D., & Hillis, S. L. (2012). Multi-Reader ROC studies with Split-Plot Designs: A Comparison of Statistical Methods. *Academic Radiology*, 19(12), 1508–1517. <https://doi.org/10.1016/j.acra.2012.09.012>
- Obuchowski, N. A., & Rockette, H. E. (1995). Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests: An ANOVA approach with dependent observations. *Commun Stat B-Simul*, 24(2), 285–308.
- Pisano, E. D., Gatsonis, C., Hendrick, E., Yaffe, M., Baum, J. K., Acharyya, S., Conant, E. F., Fajardo, L. L., Bassett, L., D’Orsi, C., Jong, R., & Rebner, M. (2005). Diagnostic performance of digital versus film mammography for breast-cancer screening. *N Engl J Med*, 353(17), 1773–1783. <https://doi.org/10.1056/NEJMoa052911>
- Roe, C. A., & Metz, C. E. (1997a). Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic (ROC) data: Validation with computer simulation. *Acad Radiol*, 4(4), 298–303.
- Roe, C. A., & Metz, C. E. (1997b). Variance-component modeling in the analysis of receiver Operating Characteristic (ROC) index estimates. *Acad Radiol*, 4, 587–600.
- Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103(2684), 677. <https://doi.org/10.1126/science.103.2684.677>
- Toledano, A., & Gatsonis, C. A. (1995). Regression analysis of correlated receiver operating characteristic data. *Acad Radiol*, 2(Suppl 1), S30–S36; discussion S61–S64, S70–S71.
- Tosteson, A. N., & Begg, C. B. (1988). A general regression methodology for ROC curve estimation. *Med Decis Making*, 8(3), 204–215.

End of Slide Show, Click to exit.