Center for Devices and Radiological Health

U.S. FOOD AND DRUG ADMINISTRATION

Protecting and Promoting Public Health

Centennial Celebration
1906 – 2006

www.fda.gov/centennial

# Reader Studies

Brandon D. Gallas
NIBIB/CDRH Lab for the Assessment of Medical Imaging Systems

# Overview

- Basic elements of a reader study
  - Signal Detection
- Types of Reader Studies
  - Psychophysics
  - System Design and Optimization
  - Clinical Study



# Types of Reader Studies

- Psychophysics = Psychology + Physics
  - Goal is to understand/model the eye-brain system
  - Images are simulated and highly stylized
  - Readers have eyes and a brain
  - Example: How do noise amplitude and noise correlations impact detection of a signal?
  - Example: spatial-frequency sensitivity
  - Example: luminance-contrast sensitivity

# Types of Reader Studies

- Clinical Study
  - Goal: evaluate technology in use
  - Images are of real patient anatomy, perhaps the patient is present with a chart of background info
  - Prevalence sampling
  - Readers are doctors, radiologists, with extensive training and experience
  - Example: Are digital mammograms as good as screen-film?



# Types of Reader Studies

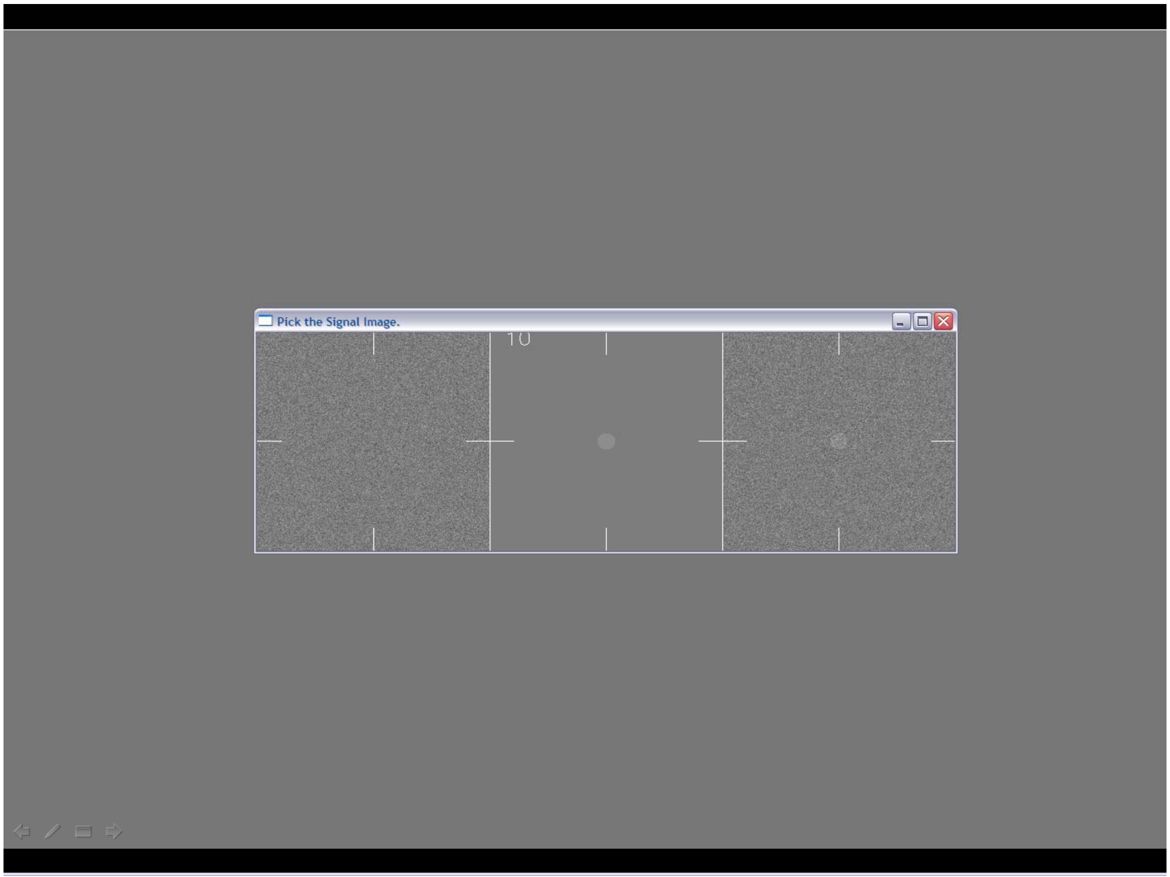
- System Optimization
  - Goal is to find imaging system parameters that yield better images
  - Images are simulated based on physics, generated in the lab with phantoms, or of real patient anatomy
  - Readers should depend on images: more clinically realistic images and signals require more clinical training and experience
  - Example: Reconstruction algorithm
  - Example: How does the display luminance curve impact detectability?

