

# Resources for Reader Studies

Brandon D. Gallas, PhD

Division of Imaging, Diagnostics, and Software Reliability

OSEL/CDRH/FDA

# Interesting Quote

- Editors of *Radiology*:
  - The audience of their journal is interested in
  - the variability between observers which
  - “*requires a sufficiently high number of observers.*”

Bankier, A. A.; Levine, D.; Halpern, E. F. & Kressel, H. Y. (2010),  
'Consensus interpretation in imaging research: is there a better way?'  
*Radiology*, **257**, (1), 14--17.

# Interesting Quote

- American College of Radiology Imaging Network (ACRIN)
  - There is a great deal of variability among readers.
  - Reader variability “*must be accounted for in the design of ACRIN trials.*”

Hillman, B. J. (2005),  
'ACRIN—lessons learned in conducting multi-center trials of imaging and cancer.'  
*Cancer Imaging*, **5 Spec No A**, S97--101.

# Phases of Evaluation

- Exploratory
    - Early
    - Pilot
  - Intermediate
    - Challenge or Stress Test
    - Lab-based
  - Advanced
    - Late
    - Clinical use
- Phase determines
    - Aims
    - Scope
    - Size
    - Analysis
    - Conclusions
  - See my workshop summary paper in Academic Radiology

Gallas, Chan, D'Orsi, Dodd, Giger, Gur, Krupinski, Metz, Myers, Obuchowski, Sahiner, Toledano, Zuley (2012), 'Evaluating Imaging and Computer-aided Detection and Diagnosis Devices at the FDA.' *Acad Radiol*, **19**, (4), 463-477.

# Outline

- Data Collection and Reader Training
- Statistical Analysis and Sizing
- Split-Plot Study Design

# Considerations

## Data Collection

- Study task is distilled version of clinical task
  - No patient information
  - Limited decision options
  - Quantitative, ready for analysis
- Test and streamline workflow

## Reader Training

- Provide Precise Written Instructions
- Discuss Instructions Face-to-Face
- Observe and assist training cases
- Cost is minimal compared to
  - Sourcing images
  - Reading time
  - Total effort

# Resources

## Data Collection

- eeDAP: Evaluation Environment for Digital and Analog Pathology
  - Retooling: integrate other image formats
  - Does not require images
  - Migrating to cloud platform
  - Can tweak to your needs
  - <https://code.google.com/p/eedap/>

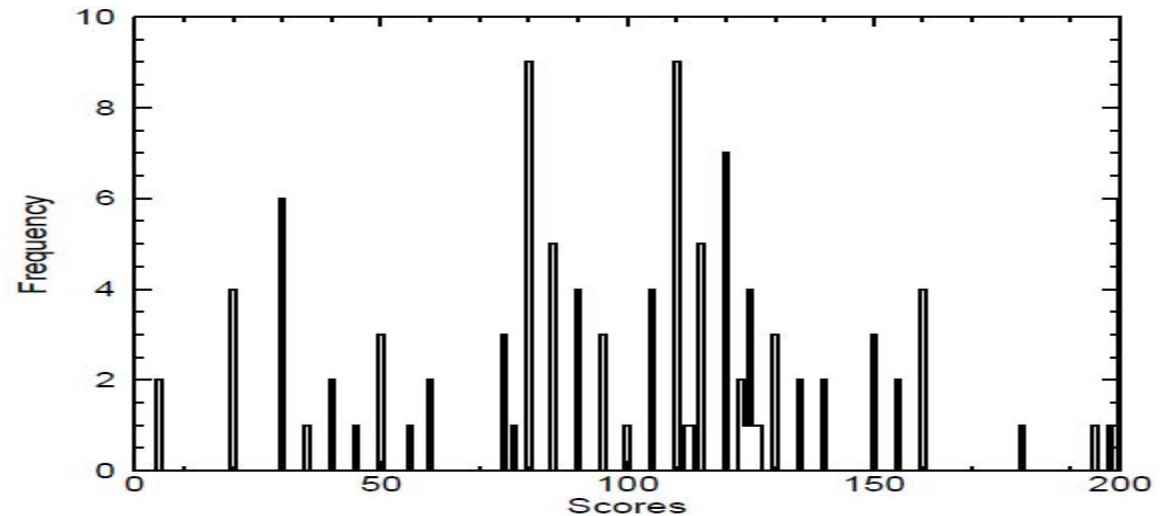
## Reader Training

- Instructions for ROC scoring
- Novel ROC data collection workflow (200 point scale)
- iMRMC website
  - <https://code.google.com/p/imrmc/wiki/iMRMCGuide>

## SFM

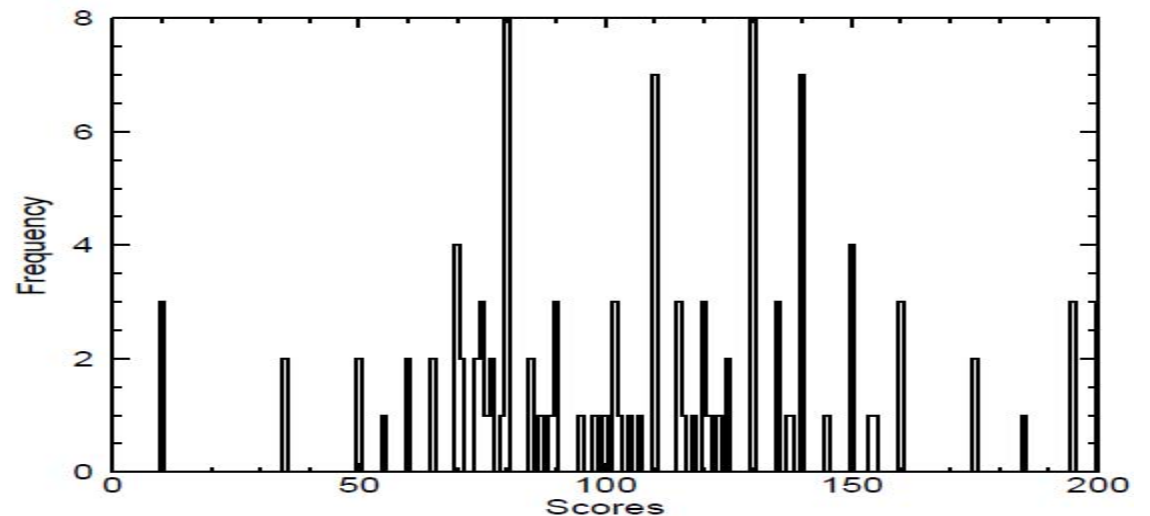
nCases = 109  
nBinsUsed = 37

Reader 02, Screening study (prevalence = 0.31)



