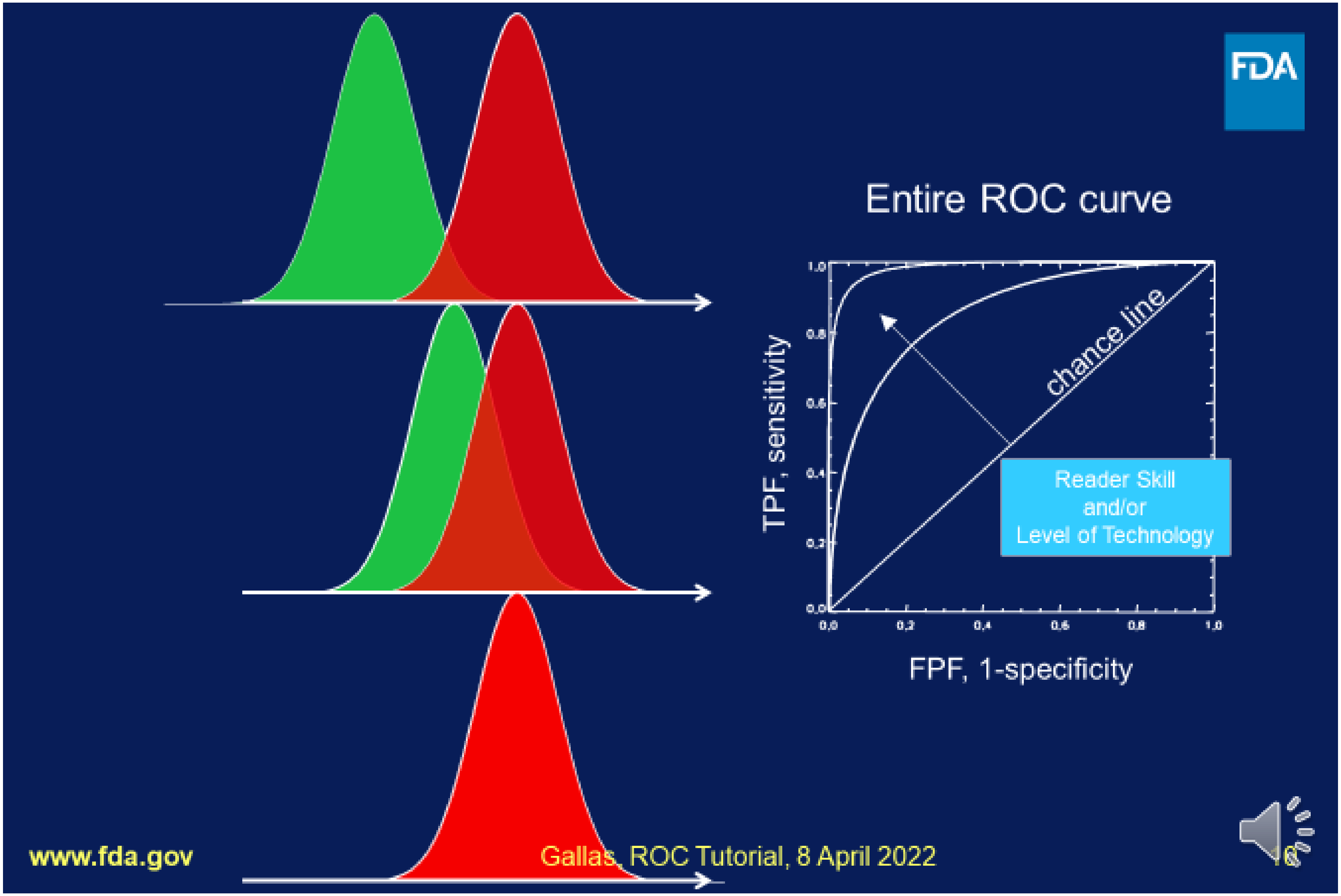
Recommended Prerequisite Training
"ROC Tutorial" by Brandon D. Gallas, 2009.



November 25, 2014

https://dilbert.com/

By Scott Adams

Tutorial on Reader Study Designs and MRMC Analysis

Brandon D. Gallas, PhD

FDA/CDRH/OSEL/DIDSR

Abstract

- Abstract: In this talk I will present the major elements to design, execute, and analyze reader studies. In a reader study, clinicians perform an objective task given medical images. The purpose of a reader study is to assess the clinicians' performance doing the task given the images or to compare performance given images from a new technology to the performance given images from the reference technology. The clinician is part of the technology assessment. The clinicians are the readers and the medical images are the cases. The objective tasks are to make diagnostic evaluations or other measurements given the images (aligned with the intended use of the images). Some refer to the evaluations as subjective because they involve humans who are biased and not fully reproducible. I prefer to refer to the evaluations as objective to distinguish them from evaluations that are, in fact, opinions that have no right answer. An objective evaluation can be compared to truth if truth is available. Of course, there is bias and variability from human evaluations. As such, analysis methods for reader performance (variance estimates, confidence intervals, hypothesis tests) need to account for such bias and variability from the readers while they also account for the bias and variability from the cases. Such an analysis is referred to as a multi-reader, multi-case MRMC analysis. An MRMC analysis can summarize average reader performance that is expected to generalize to the population of readers and the population of cases.

VIPER Paper

Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

doi: 10.1117/1.JMI.6.1.015501

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

2019

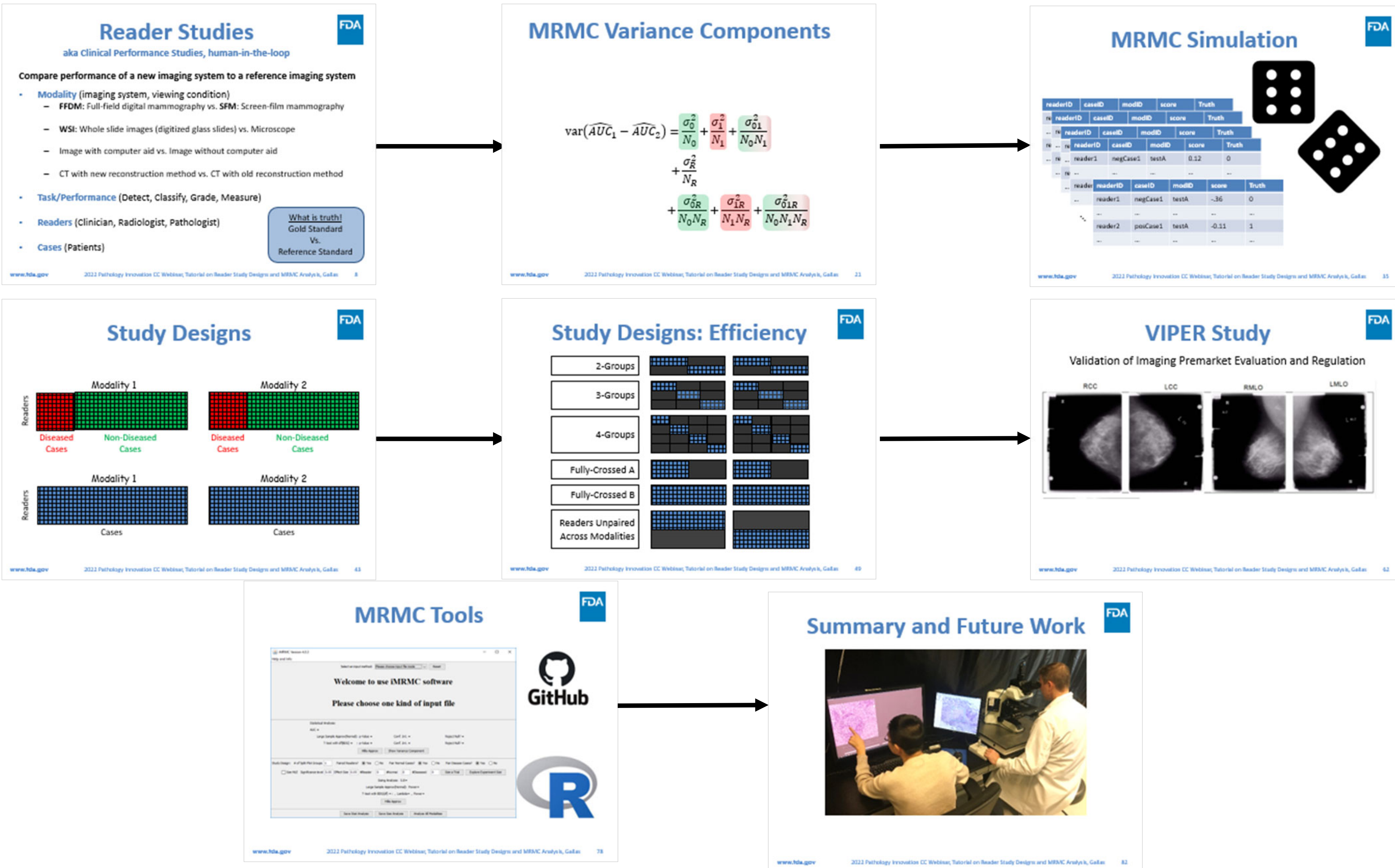
Brandon D. Gallas
Weijie Chen
Elodia Cole
Robert Ochs
Nicholas Petrick
Etta D. Pisano
Berkman Sahiner
Frank W. Samuelson
Kyle J. Myers

Disclaimer

- This is a presentation on the science of reader studies. The content does not describe regulatory requirements.

- The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services. This is a contribution of the U.S. Food and Drug Administration and is not subject to copyright.

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)
What is truth! Gold Standard Vs. Reference Standard
MRMC Variance Components
MRMC Simulation
Study Designs
Modality 1
Modality 2
Readers
Diseased Cases
Non-Diseased Cases
Cases
Study Designs: Efficiency
2-Groups
3-Groups
4-Groups
Fully-Crossed A
Fully-Crossed B
Readers Unpaired Across Modalities
VIPER Study
Validation of Imaging Premarket Evaluation and Regulation
RCC
LCC
RMLO
LMLO
MRMC Tools
Welcome to use iMRMC software
Please choose one kind of input file
GitHub
Summary and Future Work

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)

What is truth!
Gold Standard
Vs.
Reference Standard

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)

