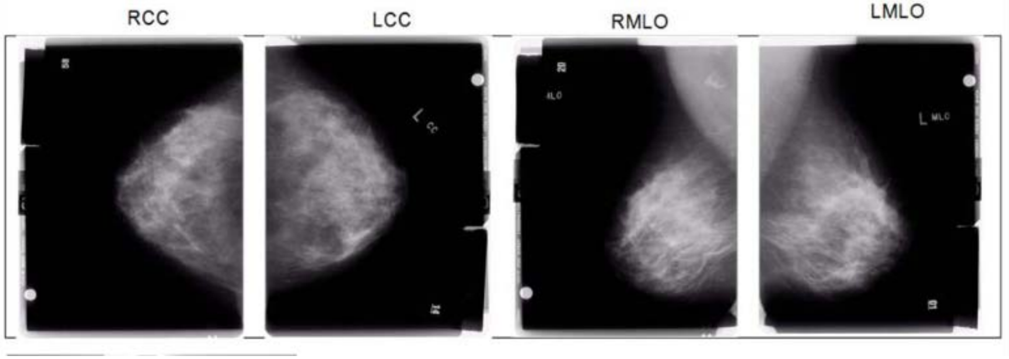


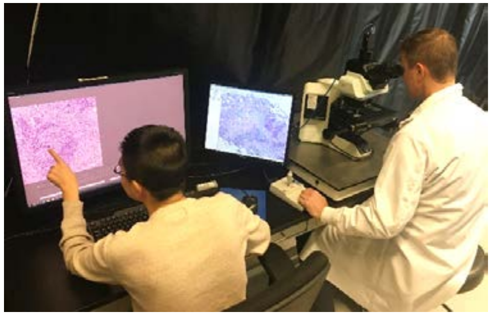
FDA U.S. FOOD & DRUG ADMINISTRATION

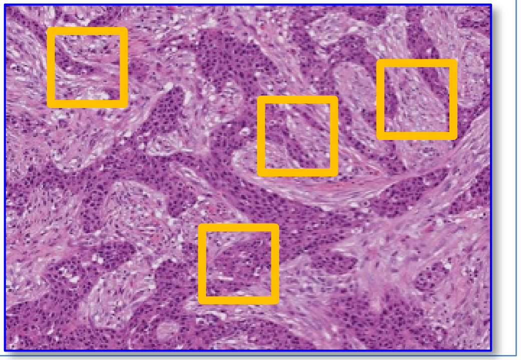
MODELING AND SIMULATING READER STUDIES TO SUPPORT THE EVALUATION OF IMAGE-BASED ALGORITHMS

Brandon D. Gallas

Division of Imaging, Diagnostics, and Software Reliability

Office of Science and Engineering Laboratories
Center for Devices and Radiological Health
U.S. Food and Drug Administration





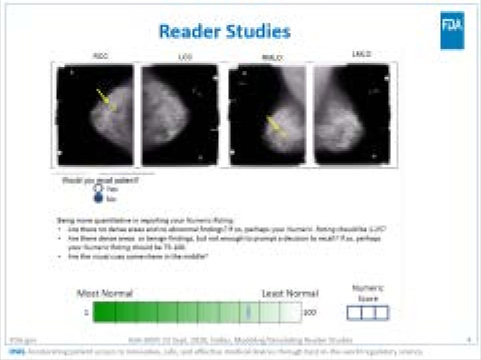
15 July 2020

Abstract

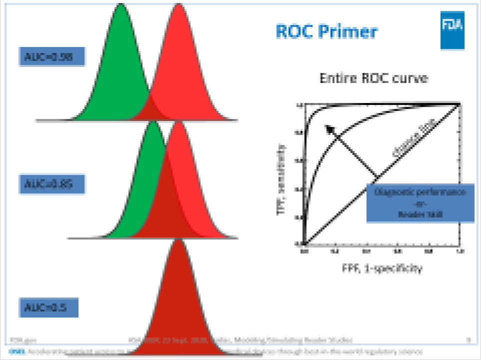
- A growing part of the medical device portfolio of CDRH includes image-based detection (e.g., find the tumor) and classification algorithms (e.g., classify an abnormality as benign or malignant). Whatever the health condition, imaging technology, or algorithm architecture (neural networks, random forests, regressions), submissions of the “software as a medical device” often include a reader study, a study in which clinicians make evaluations with and without the algorithm. Comparing the evaluations against a reference truth, we can compare the performance impact of the algorithm. The statistical analysis of such studies is not trivial since it is well known that there is a range of skill among clinicians and their evaluations are noisy. Furthermore, the studies often have multiple clinicians evaluating the same cases, leading to correlations in the data. FDA guidance recommends an MRMC (multi-reader multi-case) analysis paradigm in which a reader-averaged performance metric is analyzed (variance estimates, confidence intervals, and p-values) to account for the variability (and correlations) from readers and cases. To support such analyses, we have developed, published, and shared on GitHub statistical methods and software, data, and examples. Such development relies on simulations of MRMC studies to validate the statistical methods. In this talk, we will discuss reader studies, performance metrics, and the corresponding MRMC structures of uncertainty. We will present a simulation model that has served us well in validating MRMC analyses of detection and classification metrics. To address studies of algorithms that yield quantitative values and the within- and between-clinician agreement of such values, we have been developing new MRMC methods that analyze differences in quantitative values. To support this work, we are investigating and will present a new simulation model that better represents such data.

- 25 minutes presentation time

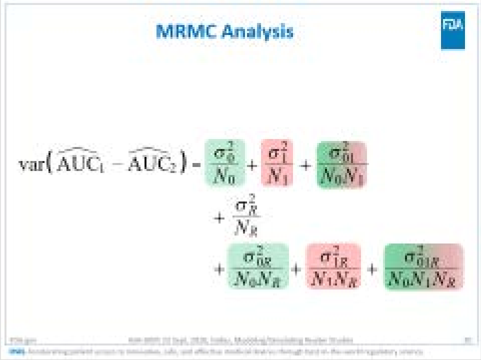
Outline



Reader Studies

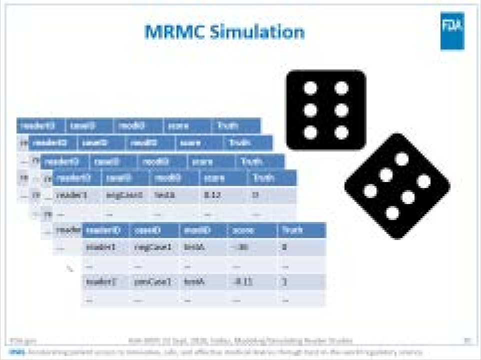


ROC Primer

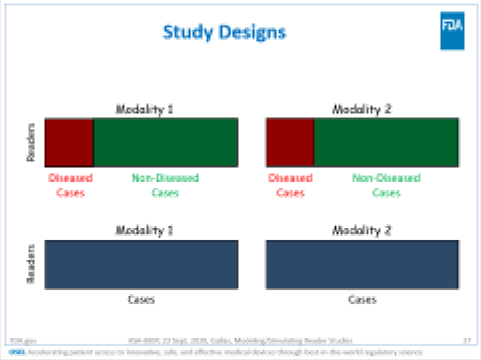


MRMC Analysis

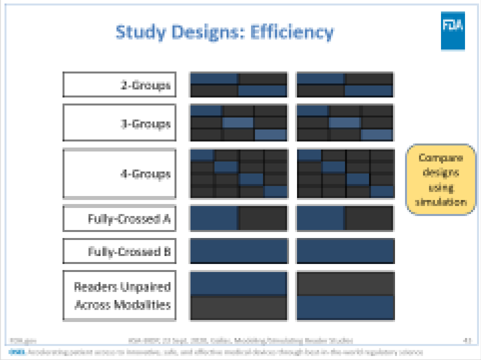
$$\mathrm{var}\left(\widehat{\mathrm{AUC}}_1 - \widehat{\mathrm{AUC}}_2\right) = \frac{\sigma_0^2}{N_0} + \frac{\sigma_1^2}{N_1} + \frac{\sigma_{01}^2}{N_0 N_1} + \frac{\sigma_R^2}{N_R} + \frac{\sigma_{0R}^2}{N_0 N_R} + \frac{\sigma_{1R}^2}{N_1 N_R} + \frac{\sigma_{01R}^2}{N_0 N_1 N_R}$$



MRMC Simulation

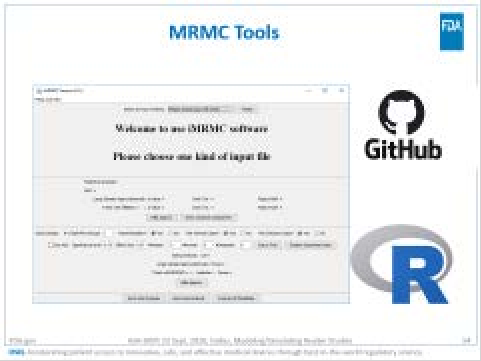


Study Designs



Study Designs: Efficiency

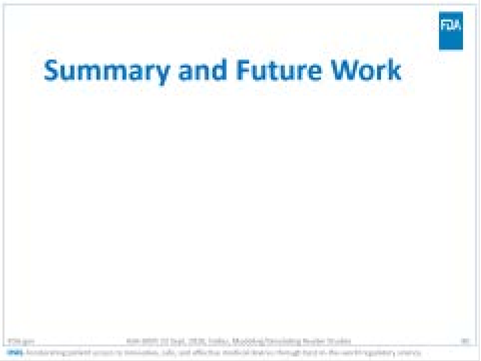
2-Groups
3-Groups
4-Groups
Fully-Crossed A
Fully-Crossed B
Readers Unpaired Across Modalities
Compare designs using simulation



MRMC Tools

Welcome to use iMRMC software

Please choose one kind of input file



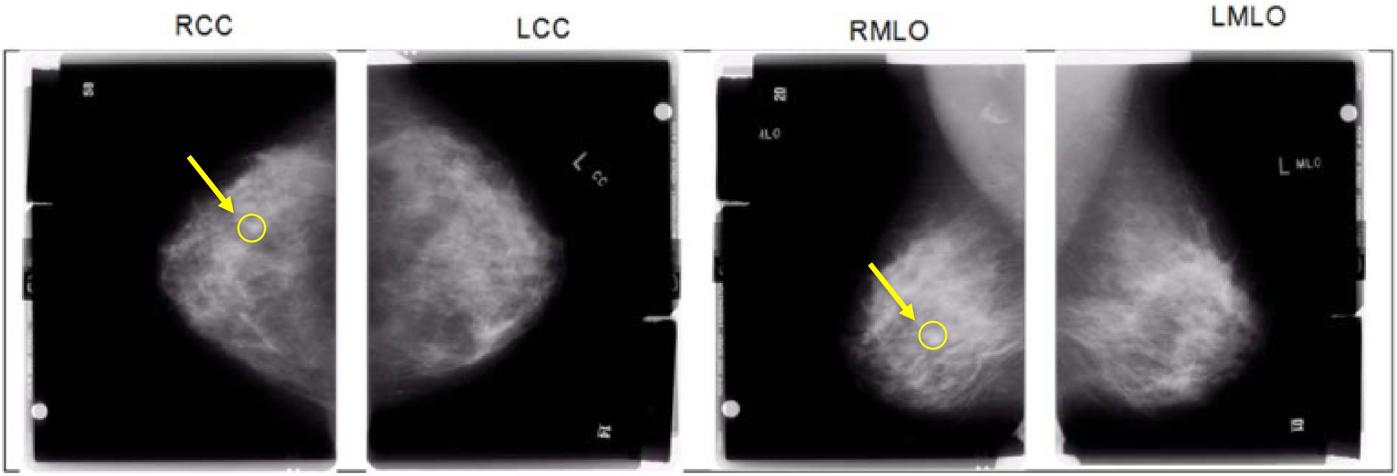
Summary and Future Work

BONUS

High-Throughput Truthing Project

HTT project

Reader Studies



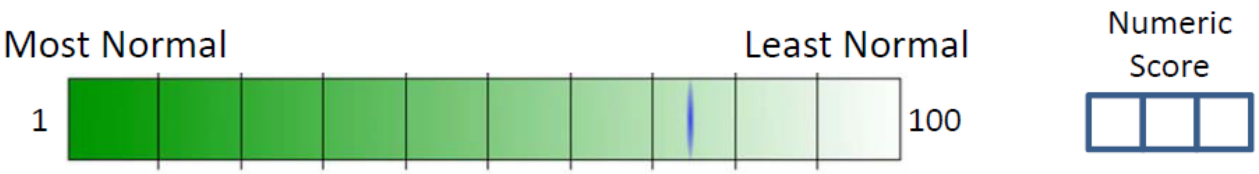
Would you recall patient?

○ 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?



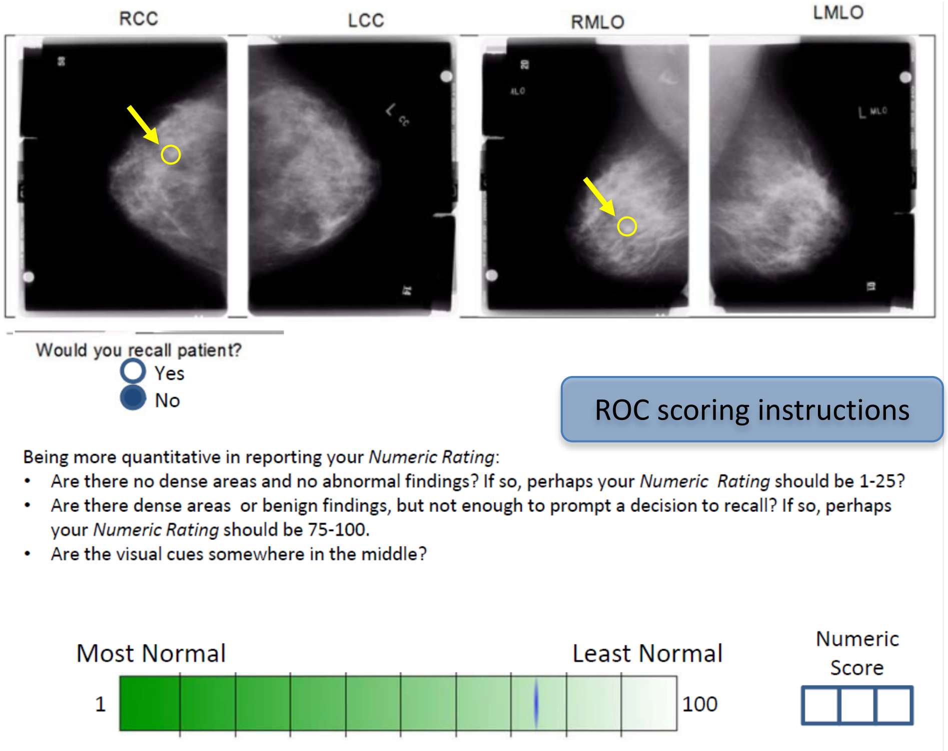
Reader Studies

Compare clinician performance with a new imaging system to a reference imaging system

Reader Studies

Example: Evaluating Computer Aids

- Modality: Images with computer aid vs. images without computer aid
- Task/Performance 1:
  Recall women with cancer
  - Binary patient management decision
  - Sensitivity, specificity
- Task/Performance 2:
  Score cancer confidence
  - More information. Goal is to rank.
  - Area under the ROC curve
- Readers: Radiologists
- Cases: Breast cancer screening population



Would you recall patient?
○ Yes
● No

ROC scoring instructions

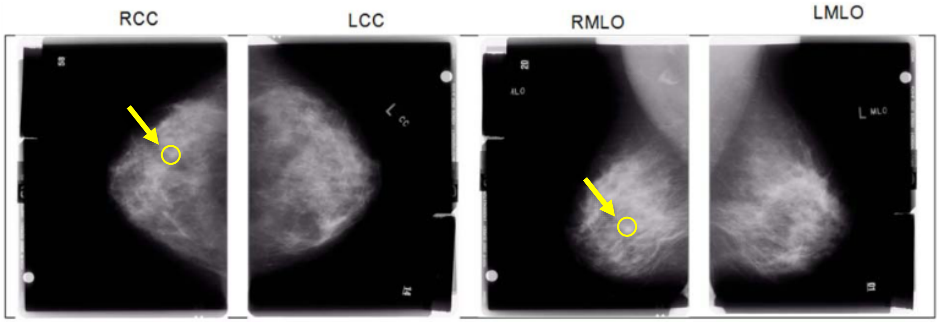
Being more quantitative in reporting your *Numeric Rating*:

- 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?

Reader Studies

Example: Evaluating Computer Aids

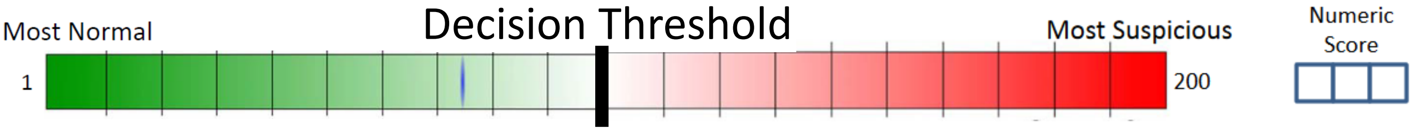
- Modality: Images with computer aid vs. images without computer aid
- Task/Performance 1:
  Recall women with cancer
  - Binary patient management decision
  - Sensitivity, specificity
- Task/Performance 2:
  Score cancer confidence
  - More information. Goal is to rank.
  - Area under the ROC curve



Would you recall patient?
● Yes
○ No

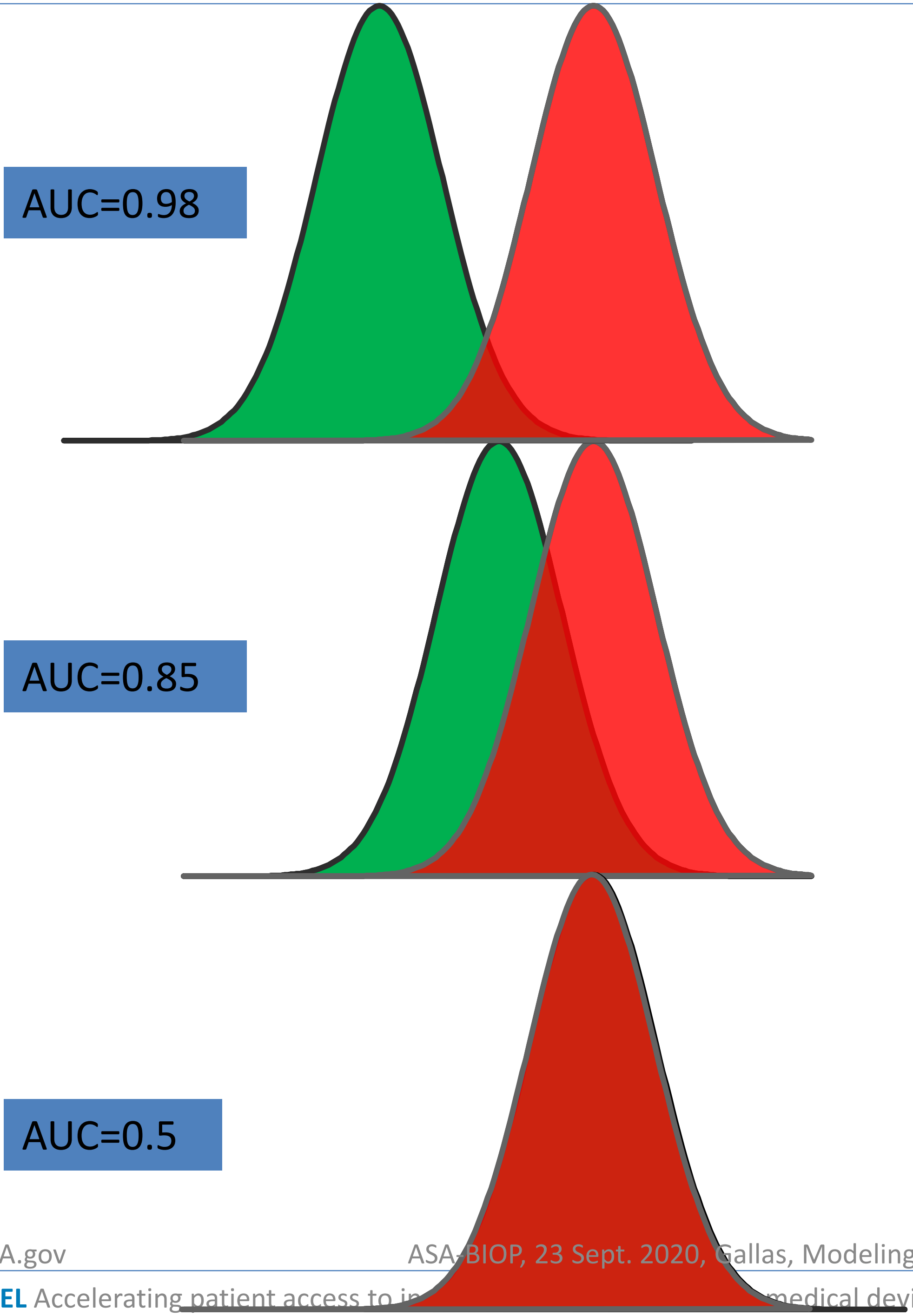
Being more quantitative in reporting your *Numeric Rating*:

- 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?