## FFDM

nCases = 109  
nBinsUsed = 47

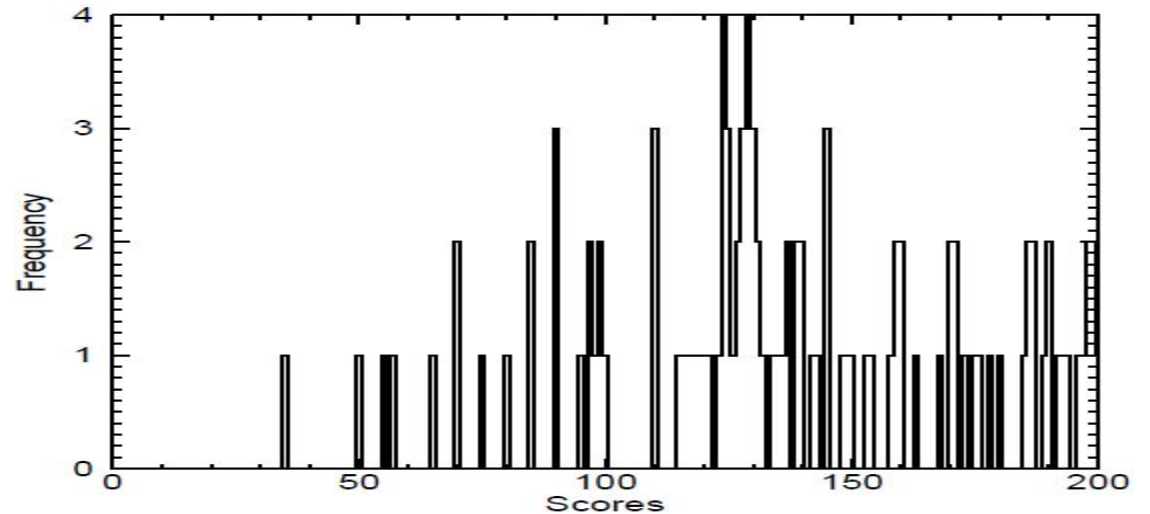




## SFM

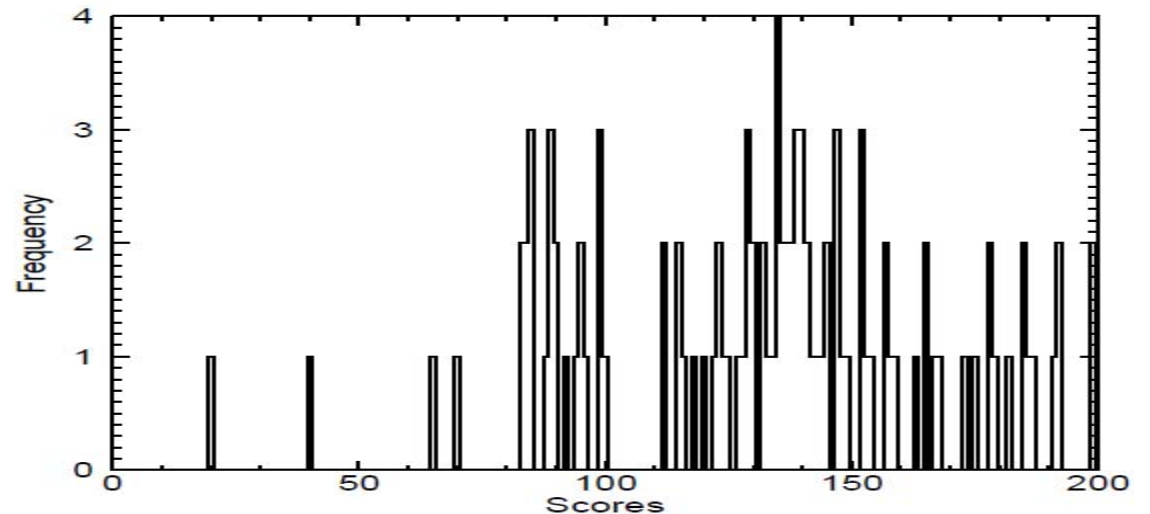
nCases = 108  
nBinsUsed = 72

Reader 03, Screening study (prevalence = 0.31)