# Simplest Experiment

- White noise images (Gaussian) with and without disk in the center
- 2 alternative forced choice (2AFC) task
- performance metric is percent correct (PC)



Experiment 1




# Simplest Experiment: Viewing Details

- Ambient lighting
- Surround
- Distance to monitor
- SKE task:
  - Reference image
  - Location cues = Fiducial Marks
- 2AFC task:
  - Feedback

Humans need to accomodate to dark images.

# Simplest Experiment: Task too easy!

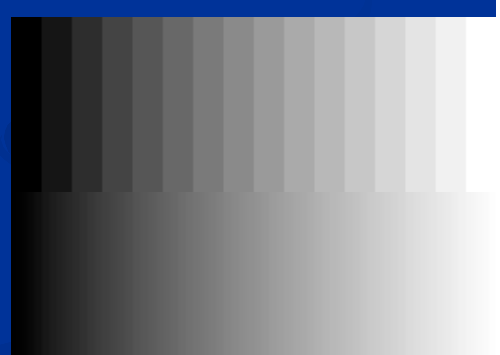
- Adjust Image Parameters
  - Background level
  - Noise level
  - Signal size/shape/intensity

- Experiment details and pitfalls



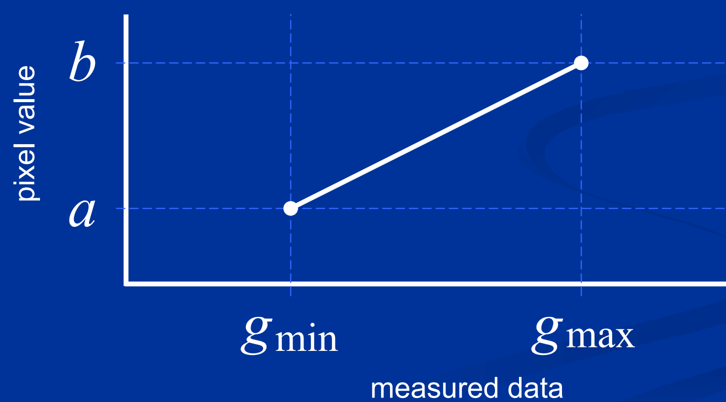
# Simplest Experiment: Pixel Values

- Image data gets mapped to pixel values p.v.
  - Most monitors display 8bit = 256 p.v.
- Know what your visualization software does
  - Display a gradient image
- Know the mapping
  - Global mapping
  - Image-dependent mapping
  - What happens to a signal?



# Simplest Experiment: Linear Mapping

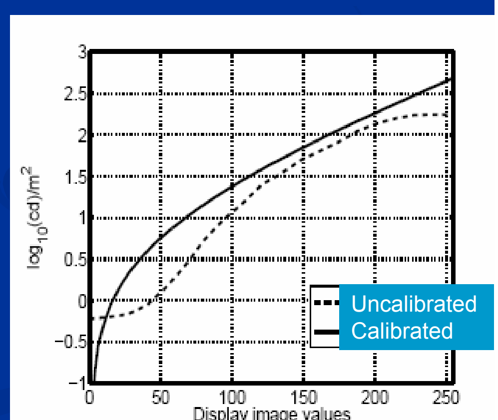
- pixel value $= \dfrac{(b-a)(g-g_{\min})}{g_{\max}-g_{\min}} + a$






# Simplest Experiment: Monitor

- Know your monitor
  - Pixel size
  - [candela/m$^2$]
  - Luminance range
  - Luminance response
- DICOM calibration standard




Courtesy S. Park JOSA (submitted 2006)

## Simplest Experiment: Task too easy!

- Adjust Image (System) Parameters
  - Background level
  - Noise level
  - Signal size/shape/intensity
- PILOT STUDY!
  - Staircase experiment