Outcome
Followup
Radiology: biopsy/pathology
Expert Panel

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

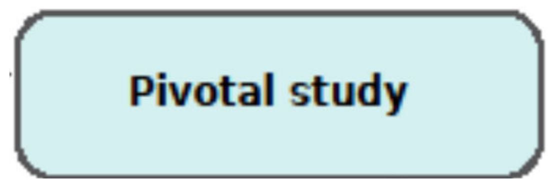
Measure
Quantitative Values

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

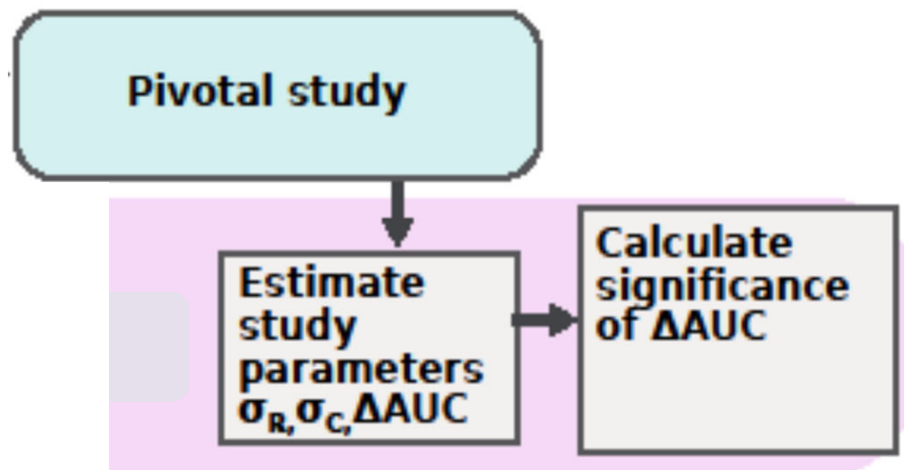
Reader Studies
Study Design



Courtesy Weijie Chen

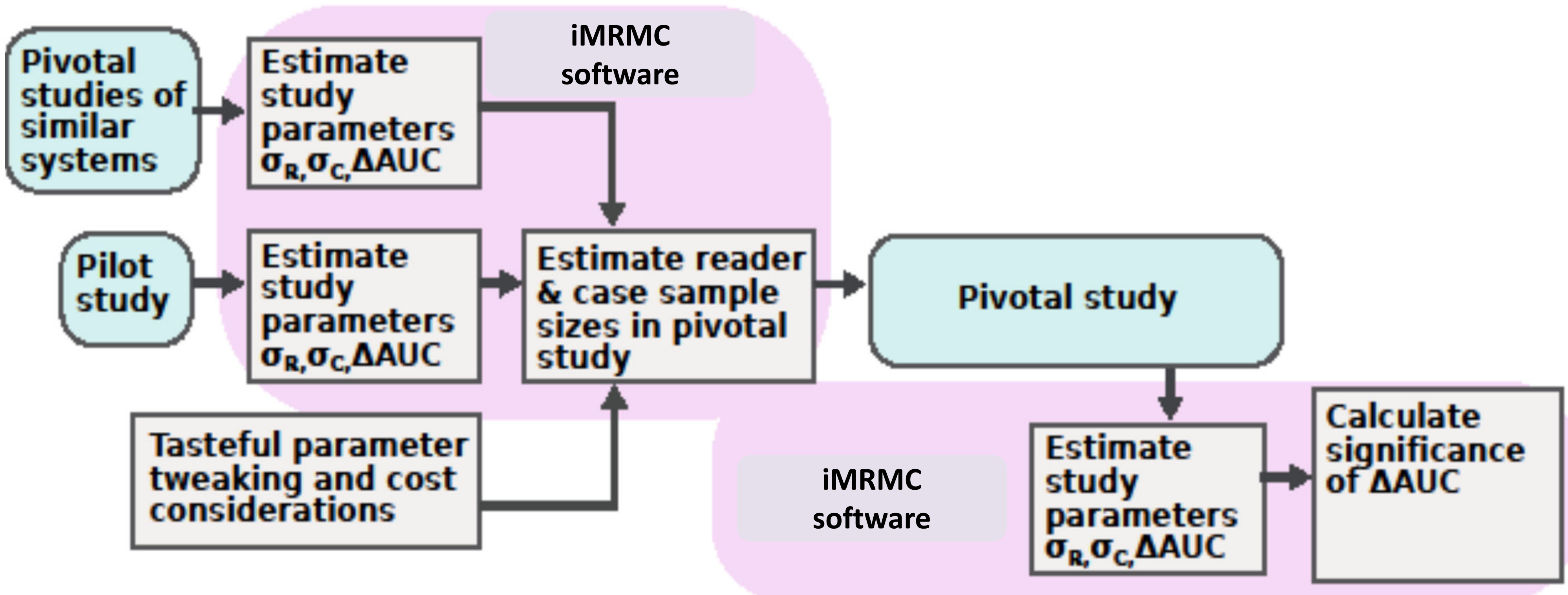
Reader Studies

Study Design



Courtesy Weijie Chen

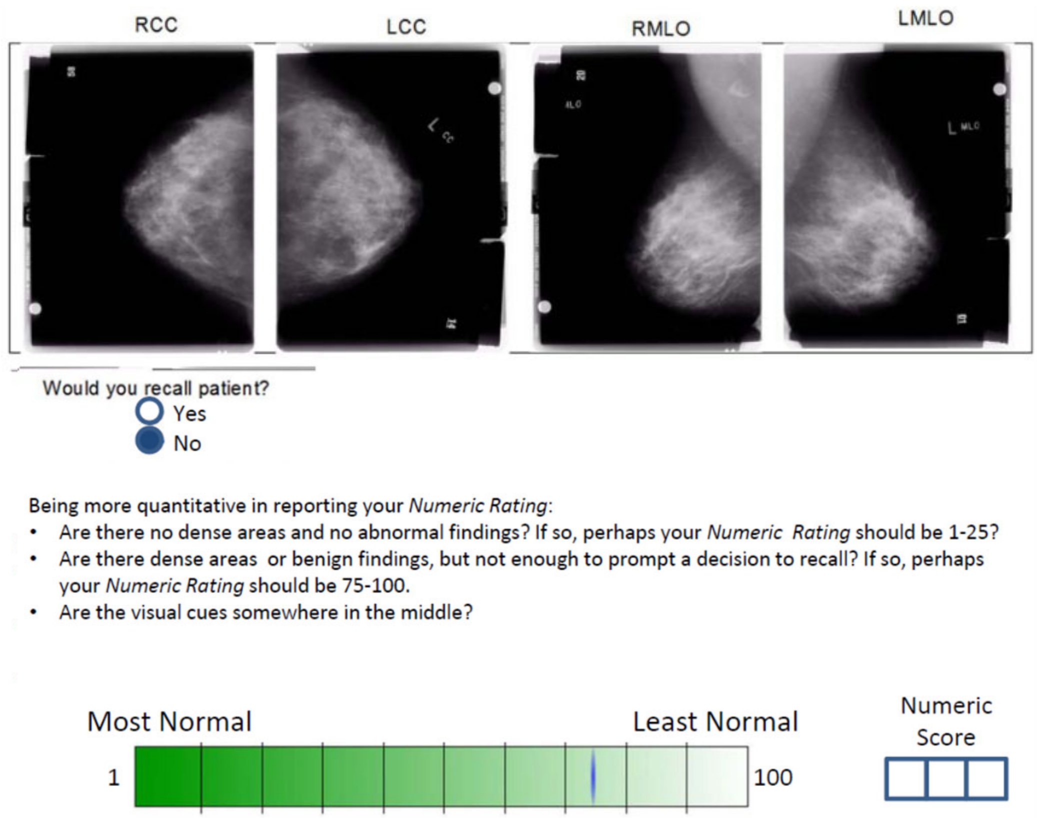
Reader Studies
Study Design



Courtesy Weijie Chen

Reader Studies
Data Collection

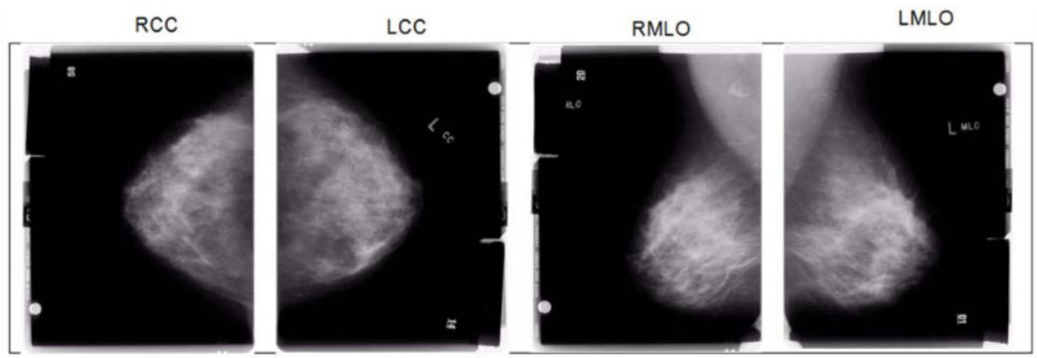
- Two steps
  - Binary patient management decision
  - More information
    (ROC scores)



Reader Studies

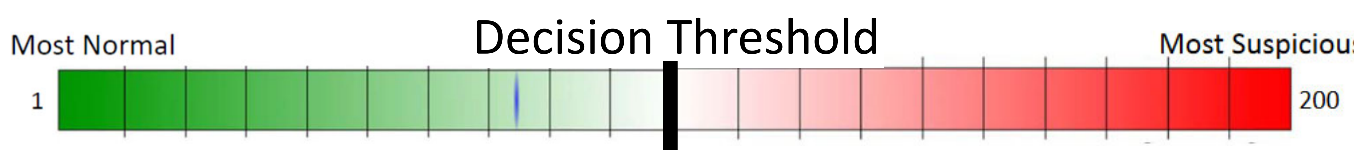
Data Collection

- 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 <u>few inconclusive</u> visual cues prompting your decision to recall? If so, perhaps your *Numeric Rating* should be 101-125?
- Are there <u>many definitive</u> 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?



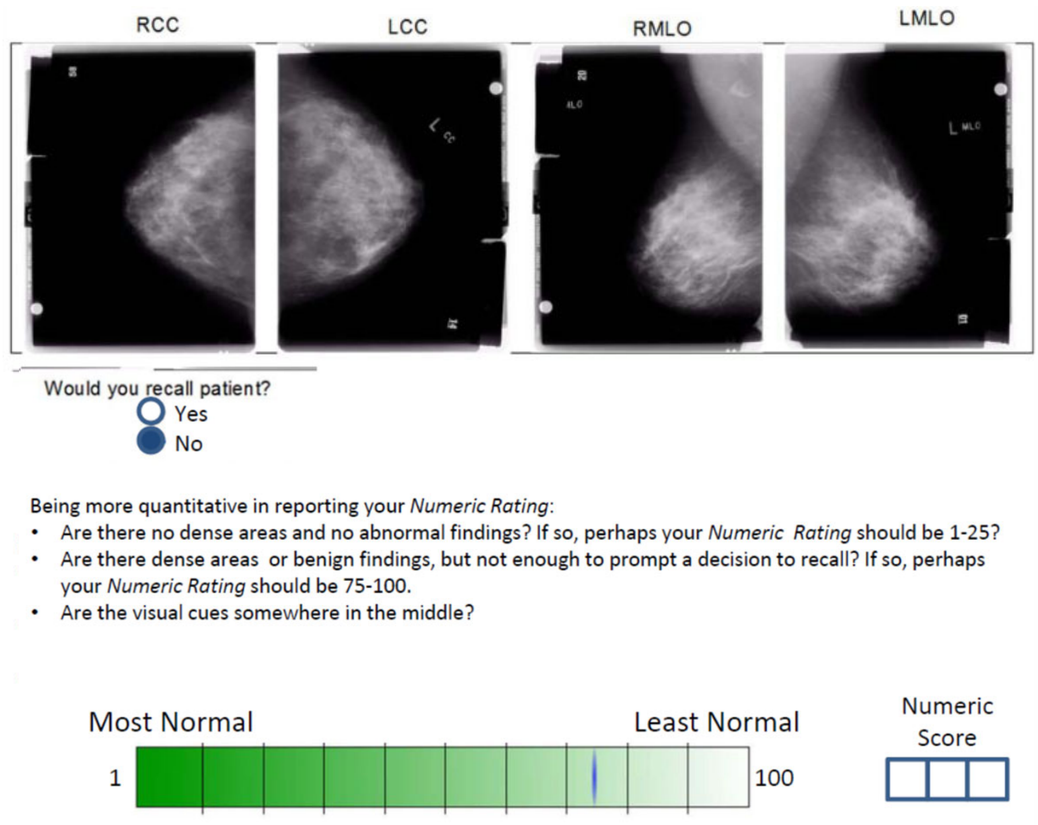
Numeric Score

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



VIPER case report form and ROC scoring instructions
https://didsr.github.io/viperData/

Diagnostic Performance
Reader-averaged AUC

- Nonparametric estimator
  - U-stats, Mann-Whitney, Wilcoxon, Trapezoid

$$\widehat{AUC}_m = \sum_{r=1}^{N_R} \frac{1}{N_R} \sum_{j=1}^{N_1} \frac{1}{N_1} \sum_{i=1}^{N_0} \frac{1}{N_0} s_{mijr}$$

$$s_{mijr} = s\left(y_{mjr} - x_{mir}\right)$$
$$= \begin{cases} 1 & y_{mjr} - x_{mir} > 0 \\ 1/2 & y_{mjr} - x_{mir} = 0 \\ 0 & y_{mjr} - x_{mir} < 0 \end{cases}$$

m: Modality
i: Non-diseased case
j: Diseased case
r: Reader

Diagnostic Performance

Nonparametric AUCs

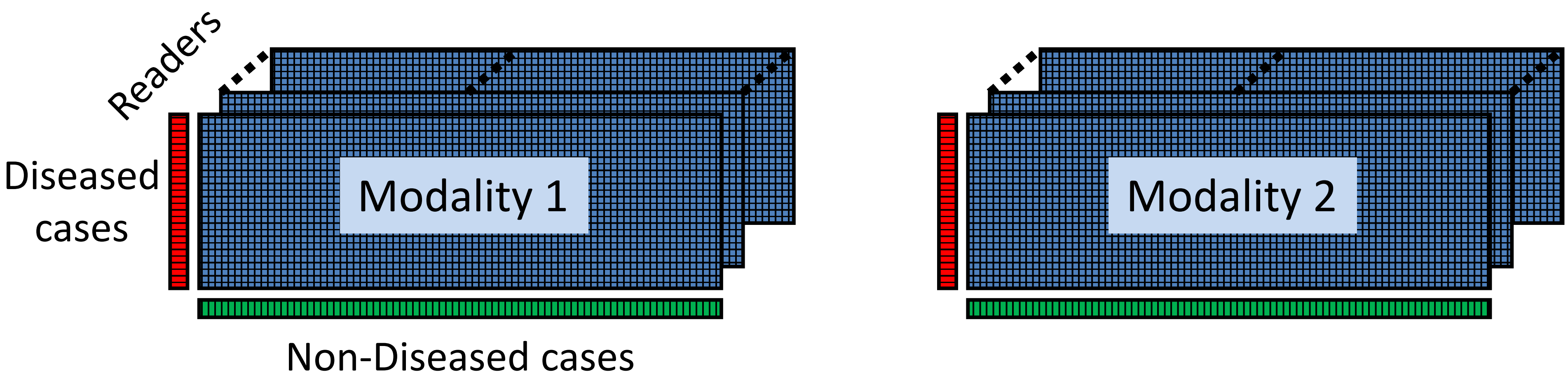
- Average elements of Success Matrix

Estimates

$$\bar{s}_{m\cdot\cdot r} = \widehat{AUC}_{mr}$$
$$\bar{s}_{m\cdots} = \widehat{AUC}_{m}$$

Population Quantities

$$E\left[s_{mijr} | mr\right] = AUC_{mr}$$
$$E\left[s_{mijr} | m\right] = AUC_{m}$$



June 20, 2020

https://dilbert.com/
By Scott Adams

MRMC Variance Components

$$\mathrm{var}\left(\widehat{AUC}_1 - \widehat{AUC}_2\right) = \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1} + \frac{\sigma_{01}^2}{N_0 N_1}$$

$$+ \frac{\sigma_R^2}{N_R}$$

$$+ \frac{\sigma_{0R}^2}{N_0 N_R} + \frac{\sigma_{1R}^2}{N_1 N_R} + \frac{\sigma_{01R}^2}{N_0 N_1 N_R}$$

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)

Non-diseased cases

Diseased cases

Interaction

$$\mathrm{var}(\widehat{AUC}_1) = \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1} + \frac{\sigma_{01}^2}{N_0 N_1}$$

Case Variability

$$+ \frac{\sigma_R^2}{N_R}$$

Reader Variability

$$+ \frac{\sigma_{0R}^2}{N_0 N_R} + \frac{\sigma_{1R}^2}{N_1 N_R} + \frac{\sigma_{01R}^2}{N_0 N_1 N_R}$$

Reader-Case 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)

Non-diseased cases

Diseased cases

Interaction

$$\mathrm{var}\left(\widehat{AUC}_1 - \widehat{AUC}_2\right) = \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1} + \frac{\sigma_{01}^2}{N_0 N_1}$$

Case Variability

$$+ \frac{\sigma_R^2}{N_R}$$

Reader Variability

$$+ \frac{\sigma_{0R}^2}{N_0 N_R} + \frac{\sigma_{1R}^2}{N_1 N_R} + \frac{\sigma_{01R}^2}{N_0 N_1 N_R}$$

Reader-Case Interaction

Different interpretation for these components
- AUC difference

MRMC Variance Components
U-statistic result

- Two Modalities
  - Gallas et al. (2009)

Non-diseased cases

Diseased cases

Interaction

$$\mathrm{var}\left(\widehat{AUC}_1 - \widehat{AUC}_2\right) = \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1} + \frac{\sigma_{01}^2}{N_0 N_1}$$

Case Variability

$$+ \frac{\sigma_R^2}{N_R}$$

Reader Variability

$$+ \frac{\sigma_{0R}^2}{N_0 N_R} + \frac{\sigma_{1R}^2}{N_1 N_R} + \frac{\sigma_{01R}^2}{N_0 N_1 N_R}$$

Reader-Case Interaction

<u>Sizing</u>
Estimate components
Explore N0, N1, NR

MRMC Variance Components
Size a Trial

- NIH/ASCCP sub-study of ALTS [2-3]
  - Atypical Squamous Cells of Undetermined Significance (ASCUS) Low-Grade Squamous Intraepithelial Lesion (LSIL) Triage Study
  - Colposcopy