## FFDM

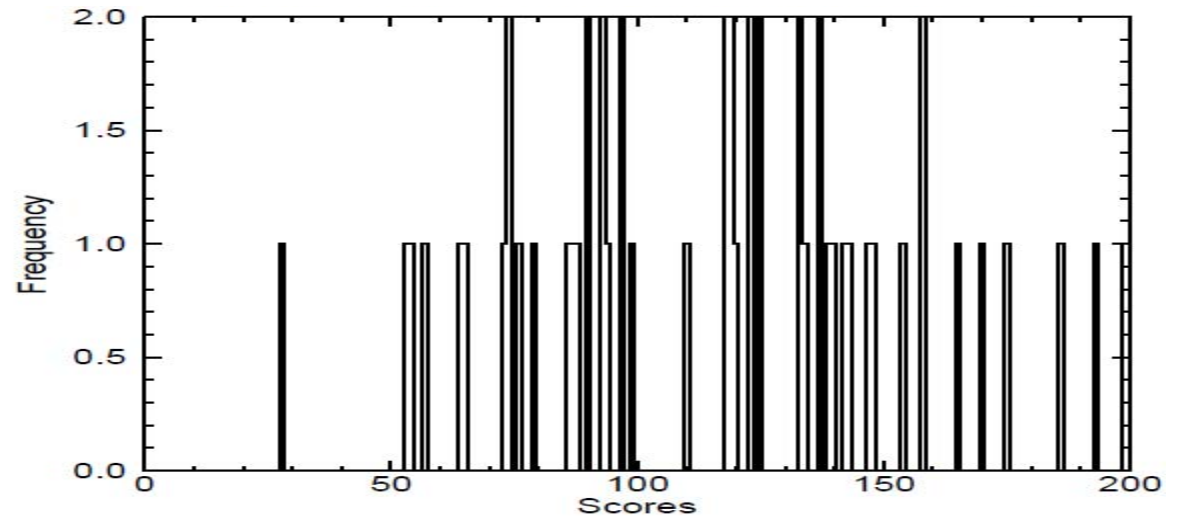
nCases = 106  
nBinsUsed = 67



## SFM

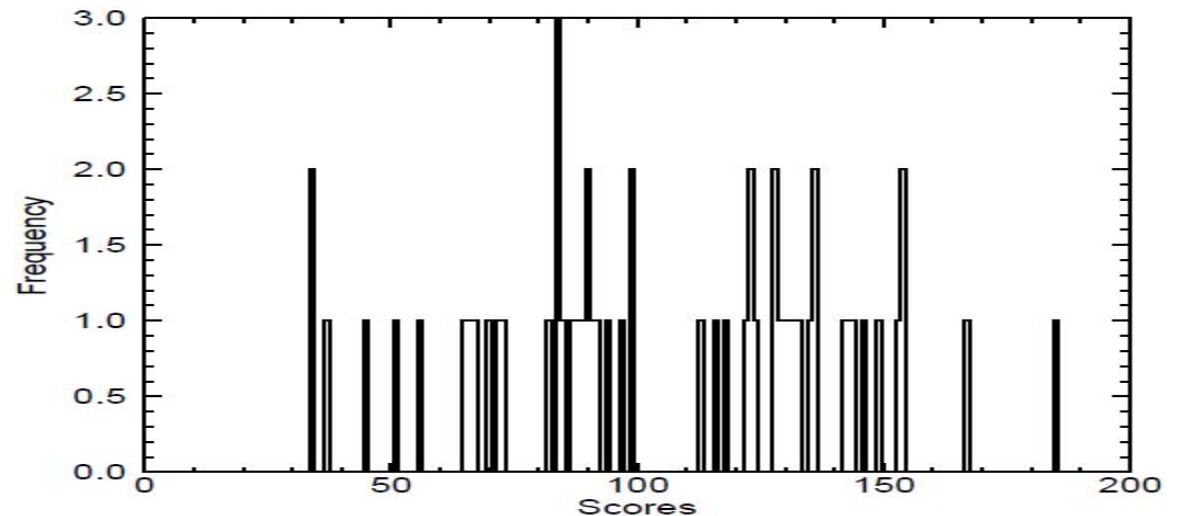
nCases = 53  
nBinsUsed = 42

Reader 04, Screening study (prevalence = 0.31)