## Staircase Experiment

- 1-up, 1-down
  - 1 Incorrect response, go up in contrast
  - 1 Correct response, go down in contrast
  - experiment converges to PC=0.5
- 1-up, N-down converges to $PC=(0.5)^{1/N}$
  - 1-up, 2-down: PC=0.707
  - 1-up, 3-down: PC=0.794
  - 1-up, 4-down: PC=0.841
  - 1-up, 5-down: PC=0.871

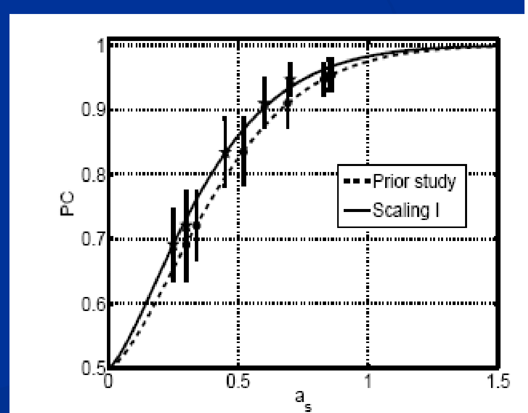
# Staircase Experiment

- Increase/Decrease Step Size (try 10%)
  - increase step size to speed convergence
  - decrease step size to increase resolution
  - fixed step sizes
- Can use as part of training
- Can use as final experiment instead of constant stimuli
  - psychometric curve
  - analysis/statistics more challenging



# Psychometric Curve: PC

- Sigmoid or S shaped curve.
- Modeled by
  - logistic function
  - gaussian function
- Fit by
  - least squares
  - maximum likelihood
- Constant stimuli




Courtesy S. Park JOSA (submitted 2006)

## Psychometric Curve: $d_A$

- Percent Correct = AUC

$$\Pr(t(g_1) > t(g_0))$$

$$= \int_{\infty} d(g_0, g_1)\, p(g_0, g_1)\, s(t(g_1) - t(g_0))$$

$$= \sum_{i=0}^{N} \frac{s_i}{N} \qquad s_i = \text{0 or 1 success for } i^{th} \text{ pair}$$



## Psychometric Curve: $d_A$

- Ideal Observer for Simplest Experiment

$$\text{AUC} = \Phi\left(\frac{1}{\sqrt{2}} d_A\right)$$

$$d_A = \frac{|\mu_1 - \mu_0|}{\sqrt{\frac{1}{2}\sigma_1^2 + \frac{1}{2}\sigma_0^2}}$$

## Psychometric Curve: $d_A$

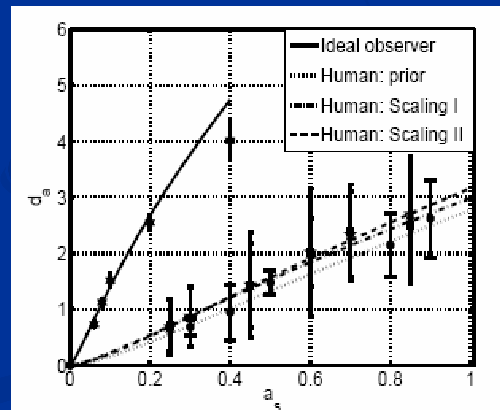
- Performance metric that is "linear" in signal amplitude (contrast)

$$d_A = \sqrt{2}\,\Phi^{-1}(\mathrm{AUC})$$

- Efficiency

$$\eta = \left(\frac{d_A(\text{human})}{d_A(\text{ideal})}\right)^2$$

$$\eta = \left(\frac{a(\text{ideal PC}=0.87)}{a(\text{human PC}=0.87)}\right)^2$$




Courtesy S. Park JOSA (submitted 2006)