- 1,000 women enrolled; 939 with evaluable Cervigrams™

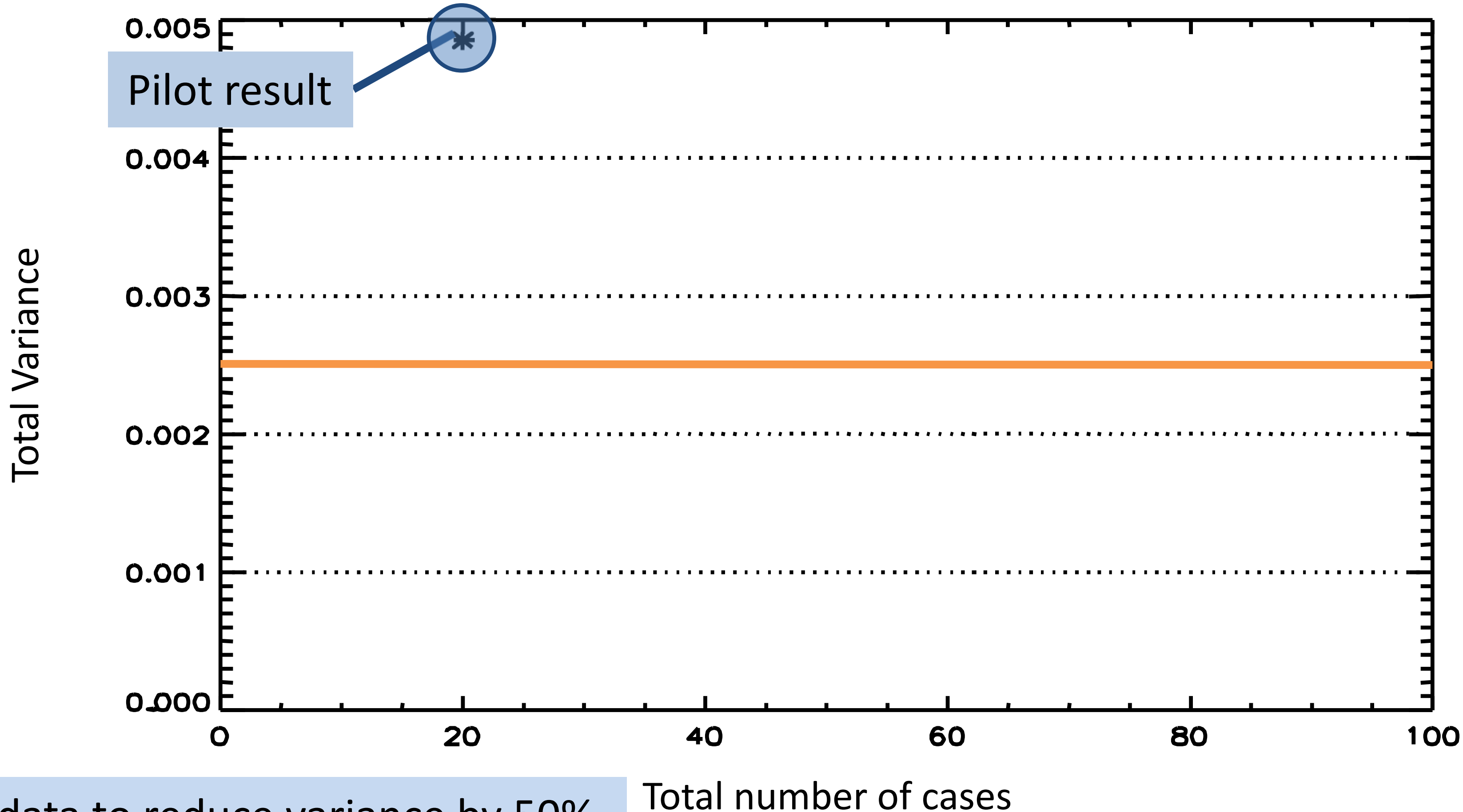
- 21 colposcopists
- 20 patients (16 normal and 4 diseased) had Cervigrams™ read by every reader (420 readings)

- Overall diagnosis for patient

4:1 sampling
-> 25% study prevalence

MRMC Variance Components

Size a Trial



Add data to reduce variance by 50%.

1:1 sampling

4:1 sampling

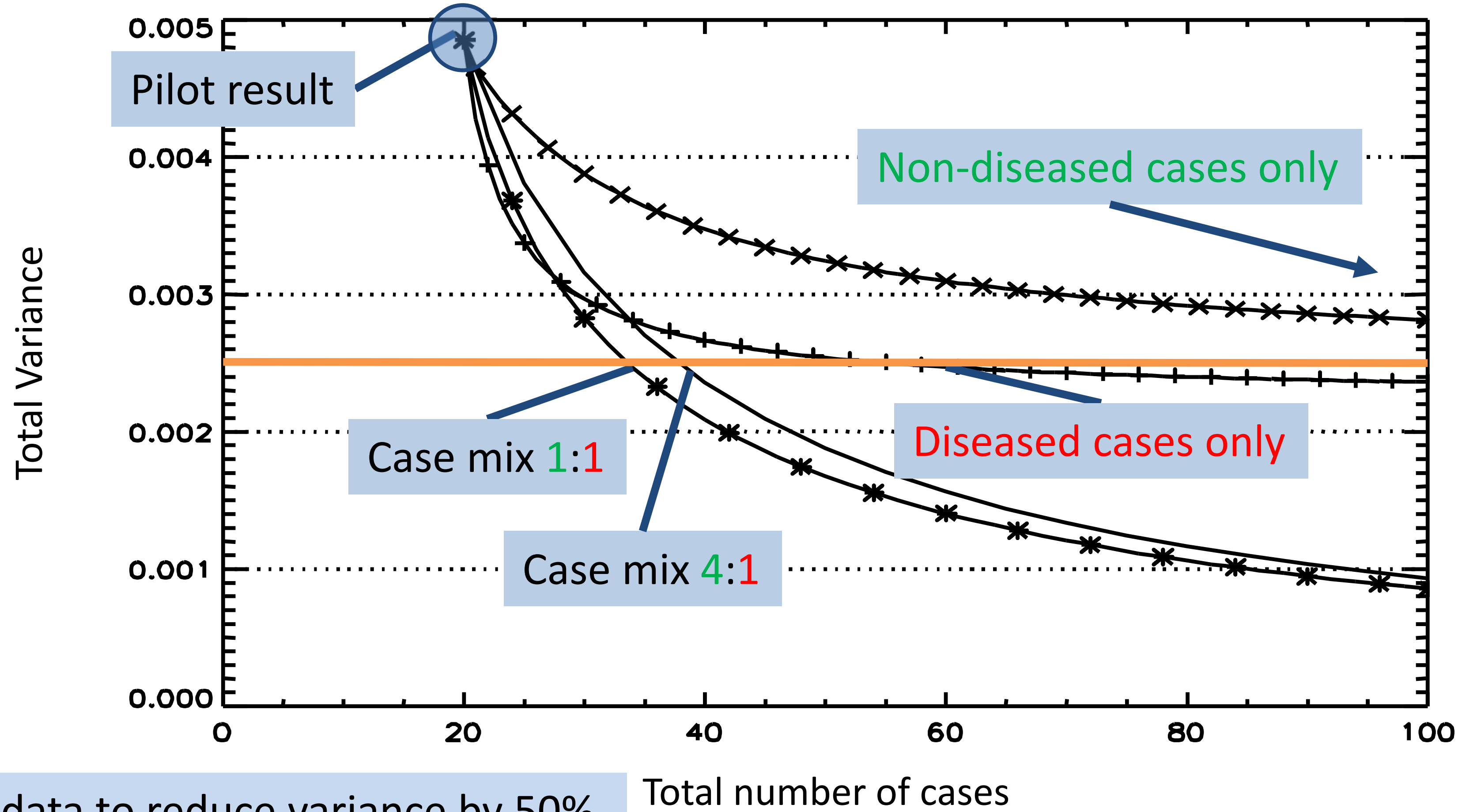
Only Non-Diseased

Only Diseased

Colposcopy Study
Plot courtesy of Hsu, NCI.

MRMC Variance Components

Size a Trial



Add data to reduce variance by 50%.

1:1 sampling	Only Non-Diseased
4:1 sampling	Only Diseased

Colposcopy Study
Plot courtesy of Hsu, NCI.

MRMC Variance Components

One-Shot Estimate of MRMC Variance: AUC[1]

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[1], ANDRIY BANDOS[2], FRANK W. SAMUELSON[1], AND ROBERT F. WAGNER[1]

[1]NIBIB/CDRH Laboratory for the Assessment of Medical Imaging Systems, Silver Spring, Maryland, USA
[2]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 Variance Components
ANOVA - model

- DBM: Dorfman, Berbaum and Metz (1992)
  - 3-way ANOVA: modality, readers, cases
  - Jackknife pseudovalues
- OR: Obuchowski & Rockette (1995)
  - 2-way ANOVA: modality, reader
  - Correlated errors
- Marginal-Mean ANOVA: Hillis (2014)
  - Hypothetical 3-way ANOVA
    no pseudovalues
  - Estimation based on OR

Given U-statistic estimator of AUC

- All representations can be written in terms of the U-statistic components of variance and obtained with simple matrix transformations

Variance in Reader Studies:
Methods & Software

- General Regression,Tosteson and Begg (1988)
- The jackknife/ANOVA, Dorfman, Berbaum and Metz (1992)
    - http://metz-roc.uchicago.edu/MetzROC
- ANOVA and correlation model, Obuchowski (1995)
    - http://www.bio.ri.ccf.org/html/rocanalysis.html
- Ordinal Regression, Toledano and Gatsonis (1995)
- Bootstrap, Beiden, Wagner, and Campbell (2000)
- U-statistics, Gallas
    - http://js.cx/~xin/index

Variance Representations

2-way ANOVA & Correlated Errors
OR & mm-ANOVA

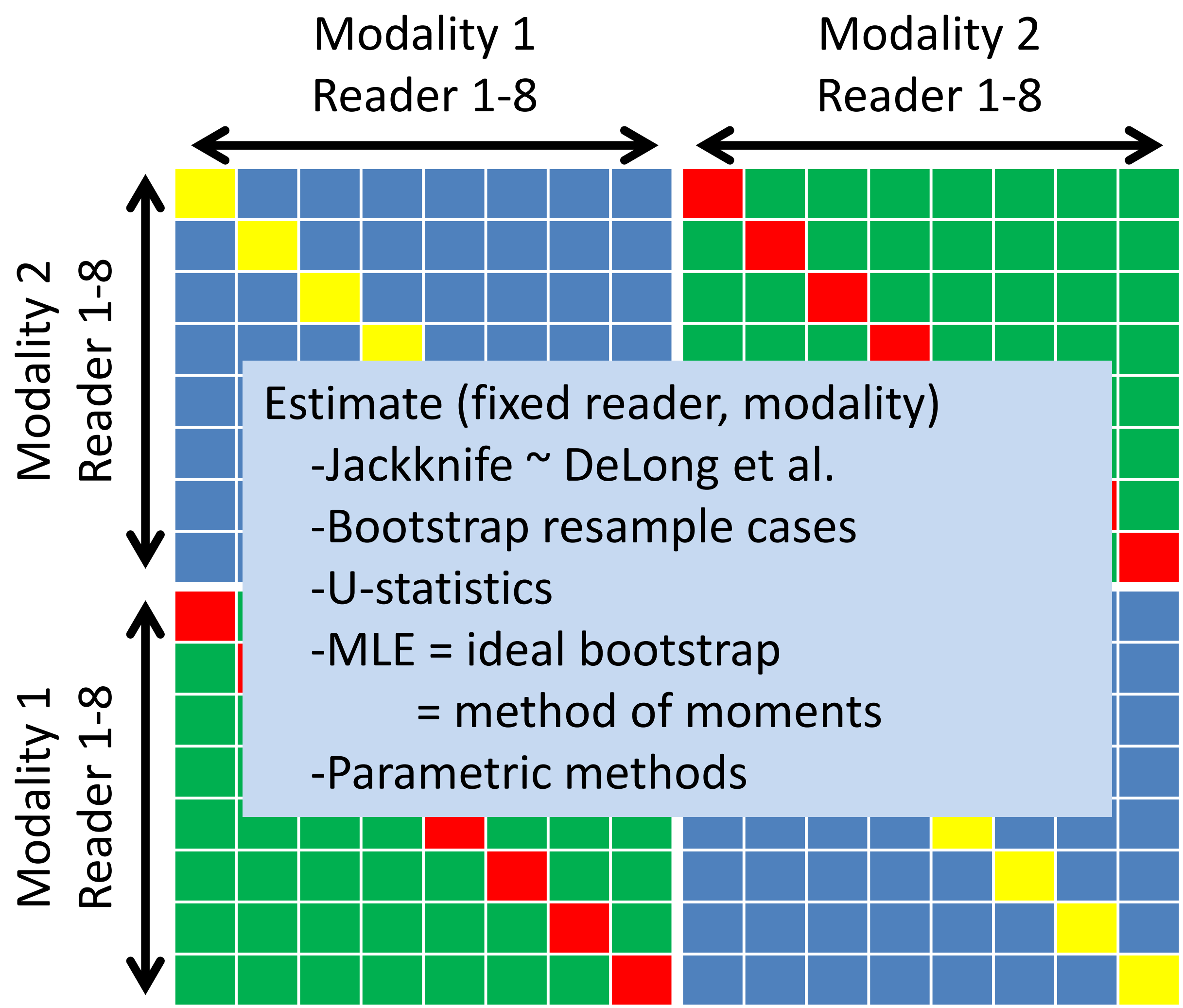
$$\mathrm{cov}\left(\widehat{\mathrm{AUC}}_{mr}, \widehat{\mathrm{AUC}}_{mr} \,|mr\right)$$

Var_ε =
Same modality, same reader

Cov1 =
Diff modality, same reader

Cov2 =
Same modality, Diff reader

Cov3 =
Diff. modality, Diff. reader



Estimate (fixed reader, modality)
- -Jackknife ~ DeLong et al.
- -Bootstrap resample cases
- -U-statistics
- -MLE = ideal bootstrap
  = method of moments
- -Parametric methods