## FFDM

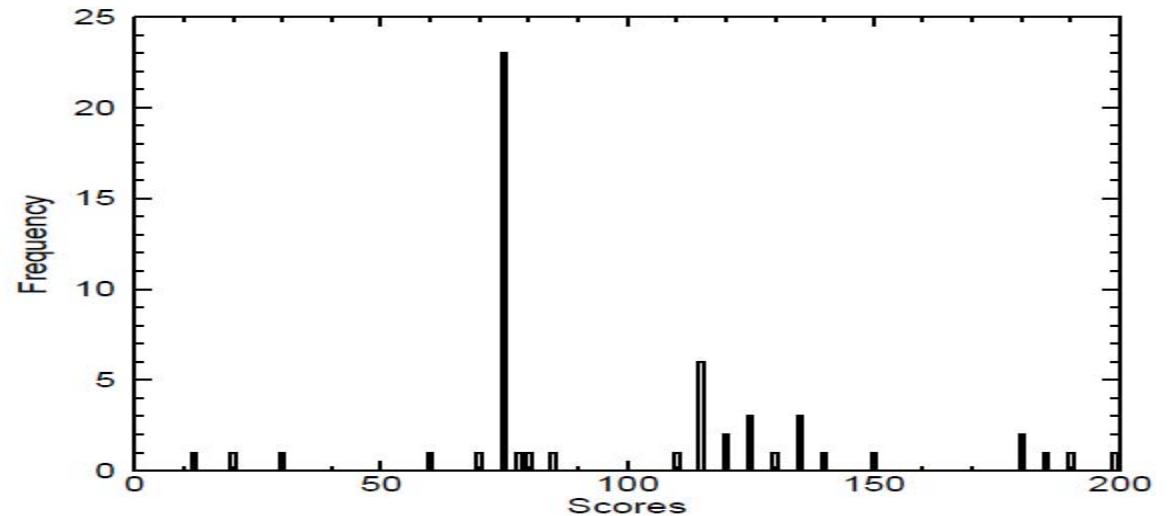
nCases = 55  
nBinsUsed = 46



## SFM

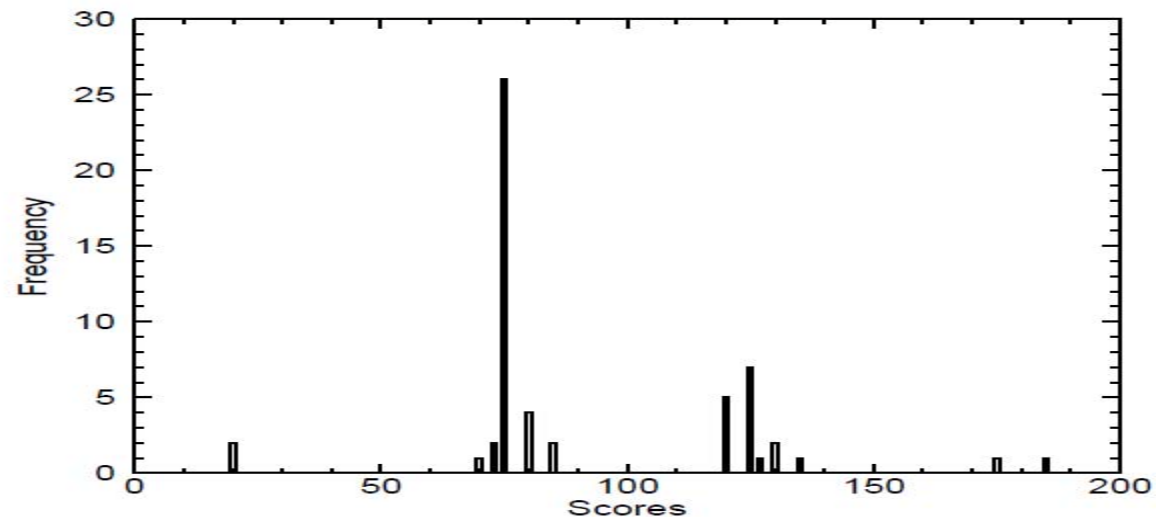
nCases = 54  
nBinsUsed = 21

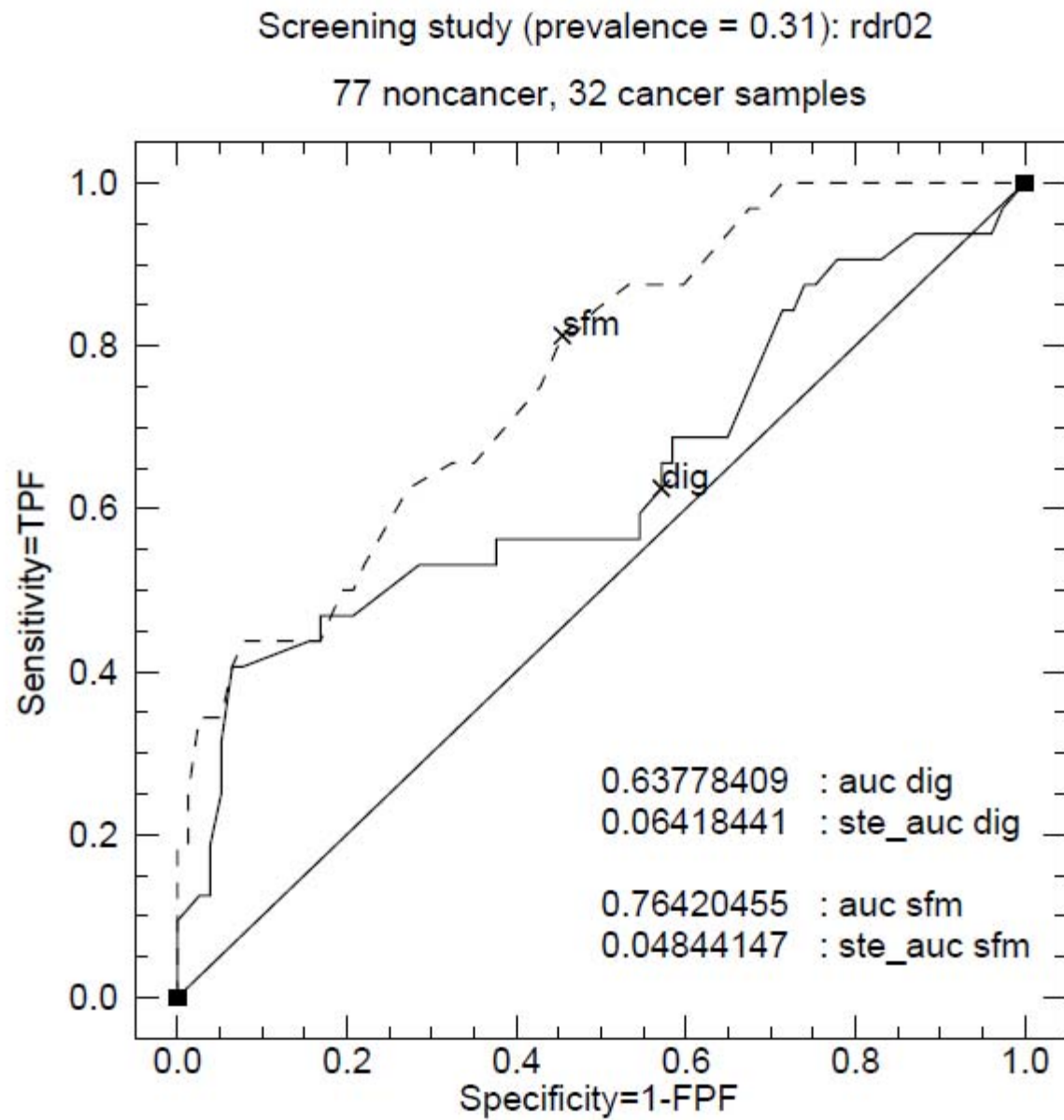
## Reader 01, Screening study (prevalence = 0.31)