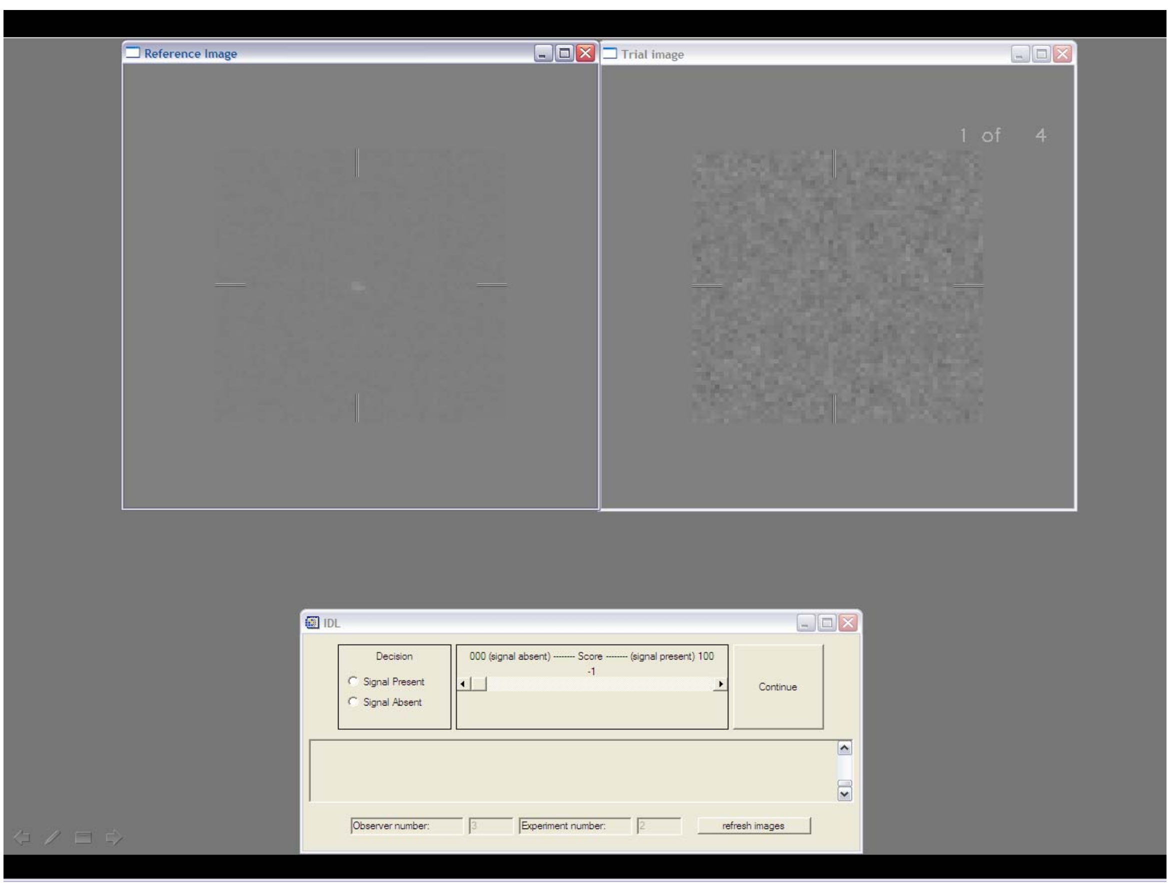


## ROC Experiment

- Images: GE2000D (24 kVp, 63mAs, Mo/Mo)
  - Regions of American College of Radiology (ACR) phantom
  - 0.1 x 0.1mm pixels, 64 x 64 ROIs (4x zoom)
  - signal: 0.24 mm specks from group 2
- Readers: Physicists in the lab
- Display: mammo-quality, 2560 x 2048 pixels
  - 0.165 mm pixel pitch
  - luminance calibration: DICOM
  - luminance range: 0.8-500 candelas/mm$^2$

# ROC Experiment

Experiment 2



# Scoring

- Yes/No
  - Doctors make yes/no decisions
  - CON: one point on the ROC curve
- 5 point scale
  - relate to clinical action items
  - same scale across readers (can pool data)
  - CON: maximum of 4 points on ROC curve
- 100-point scale
  - see the noise
  - binning loses information
  - CON: Need more training



# Reader Training

- Don’t want reader to learn during study, want to test!
- Reduce learning bias:
  - randomize case reading order
  - randomize case sets, modality
- When reading same cases in two modalities, separate the readings by as much time as possible

# Reader Training

- Psychophysics, how many cases?
  - If using simulated images...
  - If using psychology students...

  - Train extensively!



# Reader Training

- Clinical Study, how many cases?
  - Real images cost and disease is often rare
  - Doctors should know their task
  - Training important, how to score
- Essential part of a clinical study protocol
  - What signals are they looking for?
  - What does scale mean?
  - What's the population/prevalence?