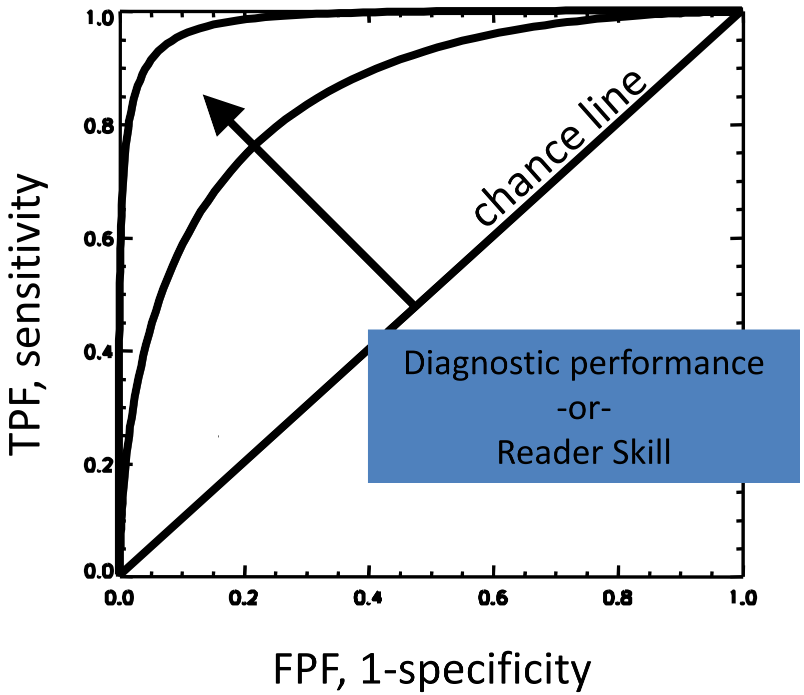


- Readers: Radiologists
- Cases: Breast cancer screening population

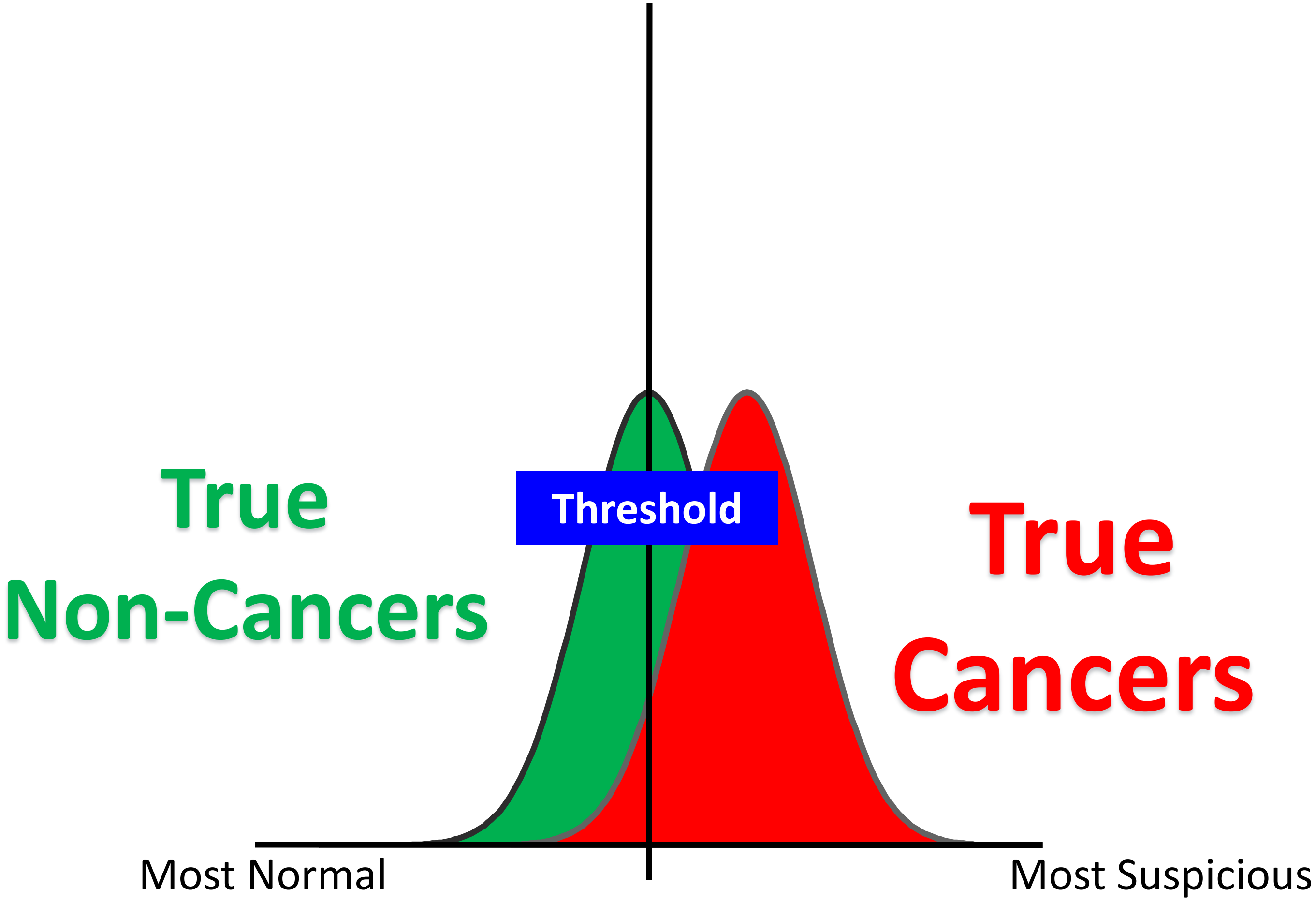
ROC Primer

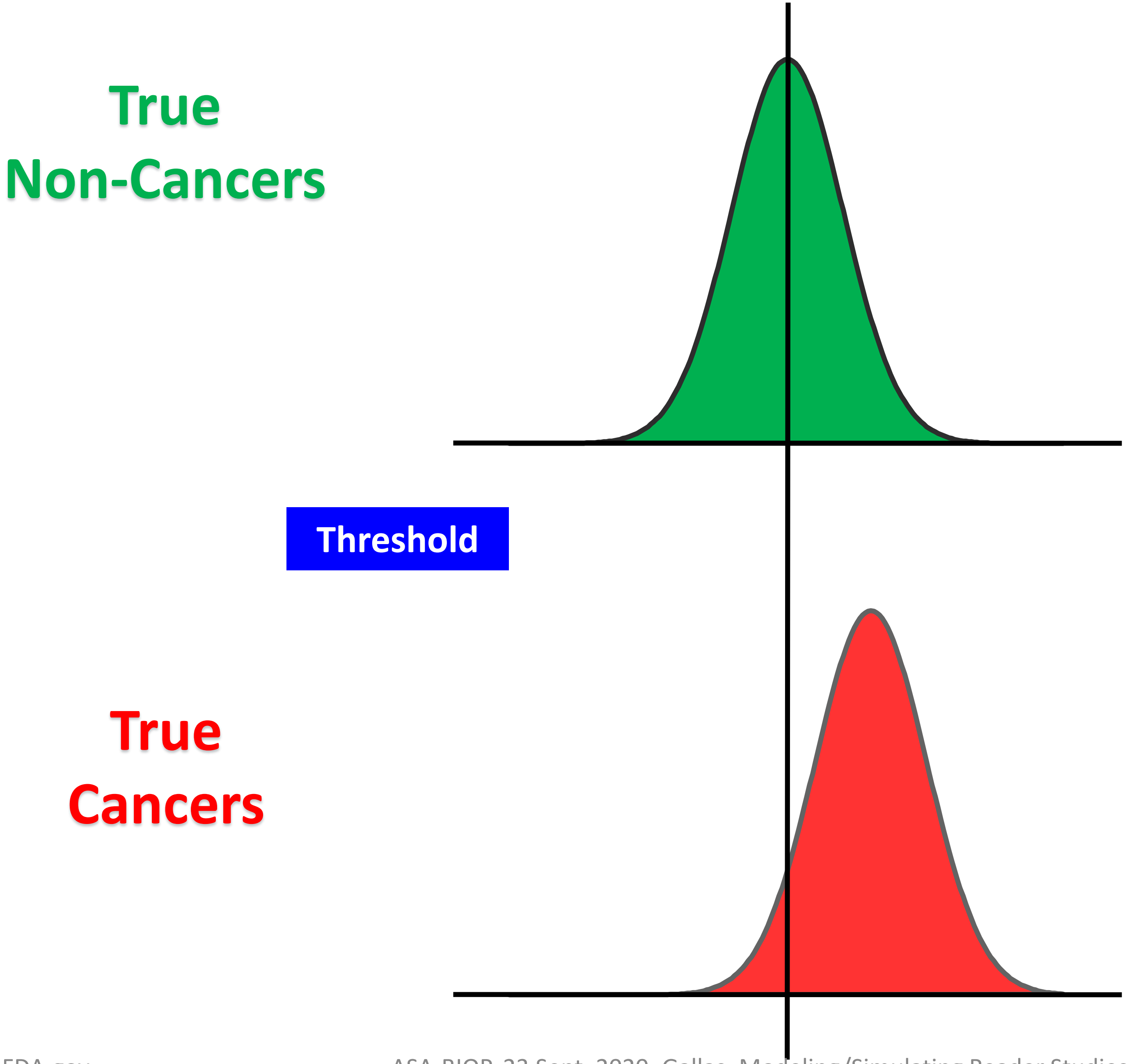


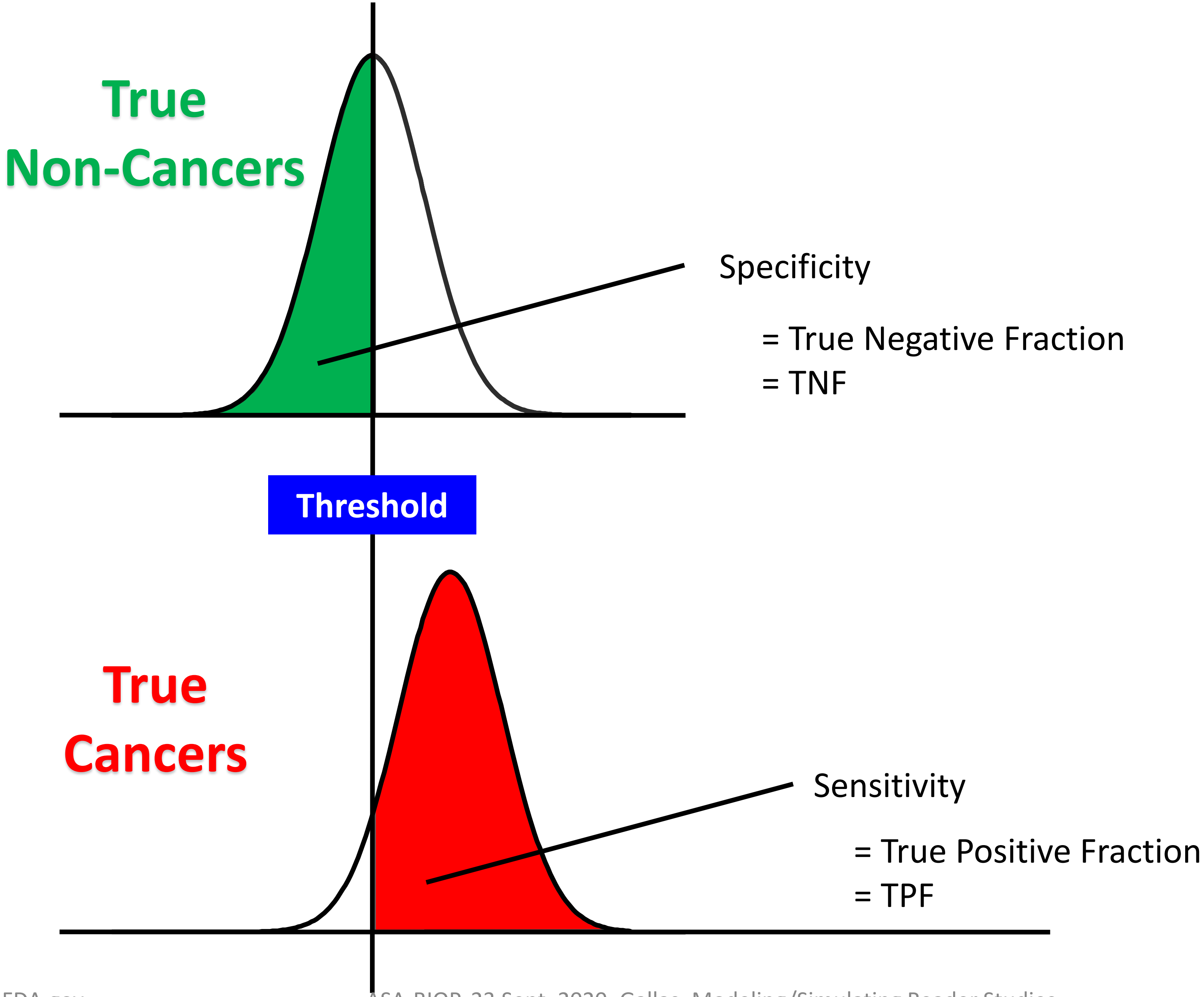
Entire ROC curve



Quick Primer on Sensitivity, Specificity, and Area Under the ROC curve


True
Non-Cancers
Threshold
True
Cancers
Most Normal
Most Suspicious


True
Non-Cancers
Threshold
True
Cancers



True Non-Cancers

Specificity

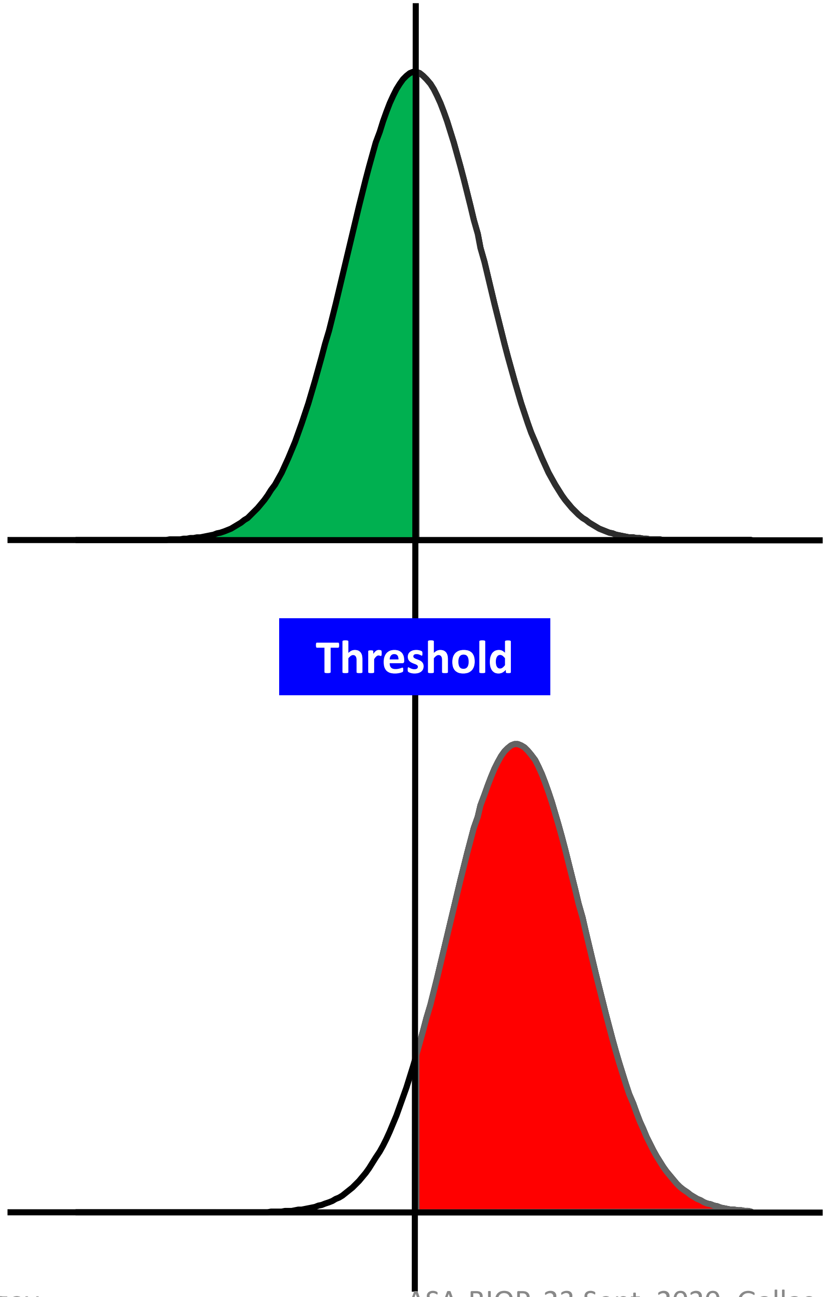
= True Negative Fraction
= TNF

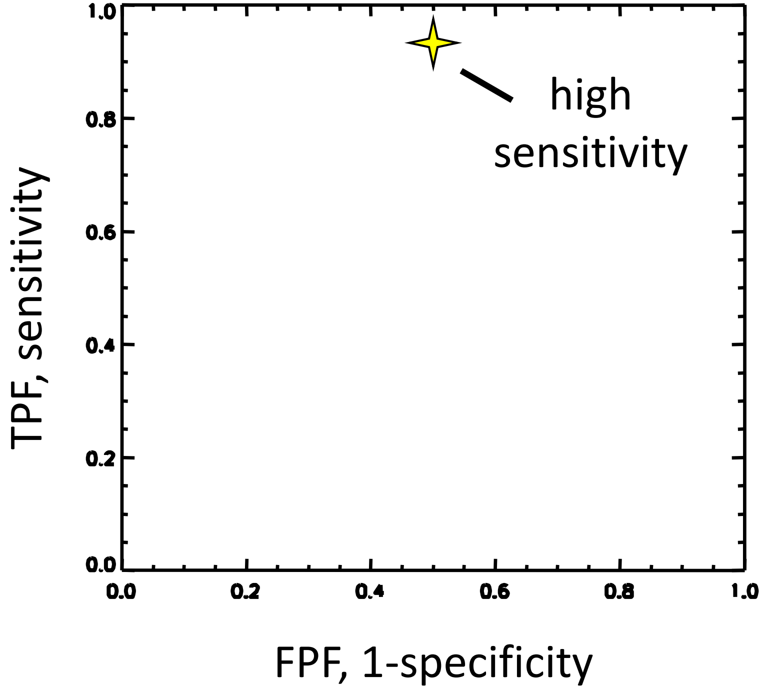
Threshold

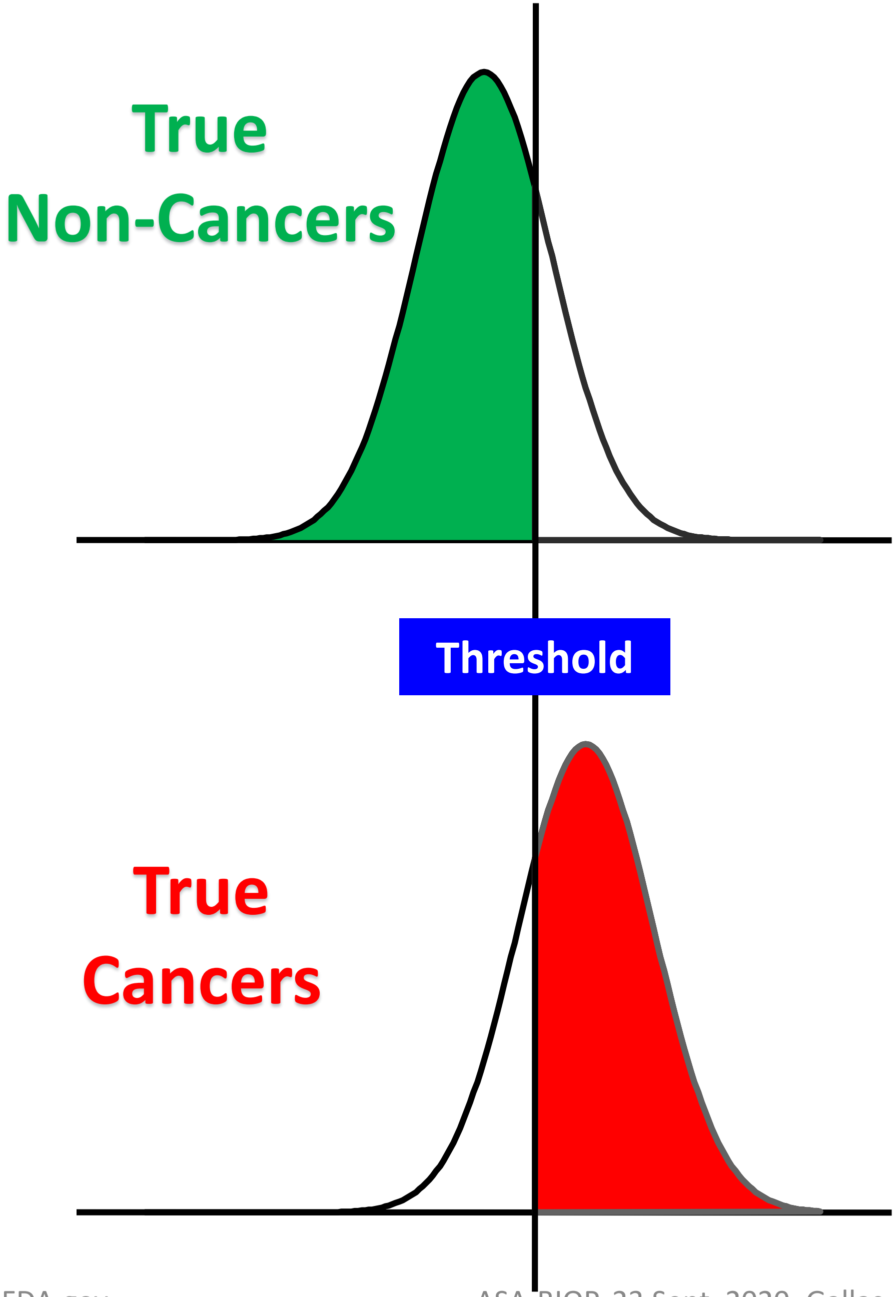
True Cancers

Sensitivity

= True Positive Fraction
= TPF


Threshold

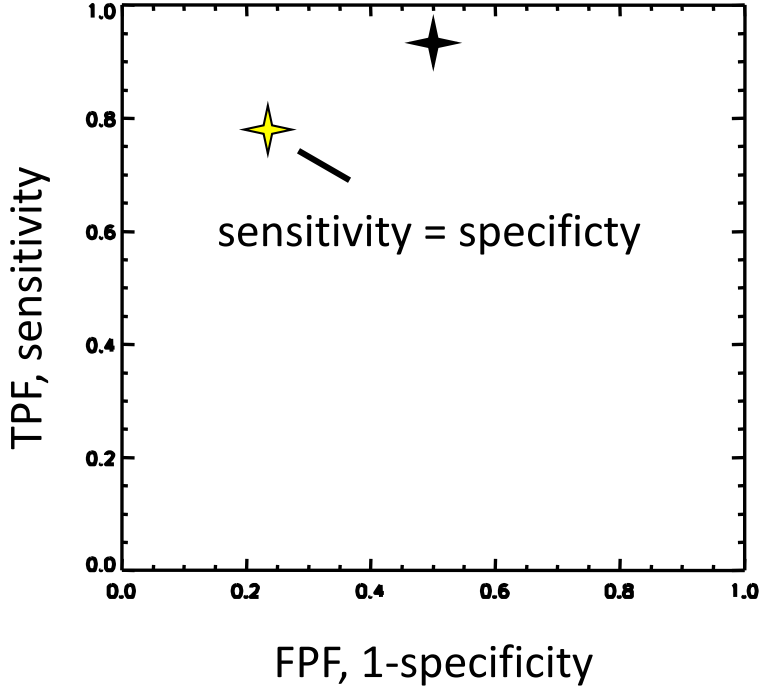

1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
TPF, sensitivity
FPF, 1-specificity
high sensitivity