February 3, 2021

https://dilbert.com/

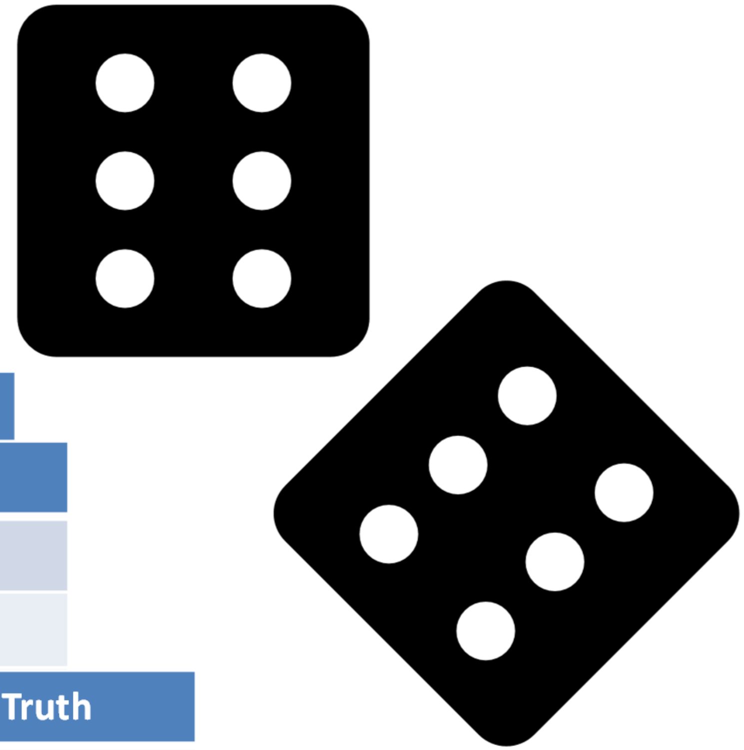
By Scott Adams

MRMC Simulation

readerID	caseID	modID	score	Truth
reader1	negCase1	testA	0.12	0
…	…	…	…	…

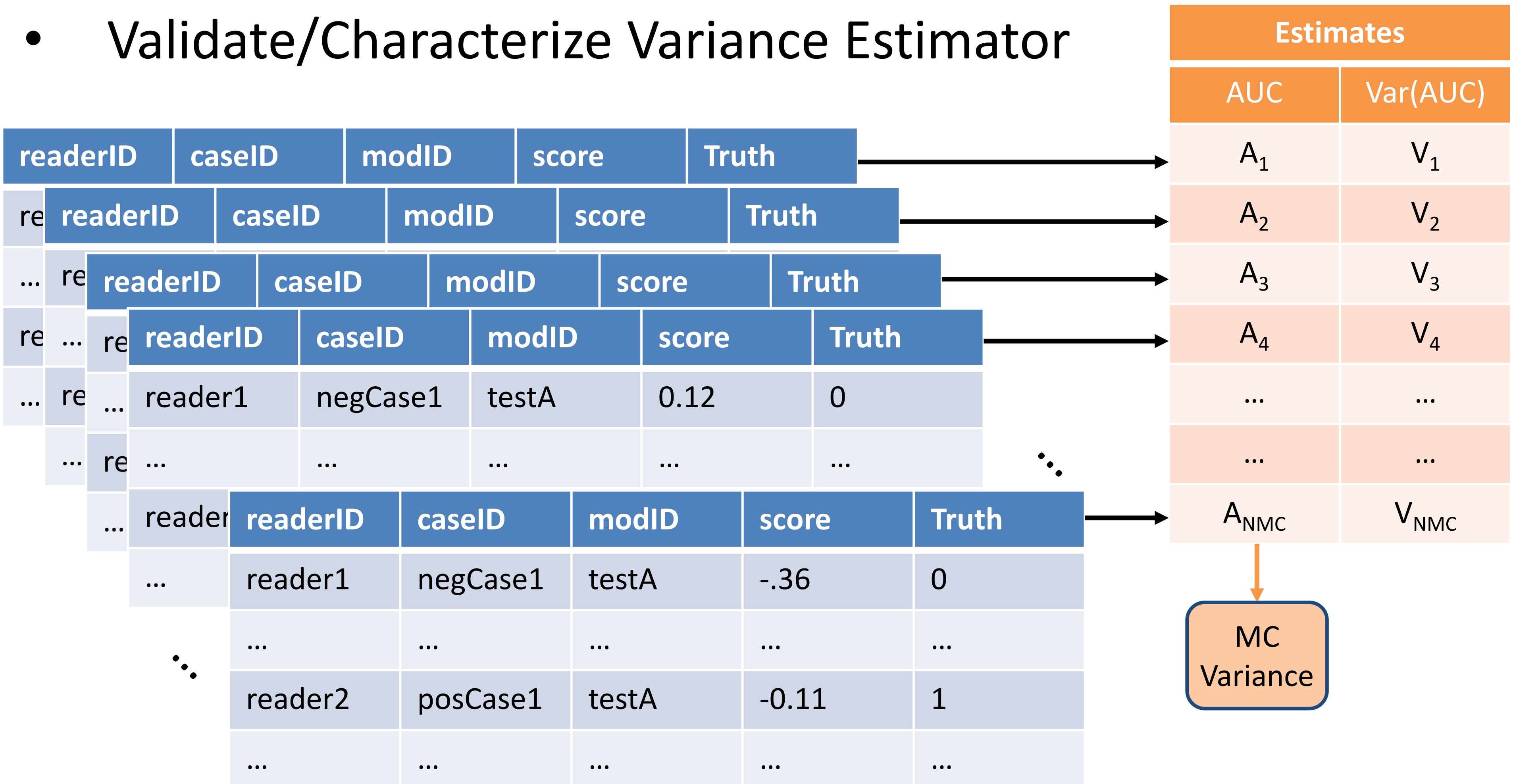
⋱

readerID	caseID	modID	score	Truth
reader1	negCase1	testA	-.36	0
…	…	…	…	…
reader2	posCase1	testA	-0.11	1
…	…	…	…	…



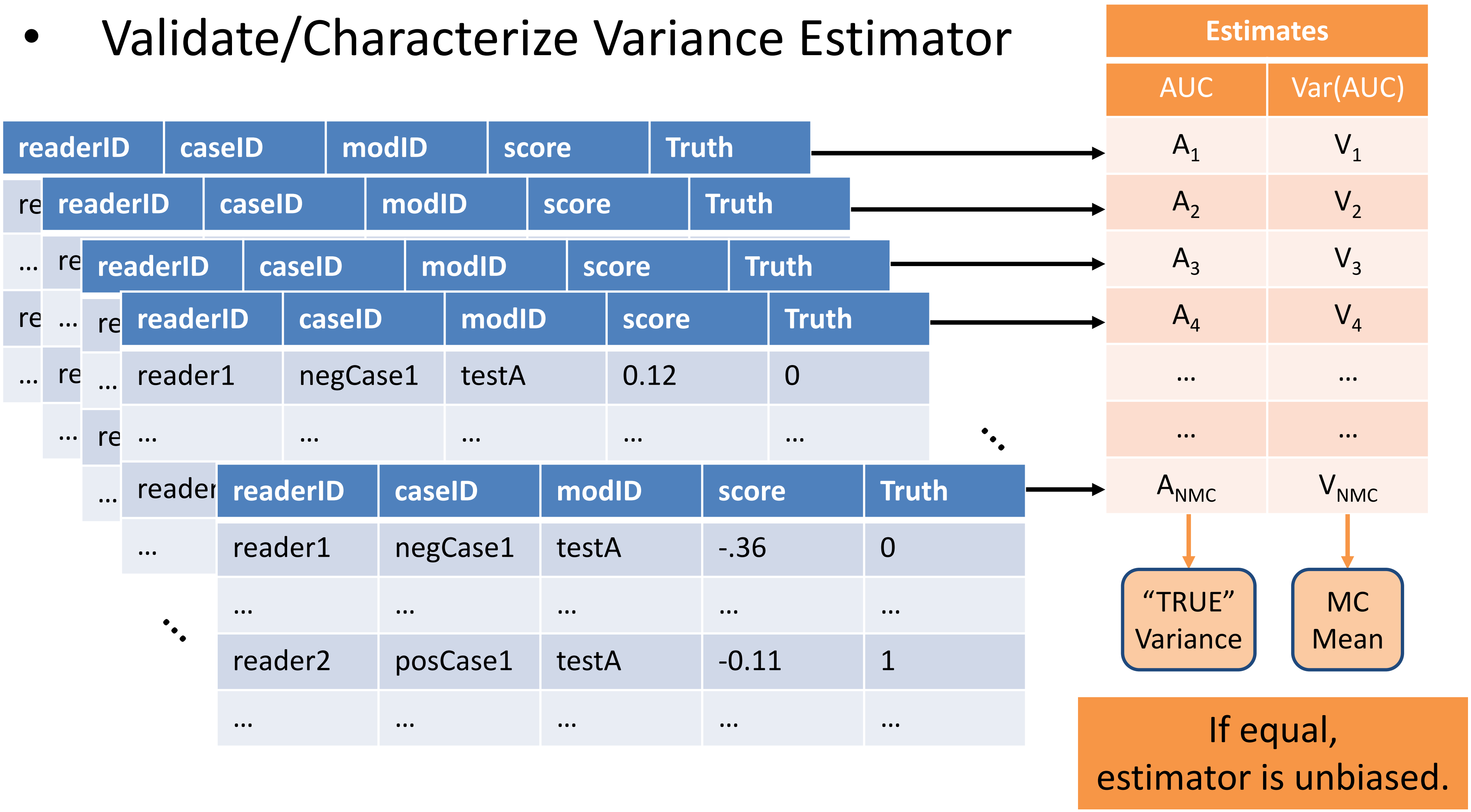
MRMC Simulation

- Validate/Characterize Variance Estimator



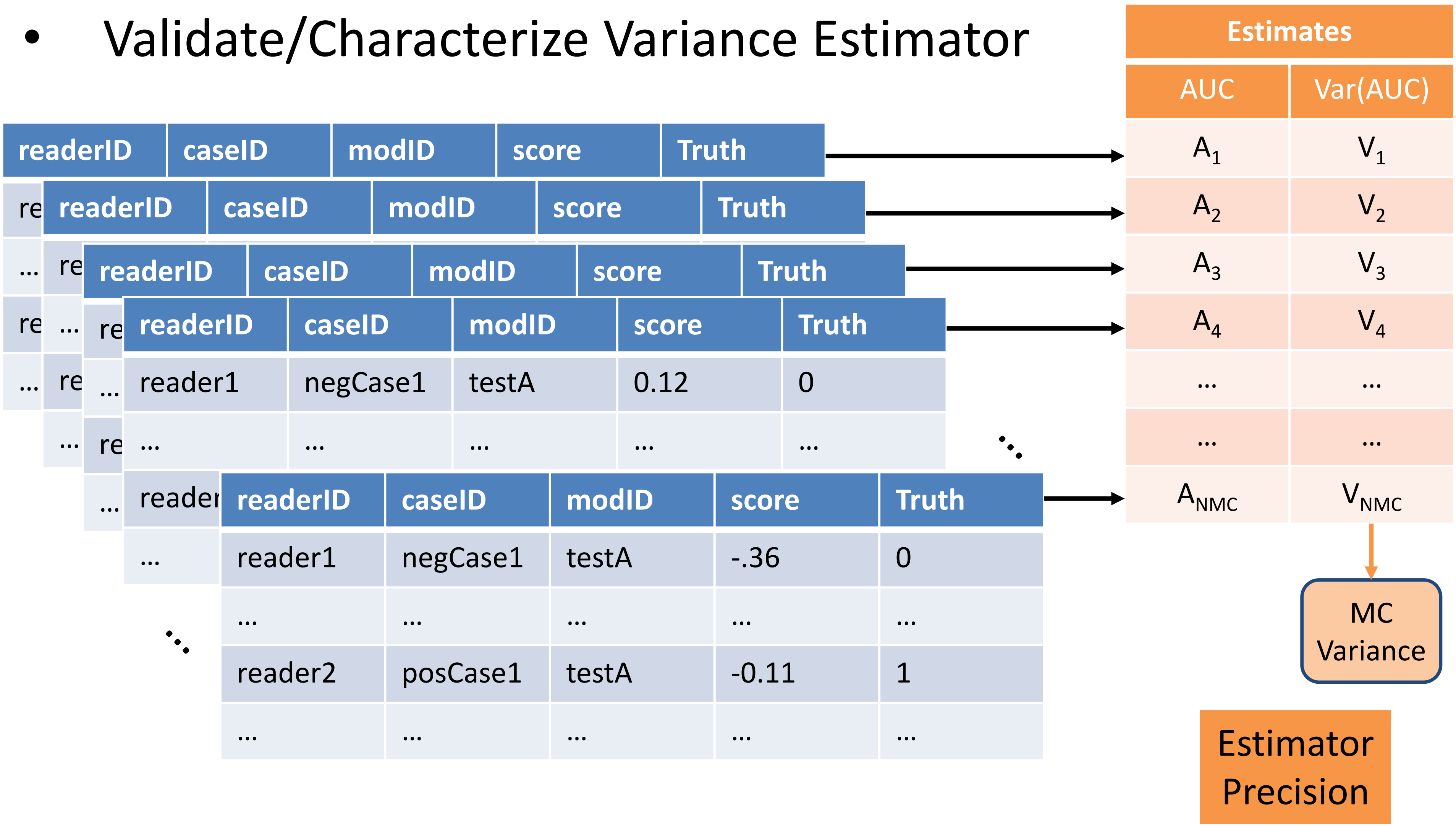
MRMC Simulation

- 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

$$X_{ijk0} = \tau_{i0}$$
$$+C_{k0} \quad + [\tau C]_{ik0}$$
$$+R_{j0} \quad + [\tau R]_{ij0}$$
$$+[RC]_{jk0} + [\tau RC]_{ijk0}$$