## FFDM

nCases = 55  
nBinsUsed = 13

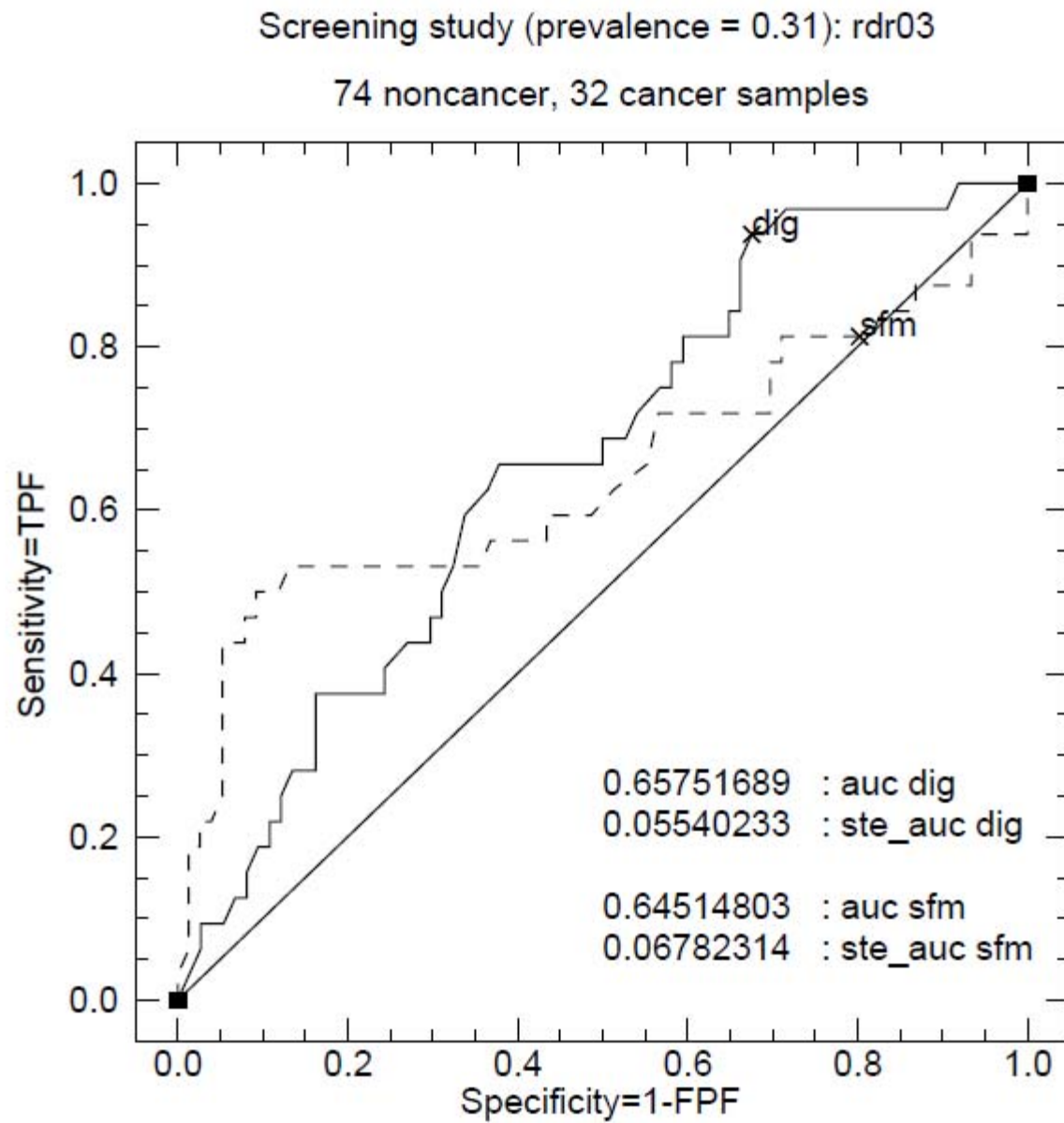




FDA VIPER data: Validation of Imaging Premarket Evaluation and Regulation

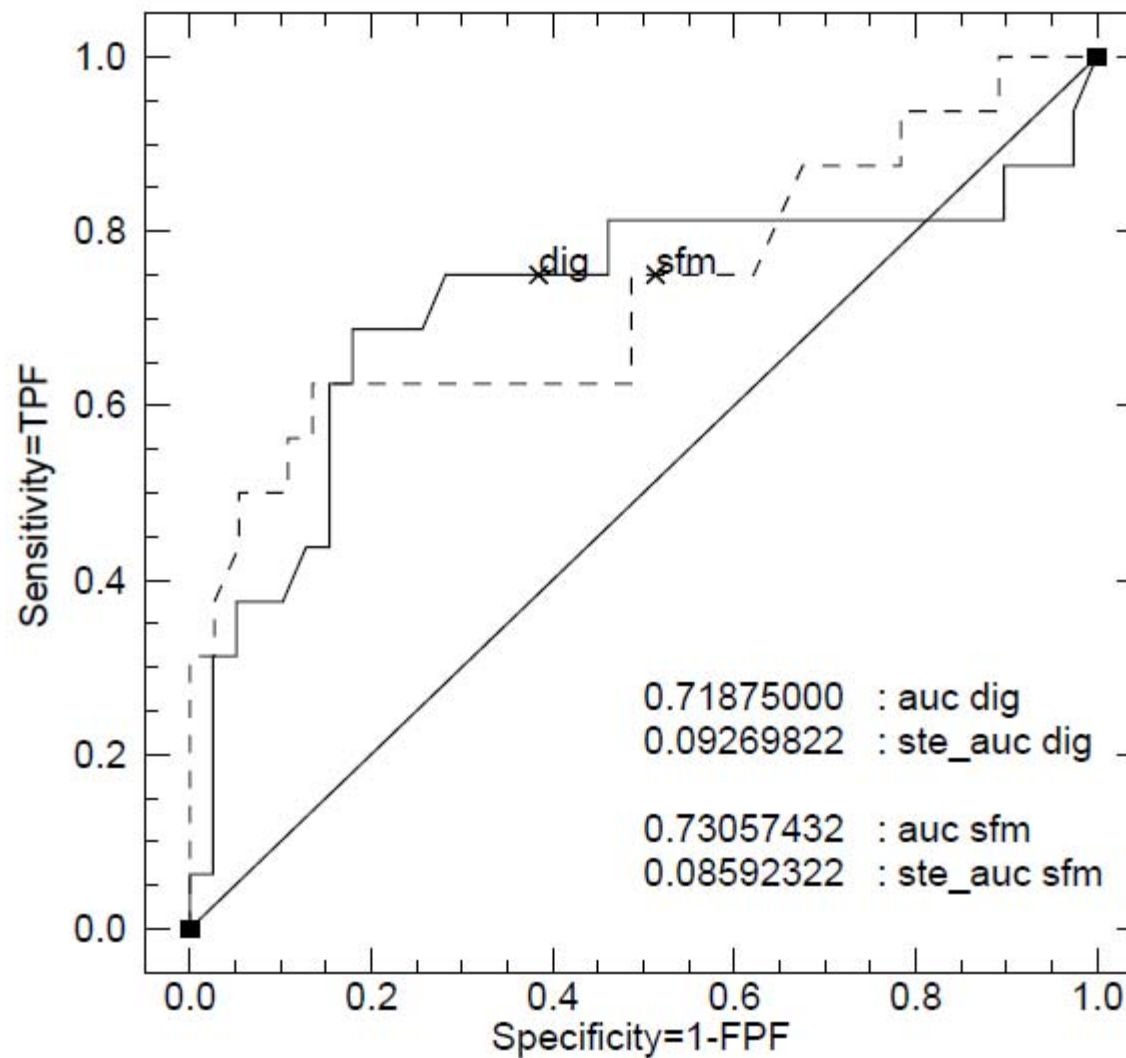
2/22/2015

SPIE MI 2015, workshop



Screening study (prevalence = 0.31): rdr04

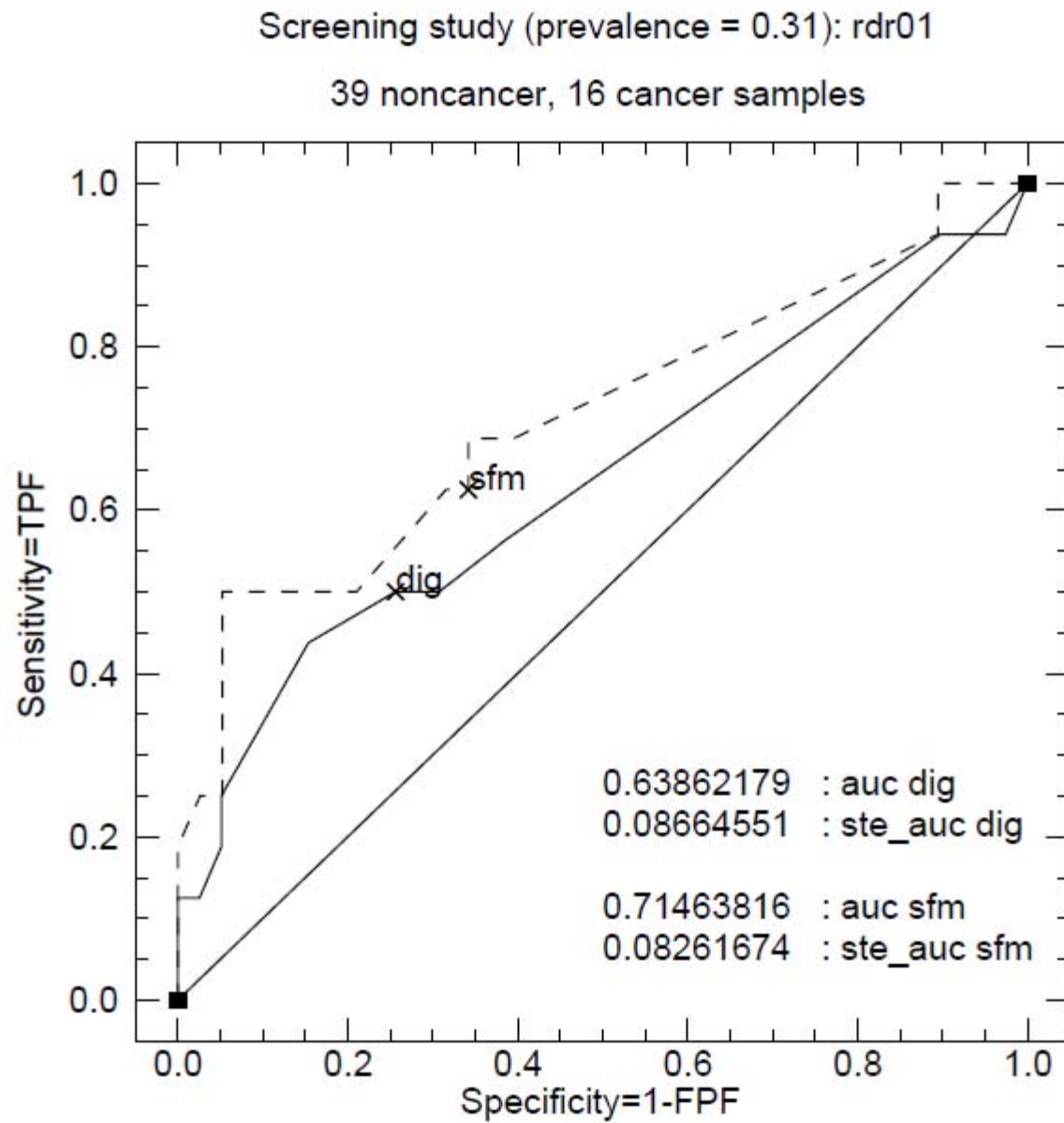
39 noncancer, 16 cancer samples



FDA VIPER data: Validation of Imaging Premarket Evaluation and Regulation

2/22/2015

SPIE MI 2015, workshop



# “MRMC” Statistical Analysis

Accounts for **M**ultiple **R**eaders, **M**ultiple **C**ases

- Components-of-variance

Variance of Difference in Endpoints →

$$V = \frac{\sigma_r^2}{R} + \frac{\sigma_c^2}{N} + \frac{\sigma_{rc}^2}{RN}$$

← internal noise, reader jitter

↑ reader    ↑ case    ↑ reader-case

The diagram shows the formula for the variance of the difference in endpoints, V. The formula is presented in a yellow box: V = (sigma\_r^2 / R) + (sigma\_c^2 / N) + (sigma\_rc^2 / RN). To the left of the box, the text 'Variance of Difference in Endpoints' is followed by a purple arrow pointing to the V. To the right, a purple arrow points from the box to the text 'internal noise, reader jitter'. Below the box, three purple arrows point upwards to the denominators R, N, and RN, which are labeled 'reader', 'case', and 'reader-case' respectively.

- Variance estimation is first step for
  - Confidence Intervals
  - Hypothesis Testing



# MRMC Statistical Analysis Tools

- University of Chicago
  - <http://metz-roc.uchicago.edu/>
- University of Iowa
  - <http://perception.radiology.uiowa.edu/Home/tabid/87/Default.aspx>
- FDA: iMRMC
  - <https://code.google.com/p/imrmc/>

MRMC Analysis of the Area Under the ROC curve (AUC)

- + Alternate study designs