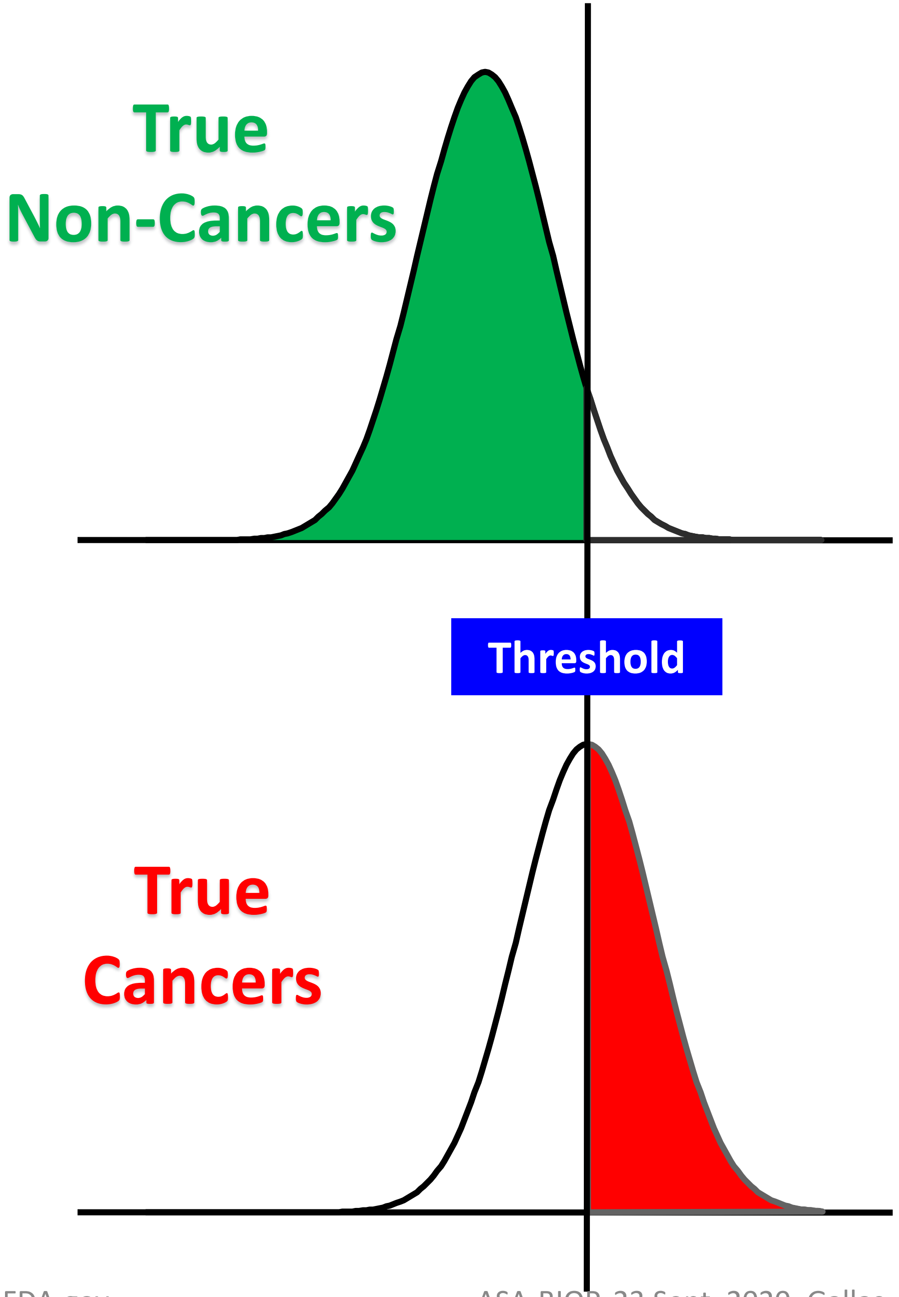
True Non-Cancers

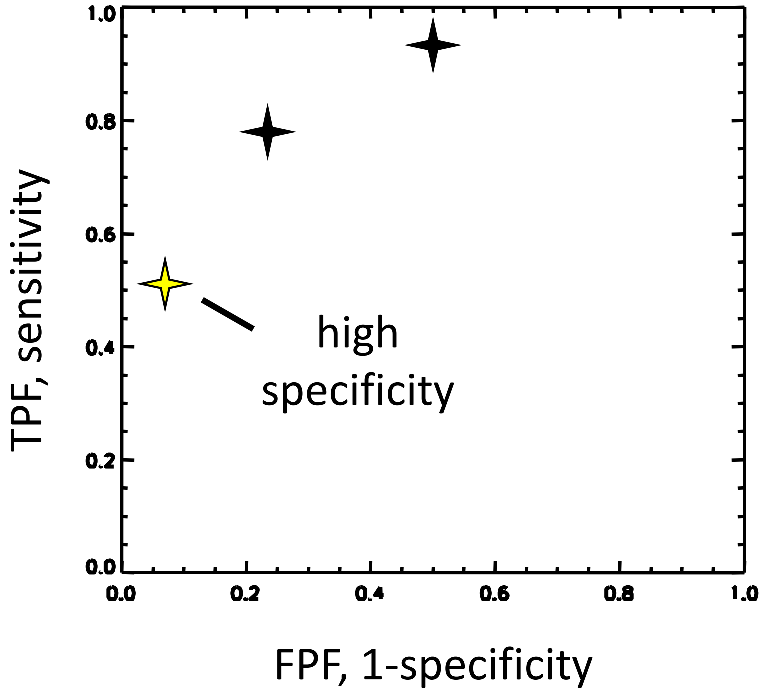
Threshold

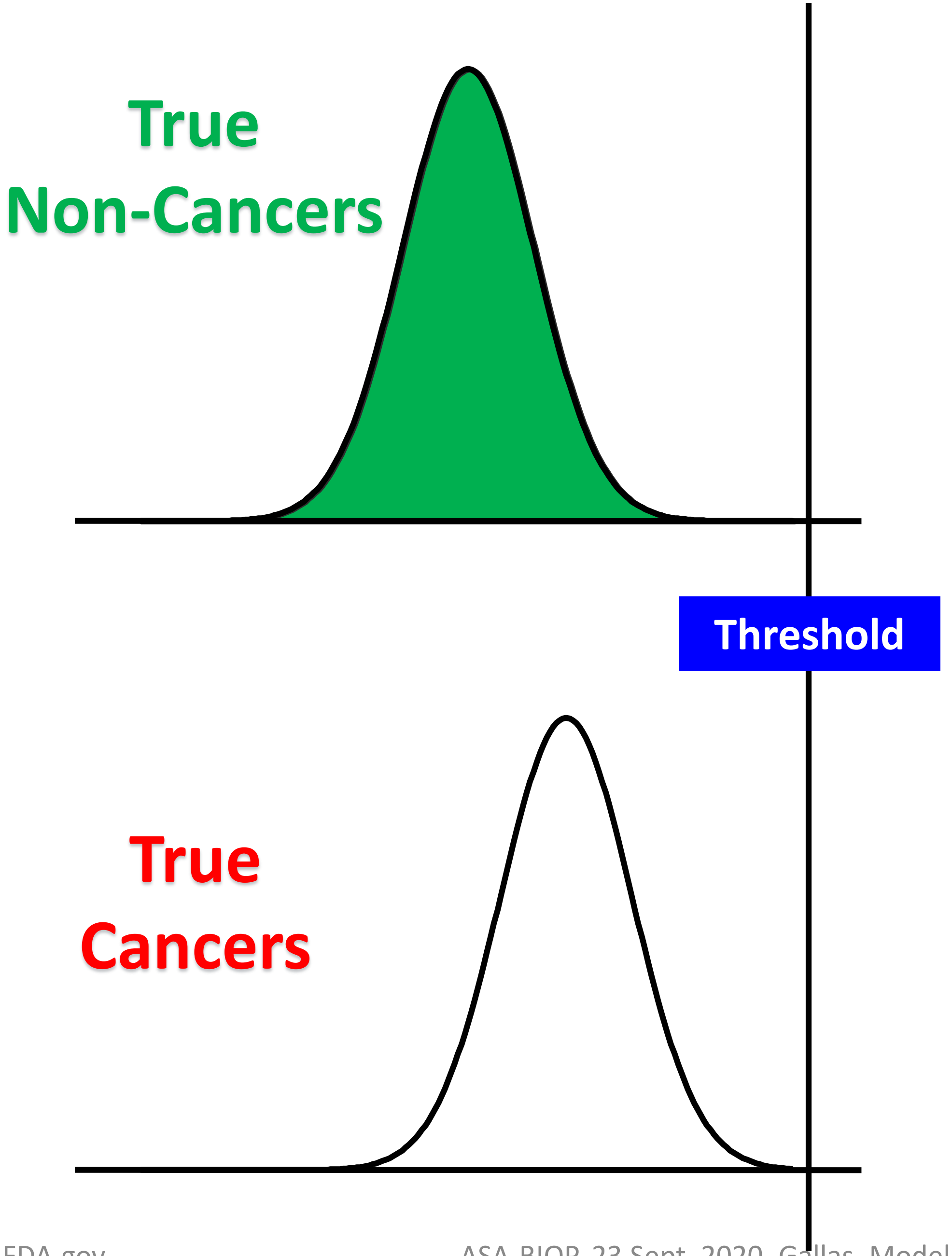
True Cancers



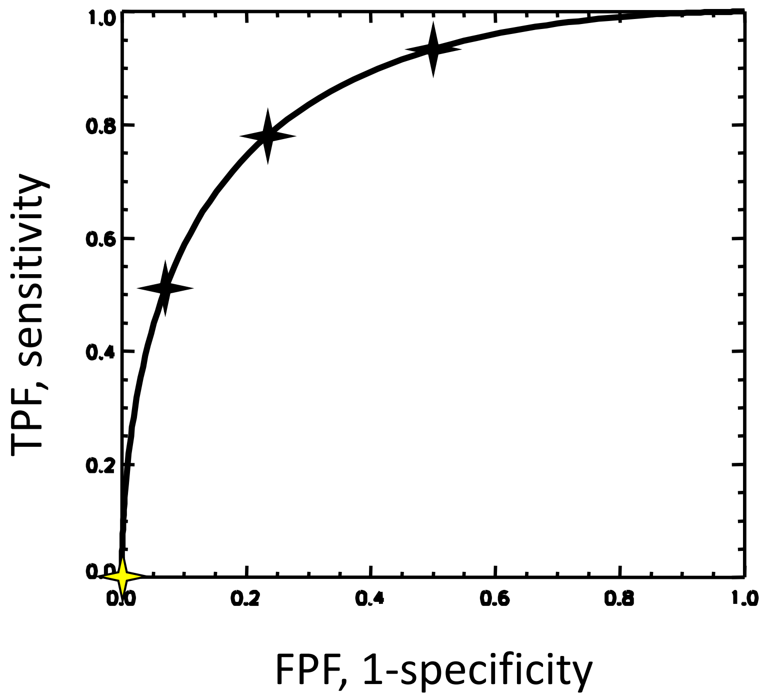
sensitivity = specificty


True
Non-Cancers
Threshold
True
Cancers


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
TPF, sensitivity
FPF, 1-specificity
high
specificity



Entire ROC curve



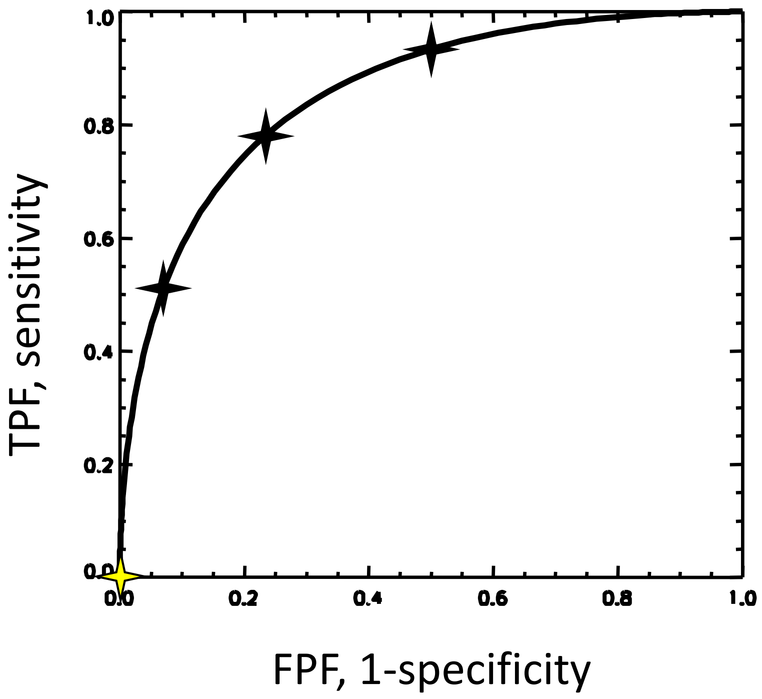
Different readers use different thresholds

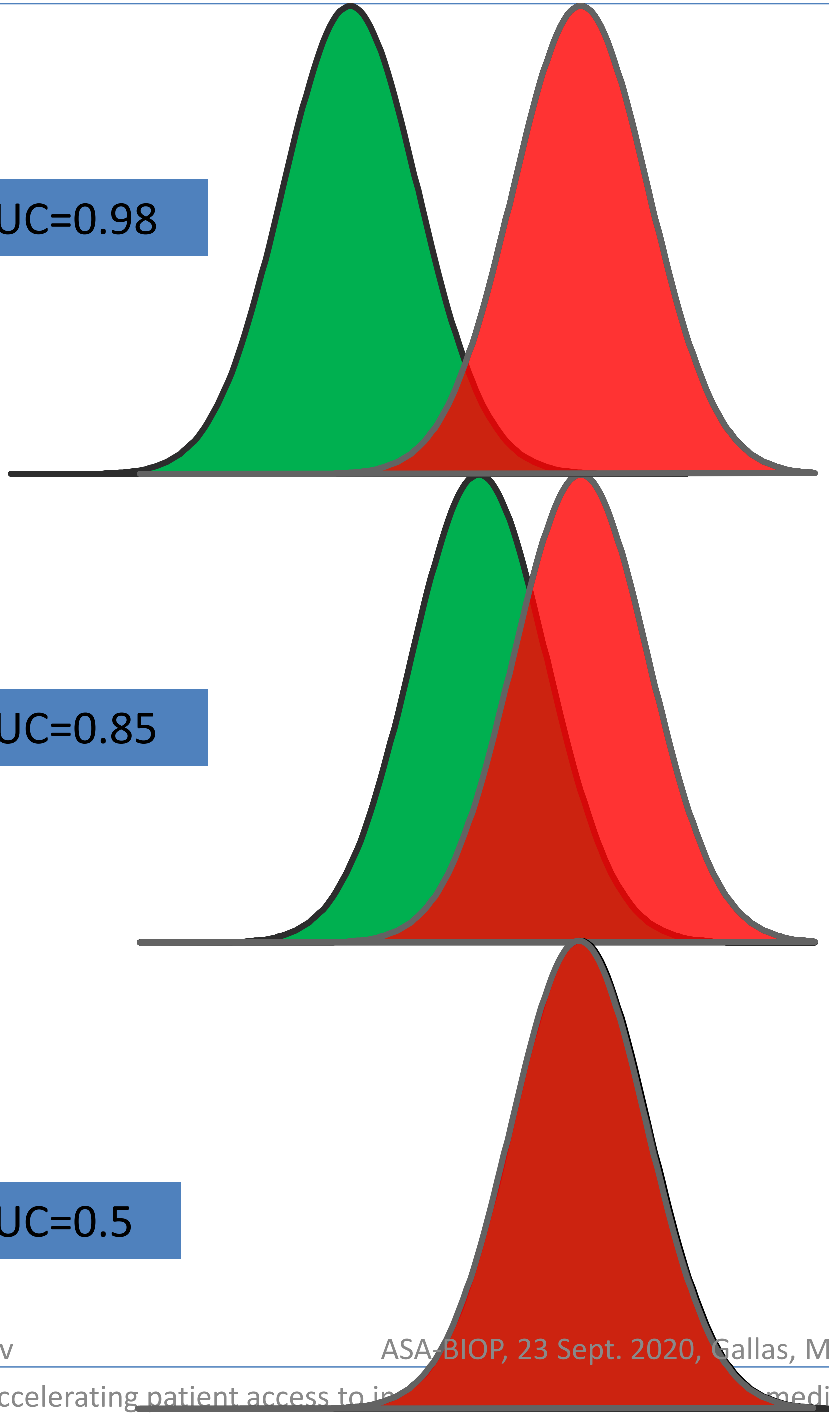
- Significant and important source of variability in sensitivity and specificity.

Prevalence effect

- As prevalence increases so does sensitivity (at the expense of specificity)
- The more cancers you see the more cases you call cancer

Entire ROC curve



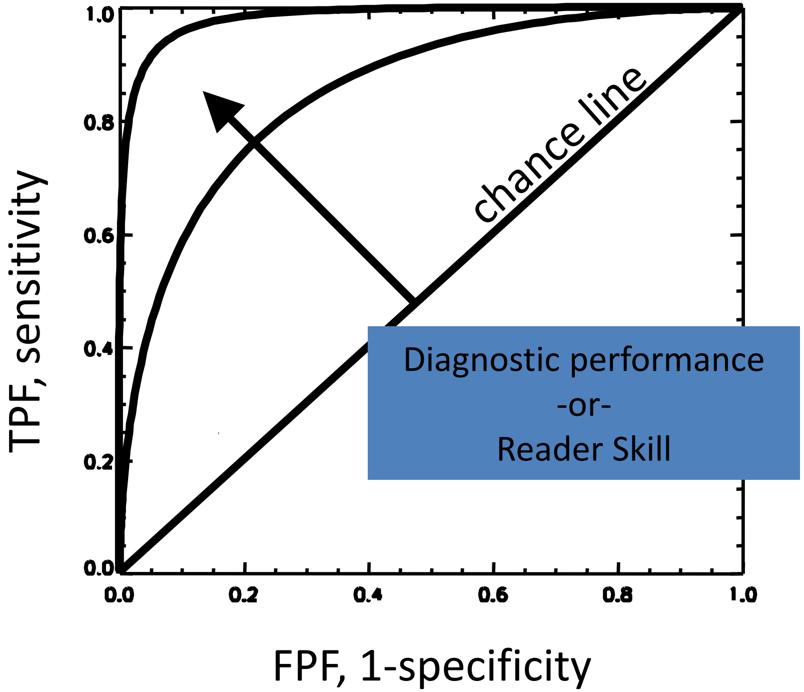


AUC=0.98

AUC=0.85

AUC=0.5

Entire ROC curve



Diagnostic performance
-or-
Reader Skill

Non-Parametric AUC estimate

- Single reader:

Signal-Present Scores (Cancer)

Signal-Absent Scores (Not Cancer)

$$\widehat{AUC}_r = \frac{1}{N_0 N_1} \sum_{k=1}^{N_1} \sum_{k'=1}^{N\_0} s({\color{red}X_{rk}} - {\color{green}Y_{rk'}})$$

$$s(x) = \begin{cases} 1.0, & x > 0 \\ 0.5, & x = 0 \\ 0.0, & x < 0 \end{cases}$$

Correct ranking

Incorrect ranking

- Average over readers:

$$\widehat{AUC}. = \frac{1}{N_R N_0 N_1} \sum_{r=1}^{N_R} \sum_{k=1}^{N_1} \sum_{k'=1}^{N_0} s(X_{rk} - Y_{rk'})$$

MRMC Analysis

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

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

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

MRMC Analysis

MRMC: Multi-reader, Multi-case Analysis

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

MRMC Analysis
Variance Components

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

MRMC Analysis

Variance Components

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

MRMC Analysis
Variance Components

- Single Modality
  - Gallas et al. (2009)

Non-diseased cases

Diseased cases

Interaction

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

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

Given U-statistic estimator of reader-averaged AUC

7 variance components
7 coefficients

No modeling

MRMC Analysis

Variance Components

- Two Modalities
  - Gallas et al. (2009)

Non-diseased cases

Diseased cases

Interaction

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

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 Analysis

Variance Components

- Two Modalities
  - Gallas et al. (2009)

Non-diseased cases

Diseased cases

Interaction

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

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

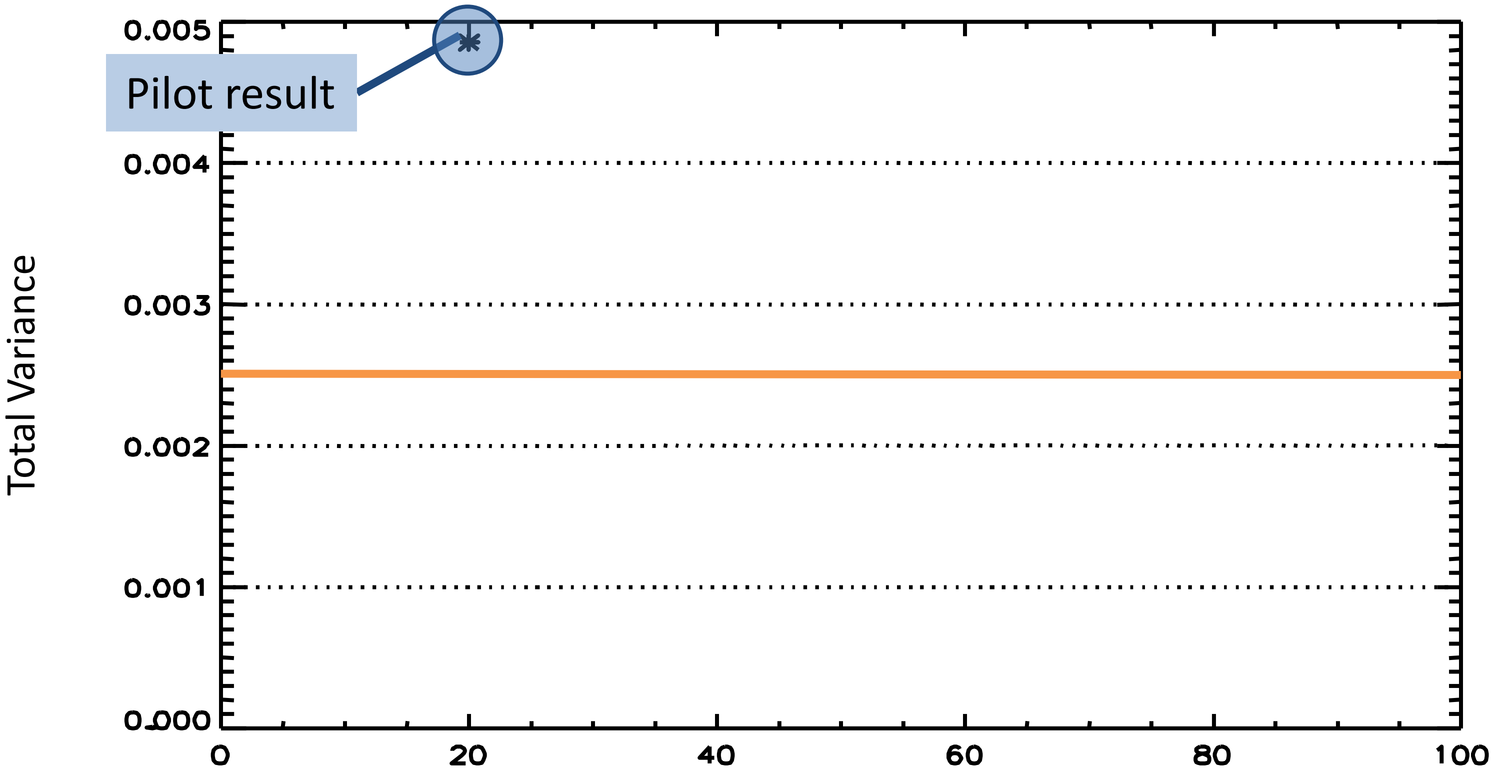
Sizing

Estimate components

Explore N0, N1, NR

MRMC Analysis

Size a Trial



Pilot result

Total number of cases

Add data to reduce variance by 50%.

1:1 sampling

4:1 sampling

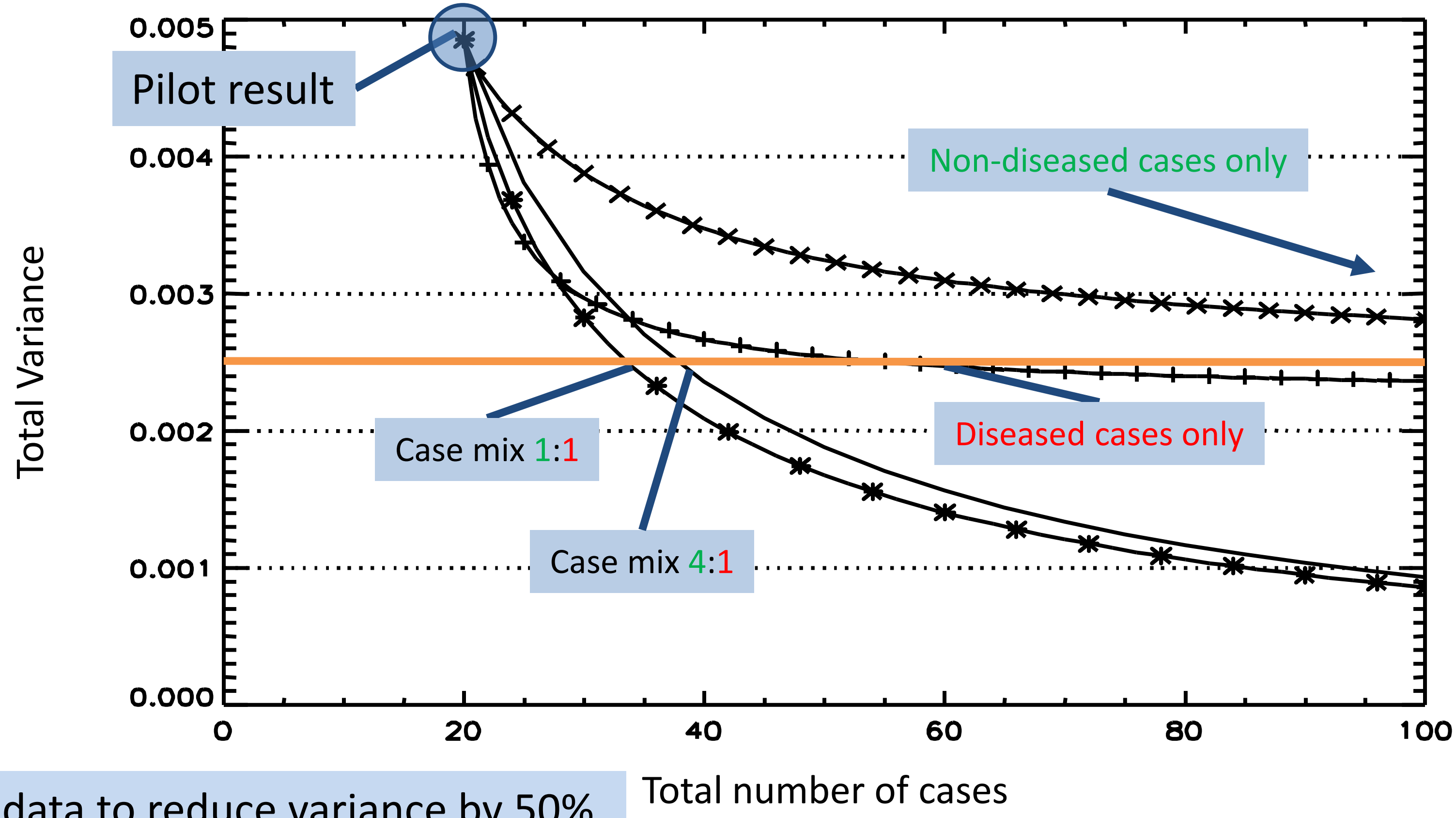
Only Non-Diseased

Only Diseased

Colposcopy Study
Plot courtesy of Hsu, NCI.

MRMC Analysis

Size a Trial



Add data to reduce variance by 50%.

1:1 sampling

4:1 sampling

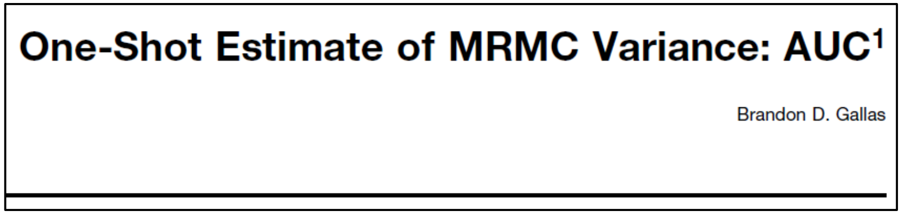
Only Non-Diseased

Only Diseased

Colposcopy Study
Plot courtesy of Hsu, NCI.