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

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

Looks like
3-way ANOVA

Warning
Simulation for scores not AUC

MRMC Simulation

Build on Roe and Metz model

- Binary Data

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

Multireader multicase variance analysis for binary data

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

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

Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

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

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

2014

Weijie Chen
Adam Wunderlich
Nicholas Petrick
Brandon D. Gallas

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

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

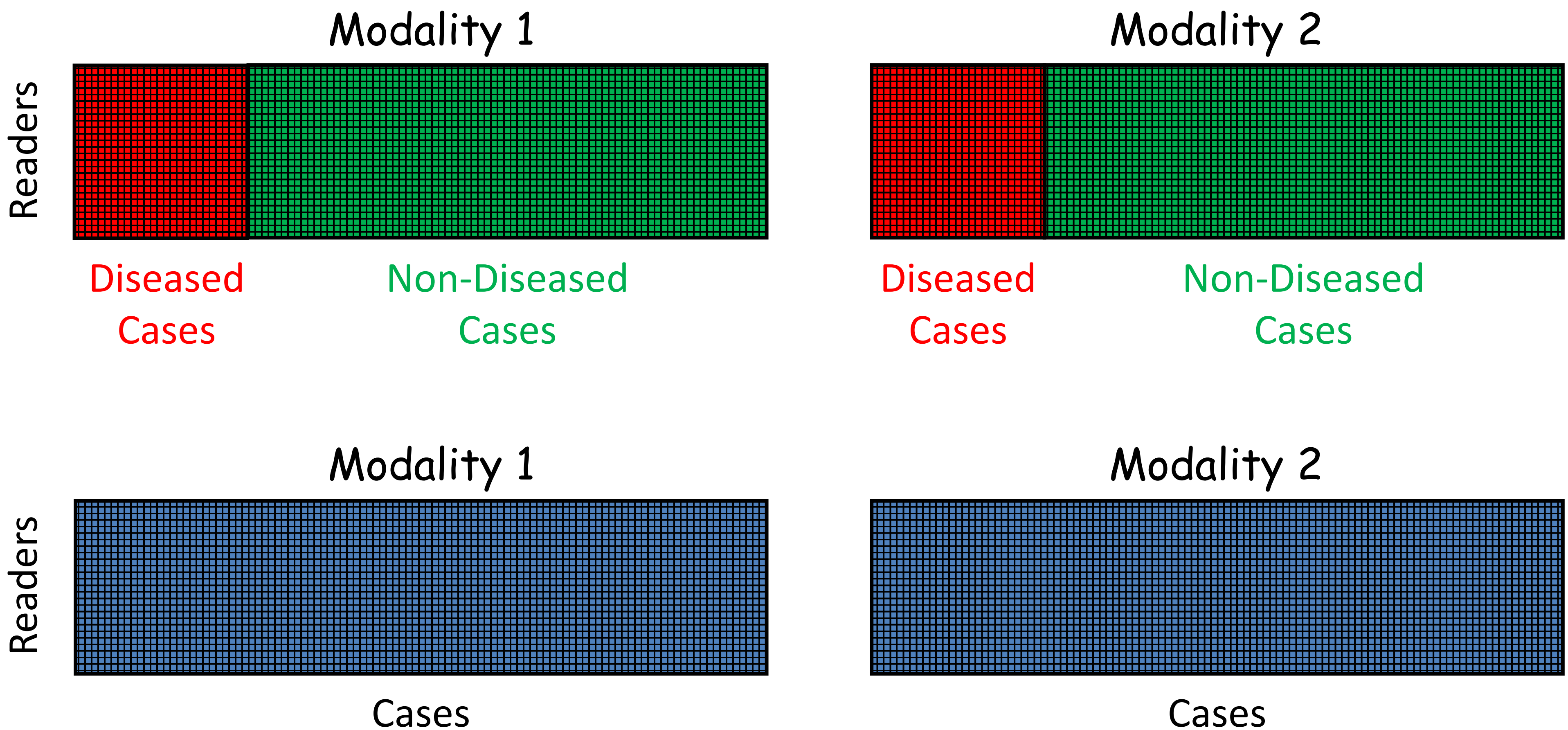
Brandon D. Gallas
Stephen L. Hillis

2014

March 24, 2015

https://dilbert.com/
By Scott Adams

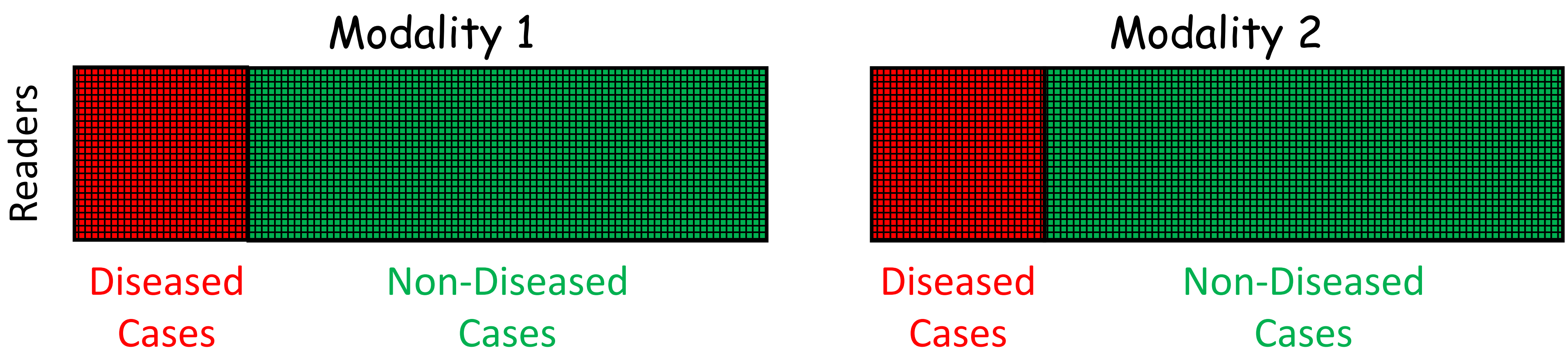
Study Designs



Study Designs
Fully-Crossed

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

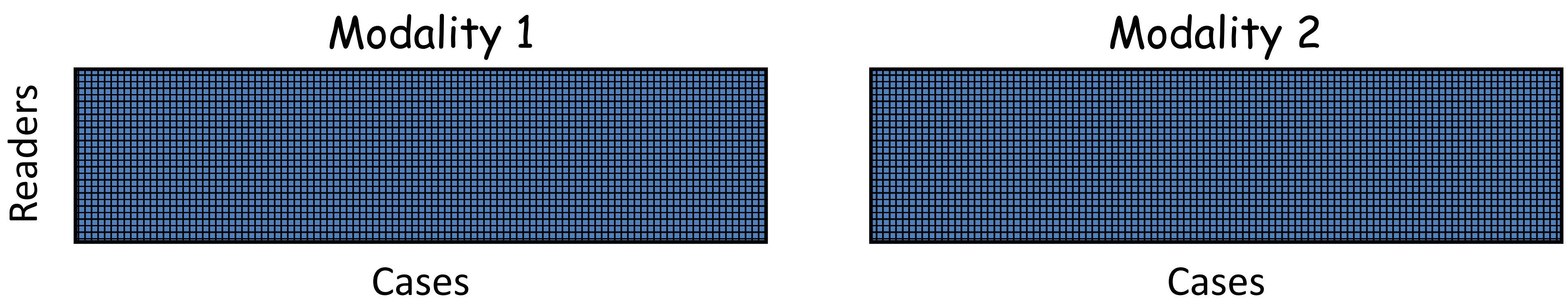
Data Array
Rows = readers
Cols = cases



Study Designs
Fully-Crossed

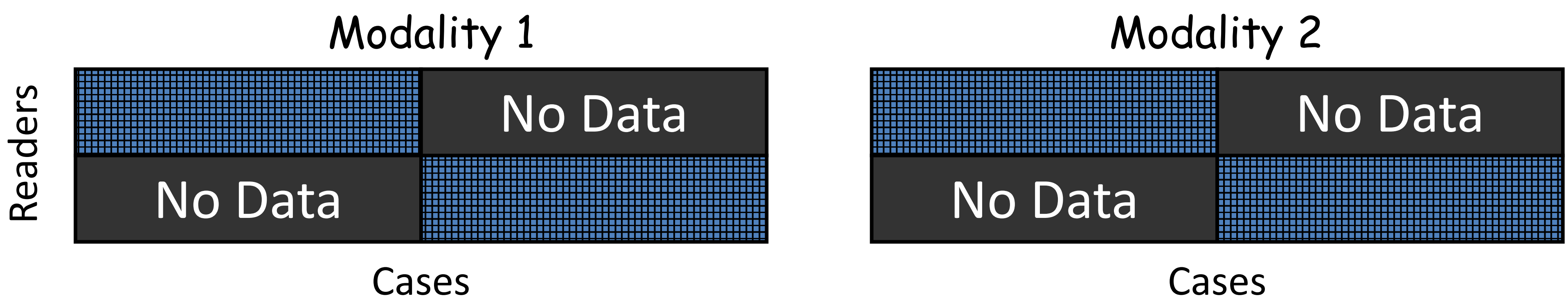
- Fully-crossed study
  - All readers read all cases
  - Readers and cases are paired across modalities

Remove truth labels to unclutter study design concepts.



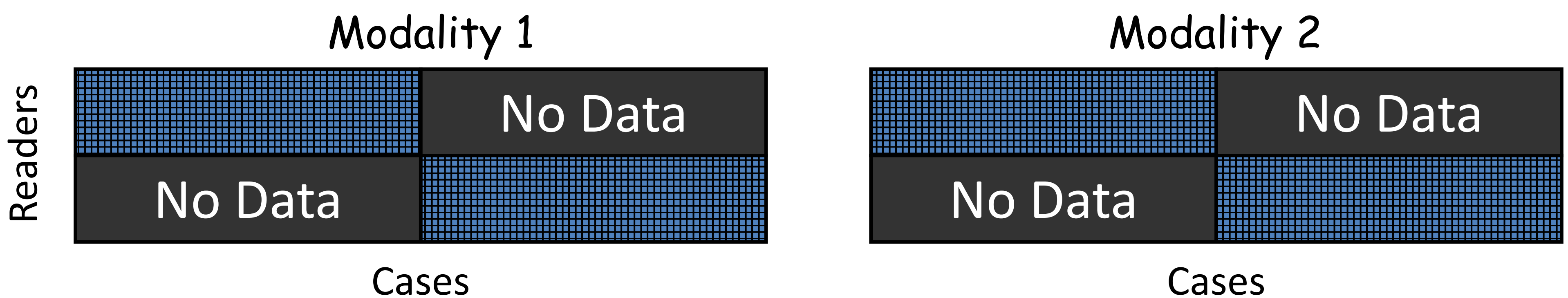
Study Designs
Split-Plot

- 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

ScienceDirect

ELSEVIER

Neural Networks 21 (2008) 387–397

Neural Networks

www.elsevier.com/locate/neunet

2008 Special Issue

Reader studies for validation of CAD systems☆

Brandon D. Gallas*, 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:

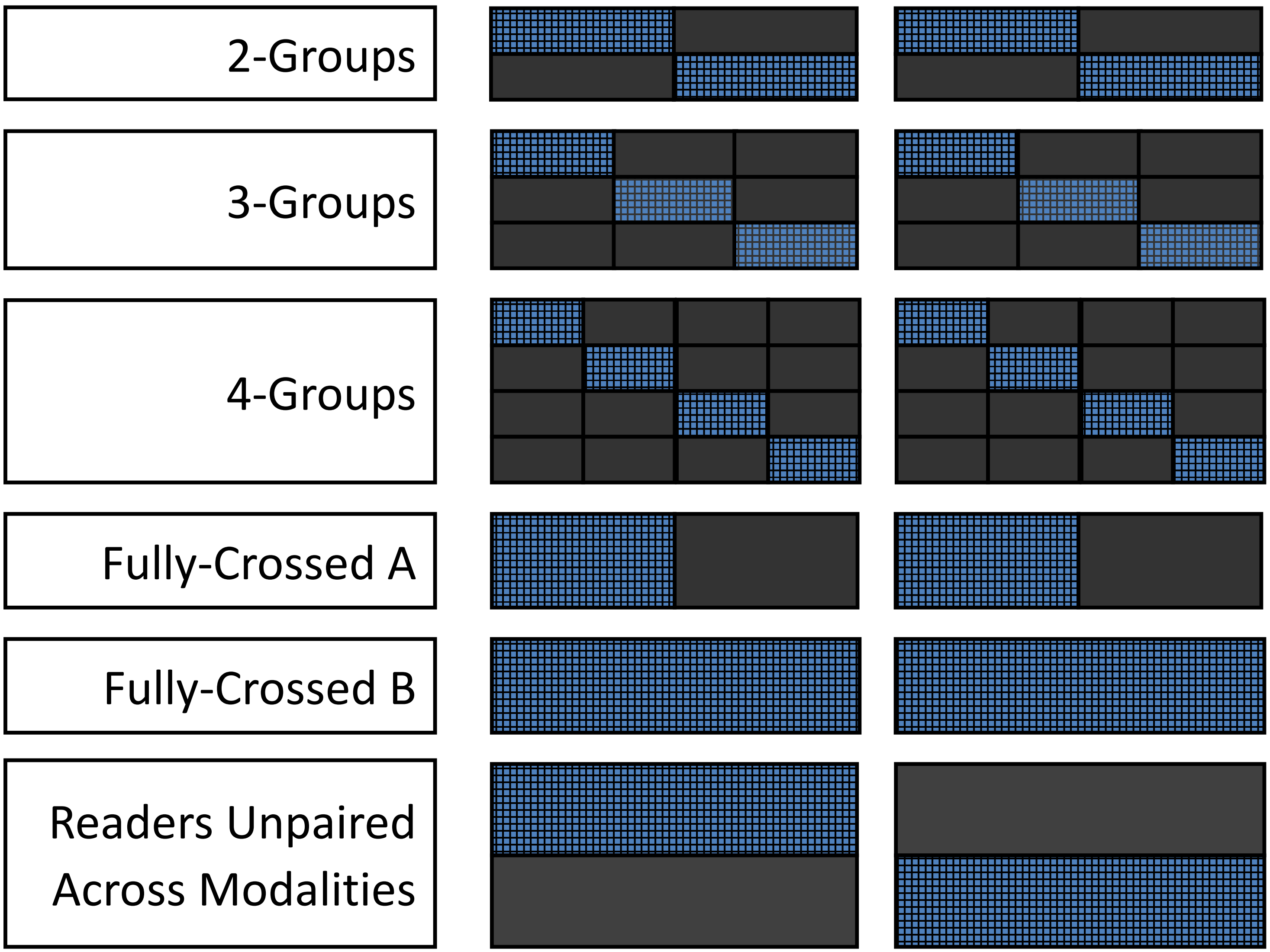
A Comparison of Statistical Methods

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

Var_r	Var_c	Var_rc	Var_tr	Var_tc	Var_trc
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†
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

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

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†
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 (J)	Number of Patients*	Total Number of Image Interpretations	Number of Image Interpretations per Reader	Statistical Efficiency†
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†
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†
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 (J)	Number of Patients*	Total Number of Image Interpretations	Number of Image Interpretations per Reader	Statistical Efficiency†
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†
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.