- + Sizing ROC studies given a pilot study
- + Simulate ROC studies
- + Can treat binary performance

# iMRMC Demo?

<https://code.google.com/p/imrmc/>

Help and Info

Select an input method:

☐ MLE (avoid negatives)

H0: AUC\_A - AUC\_B = 0.00, two-sided alternative, 95% significance, 5 Readers, 69 Normal cases, 45 Disease cases.  
AUC\_A = 0.897, AUC\_B = 0.941, AUC\_A - AUC\_B = -0.044, sqrt(total var) = 2.067E-2, T Statistic = 2.119E0  
T-stat df(Normal Approx) = ∞ p-Value = 0.0341 Conf. Int. = (-0.0843, -0.0033) Reject Null? = 1.0000  
df(BDG) = 12.81 p-Value = 0.0556 Conf. Int. = (-0.0888, 0.0012) Reject Null? = 0.0000  
df(Hillis 2008) = 15.03 p-Value = 0.0512 Conf. Int. = (-0.0879, 0.0003) Reject Null? = 0.0000

BDG BCK DBM OR MS

	M1	M2	M3	M4	M5	M6	M7	M8
comp M0	8.66715E-1	8.45049E-1	8.15017E-1	8.06144E-1	8.26498E-1	8.24314E-1	8.04694E-1	8.03582E-1
coeff M0	6.44122E-5	4.38003E-3	2.83414E-3	1.92721E-1	2.57649E-4	1.75201E-2	1.13366E-2	-2.29114E-1
comp M1	9.23205E-1	9.09305E-1	8.89107E-1	8.85639E-1	8.95853E-1	8.93113E-1	8.85659E-1	8.84713E-1
coeff M1	6.44122E-5	4.38003E-3	2.83414E-3	1.92721E-1	2.57649E-4	1.75201E-2	1.13366E-2	-2.29114E-1
comp product	8.61643E-1	8.59748E-1	8.45782E-1	8.44938E-1	8.54915E-1	8.53633E-1	8.43955E-1	8.43402E-1
- coeff product	1.28824E-4	8.76006E-3	5.66828E-3	3.85443E-1	5.15298E-4	3.50403E-2	2.26731E-2	-4.58229E-1
total	4.29207E-6	1.52679E-4	3.55958E-5	3.67457E-4	3.22580E-6	1.78023E-4	2.76859E-5	-3.41645E-4

total var=4.273E-4

Significance level  Effect Size  #Reader  #Normal  #Diseased

Sizing Analysis: SqrtVar=1.781E-2, Stat=2.808E0  
Normal Approx: df= ∞, Power= 0.80  
BDG: df= 31.71, Lambda= 7.88, Power= 0.78  
Hillis 2011: df= 74.51, Lambda= 7.88, Power= 0.80