MRMC Analysis

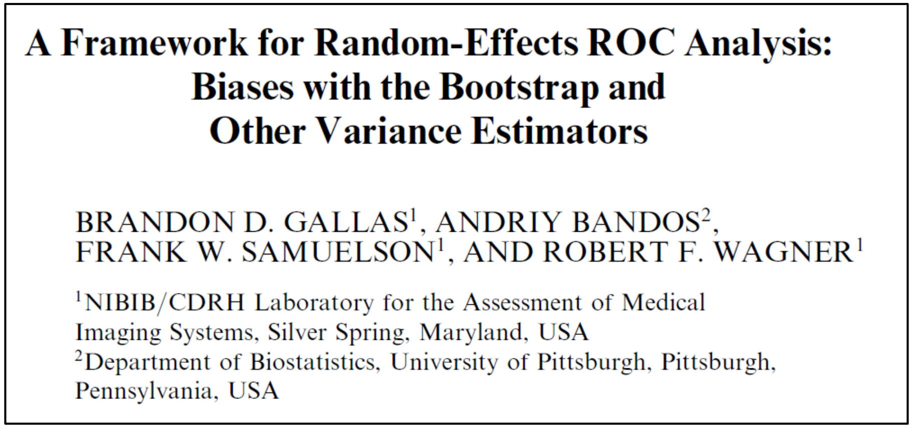
Publications and Software



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]

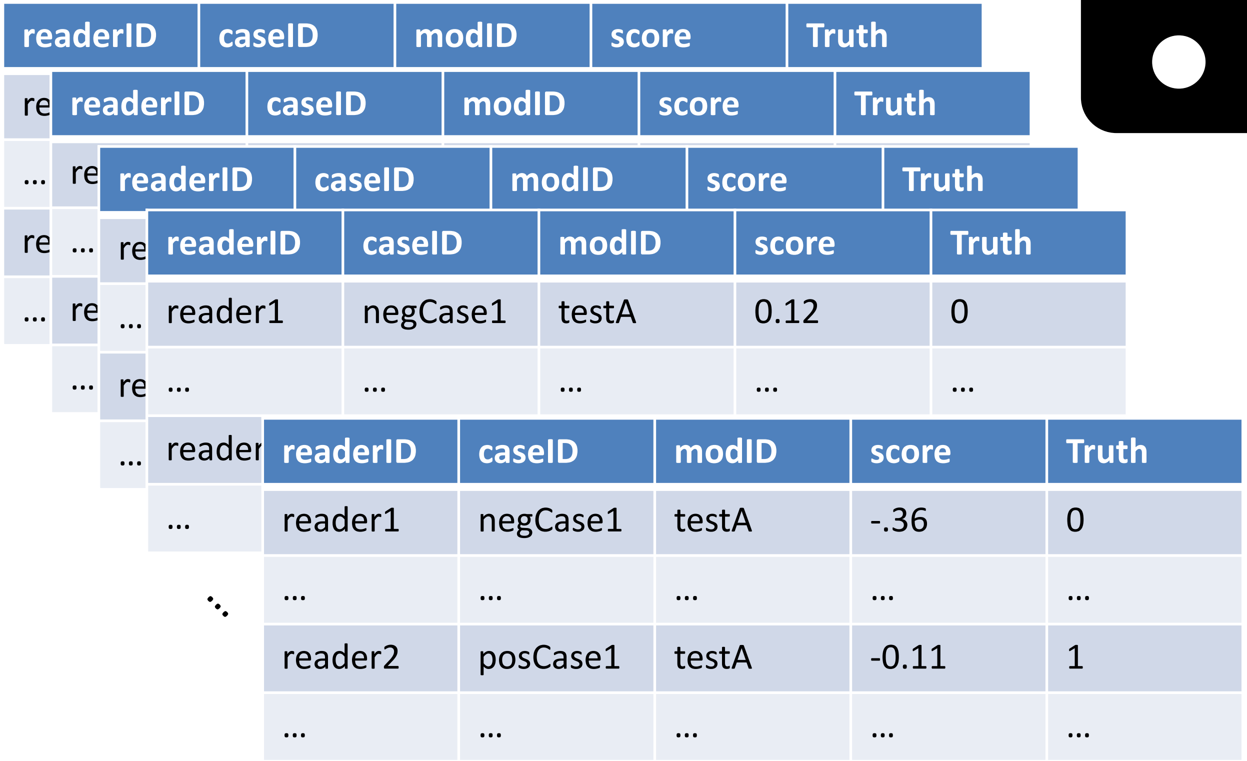
[1]NIBIB/CDRH Laboratory for the Assessment of Medical Imaging Systems, Silver Spring, Maryland, USA
[2]Department of Biostatistics, University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Communications in Statistics - Theory and Methods, 2009
https://doi.org/10.1080/03610920802610084

Published iMRMC Software

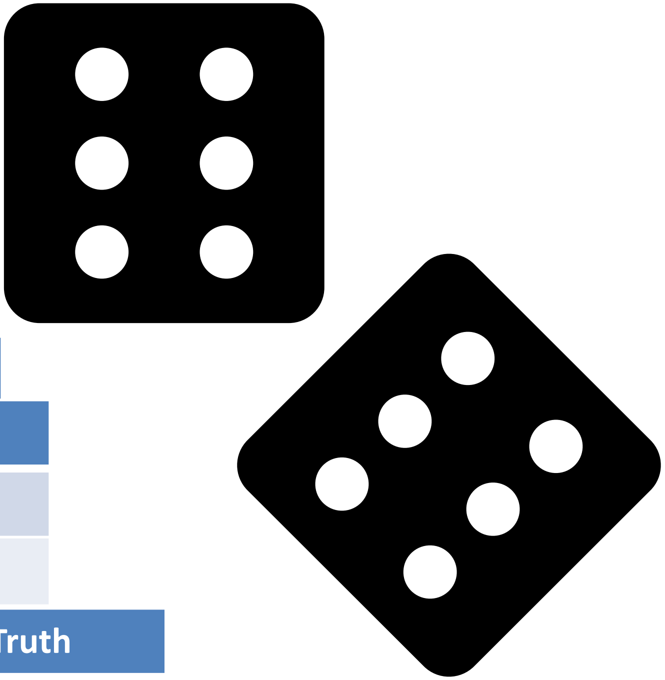
- 2013: Java Application - Google Code
  - Retired
- 2015: Java Application – GitHub
  - https://github.com/DIDSR/iMRMC
- 2017: R Package – CRAN
  - https://cran.r-project.org/web/packages/iMRMC/index.html

MRMC Simulation



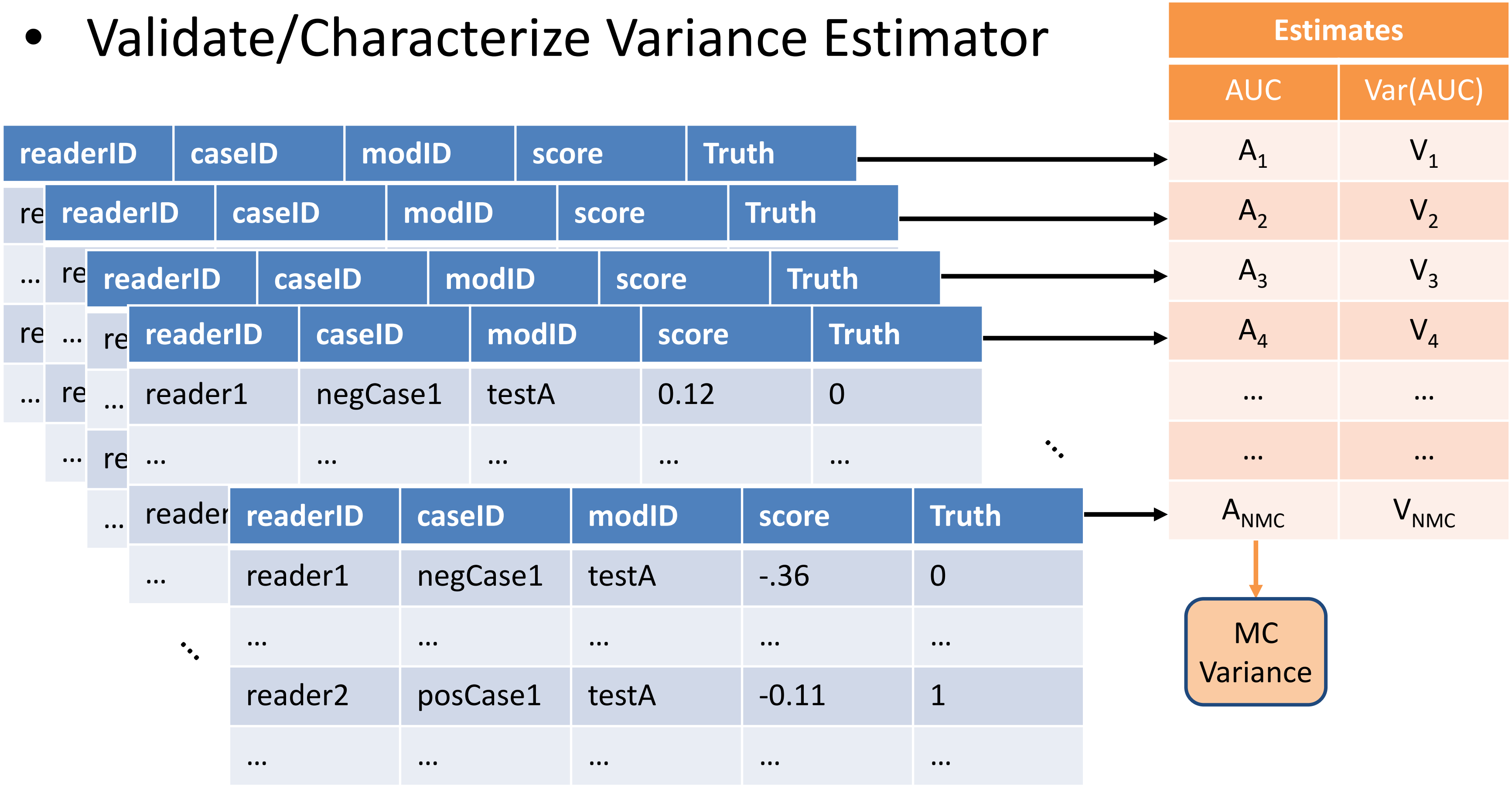
readerID	caseID	modID	score	Truth
reader1	negCase1	testA	0.12	0
…	…	…	…	…

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



MRMC Simulation

- Validate/Characterize Variance Estimator



readerID	caseID	modID	score	Truth

readerID	caseID	modID	score	Truth

readerID	caseID	modID	score	Truth

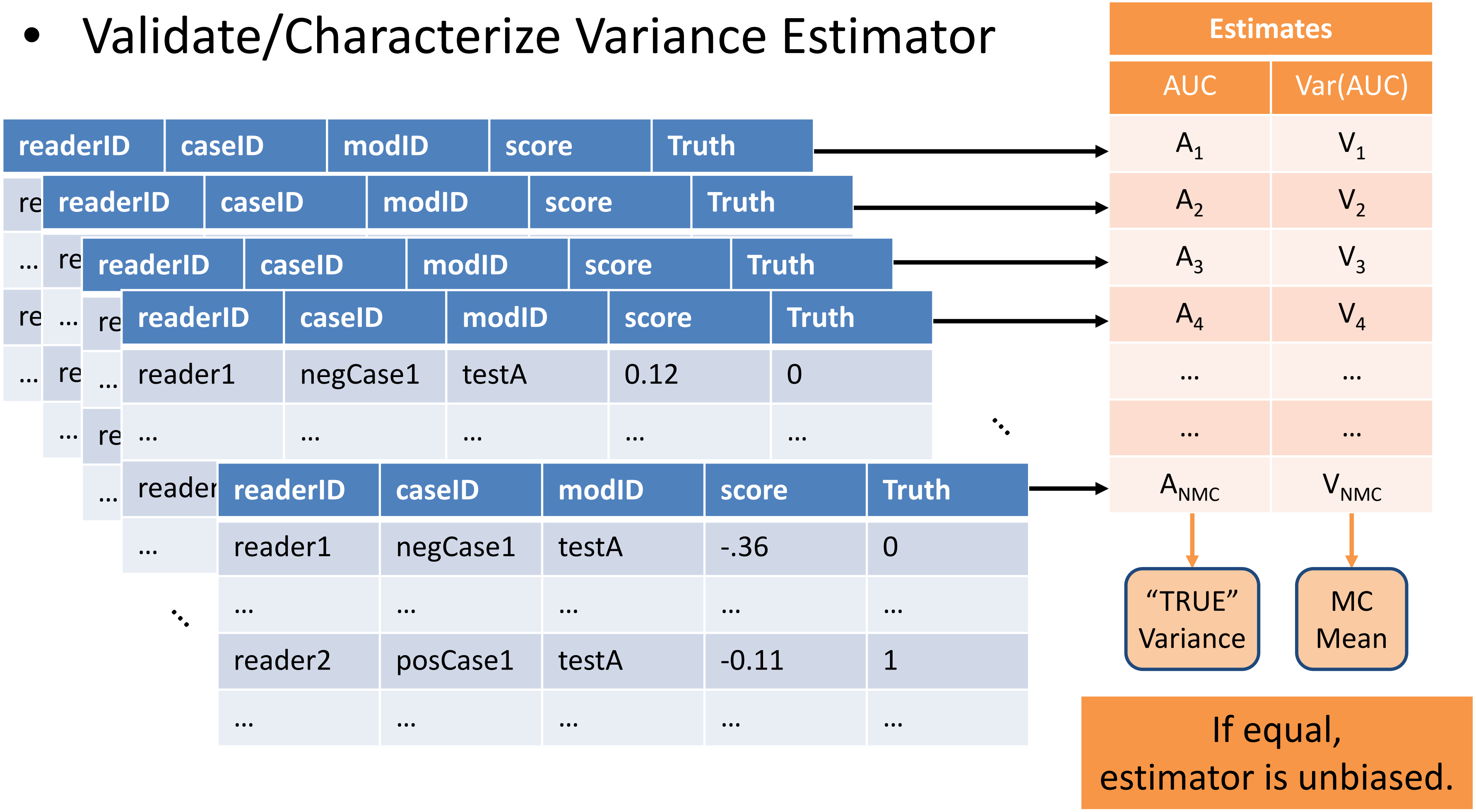
readerID	caseID	modID	score	Truth
reader1	negCase1	testA	0.12	0
…	…	…	…	…

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

Estimates	
AUC	Var(AUC)
A_1	V_1
A_2	V_2
A_3	V_3
A_4	V_4
…	…
…	…
A_{NMC}	V_{NMC}

MRMC Simulation

- Validate/Characterize Variance Estimator



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
…	…	…	…	…

Estimates	
AUC	Var(AUC)
A_1	V_1
A_2	V_2
A_3	V_3
A_4	V_4
…	…
…	…
A_{NMC}	V_{NMC}

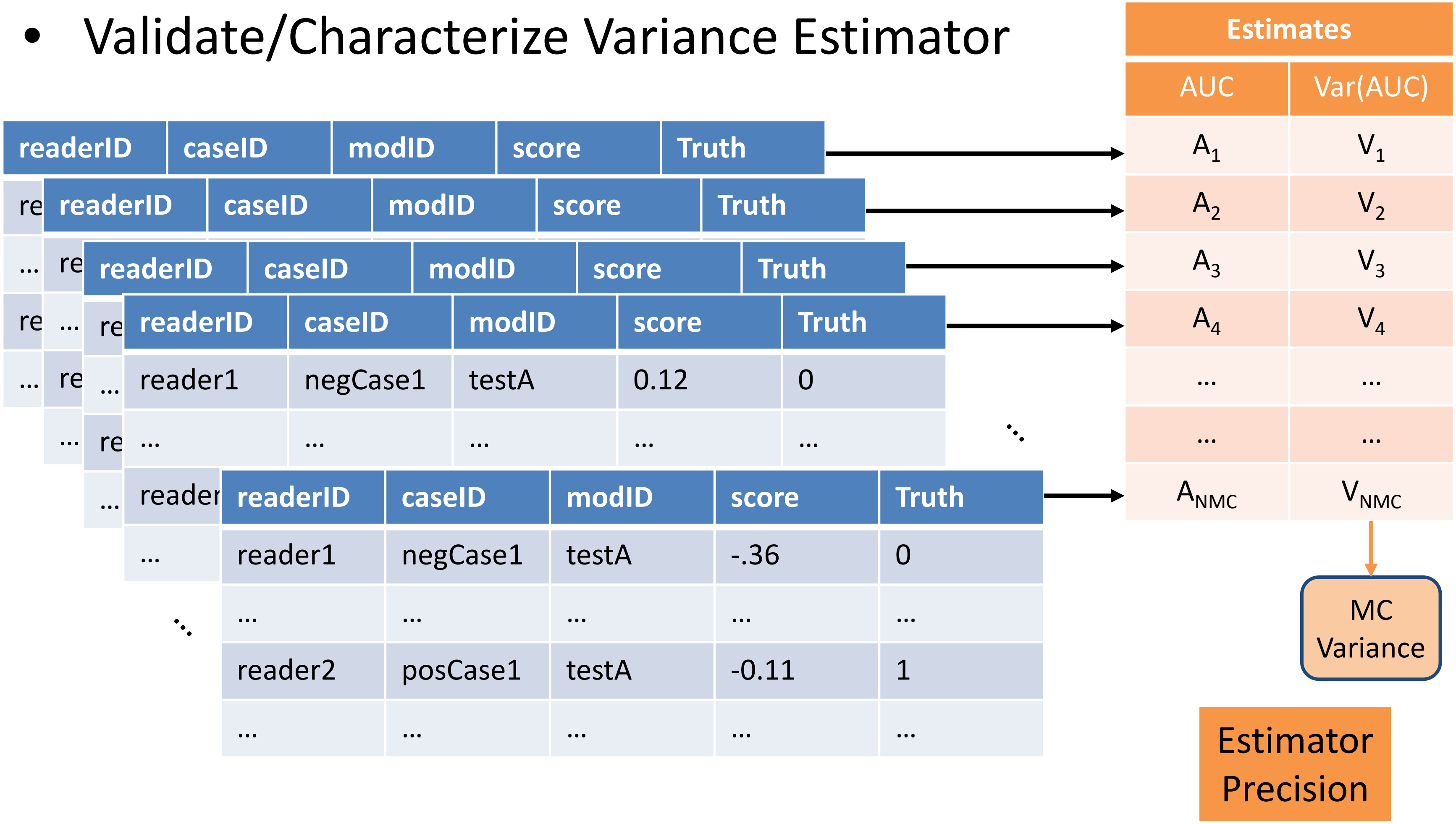
"TRUE" Variance

MC Mean

If equal,
estimator is unbiased.

MRMC Simulation

- Validate/Characterize Variance Estimator



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
…	…	…	…	…

Estimates	
AUC	Var(AUC)
A_1	V_1
A_2	V_2
A_3	V_3
A_4	V_4
…	…
…	…
A_{NMC}	V_{NMC}

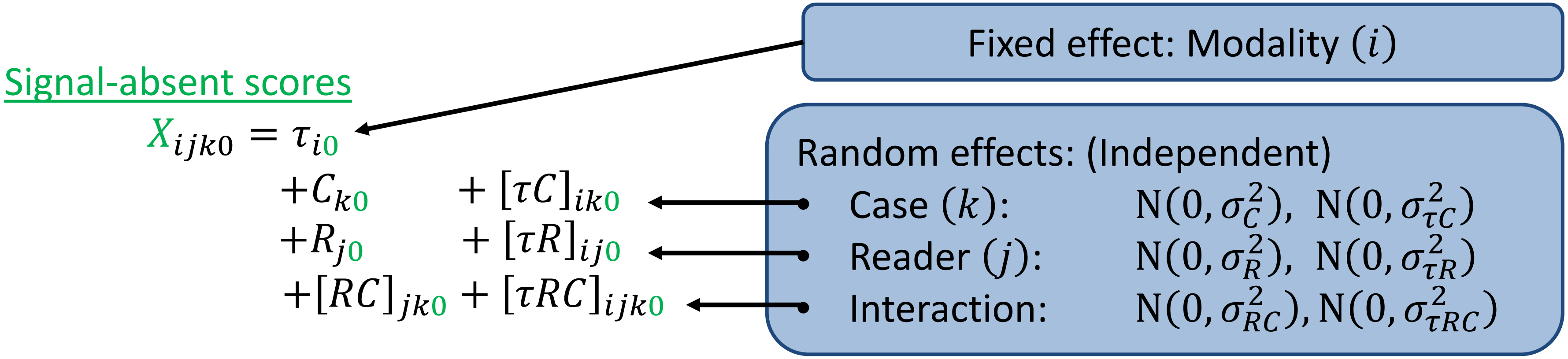
MC Variance

Estimator Precision

MRMC Simulation

Roe and Metz Model (1997)

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



Signal-absent scores

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

Fixed effect: Modality (i)

Random effects: (Independent)

- Case (k): $\mathrm{N}(0,\sigma_C^2)$, $\mathrm{N}(0,\sigma_{\tau C}^2)$
- Reader (j): $\mathrm{N}(0,\sigma_R^2)$, $\mathrm{N}(0,\sigma_{\tau R}^2)$
- Interaction: $\mathrm{N}(0,\sigma_{RC}^2)$, $\mathrm{N}(0,\sigma_{\tau RC}^2)$

MRMC Simulation
Roe and Metz Model (1997)

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

Signal-present scores

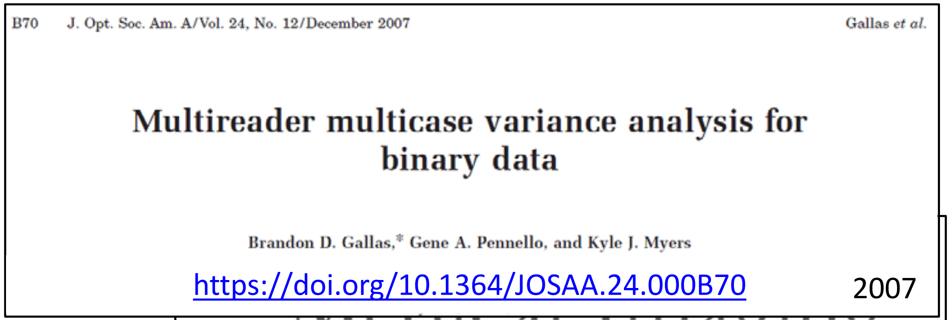
$$\begin{aligned} Y_{ijk1} = \tau_{i1} & \\ +C_{k1} & \quad + [\tau C]_{ik1} \\ +R_{j1} & \quad + [\tau R]_{ij1} \\ +[RC]_{jk1} & \quad + [\tau RC]_{ijk1} \end{aligned}$$

Looks like
3-way ANOVA

Warning
Simulation for scores not AUC

MRMC Simulation
Roe and Metz Model: Updates

- Binary Data

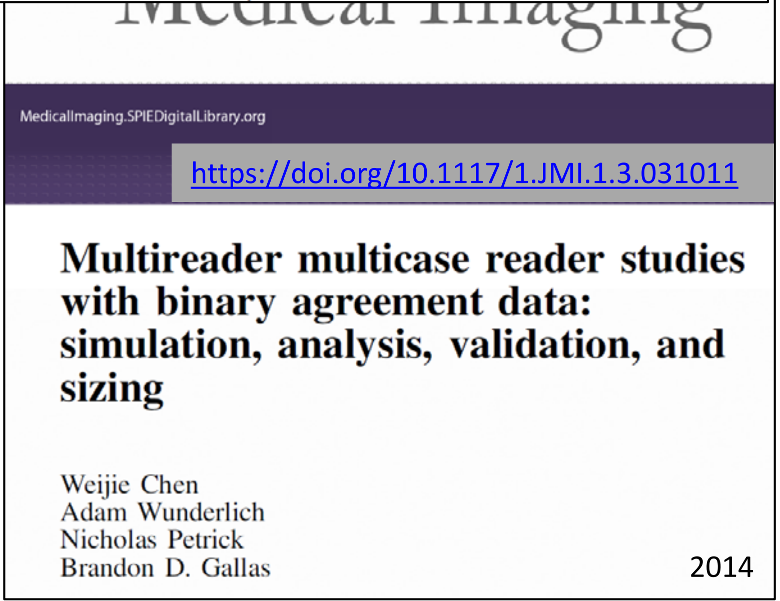


B70 J. Opt. Soc. Am. A/Vol. 24, No. 12/December 2007 Gallas *et al.*

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 2007



Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

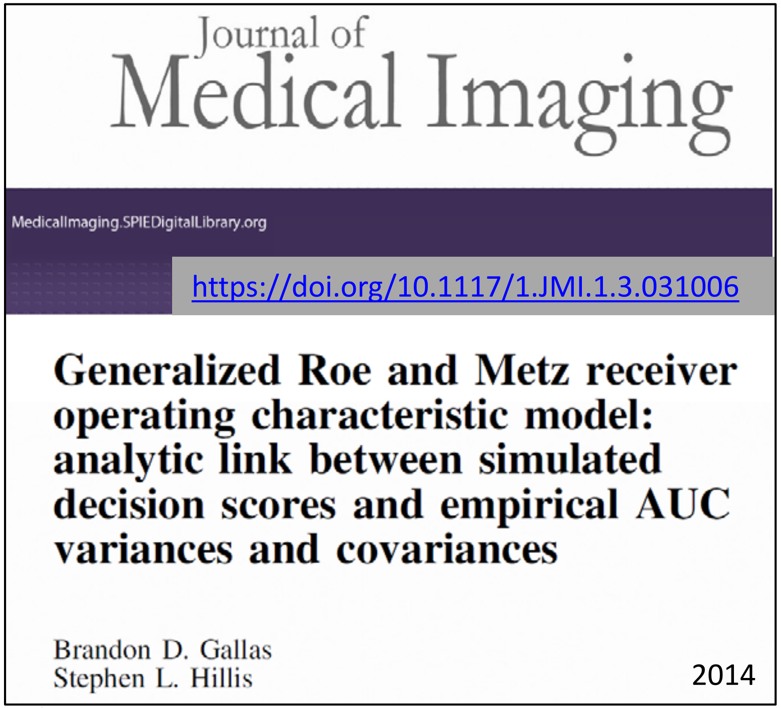
https://doi.org/10.1117/1.JMI.1.3.031011