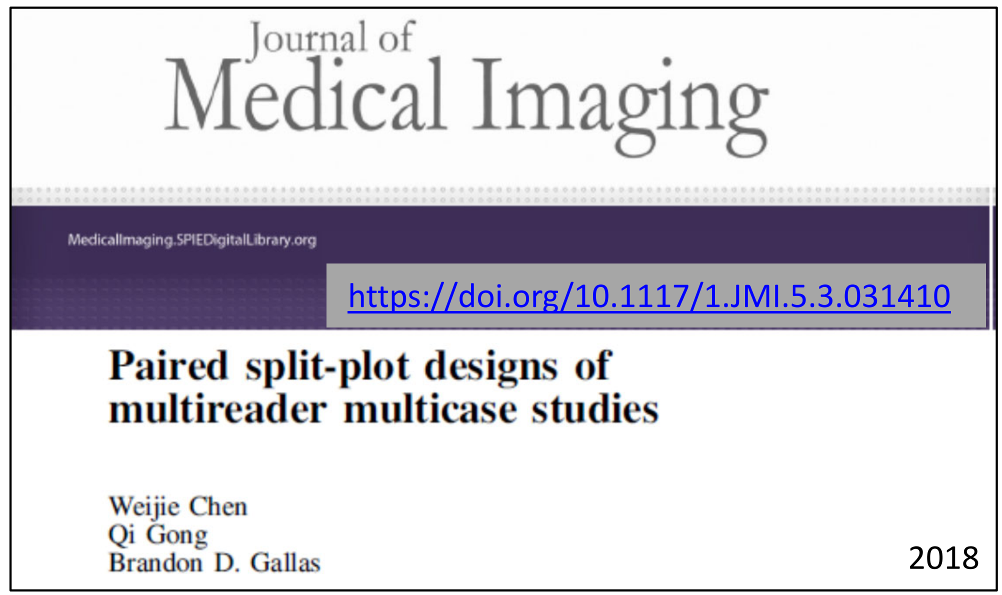
Study Designs:
Efficiency

- Simulation informed theory
  - More groups = less variance

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

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

Re-organize components


Journal of
Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

https://doi.org/10.1117/1.JMI.5.3.031410

Paired split-plot designs of multireader multicase studies

Weijie Chen
Qi Gong
Brandon D. Gallas

2018

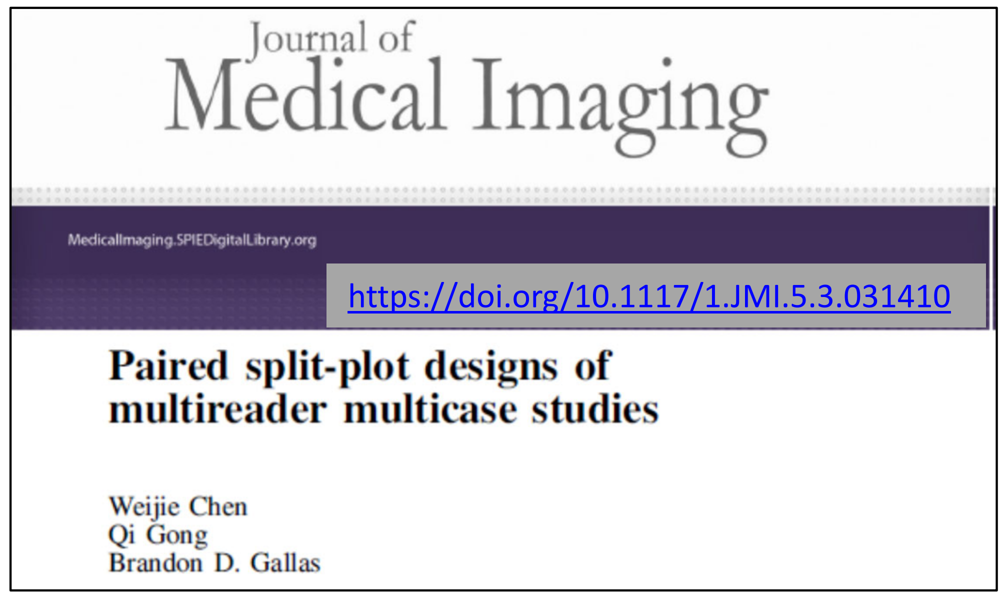
Study Designs:
Efficiency

- Simulation informed theory
  - More groups = less variance

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

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

More groups
= less variance


Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

https://doi.org/10.1117/1.JMI.5.3.031410

Paired split-plot designs of multireader multicase studies

Weijie Chen
Qi Gong
Brandon D. Gallas

Study Designs:
Efficiency

- Simulation informed theory
  - Observations per reader is fixed
  - More groups requires more cases

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

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

More groups
= less variance

Journal of
Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

https://doi.org/10.1117/1.JMI.5.3.031410

Paired split-plot designs of multireader multicase studies

Weijie Chen
Qi Gong
Brandon D. Gallas

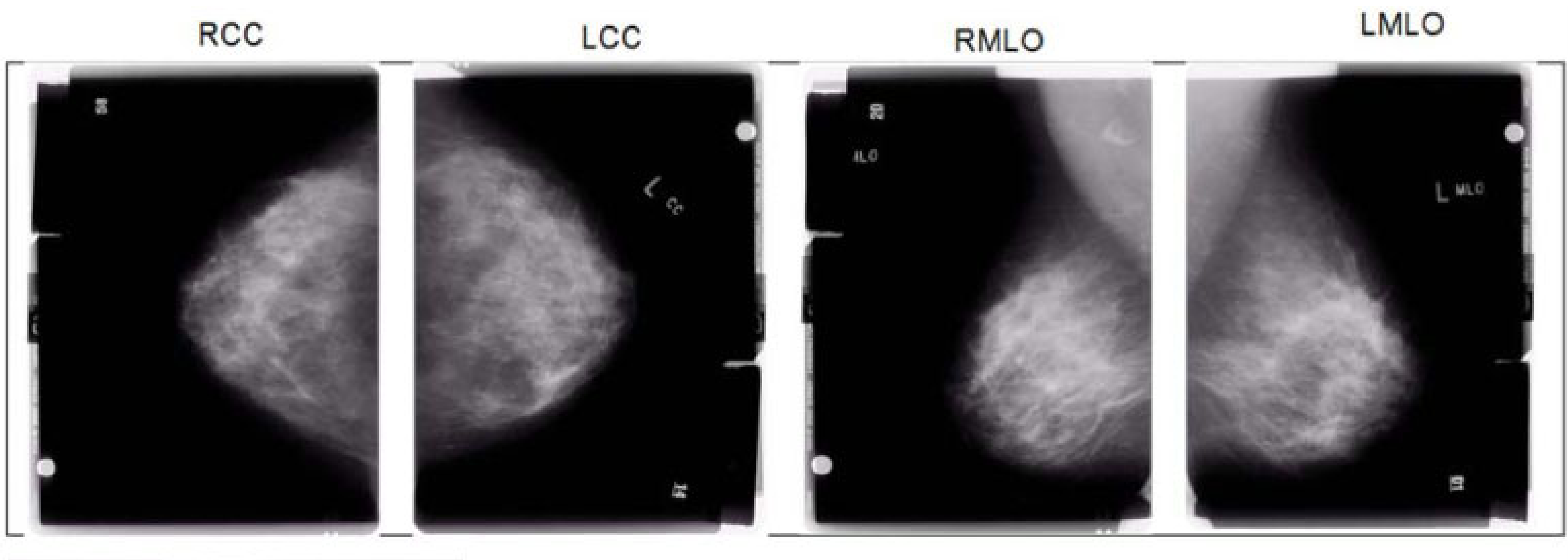
December 21, 2007

https://dilbert.com/

By Scott Adams

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

- ~~11% “Low” prevalence~~

4. 29% “Moderate” prevalence
5. 50% “High” prevalence

VIPER Study Design

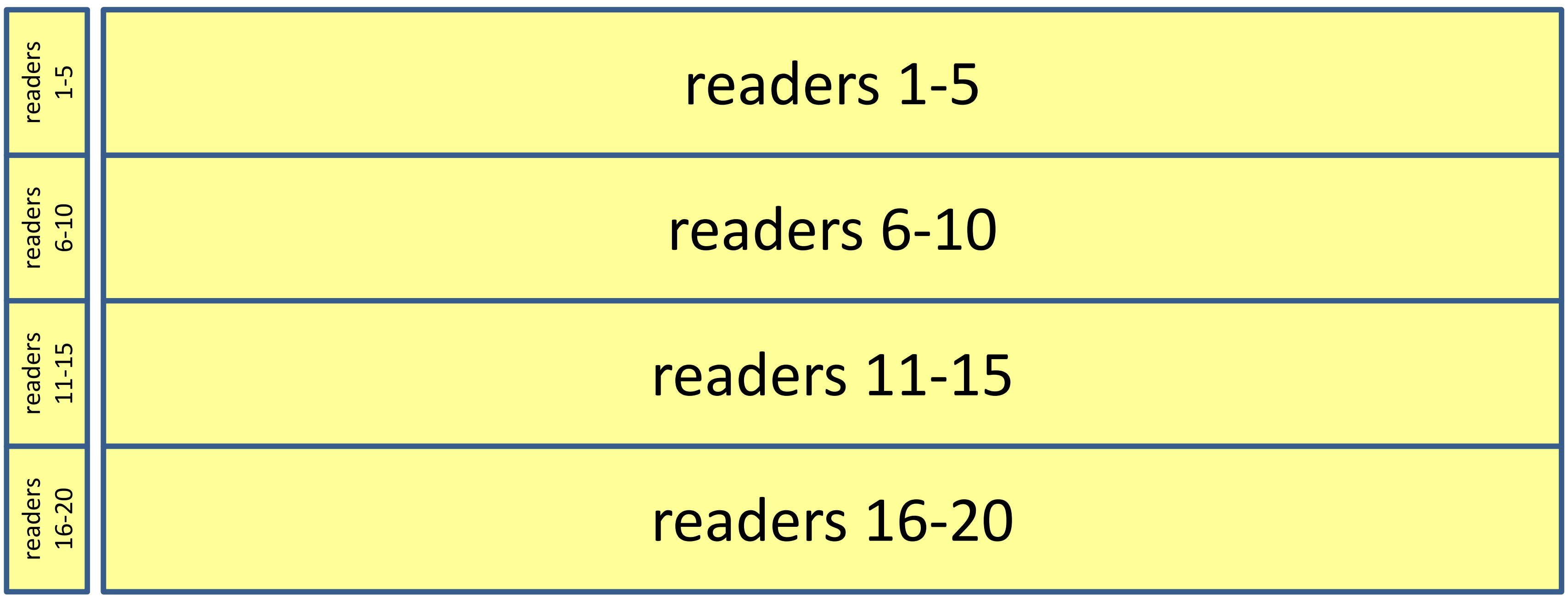
- 20 readers (rows)
- 716 cases (columns)

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

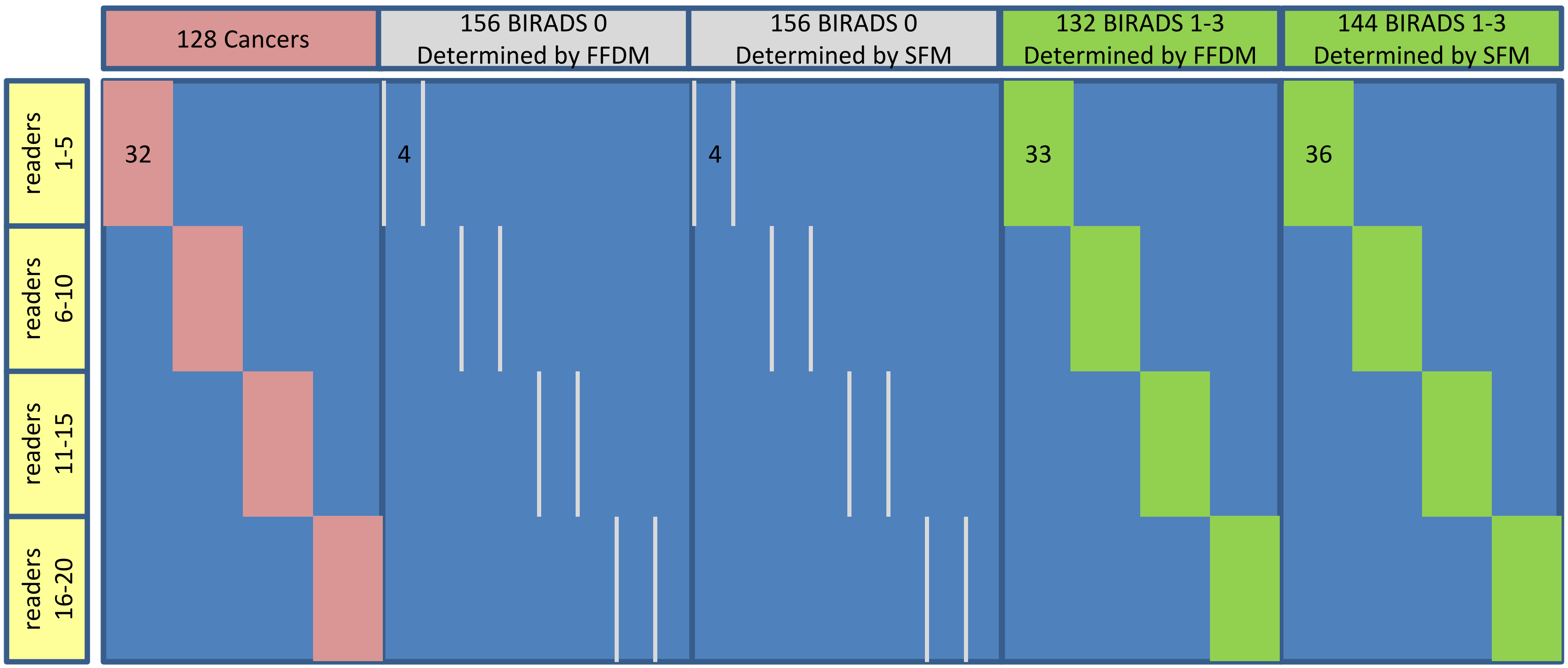
- 20 readers
- 4 split-plot groups



VIPER Study

Screening Study: 29% “Moderate” Prevalence

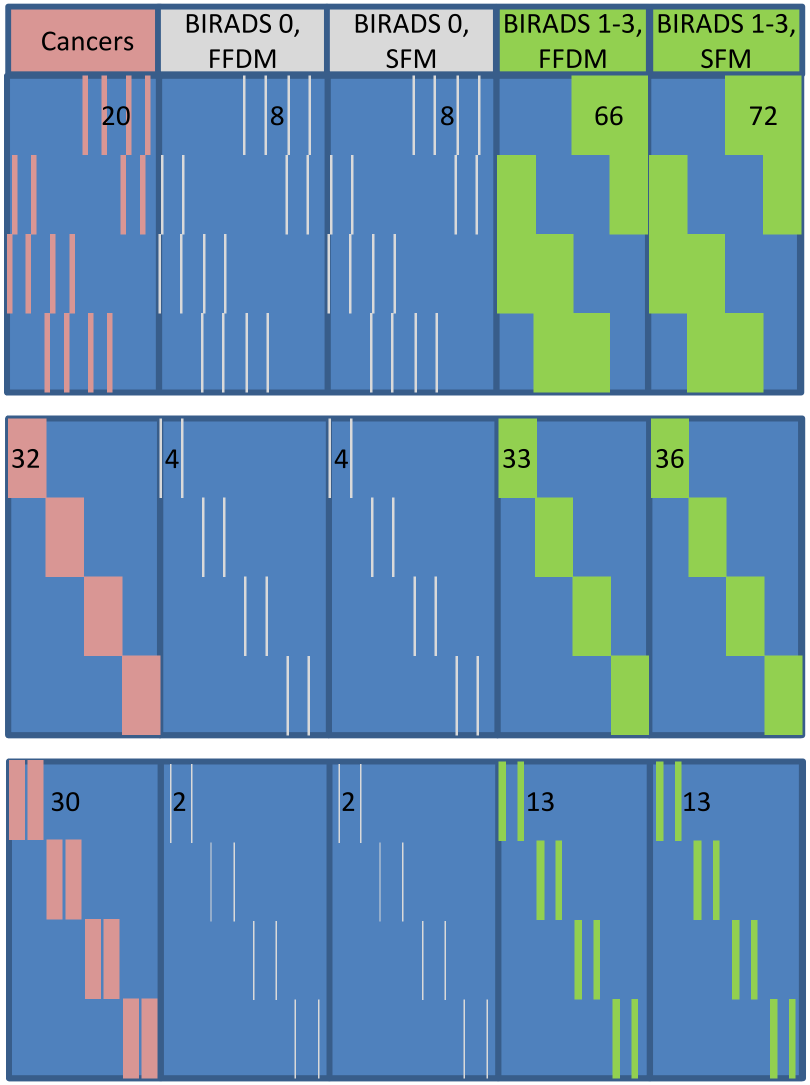
- Per modality we have
  - 109 cases per reader
- Total per modality we have
  - 436 cases total, 2180 obs.