Data Input

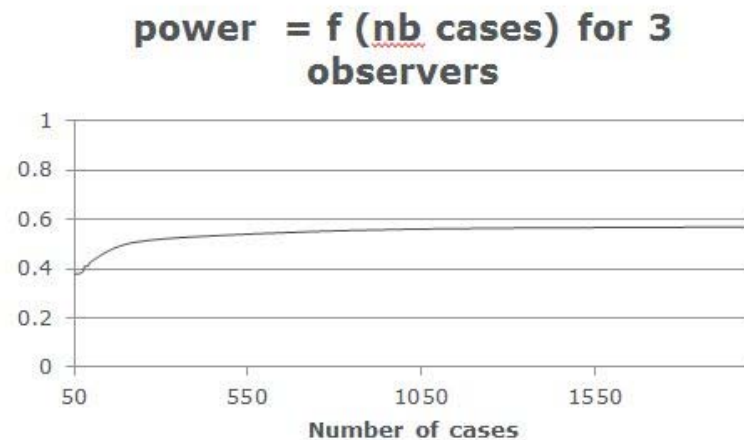
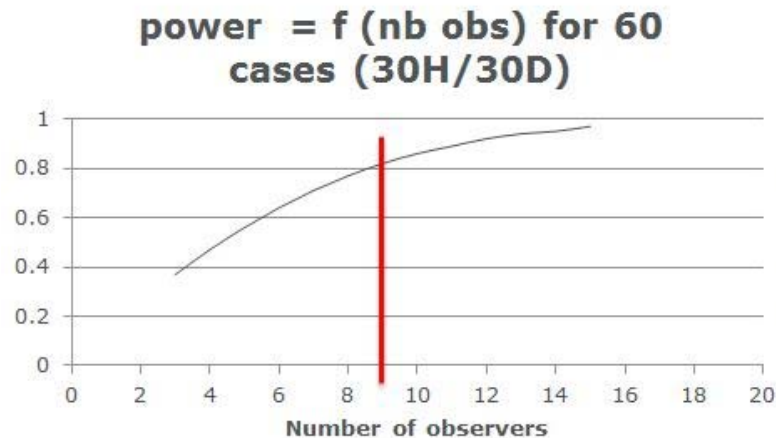
Data Analysis

Size a Study

# Statistical Analysis and Sizing

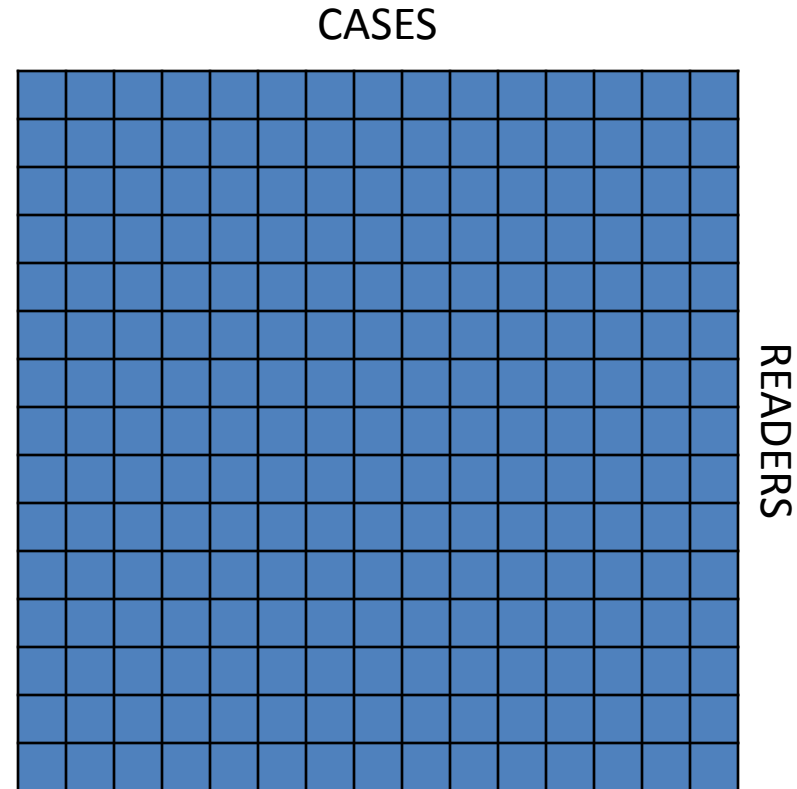
- Sizing results example

number of cases	60	60	70	70	80	80	100	100	120	120
number of observers	8	10	8	10	8	10	8	10	8	10
Power	0.77	0.86	0.79	0.88	0.8	0.89	0.84	0.91	0.86	0.93



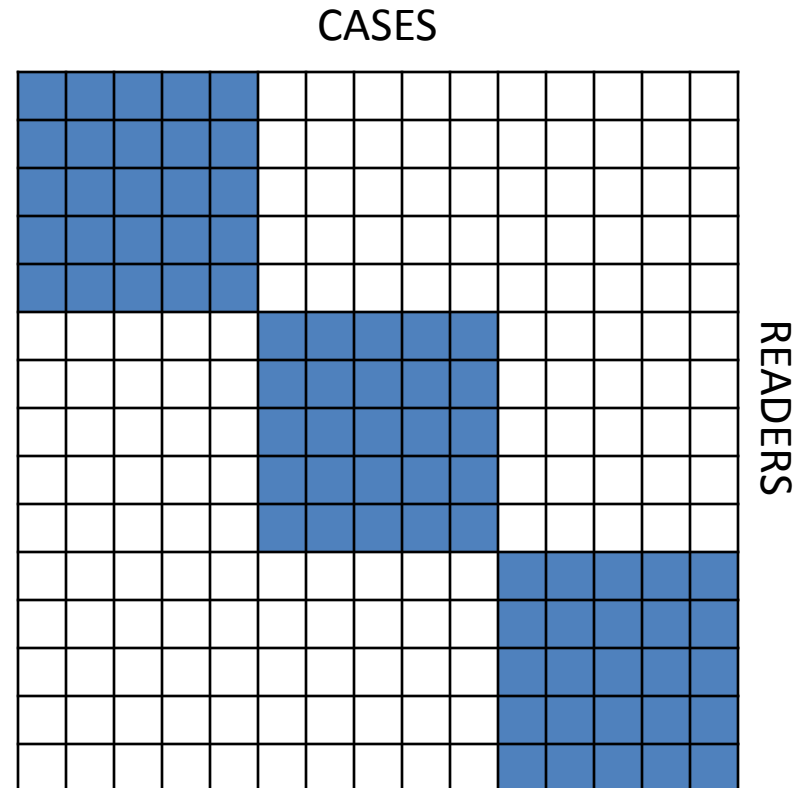
# Fully-crossed study design

- Every column indicates a case
- Every row indicates a reader
- Fully-crossed study
  - Every reader reads every case
- Takes a lot of reading time per reader and total
- Lots of redundant information