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

Weijie Chen
Adam Wunderlich
Nicholas Petrick
Brandon D. Gallas

2014

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



Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

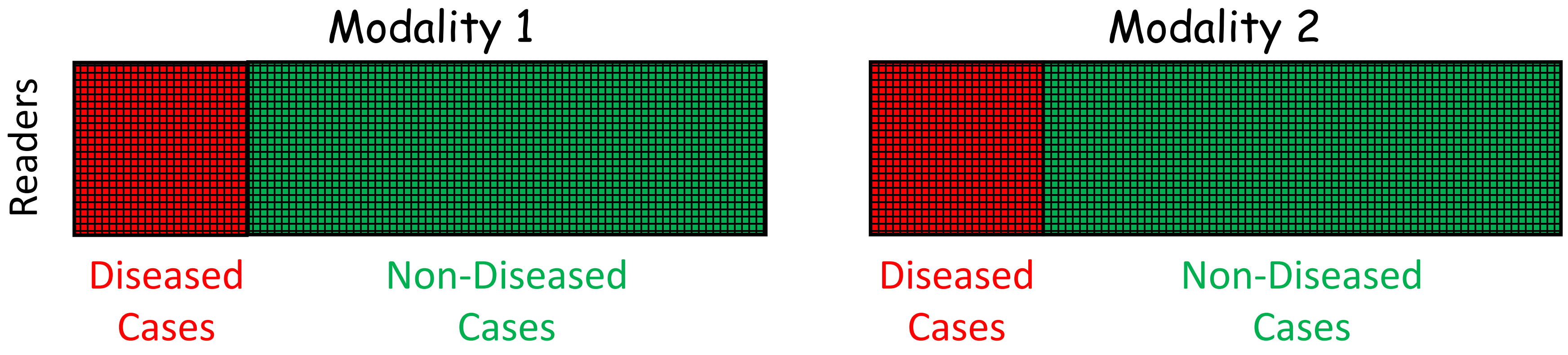
https://doi.org/10.1117/1.JMI.1.3.031006

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

Study Designs



Modality 1

Readers

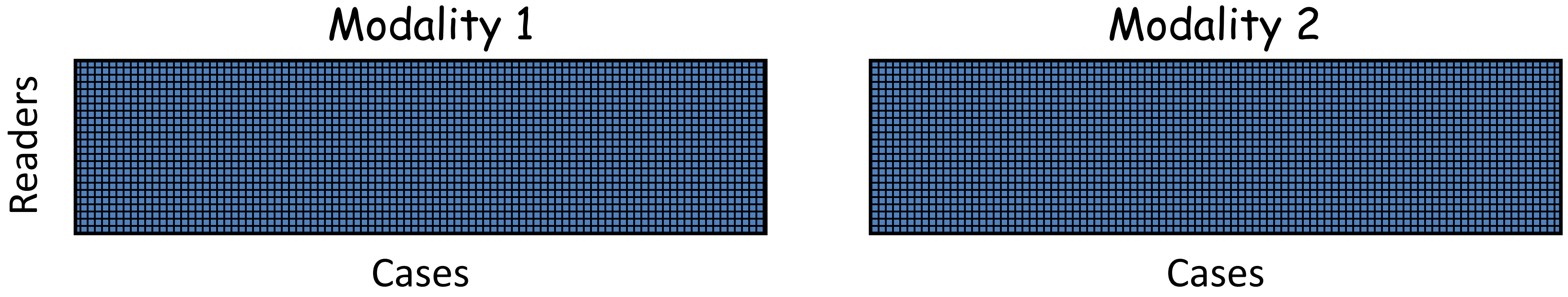
Diseased Cases

Non-Diseased Cases

Modality 2

Diseased Cases

Non-Diseased Cases



Modality 1

Readers

Cases

Modality 2

Cases

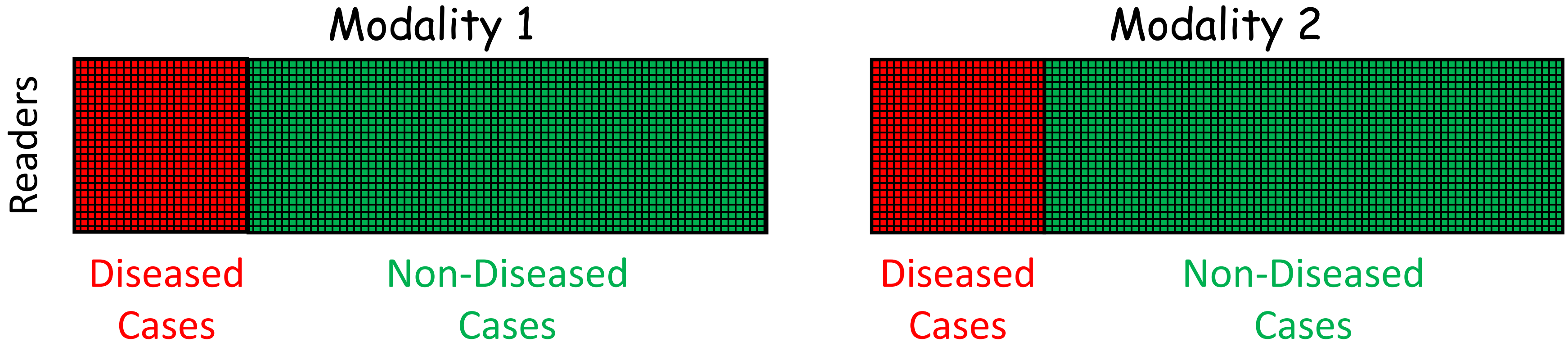
Study Designs
Fully-Crossed

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

Data Array

Rows = readers

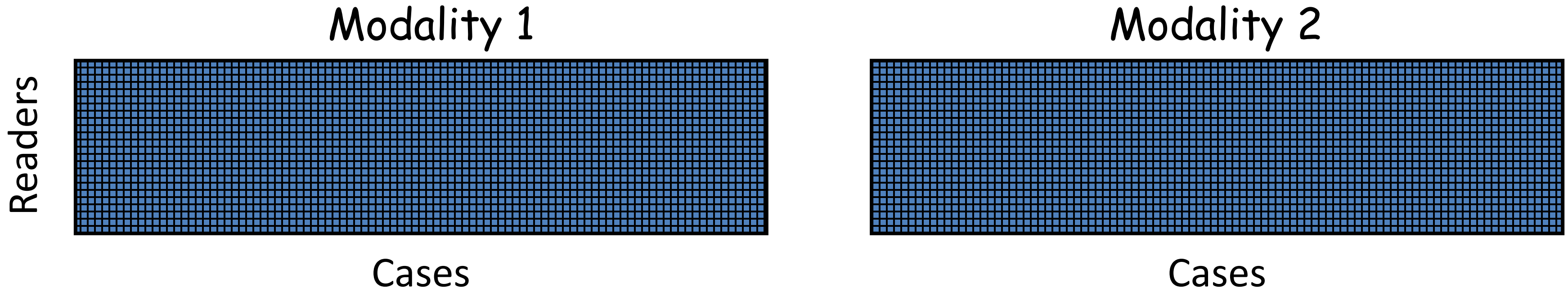
Cols = cases



Study Designs
Fully-Crossed

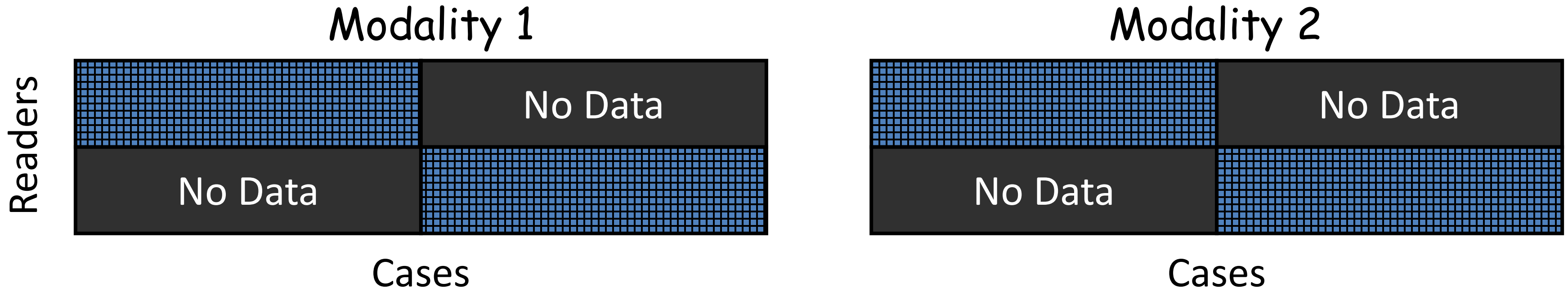
- Fully-crossed study
  - All readers read all cases
  - Readers and cases are paired across modalities

Remove truth labels to unclutter study design concepts.



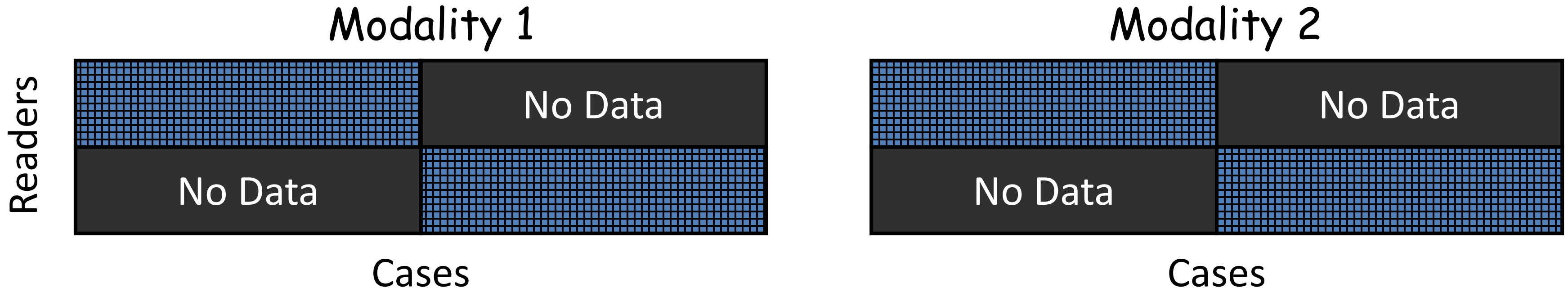
Study Designs
Split-Plot

- 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



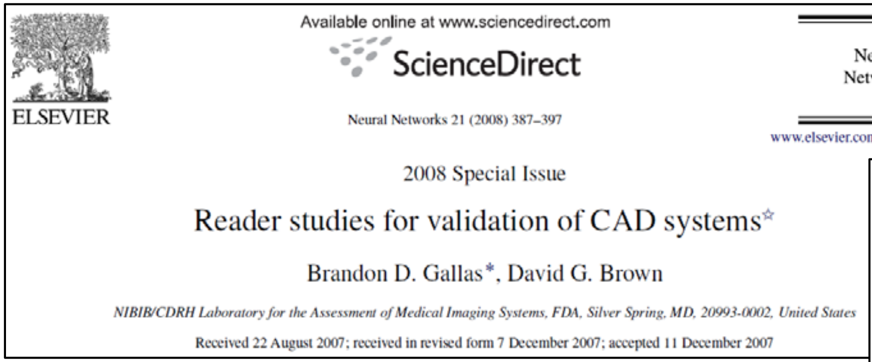
Study Designs
Split-Plot

- 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

Neural Networks 21 (2008) 387–397

Neural Networks

www.elsevier.com/locate/neunet

ELSEVIER

2008 Special Issue

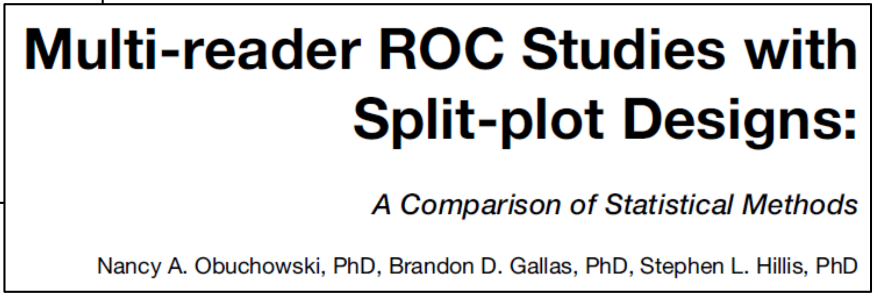
Reader studies for validation of CAD systems☆

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:

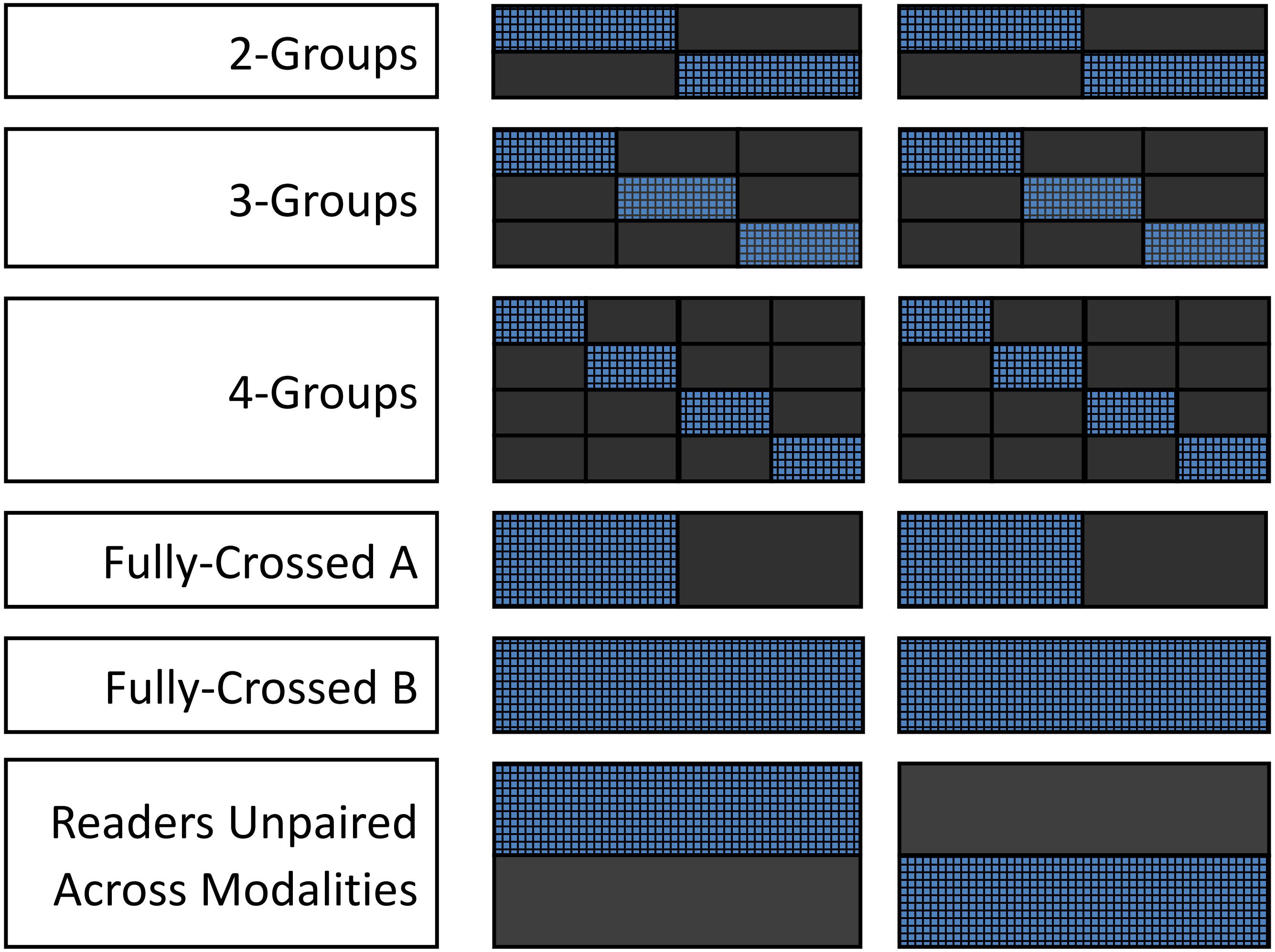
A Comparison of Statistical Methods

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



2-Groups

3-Groups

4-Groups

Fully-Crossed A

Fully-Crossed B

Readers Unpaired Across Modalities

Compare designs using simulation

Study Designs: Efficiency

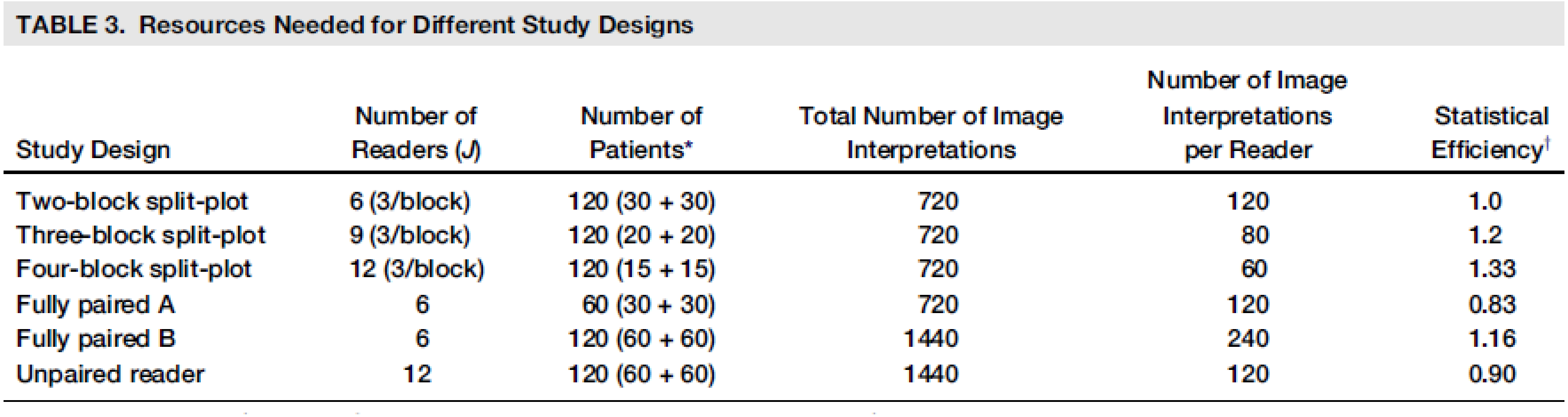


TABLE 3. Resources Needed for Different Study Designs

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

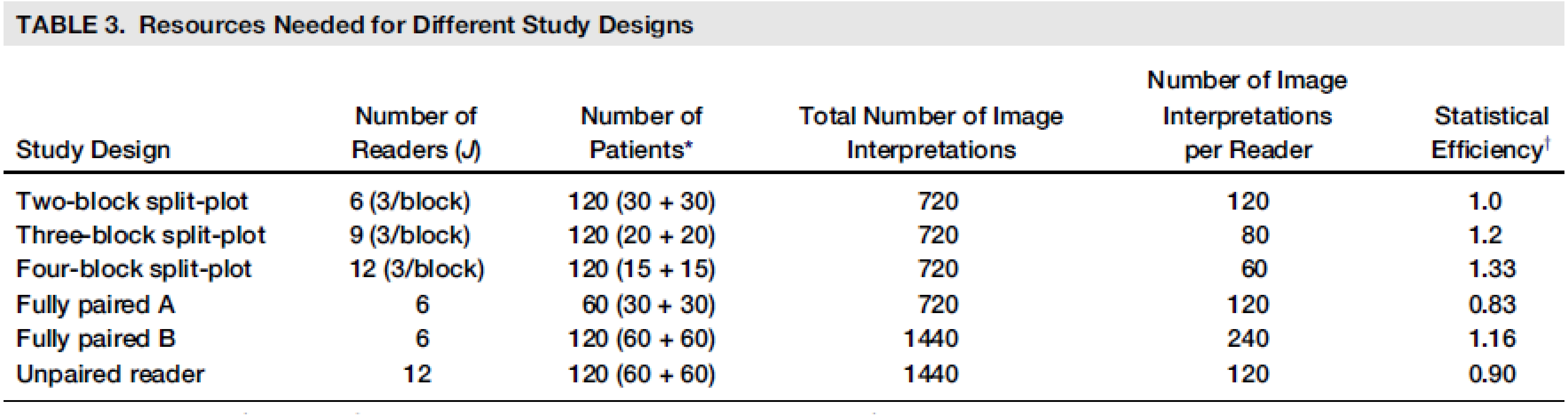


TABLE 3. Resources Needed for Different Study Designs

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

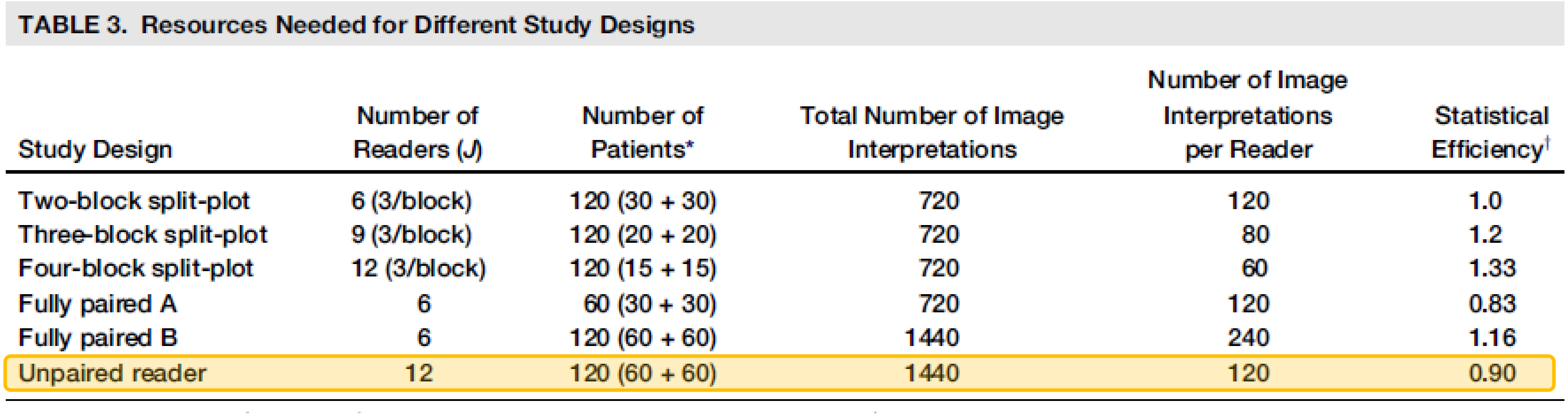


TABLE 3. Resources Needed for Different Study Designs

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

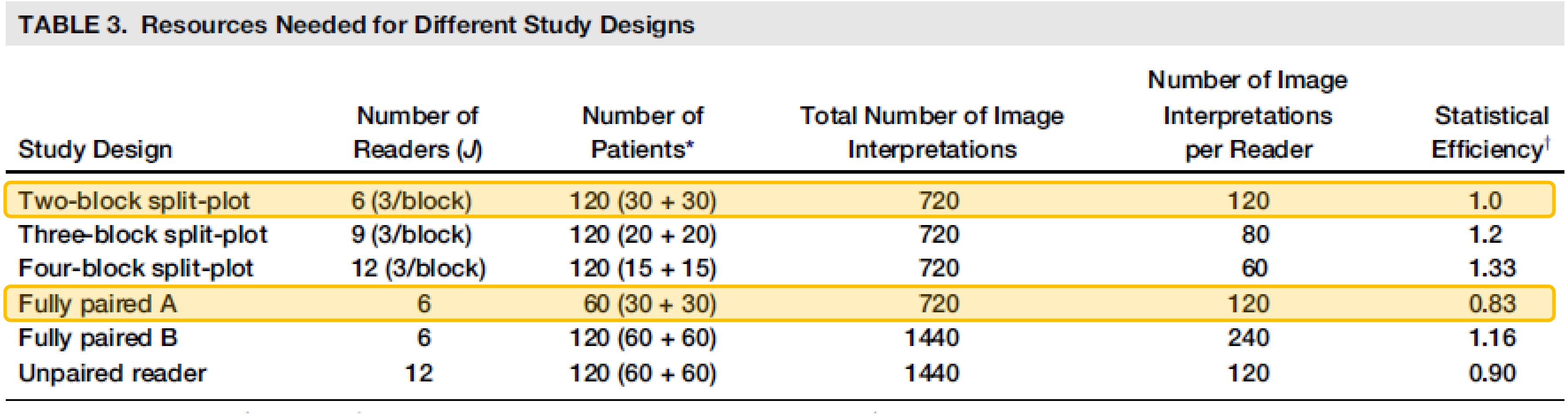


TABLE 3. Resources Needed for Different Study Designs

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

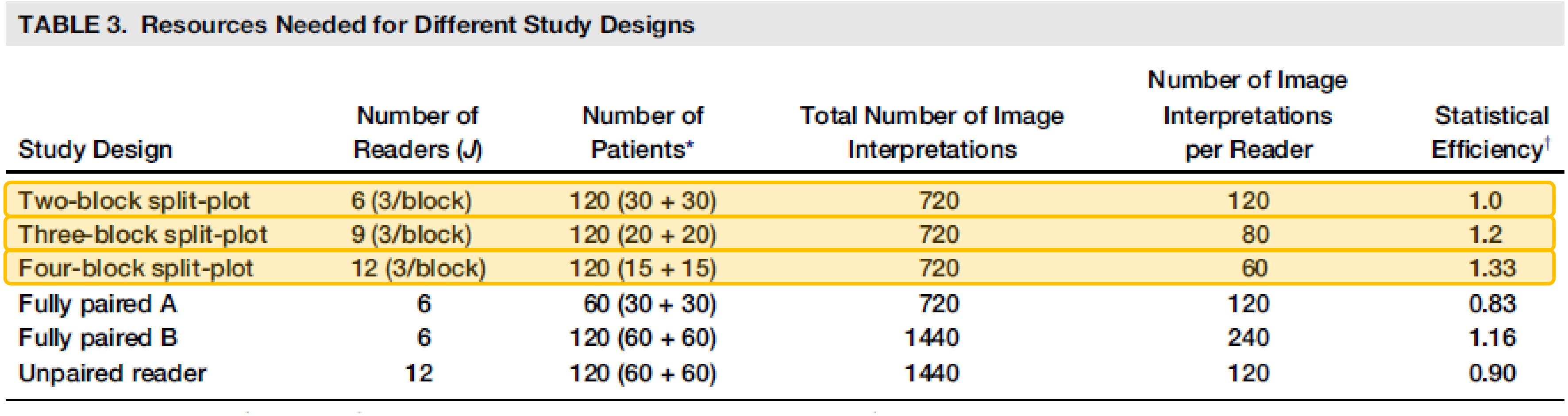


TABLE 3. Resources Needed for Different Study Designs

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

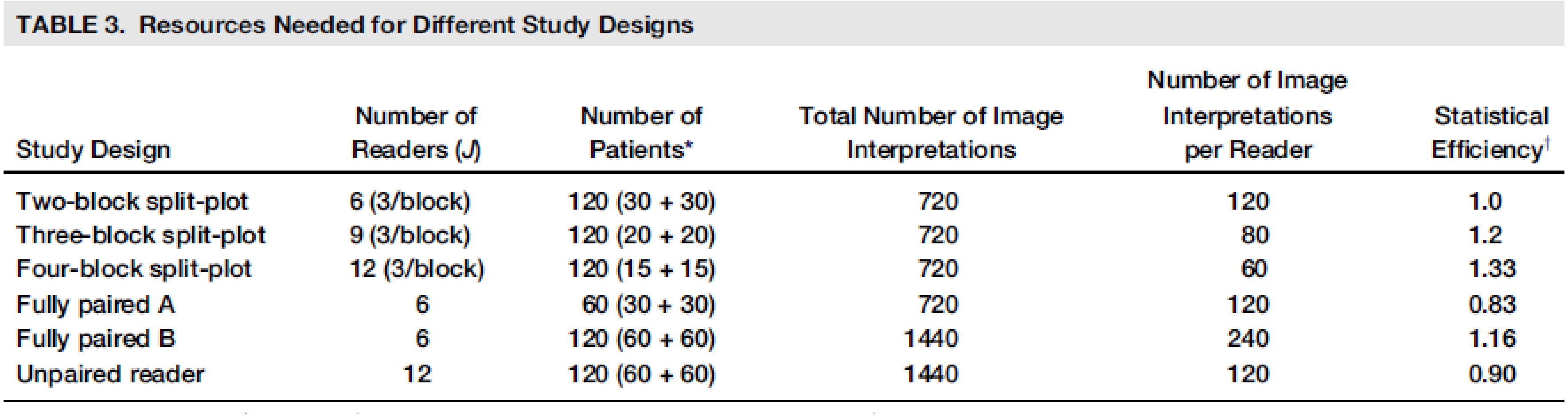


TABLE 3. Resources Needed for Different Study Designs

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

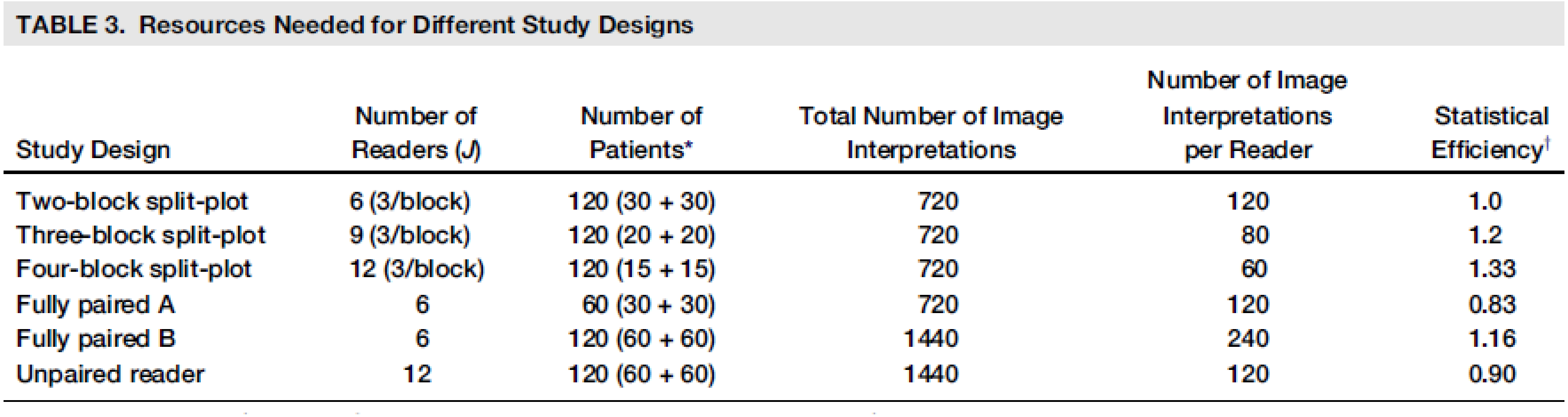


TABLE 3. Resources Needed for Different Study Designs

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.

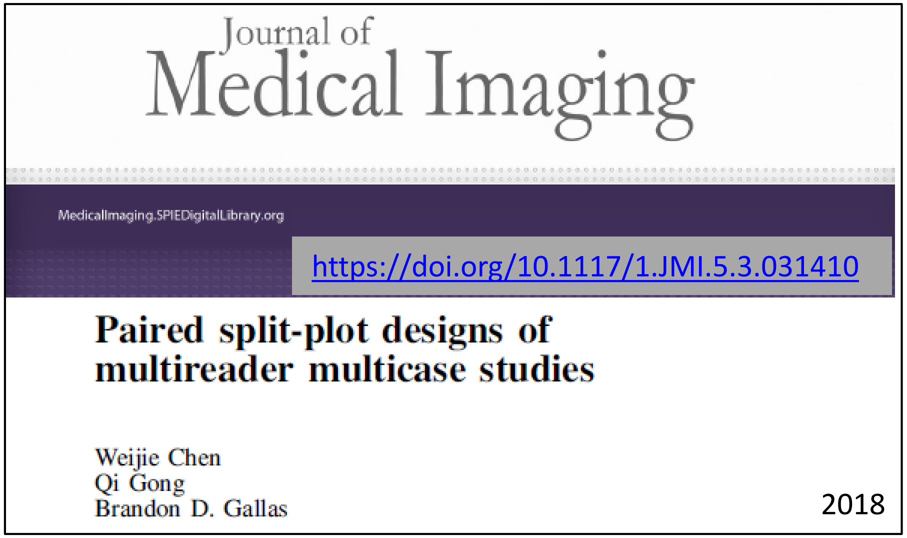
Study Designs:
Efficiency

- 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:

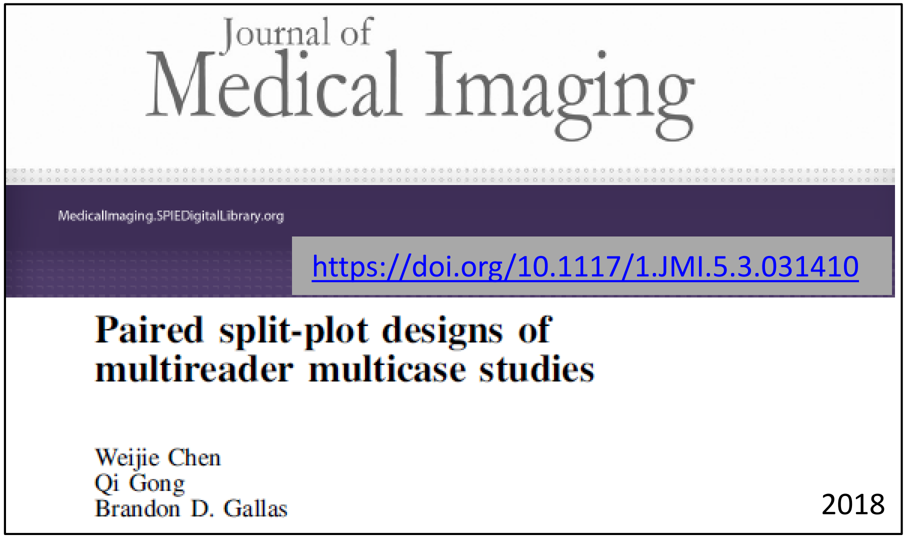
Efficiency

- 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

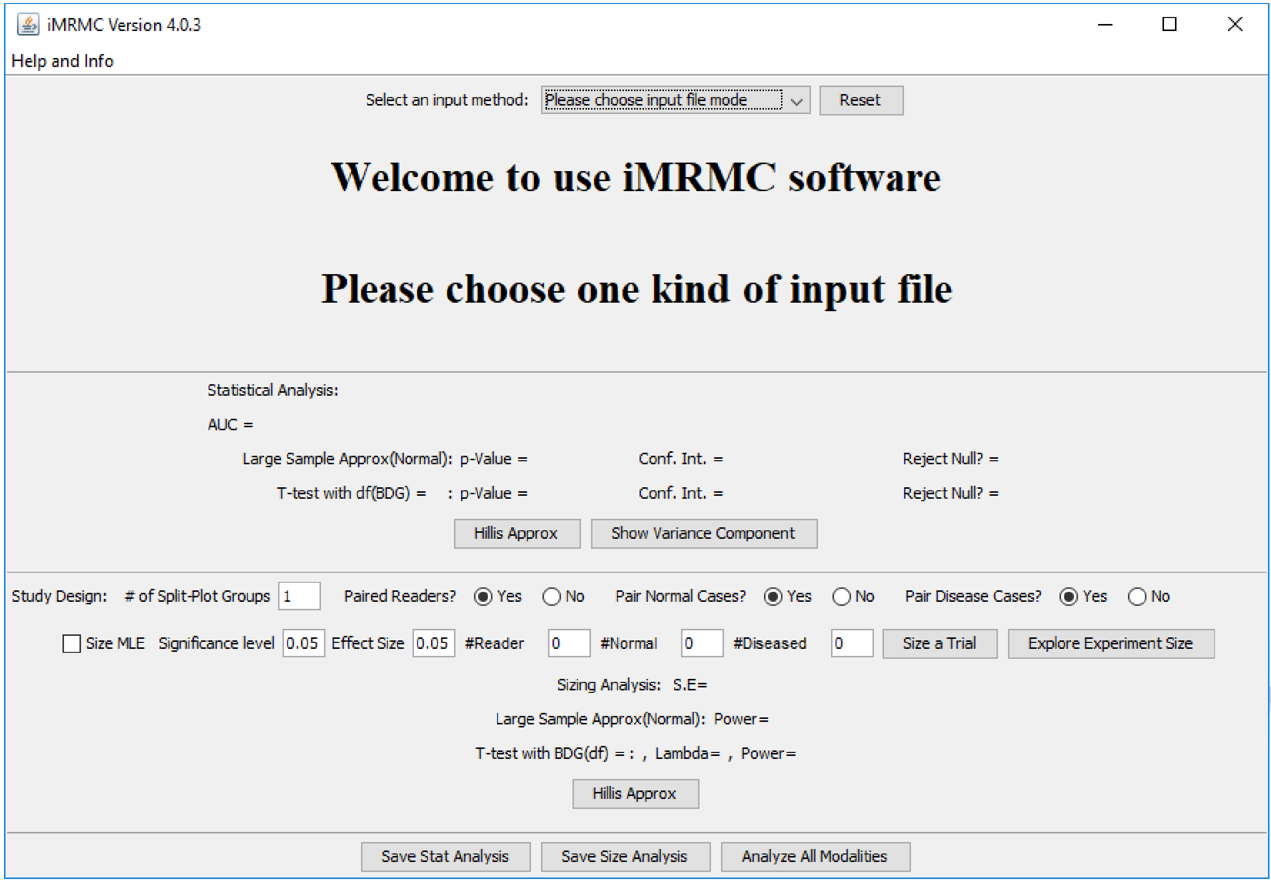
https://doi.org/10.1117/1.JMI.5.3.031410

Paired split-plot designs of multireader multicase studies

Weijie Chen
Qi Gong
Brandon D. Gallas

2018

MRMC Tools



Select an input method: Please choose input file mode

Welcome to use iMRMC software

Please choose one kind of input file

Statistical Analysis:

AUC =

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

T-test with df(BDG) = : p-Value = Conf. Int. = Reject Null? =

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

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

Sizing Analysis: S.E=

Large Sample Approx(Normal): Power=

T-test with BDG(df) = : , Lambda= , Power=

MRMC Tools

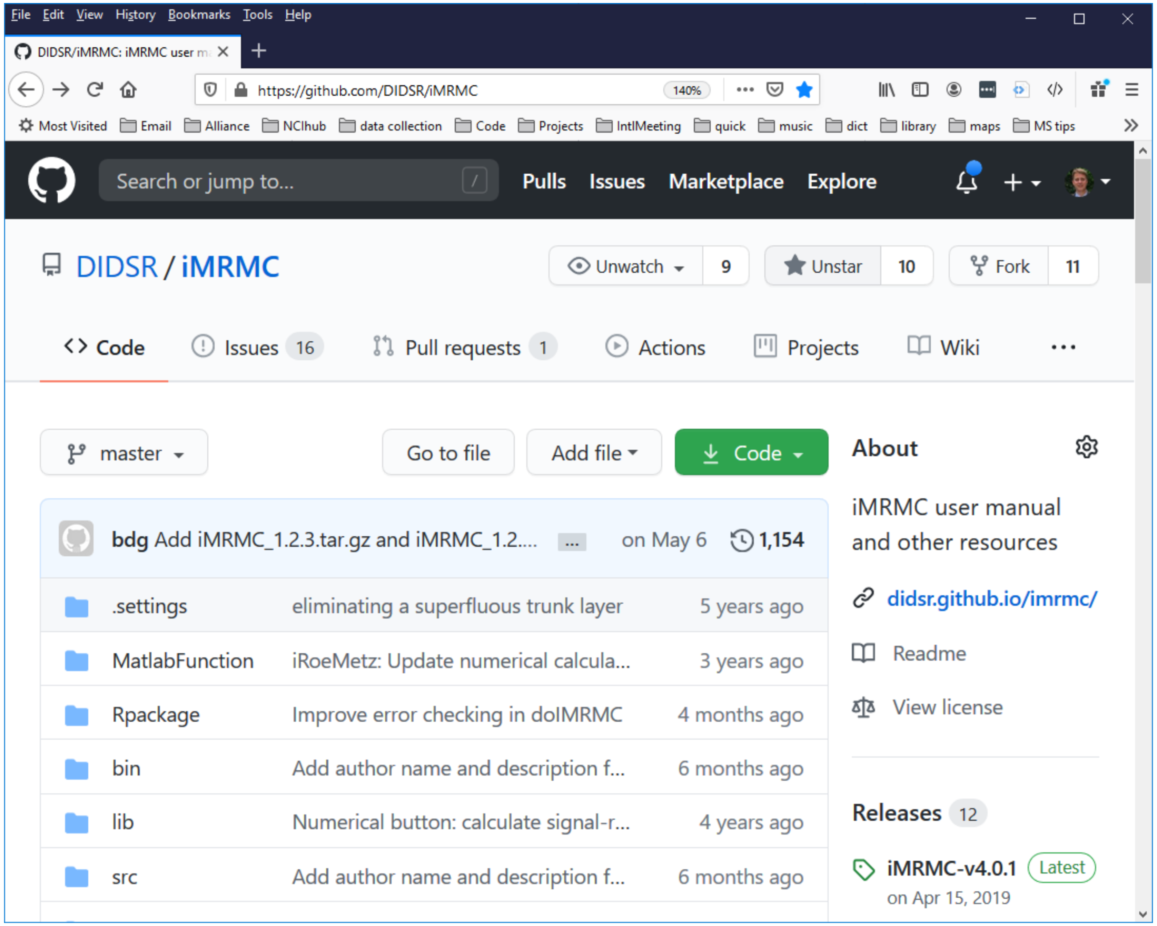
iMRMC Software, GitHub Repository

- 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



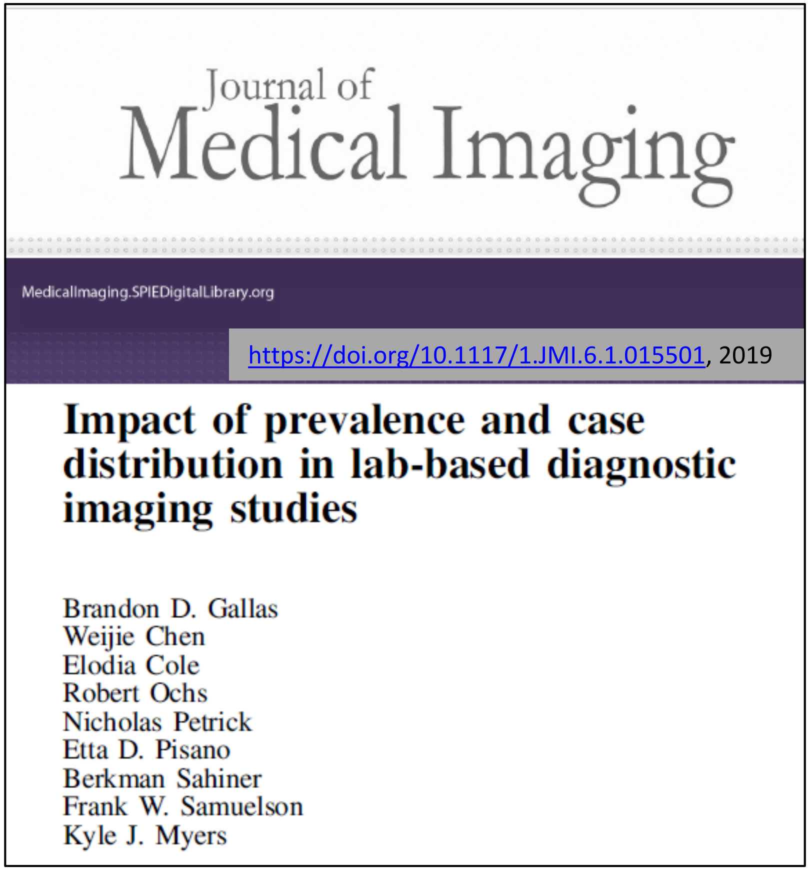
iMRMC user manual and other resources

didsr.github.io/imrmc/

bdg Add iMRMC_1.2.3.tar.gz and iMRMC_1.2.... on May 6 1,154

.settings	eliminating a superfluous trunk layer	5 years ago
MatlabFunction	iRoeMetz: Update numerical calcula...	3 years ago
Rpackage	Improve error checking in doIMRMC	4 months ago
bin	Add author name and description f...	6 months ago
lib	Numerical button: calculate signal-r...	4 years ago
src	Add author name and description f...	6 months ago

MRMC Analysis
Publications and Software



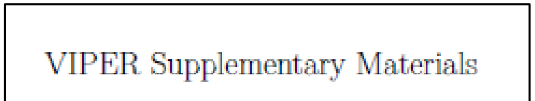
Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

https://doi.org/10.1117/1.JMI.6.1.015501, 2019

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



VIPER Supplementary Materials

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

GitHub Wiki Page: iMRMC-Datasets

- https://github.com/DIDSR/iMRMC/wiki/iMRMC-Datasets

viperData R package

- Data
- R scripts
- R Markdown Files

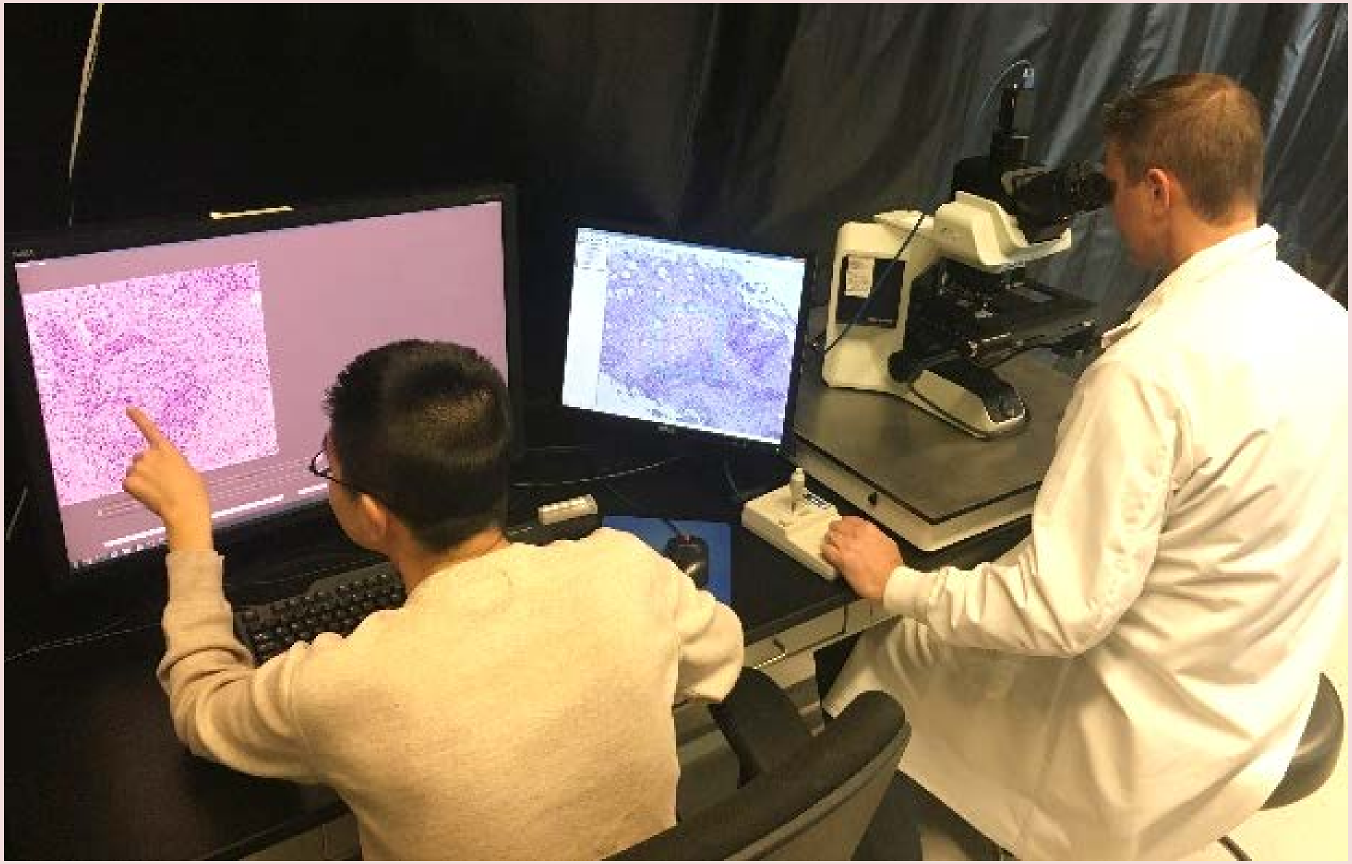
Supplementary Materials

- Study Designs (Split-Plot)
- Sizing analysis
- Histograms of reader scores and ROC curves

All analyses fully reproducible

BONUS

High-Throughput Truthing Project
HTT project



- Slides originally presented at
- SIIM: Society for Imaging Informatics in Medicine

- Play recorded audio (with fingers crossed)

High-throughput truthing (HTT) Collaborators

Collaboration of Volunteers

Engage stakeholders through the Alliance for Digital Pathology

Pathologists

Academia

Health Systems

Associations

Industry

Involve experts & the community.