VIPER Study
5 sub-studies

20,382 observations!
20 readers per sub-study

Screening Studies

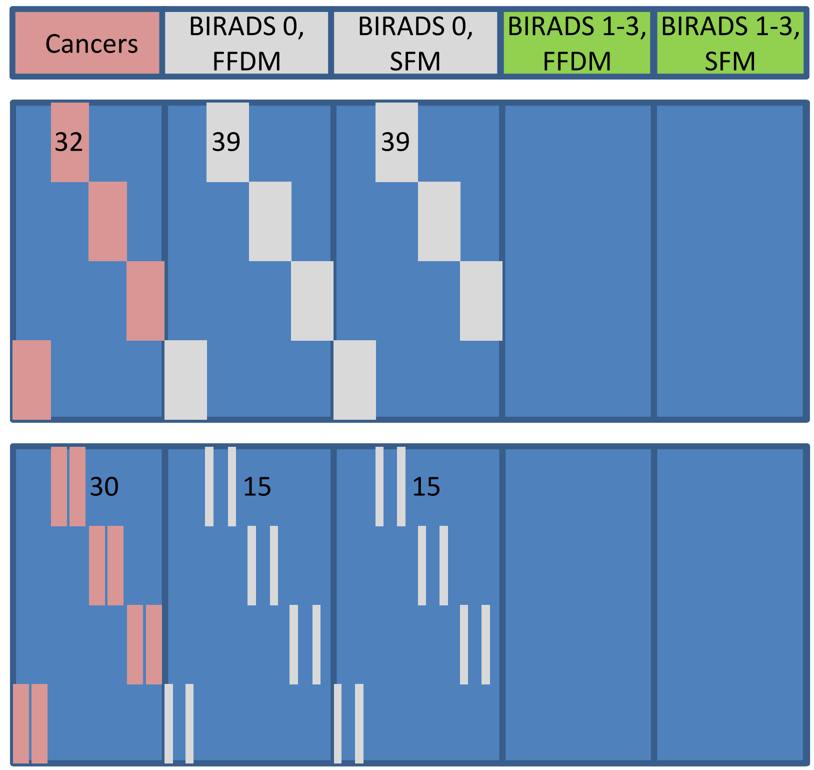


Low Prevalence P = 0.11

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)

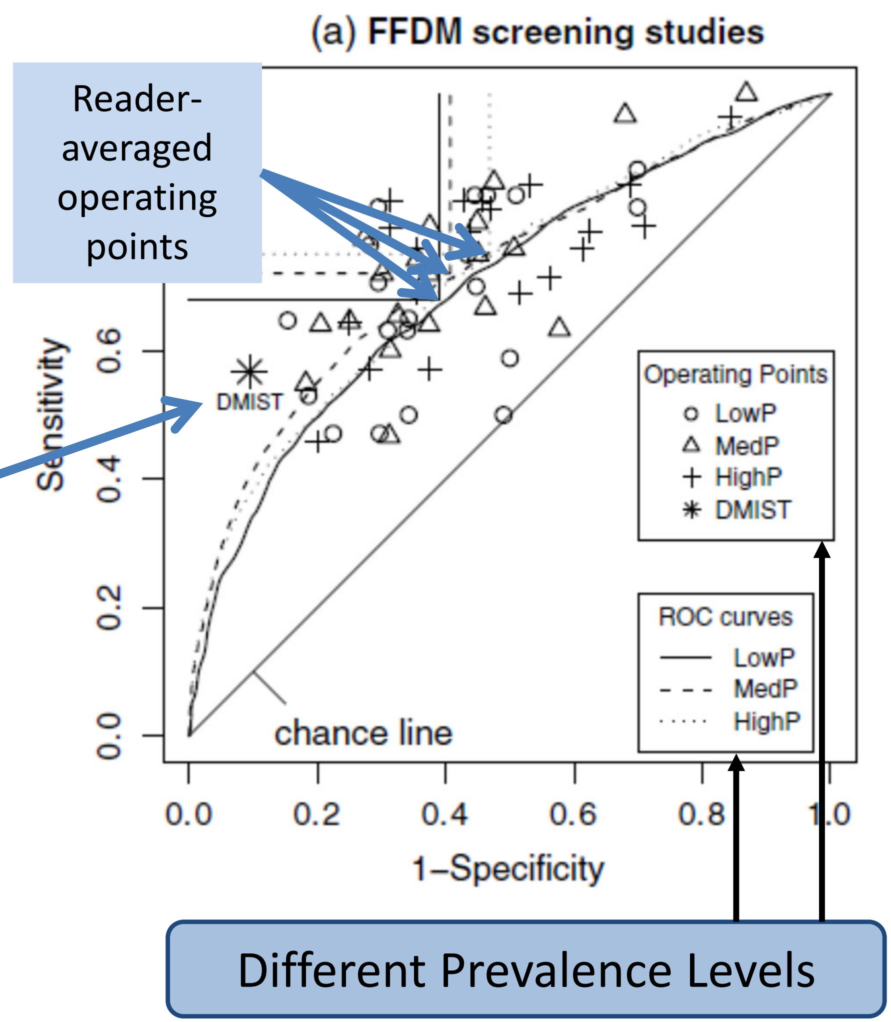
SPIE 2015
https://github.com/DIDSR/iMRMC/blob/gh-pages/000_resources/2015SPIE-MIworkshopBDG-4.pdf
Instructions and eCRF
https://github.com/DIDSR/iMRMC/tree/gh-pages/000_resources/VIPER

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

VIPER Study
Results

Screening Studies

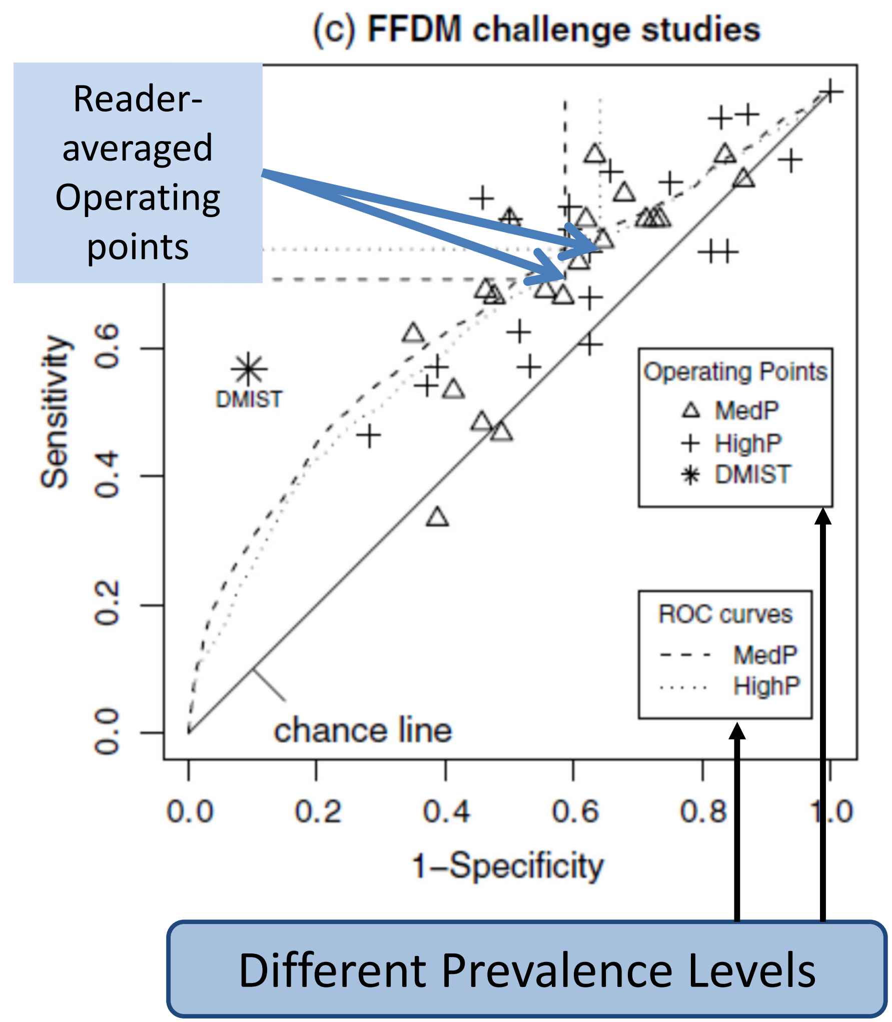
- Nearly identical ROC curves
- Wide range of operating points
  - Appear overlapping with respect to prevalence
- DMIST operating point furthest to left
  - Lowest prevalence
  - Highest specificity (behavior)
- Reader-averaged operating points move up and to the right with increased prevalence.



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
Operating Points Depend on Prevalence

- Trend is consistent with the expected behavior of a decision-maker that is maximizing a risk-benefit relationship between the true and false positives, and the true and false negatives.

- C. E. Metz, “Basic principles of ROC analysis,” *Semin Nucl Med*, vol. 8, no. 4, pp. 283–298, 1978.

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 (model-based)

- **Less than half the observations**
- **Better precision**

	Viper Split-Plot, 4 groups Low Prevalence (13%)	Fully-crossed Low Prevalence (10%)
20 readers	SE (# of observations) # observations per reader # cases	SE (# of observations) # observations per reader # cases
Standard Error: AUC	0.023 (3480 obs.)	0.041 (8700 obs.)
Standard Error: Sensitivity (more cancers in split-plot)	0.038 (400 obs.) 20 cancers per reader 80 total	0.056 (2540 obs.) 30 cancers per reader 30 total
Standard Error: Specificity (constraint: same # of cases across studies)	0.040 (3080 obs.) 154 non-cancers per reader 308 total	0.039 (6160 obs.) 308 non-cancers per reader 308 total

VIPER Paper

Journal of Medical Imaging

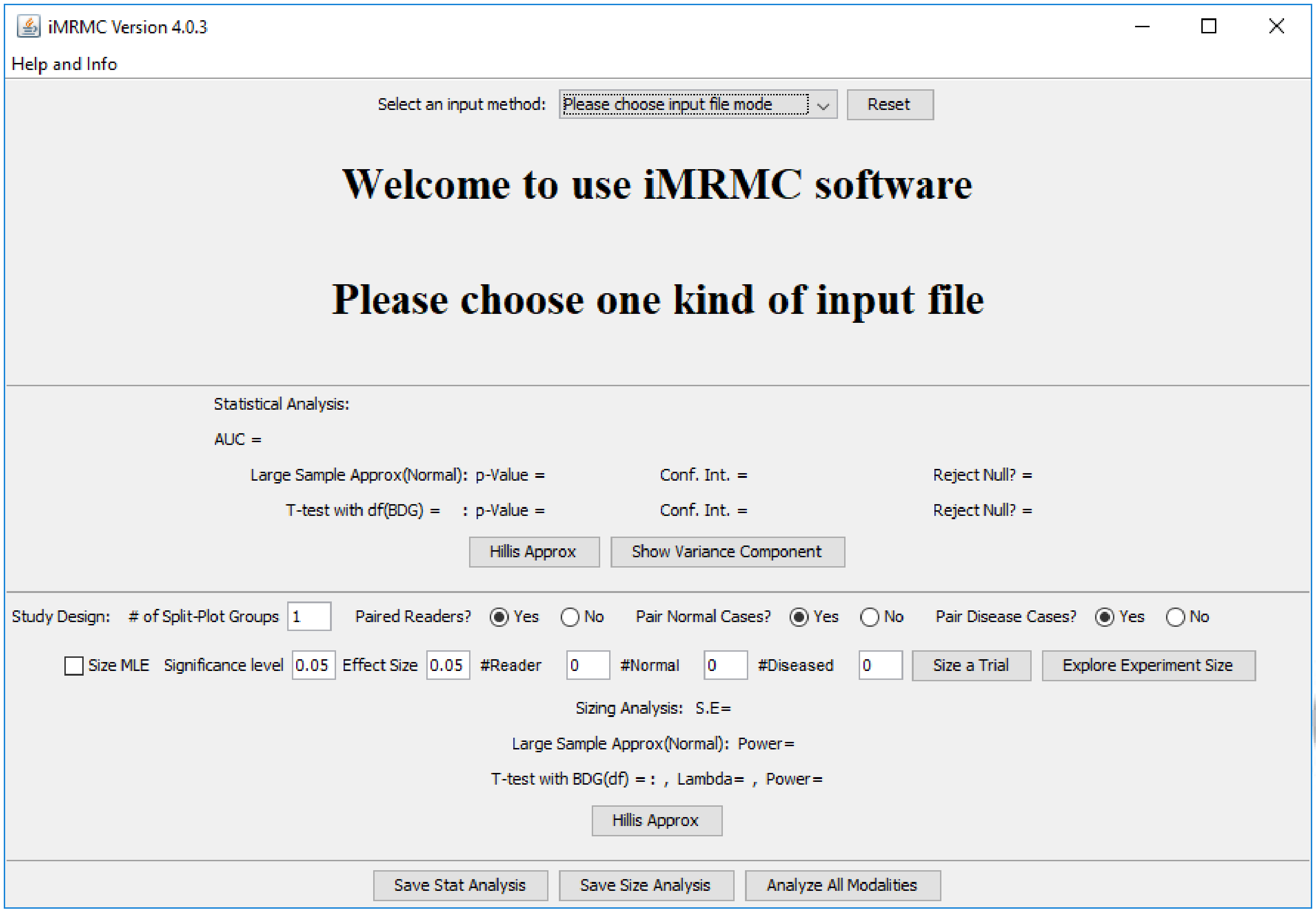
MedicalImaging.SPIEDigitalLibrary.org

doi: 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


iMRMC Version 4.0.3
Help and Info
Select an input method: Please choose input file mode
Reset
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-test with df(BDG) = : p-Value = Conf. Int. = Reject Null? =
Hillis Approx
Show Variance Component
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
Size a Trial
Explore Experiment Size
Sizing Analysis: S.E=
Large Sample Approx(Normal): Power=
T-test with BDG(df) = : , Lambda= , Power=
Hillis Approx
Save Stat Analysis
Save Size Analysis
Analyze All Modalities