# Reader Training

## Personal Rules of Thumb

- Pretest train
  - Provide examples of all image types
  - Range of contrasts
  - Incorporate staircase and pilot studies
  - provide feedback
- Warmup train
  - Need to accomodate to specific task
- Number of training samples
  - 25% number of testing samples



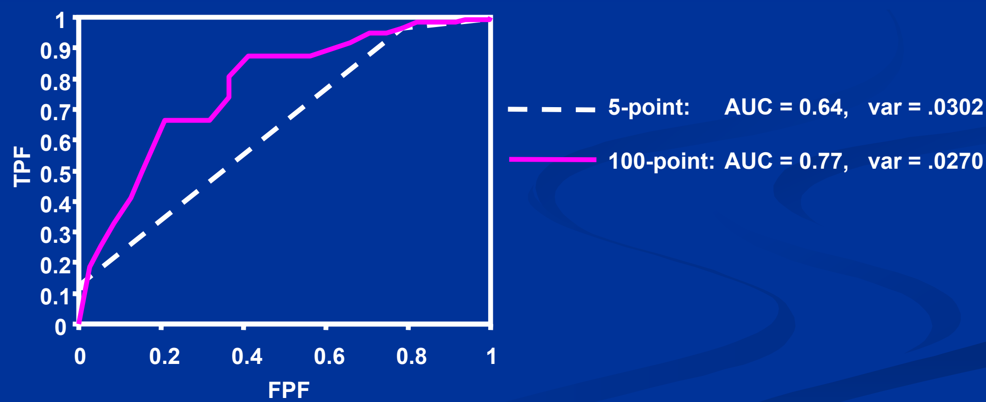
# Nonparametric ROC analysis

| | scores | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # cases | total | 1 | 2 | 3 | Threshold | 4 | 5 | |
| signal-absent | 100: | 0 | 2 | 20 | | 77 | 1 | 78 |
| signal-present | 100: | 1 | 0 | 3 | | 83 | 13 | 96 |

| | | | | | | |
|---|---|---|---|---|---|---|
| FPF (%) | 100 | 100 | 98 | 78 | 1 | 0 |
| TPF (%) | 100 | 99 | 99 | 96 | 13 | 0 |

# 5-point vs 100-point scale Empirical ROC cuves

- Same Simulated Data, Different Bins






# Nonparametric AUC

- Wilcoxon-Mann-Whitney Statistics

$$\Pr(t(g_1) > t(g_0))$$

$$= \int_{\infty} dg_0\, p_0(g_0) \int_{\infty} dg_1\, p_1(g_1)\, s(t(g_1) - t(g_0))$$

$$= \sum_{i=0}^{N_0} \sum_{j=0}^{N_1} \frac{s_{ij}}{N_0 N_1}$$

$s_{ij}$ = 0 or 1 success comparing
$i^{th}$ signal-absent score
$j^{th}$ signal-present score

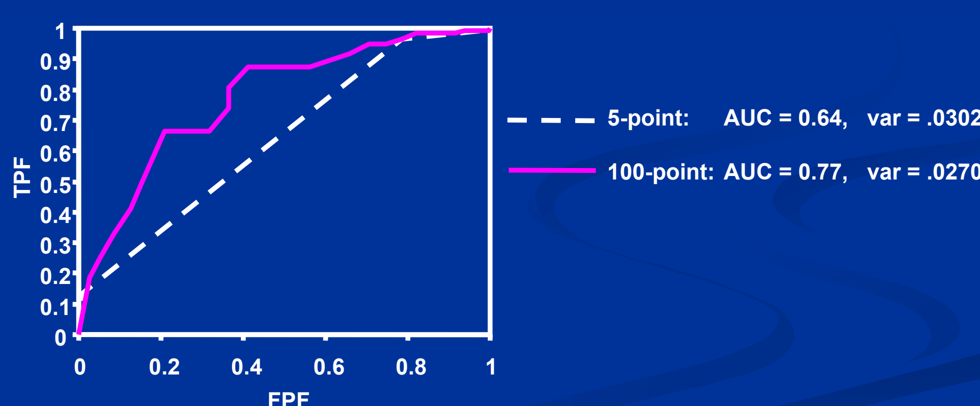
# Parametric ROC analysis

- Smooth ROC curve
- ML-EM
  - Maximum Likelihood Expectation Maximization
  - Semi-parametric, Binormal Model
  - Dorfman and Alf 1968
  - software: www-radiology.uchicago.edu/krl/index.htm