HTT Core Collaborators

Project mgmt.

Sarah Dudgeon, MPH
FDA/CDRH/OSEL/DIDSR

caMicroscope team

Ashish Sharma, PhD
Emory University Department of Biomedical Informatics

Joel Saltz, MD PhD
Dept. of Biomedical Informatics, Stony Brook Medicine

Nan Li, MS
Dept. of Biomedical Informatics, Stony Brook Medicine

PathPresenter team

Matthew Hanna, MD
Memorial Sloan Kettering, New York, NY

Rajendra Singh, MD
Icahn School of Medicine at Mt Sinai

Krushnavadan Acharya, MCA
PathPresenter

Slides and Clinical

Roberto Salgado
Peter Mac Callum Cancer Center; GZA-ZBA Hospitals
International Working Group for TILs in Breast cancer

Denis Larismont
Institut Jules Bordet

Statistics

Si Wen
FDA/CDRH/OSEL/DIDSR

Manasi Sheth
FDA/CDRH/OPEQ/OCEA/Biostatistics

Chava Zibman
FDA/CDRH/OPEQ/OCEA/Biostatistics

Weijie Chen, PhD
FDA/CDRH/OSEL/DIDSR

Committee

Mohamed Amgad, MSc
Emory University School of Medicine, Atlanta, GA

Rajarsi Gupta, MD, PhD
Renaissance School of Medicine and Dept. of Biomedical Informatics, Stony Brook Medicine

Steven N. Hart, PhD
Mayo Clinic, Rochester, MN

Joe Lennerz, MD, PhD
Massachusetts General Hospital, Boston, MA

Richard Huang, MD, MS
Massachusetts General Hospital, Boston, MA

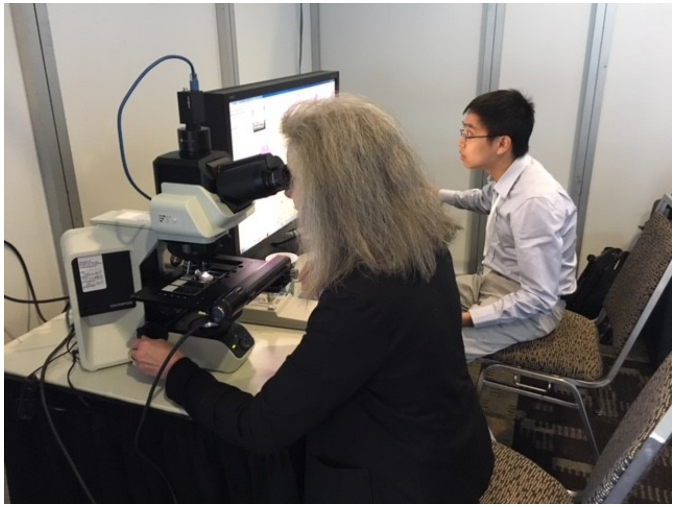
Anant Madabhushi, PhD
Case Western Reserve University

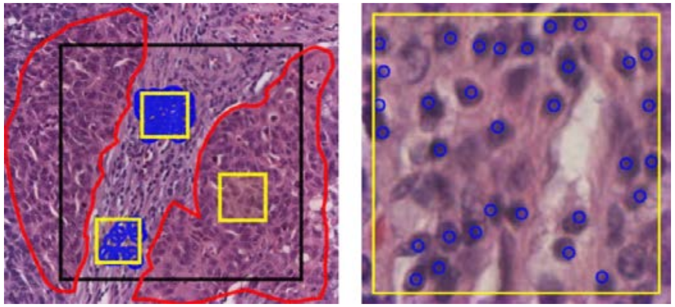
Kyle J. Myers, PhD
FDA/CDRH/OSEL/DIDSR

High-throughput truthing (HTT) Project

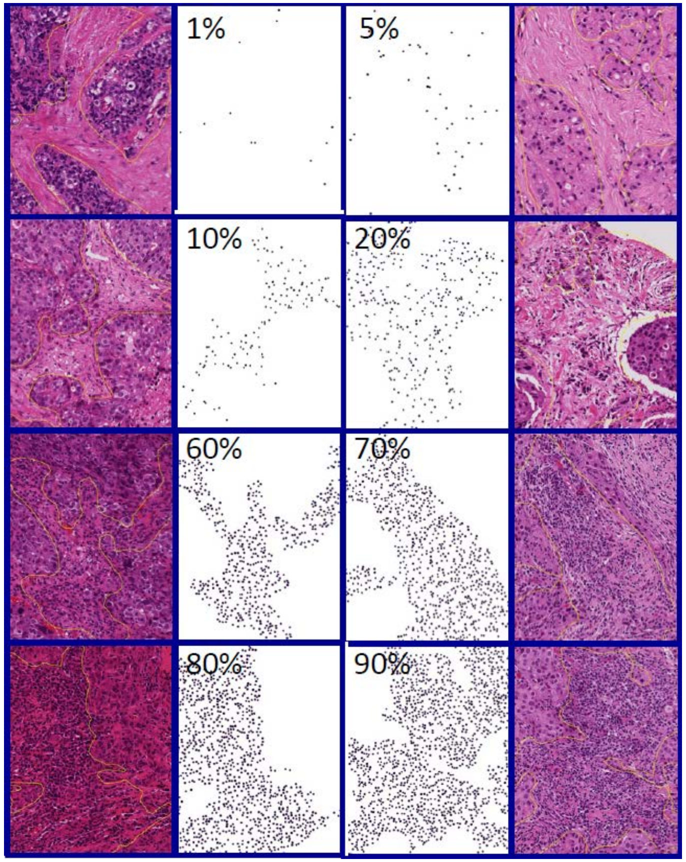
Demonstration project

- Collect multi-reader image annotations to establish biomarker truth
- Annotations support validation of an algorithm
- Pursue FDA qualification of a [Medical Device Development Tool]
- Application: Stromal Tumor Infiltrating Lymphocytes (sTILs) are prognostic in breast cancer



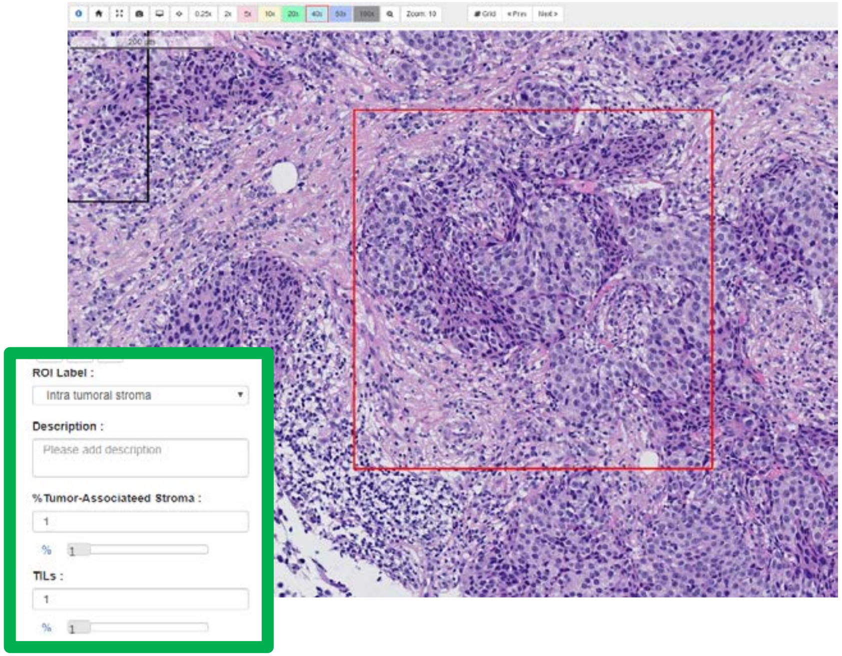


Standardized Annotations Yield a Biomarker



- Quantitative Biomarker
- Density of sTILs: 0-100

Standardized Annotations Yield a Biomarker



ROI Label : Intra tumoral stroma

Description : Please add description

%Tumor-Associateed Stroma : 1

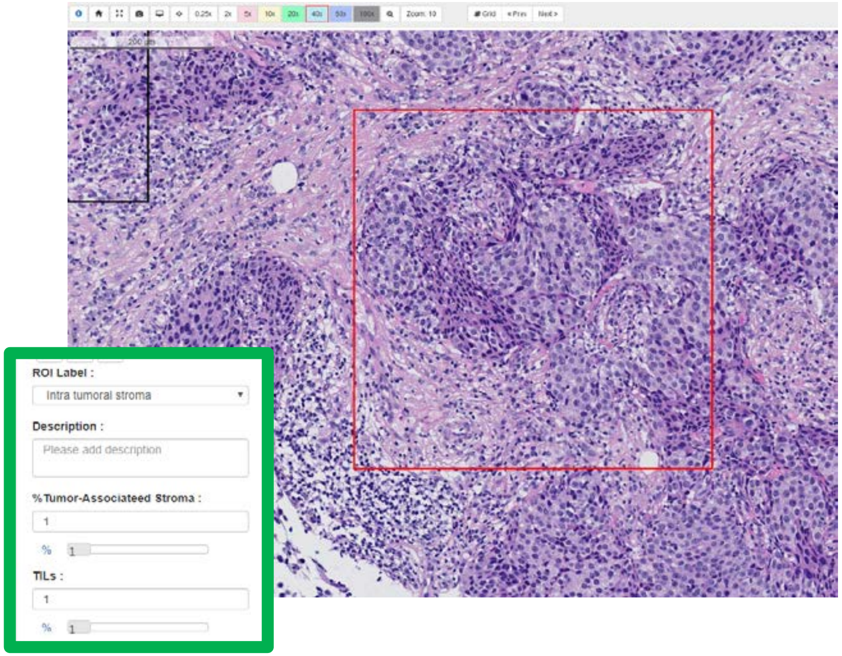
% 1

TILs : 1

% 1

- Pathologist
  - Takes time
  - Requires training
  - Noisy
  - Board Certification
- Algorithm
  - Fast
  - Requires training
  - Reproducible
  - Regulatory permission

Standardized Annotations Yield a Biomarker



ROI Label :

Intra tumoral stroma ▾

Description :

Please add description

%Tumor-Associateed Stroma :

1

% 1

TILs :

1

% 1

- Pathologist

Literature
Examples with truth
(feedback)

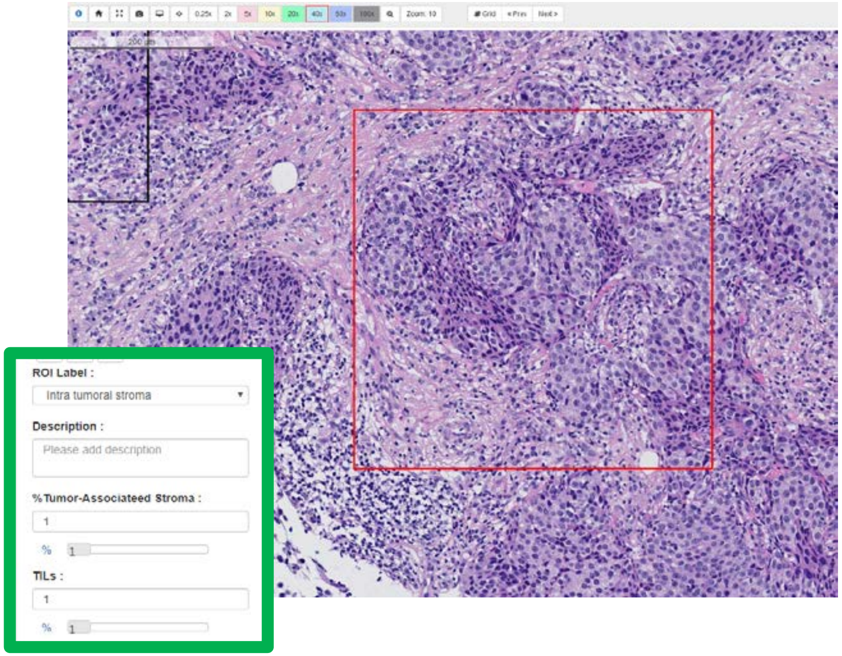
  - Takes time
  - Requires training
  - Noisy
  - Board Certification

- Algorithm

~~Literature~~
Examples with truth
(feedback)

  - Fast
  - Requires training
  - Reproducible
  - Regulatory permission

Standardized Annotations Yield a Biomarker



ROI Label :
Intra tumoral stroma ▾

Description :
Please add description

%Tumor-Associateed Stroma :
1
% 1

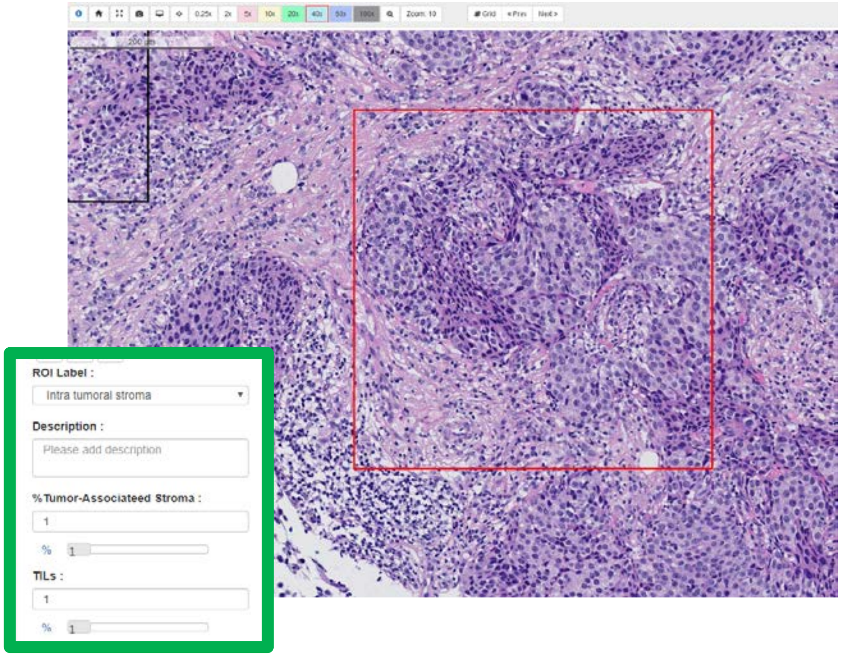
TILs :
1
% 1

- Pathologist
  - Takes time
  - Requires training
  - Noisy
  - Board Certification
- Algorithm
  - Fast
  - Requires training
  - Reproducible
  - Regulatory permission

<u>"Truth by pathologist"</u>
Reduce and Account for
Pathologist Variability

Evaluate performance
Requires truth

Standardized Annotations Yield a Biomarker



ROI Label :
Intra tumoral stroma

Description :
Please add description

%Tumor-Associateed Stroma :
1
% 1

TILs :
1
% 1

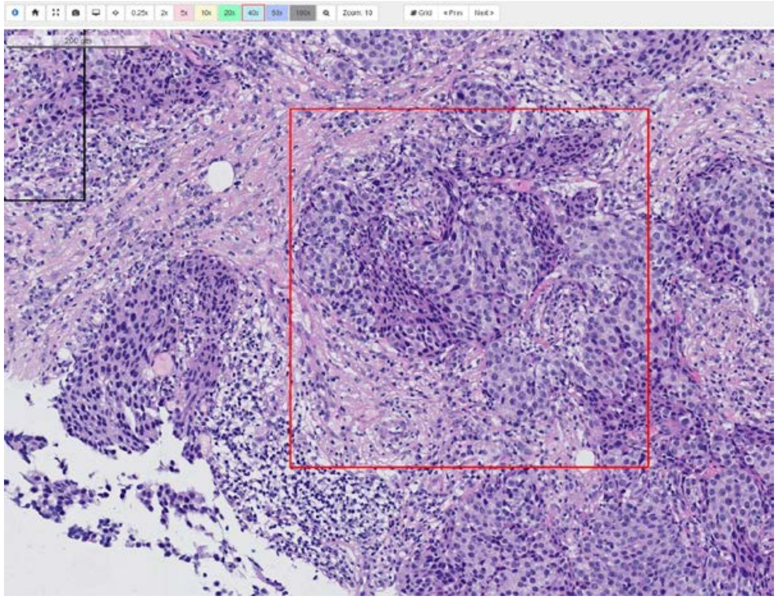
- Pathologist
  - Takes time
  - Requires training
  - Noisy
  - Board Certification
- Algorithm
  - Fast
  - Requires training
  - Reproducible
  - Regulatory permission

Evaluate performance
Requires truth

<u>“Truth by pathologist”</u>

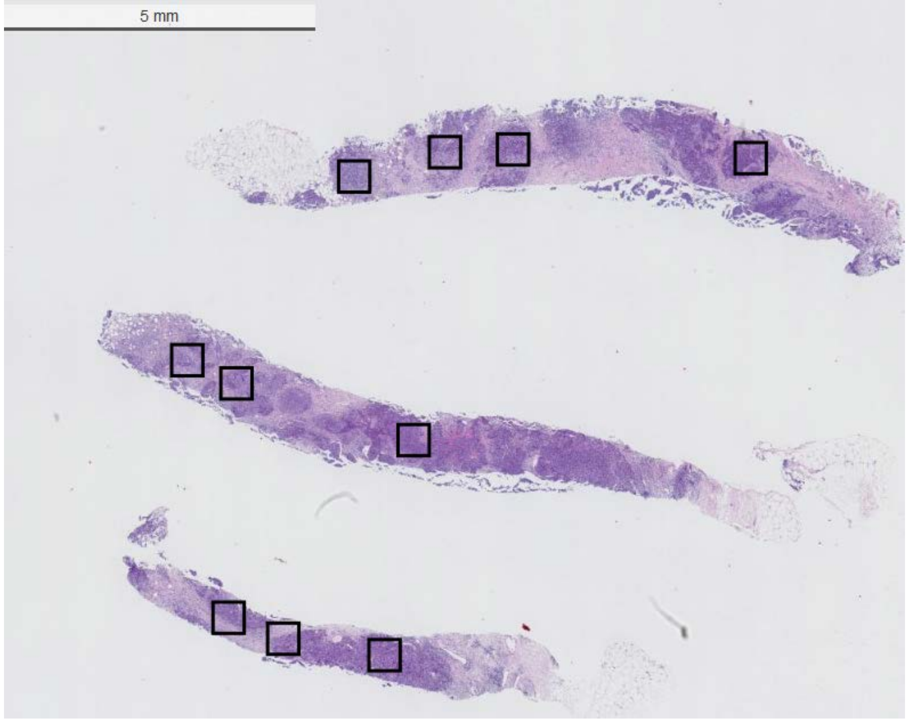
- Additional training
- Multiple pathologists per region / image
- Sophisticated analysis

Patch to Whole Slide Image



- Zoom Out

Whole Slide Images: Digital Scans of Glass slides

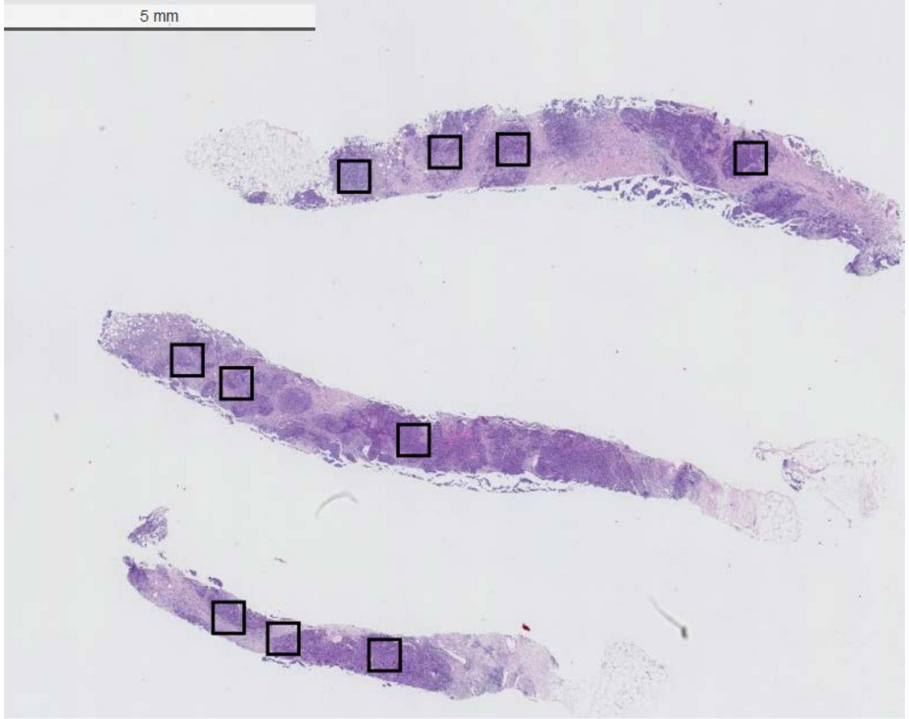


- Breast Cancer Biopsies
- Square Regions of Interest control the evaluation areas

<u>Current selection by pathologist:</u>

- Areas in tumor (~50%)
- Areas in tumor margin (~20%)
- Other (~30%)

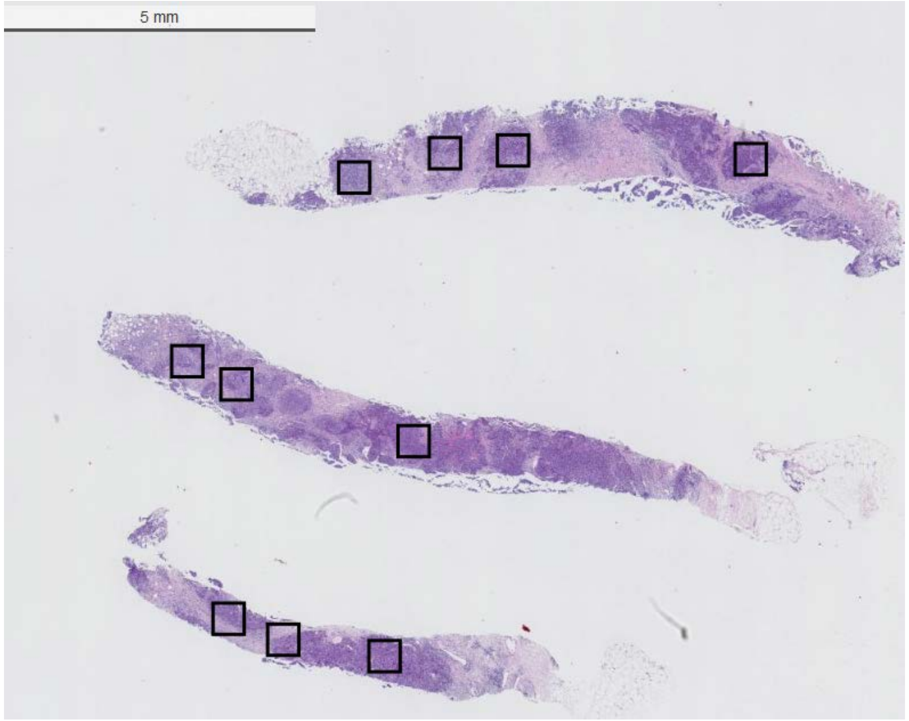
Whole Slide Images: Digital Scans of Glass slides



- Breast Cancer Biopsies
- Square Regions of Interest control the evaluation areas

Study to prepare the study.
Cover the range of scores.