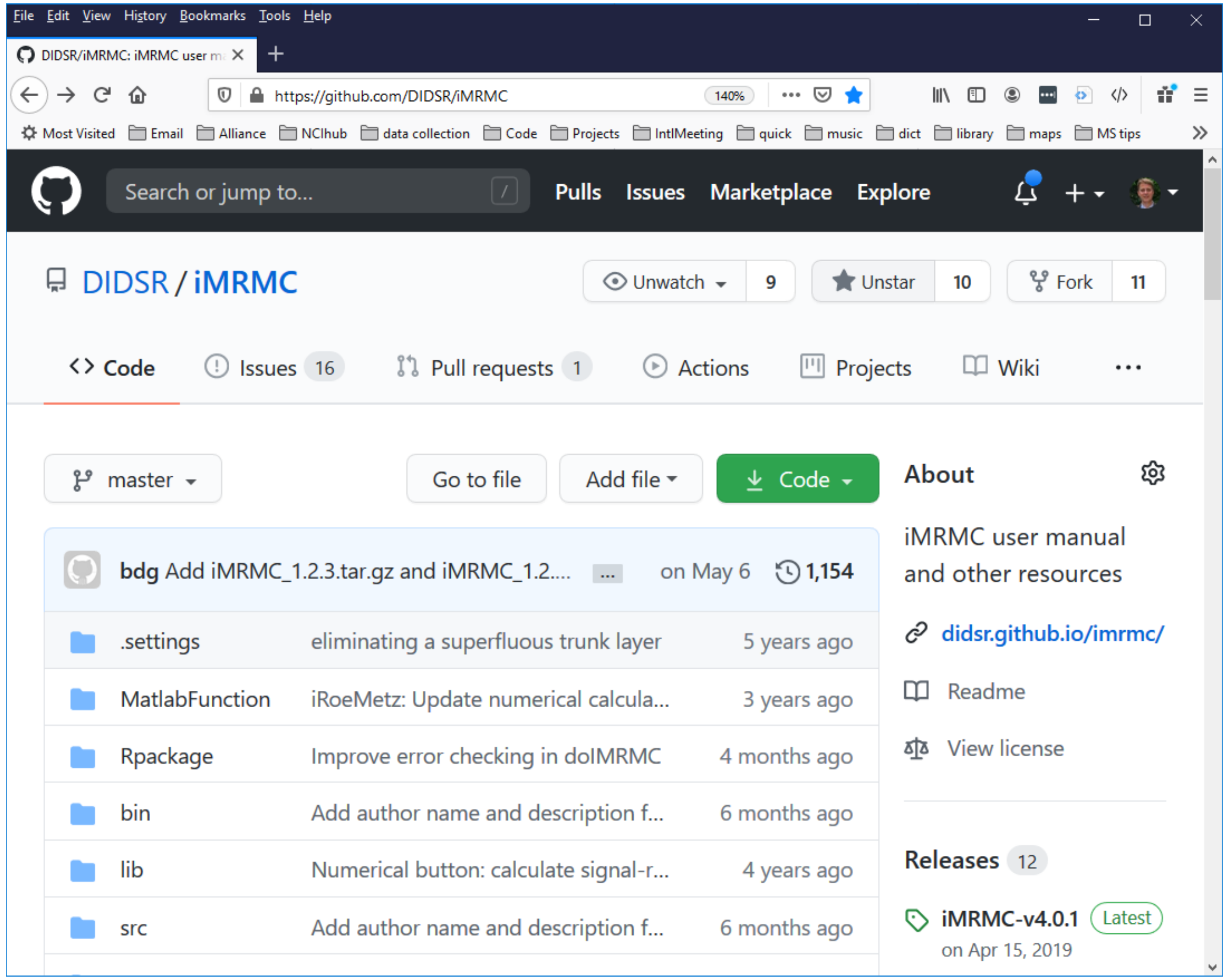
MRMC Tools
iMRMC Software, GitHub Repository

- 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

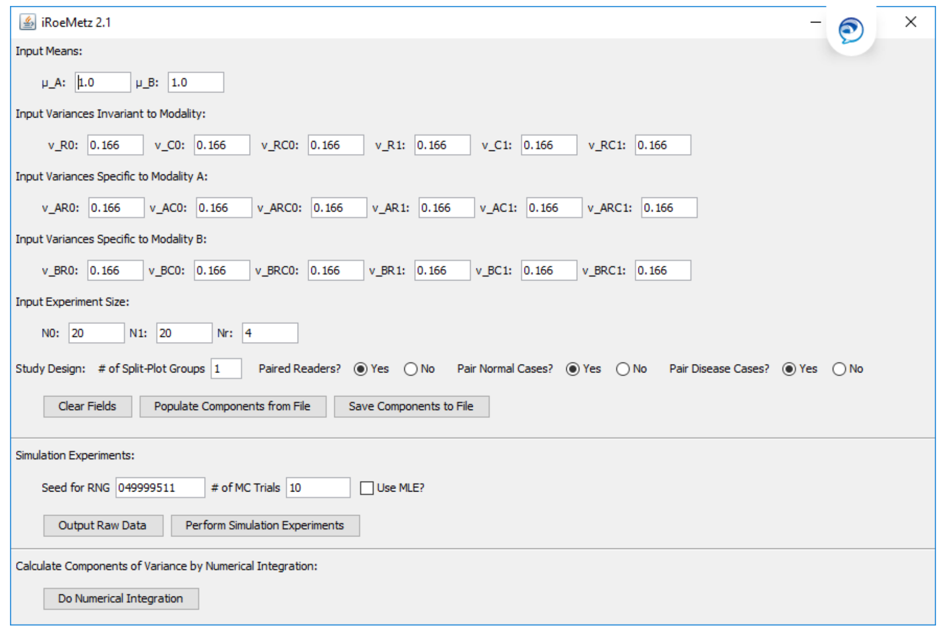
- Download java app
  - https://github.com/DIDSR/iMRMC
  - Backup on desktop

- Check out Wiki
  - Explore VIPER data package
  - Backup in Zotero

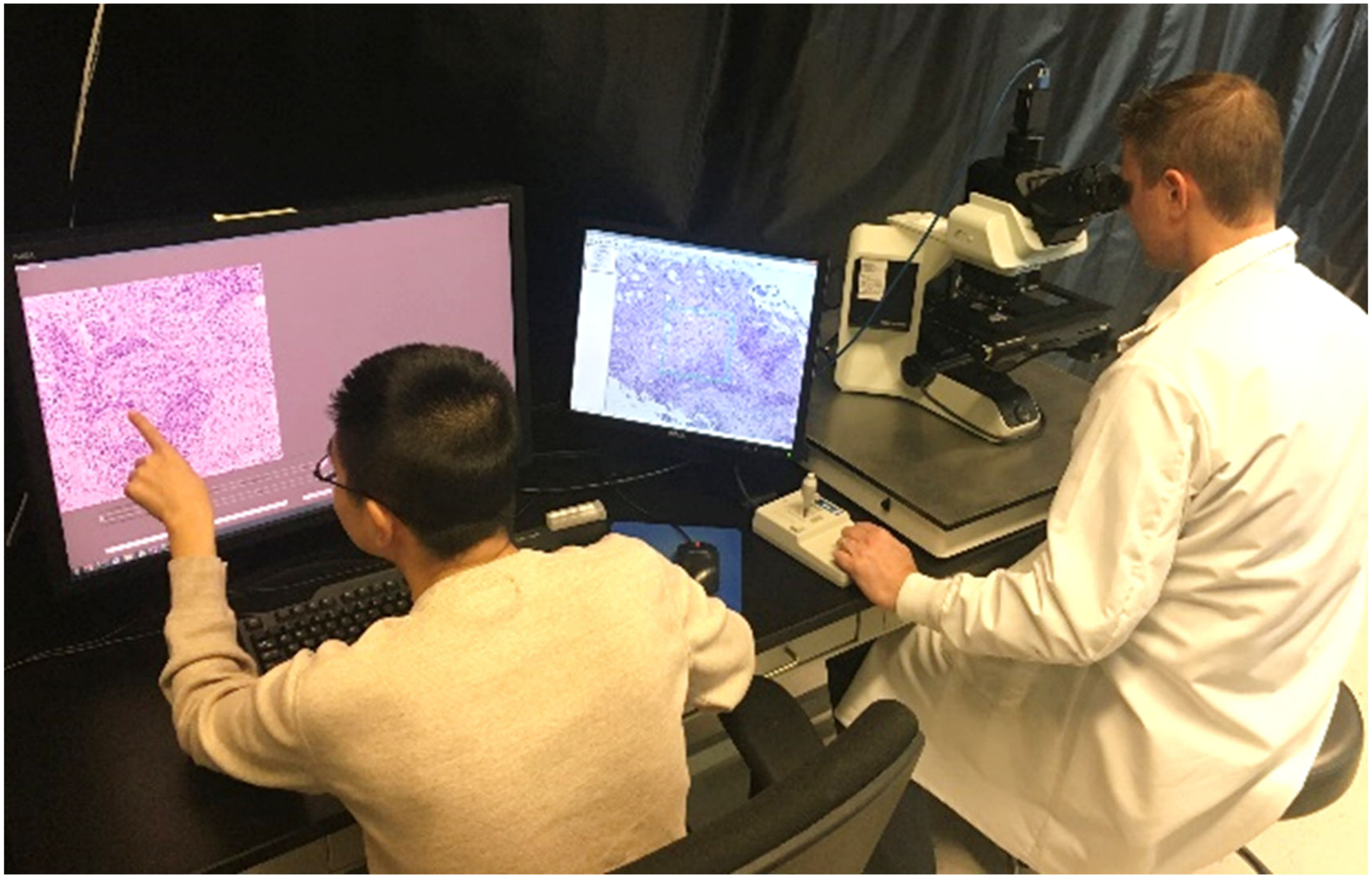
MRMC Tools
iRoeMetz simulation Demo

- Download java app
  - https://github.com/DIDSR/iMRMC
  - Backup on desktop

- Simulation configurations
  - https://github.com/DIDSR/iMRMC/tree/master/Rpackage/iMRMC/inst/data-raw
  - Backup on desktop



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

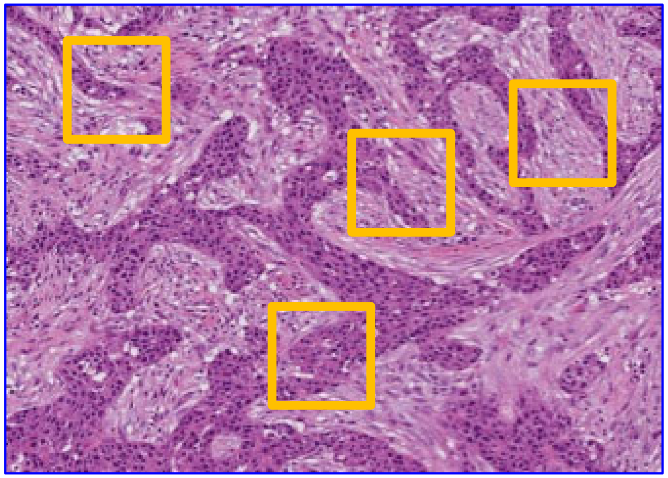
- MRMC variance of AUC framework allows study sizing
  - Variance components
  - Coefficients that correspond to experiment size
- Framework (and simulation) allow study of tradeoffs
  - Resources (Number of readers, cases, and observations)
  - Statistical efficiency
- Split-plot studies are less burdensome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

Summary

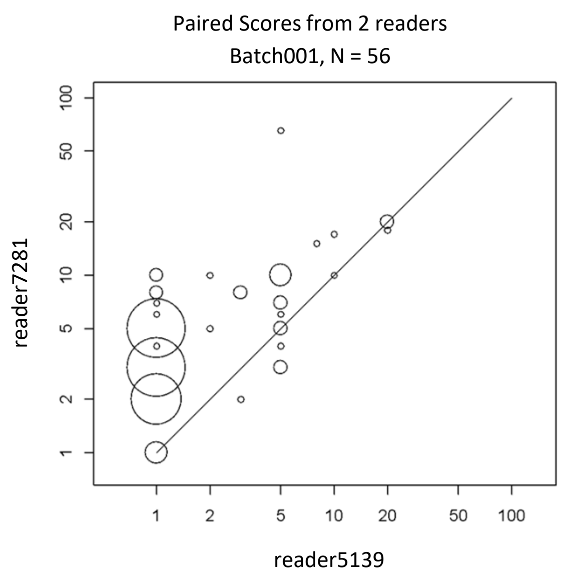
- VIPER study collected 20,382 observations
  - Real radiologists
  - Clinical images
  - Five sub-studies
    - Explore enrichment
    - Explore changes to study population
  - Demonstrated modeling and theory concepts
  - Found AUC to be
    - Robust to enrichment
    - Moderately robust to differences 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
  - Du, Gallas (2022) Stat Methods Med Res https://doi.org/10.1177/09622802221111539



- Quantitative Measurements
  - Guidance Document: "…Quantitative Imaging…"
  - Between-reader agreement
  - Within-reader agreement
  - Algorithm-reader agreement
  - Within and between modalities
  - 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

- H. Du, S. Wen, Y. Guo, F. Jin, and B. D. Gallas, "Single reader between-cases AUC estimator with nested data," *Stat Methods Med Res*, p. 9622802221111540, Jul. 2022, doi: 10.1177/09622802221111539.
- 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
- 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.

Measurement Scales

- Measurement classes
  - Nominal
  - Ordinal
  - Interval
  - Ratio
- Class depends on the measurement process
  - And truth (stimulus intensity)
- Class determines legitimate mathematical / statistical operations

- Seven year debate / committee
- Is it possible to measure sensation?
  - Sensation intensity
    = function of stimulus intensity
  - No agreement
  - What is the meaning of “measurement”?
  - How do you “add” sensory measurements?
  - How define equality of sensory measurements?

SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

Measurement Scales

TABLE 1

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

Quantitative