# Fully-crossed study design

- Every column indicates a case
- Every row indicates a reader
- Split-plot study
  - Reader only read cases in their group
- BIG REDUCTION IN READING TIME
- MINIMAL IMPACT ON STATISTICAL PRECISION



Obuchowski, N.; Gallas, B. D. & Hillis, S. L. (2012),  
'Multi-Reader ROC Studies with Split-Plot Designs: A Comparison of Statistical Methods.'  
*Acad Radiol*, **19**, (12), 1508-1517.

# Split-plot study design

Simulation example where variances are equal

## Fully-crossed

- 120 cases
- 6 readers
- 1440 evaluations total
- 240 evaluations/reader

- Fewer readers to recruit

## Split-plot (3 groups)

- 120 cases
- 9 readers
- 720 evaluations total
- 80 evaluations/reader

- 50% Fewer evaluations overall
- 33% Fewer evaluations per reader

# Closing Remarks

- Reader training can determine study success
  - Resources provided
- MRMC analysis tools are available
  - We want to help: tutorials, features, workflows
- Split-plot study design is revolutionary
  - More bang for the buck
  - VIPER study: 5 study conditions for price of 2
  - Digital Pathology: 3 reading sessions instead of 6





# Phases of Evaluation

- Exploratory
    - Early
    - Pilot
  - Intermediate
    - Challenge or Stress Test
    - Lab-based
  - Advanced
    - Late
    - Clinical use
- Ideal and controlled setting
    - Narrow task
    - Study cases, study readers
  - Hypothesis:  
Superiority or Non-Inferiority
    - Compare technologies
  - Analysis generalizes to
    - Population of Cases
    - Population of Readers
  - Not trying to estimate real-world clinical performance

# Phases of Evaluation

- Exploratory
  - Early
  - Pilot
- Intermediate
  - Challenge or Stress Test
  - Lab-based
- Advanced
  - Late
  - Clinical use

- Exploratory
  - Size: Small
  - Scope: Narrow
  - Results: Limited
  - Impact: New Hypothesis
- Advanced
  - Size: Large
  - Scope: Intended Use
  - Results: Definitive
  - Impact: Policy

# Phases of Evaluation

- Exploratory
  - Early
  - Pilot
- Intermediate
  - Challenge or Stress Test
  - Lab-based
- Advanced
  - Late
  - Clinical use

## Proof of Concept

- Sample size small
- Convenient samples
  - vs. Representative
    - Simulations, phantoms, animal models, excised tissue and organs, patients
- PI and collaborator are sole study readers
- Answers
  - What does disease look like?
  - What does normal look like?
  - What can be measured?
  - How might patients benefit?

# Phases of Evaluation

- Exploratory
  - Early
  - Pilot
- Intermediate
  - Challenge or Stress Test
  - Lab-based
- Advanced
  - Late
  - Clinical use

## Prospective

- Sample size large
- Real world
  - Screening or diagnostic patient population
  - Patient care clinicians
  - Clinical reports, outcomes
  - Absolute performance
  - Meta and cost-benefit analyses
- Supports
  - Current practice guidelines
  - Policy and payment

# Intermediate Phase Challenge Test, Stress Test

- *Compare* new modality to current practice
  - Better (Superiority hypothesis)
  - Equivalent (Non-Inferiority hypothesis)
- Ideal, tightly-controlled, lab-based study
  - Readers blind to patient info
  - Task is narrow, quantitative
  - Calibrated equipment and tight protocols



# Intermediate Phase



## Challenge Test, Stress Test

- Challenging cases with and without disease
  - May not mimic intended use population
  - Creative sourcing to build study population
  - Enrich representation of key subgroups
    - Known or suspected differences
    - Statistical sub group analysis warranted?
  - Eliminate cases that don't help comparison
    - Obvious disease, obvious normal
- Technology Evaluation Not Clinical Performance

# Aims of a Reader Study: Intermediate Phase

- Task: Doctor Evaluates Patient Image
  - Narrow, Singular, Clear vs. Clinically Relevant
  - Data: Categorical, Ordinal, Interval, Ratio



- Data and Reference Standard Dictate Endpoint
  - Binary reference  Se/Sp, Area under ROC
  - Multi-level reference  Concordance

# Aims of a Reader Study: Intermediate Phase

Hypothesis testing: Assume what you want to disprove

- Better = Superiority

– H0: 

Endpoint New Modality
--------------------------

 – 

Endpoint Current Practice
------------------------------

 = 0

– H1: 

Endpoint New Modality
--------------------------

 – 

Endpoint Current Practice
------------------------------

 > 0

- Equivalent = Non-Inferiority

– H0: 

Endpoint New Modality
--------------------------

 – 

Endpoint Current Practice
------------------------------

 = 

Non-Inferiority Margin
---------------------------

– H1: 

Endpoint New Modality
--------------------------

 – 

Endpoint Current Practice
------------------------------

 > 

Non-Inferiority Margin
---------------------------



# Sample eCRF: QC check

Link key info to each observation

Reader ID

Modality

Case ID

View Box #

RCC

LCC

RMLO

LMLO

Thumbnail images confirm correct case

# Sample eCRF: ROC in 2 steps

Would you recall patient?

- ☐ Yes  
☒ No

Simple question with clinical relevance  
== Operating point on ROC curve

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?

Reiterate  
Instructions

Most Normal

Least Normal



Numeric  
Score

	7	5
--	---	---

Allow Quantitation  
Allow Control

# Sample eCRF: ROC in 2 steps

Would you recall patient?

- ☒ Yes  
☐ No

Different “clinical” decision  
== Threshold @ 100

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?

Least Suspicious

Most Suspicious

101



200

Numeric  
Score

--	--	--