Whole Slide Image to Patient



- Zoom Out

Patients

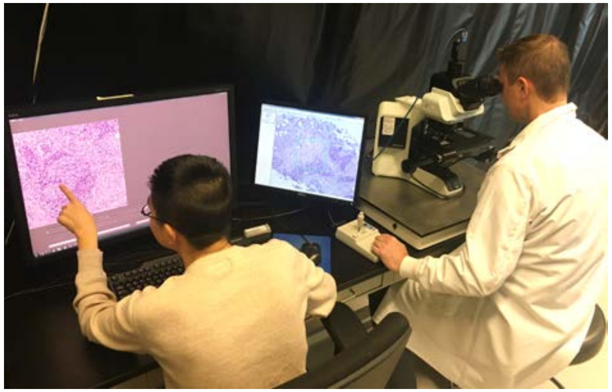
Subgroup Description		Planned for MDDT?
Age	<40 years old	Yes
	40-60 years old	Yes
	>60 years old	Yes
Breast Cancer Subtypes	Luminal A	Maybe
	Luminal B	Maybe
	Triple-negative	Yes
	HER2 positive	Maybe
	Normal-like	Maybe
Breast Cancer Stages	0	Yes
	I	Yes
	II	Yes
	III	Yes
	IV	Yes
Patients After Therapy	Therapy 1	No
	Therapy 2	No
	Therapy 3	No

- Define the patient population

TILs always look the same.
Background “context” looks different.

Update: Choices & Challenges

	Digital Modes		Microscope Mode
	PathPresenter	caMicroscope	eeDAP
nReaders	7	8	7
nObs at USCAP	850	300	440
nObs post USCAP	232	572	0
nObs Total	1082	872	440



Total Obs
2394

Data-collection test run

- Alliance Meeting
- USCAP Annual Meeting
- Feb. 28, 2020

Four workstations

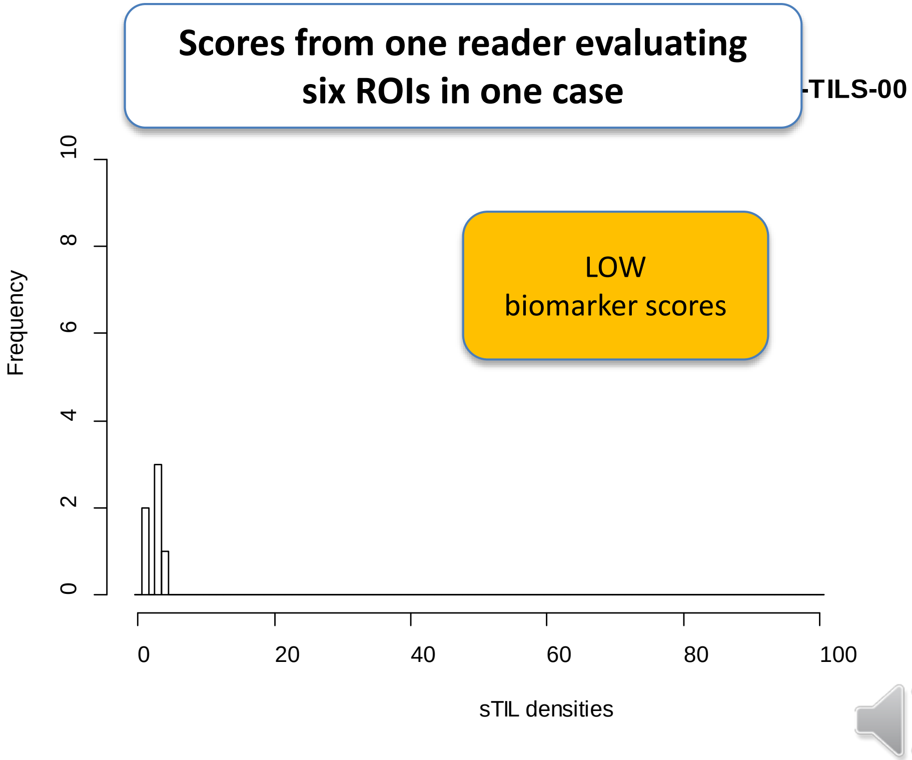
- 2 microscopes
- 2 digital platform

64 slides (balance sampling within and across specimens)

- 8 batches of 8 slides
- 10 ROIs per slide
- 30 minute sessions

What does the data look like?

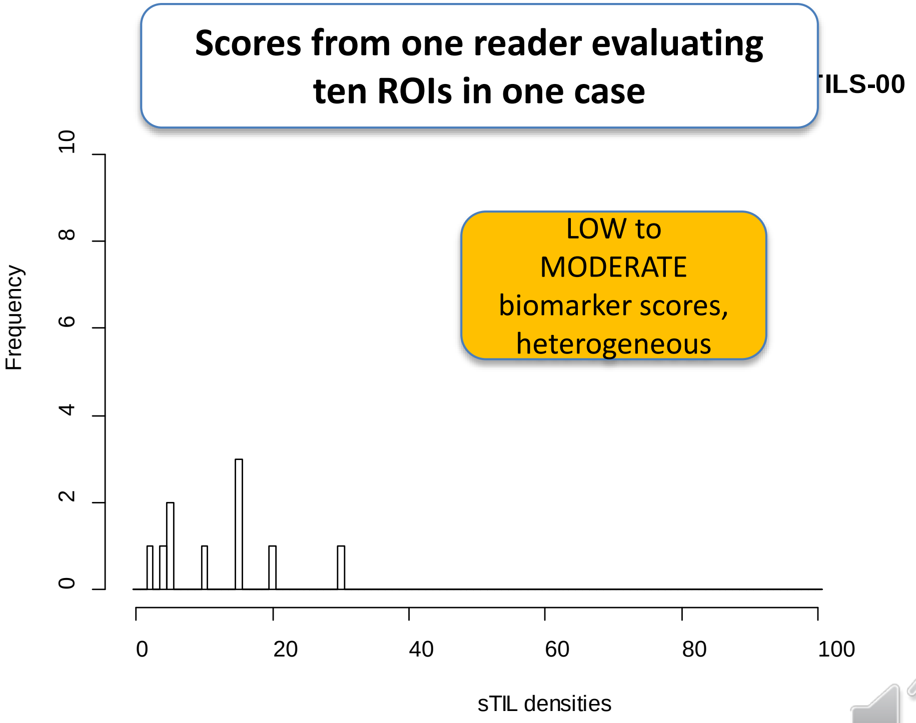
- Histogram of Biomarker Scores
- Many slides yield LOW biomarker scores


Scores from one reader evaluating six ROIs in one case

LOW biomarker scores

What does the data look like?

- Histogram of Biomarker Scores
- Many slides yield LOW biomarker scores
- Some slides yield LOW to MODERATE biomarker scores

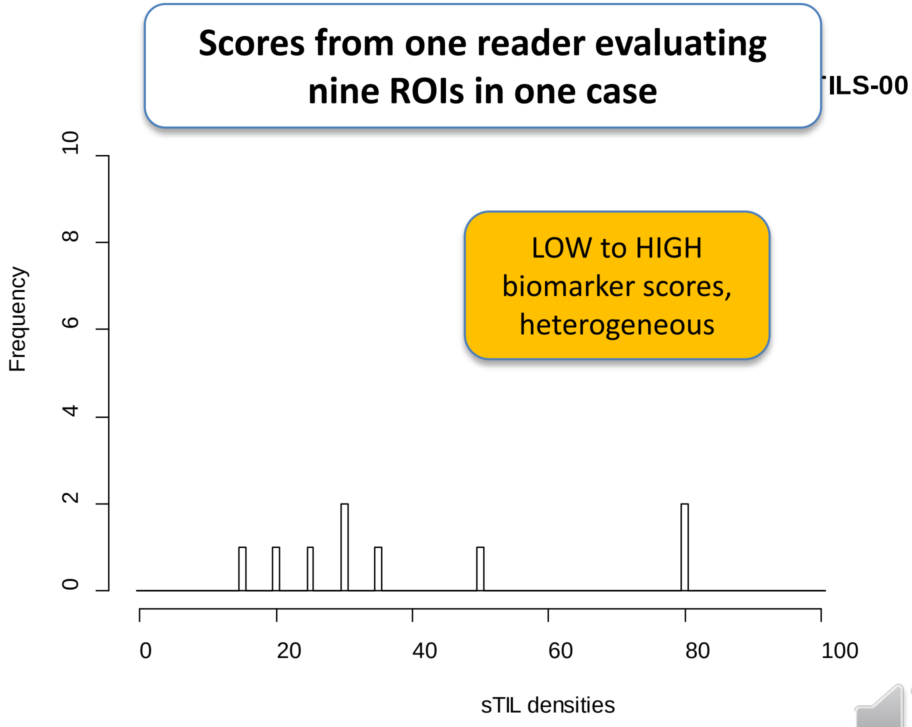


Scores from one reader evaluating ten ROIs in one case

LOW to MODERATE biomarker scores, heterogeneous

What does the data look like?

- Histogram of Biomarker Scores
- Many slides yield LOW biomarker scores
- Some slides yield LOW to MODERATE biomarker scores
- Some slides yield LOW to HIGH biomarker scores



Scores from one reader evaluating nine ROIs in one case

LOW to HIGH biomarker scores, heterogeneous

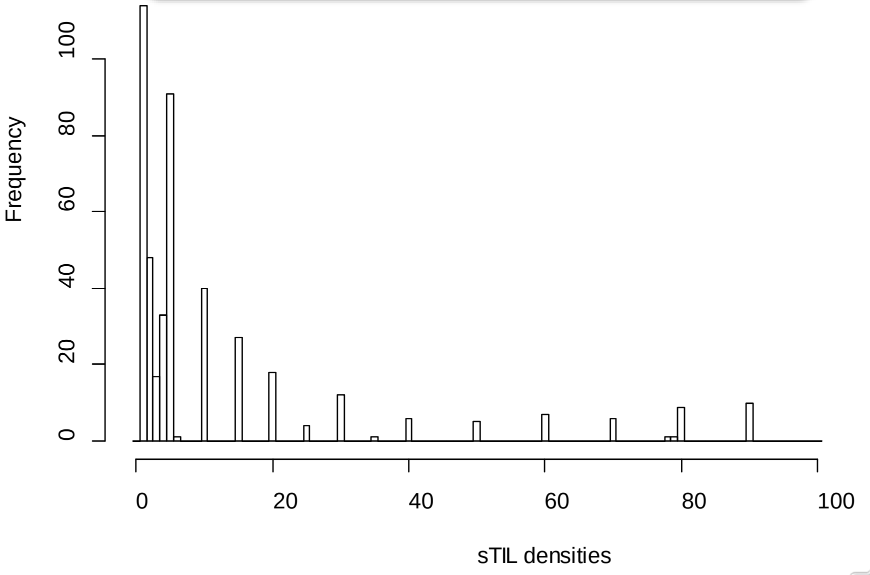
What does the data look like?

- Histogram of Biomarker Scores

- One reader
- All 64 slides
- 10 ROIs per slide

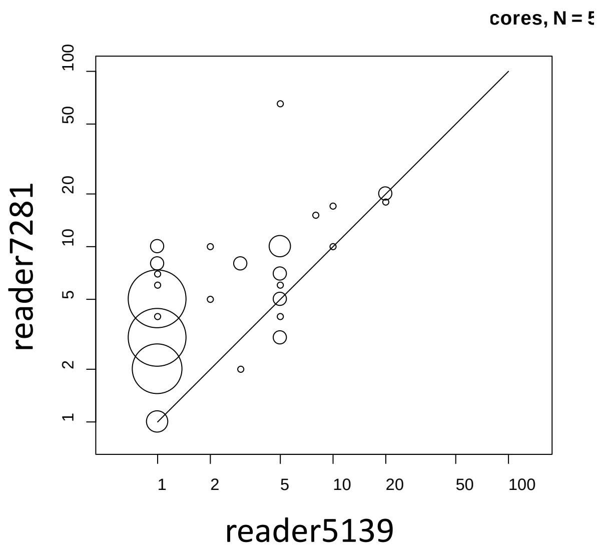
- Oversampling low scores

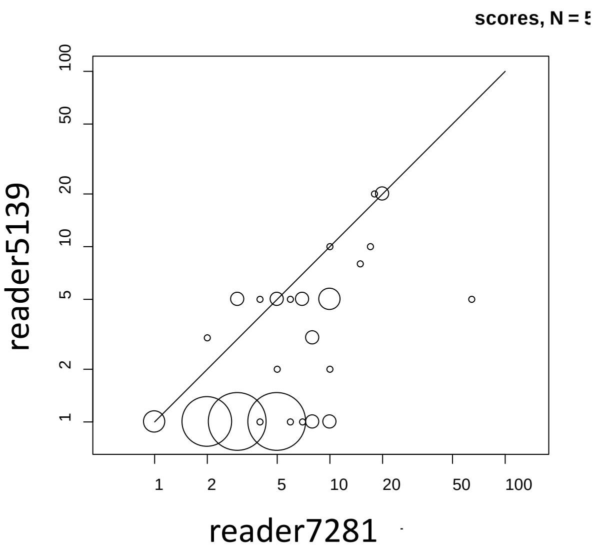
Scores from one reader evaluating all ROIs in 64 cases



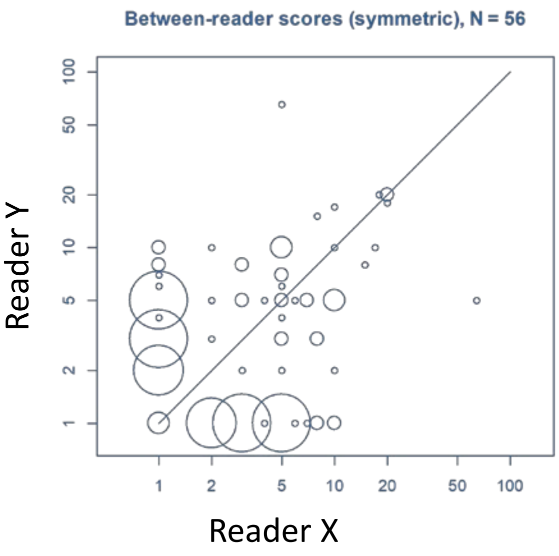
Agreement: Start with a scatter plot

- Two readers, batch001
- Plot axes scaled log base 10
- Circle size proportional with number of observations
- Flip reader7281 <-> reader5139 == Flip x <-> y



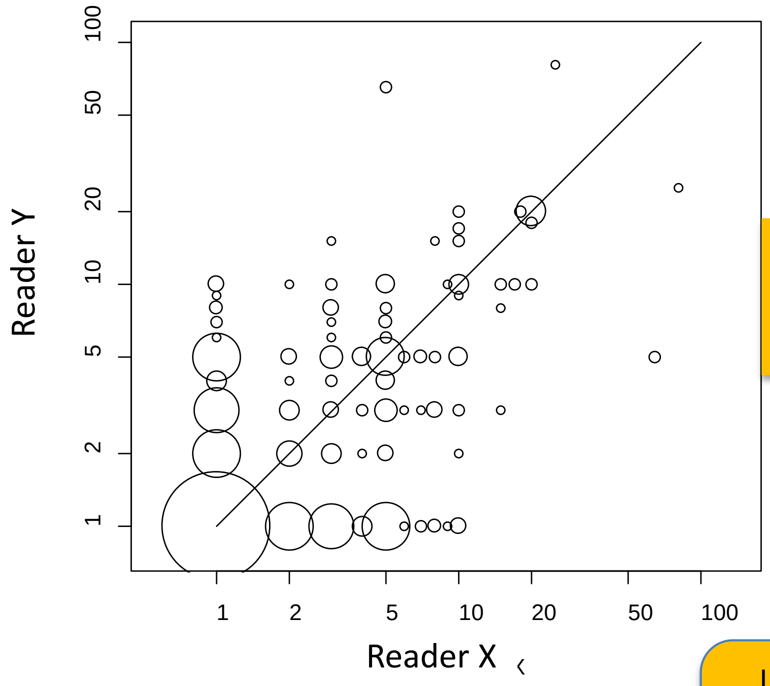


Combine to Symmetrize



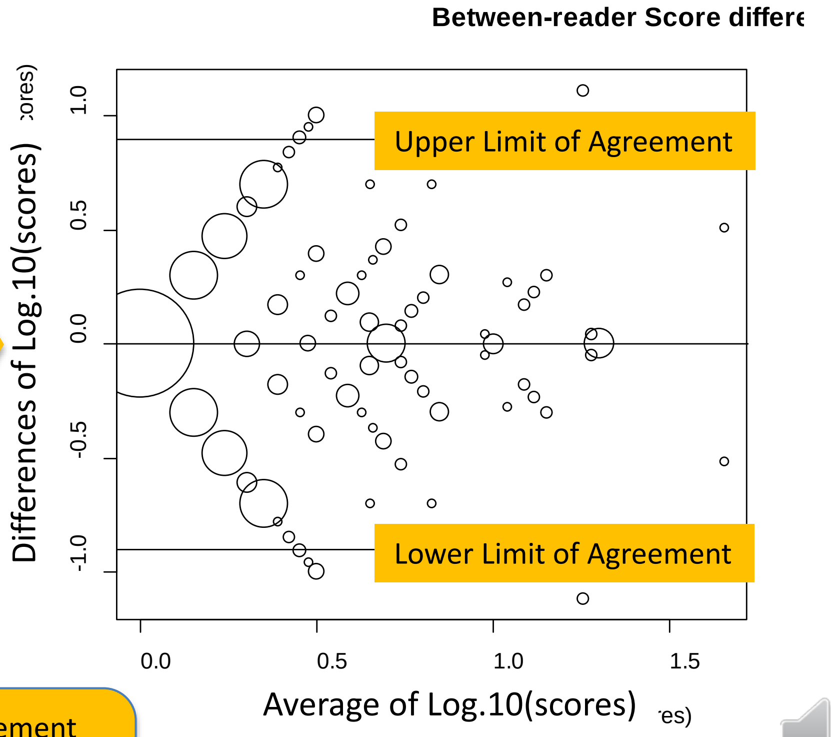
Agreement: Consider all pairs of readers

Score differences from 3 readers
Batch001, N = 346



Rotate
45°

Bland-Altman Plot



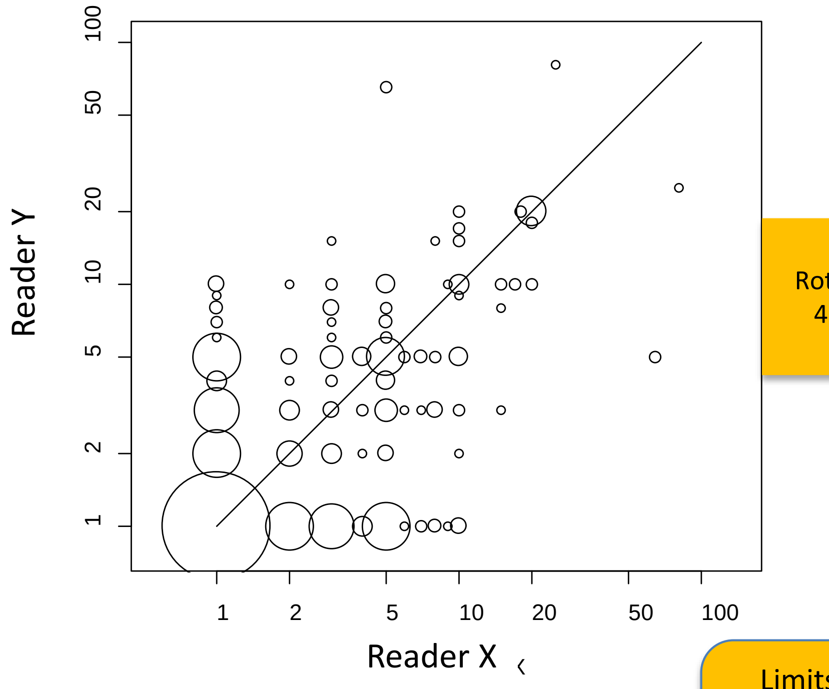
Upper Limit of Agreement

Lower Limit of Agreement

Limits of Agreement should account for reader and case variability

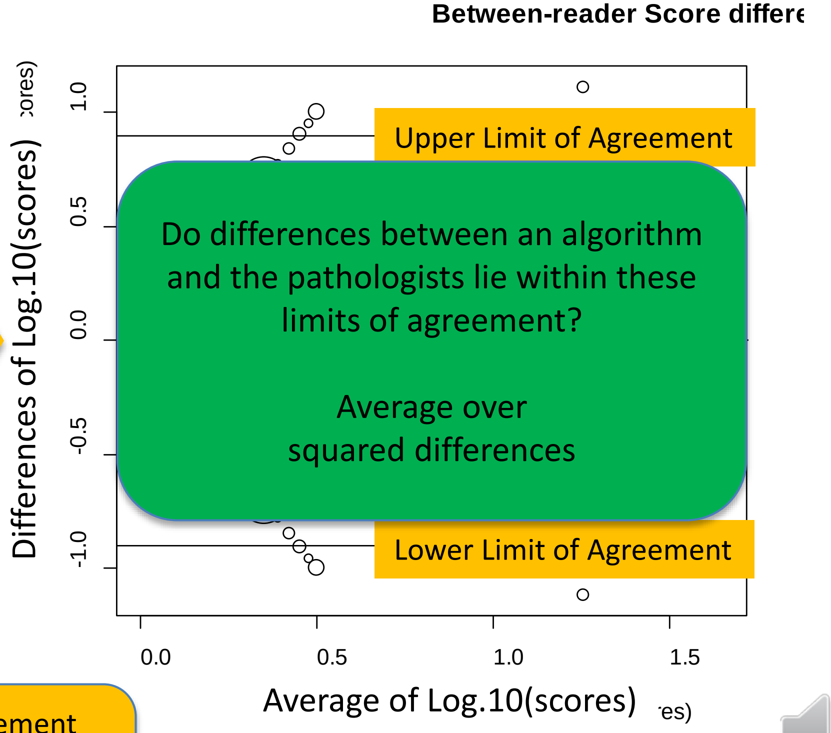
Agreement: Consider all pairs of readers

Score differences from 3 readers
Batch001, N = 346



Rotate 45°

Bland-Altman Plot



Upper Limit of Agreement

Do differences between an algorithm and the pathologists lie within these limits of agreement?

Average over squared differences

Lower Limit of Agreement

Limits of Agreement should account for reader and case variability

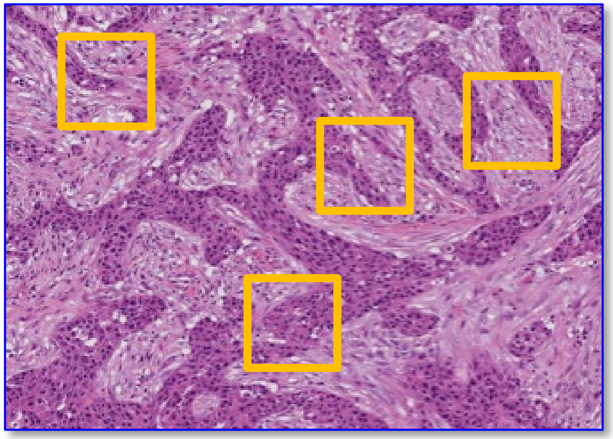
Summary and Future Work

Summary

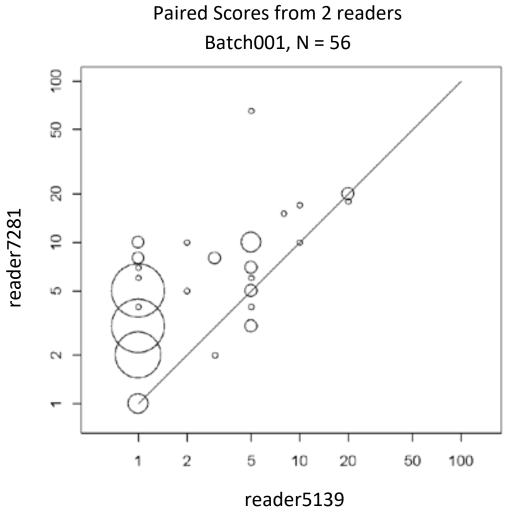
- MRMC variance of AUC framework allows study sizing
  - Variance components
  - Coefficients that correspond to experiment size
- Framework (and simulation) allow study of tradeoffs
  - Resources (Number of readers, cases, and observations)
  - Statistical efficiency
- Split-plot studies are less burdensome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

Future (Current) Work
(to support the HTT project)

- Cluster / Nested Data
  - Multiple regions per case
  - Building simulation



- Quantitative Measurements
  - Between-reader agreement
  - Within-reader agreement
  - Algorithm-reader agreement
  - Generalizing MRMC methods and simulation
  - Correlation, Mean-squared error


Paired Scores from 2 readers
Batch001, N = 56
reader7281
reader5139