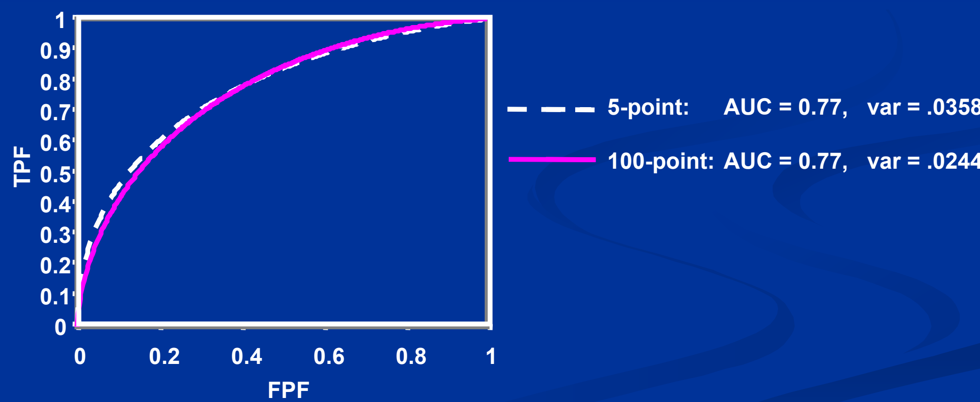
# 5-point vs 100-point scale Empirical ROC cuves

- Same Simulated Data, Different Bins



# 5-point vs 100-point scale MLEM ROC cuves

- Same Simulated Data, Different Bins



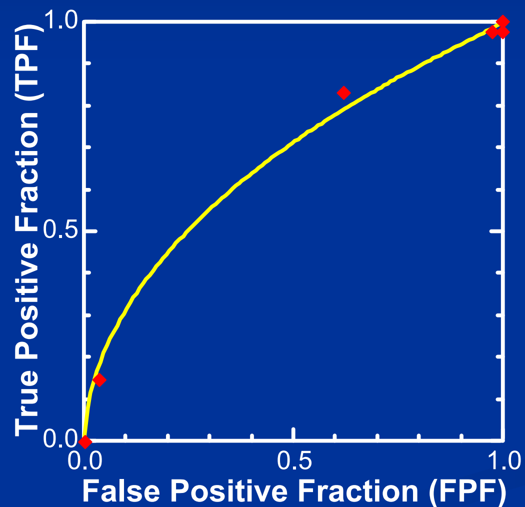



# BIRADS Action Item Scale

## Breast Imaging Reporting and Data System

- BIRADS 0 - Need Additional Imaging Evaluation and/or Prior Mammograms For Comparison
- BIRADS 1 – Negatives, One year routine follow-up
- BIRADS 2 – Benign finding(s), One year routine follow-up
- BIRADS 3 – Probably Benign Finding Initial Short-Interval Follow-Up Suggested
- BIRADS 4 – Suspicious Abnormality Biopsy Should Be Considered
- BIRADS 5 – Highly Suggestive of Malignancy Appropriate Action Should Be Taken, Biopsy,…
- BIRADS 6 – Known Biopsy – Proven Malignancy, Appropriate ActionShould Be Taken, Biopsy,…

## Empirical and MLE-fitted ROC Curves
## BI-RADS Scores (2,3,4,5)




**Empirical AUC = 0.6327**

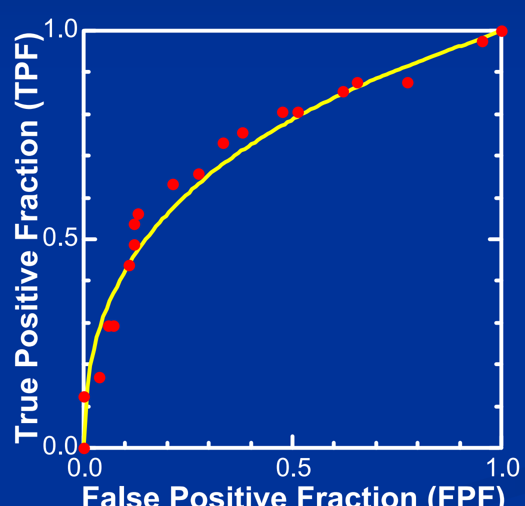
**Fitted AUC = 0.6676**

- **125 total cases**
  - **84 benign**
  - **41 malignant**

Courtesy Y. Jiang, U. Chicago



## Empirical and MLE-fitted ROC Curves
## 100-Point Likelihood Estimates




**Empirical AUC = 0.7443**

**Fitted AUC = 0.7365**

- **125 total cases**
  - **84 benign**
  - **41 malignant**

Courtesy Y. Jiang, U. Chicago

## Variance: Single Reader PC

- Successes are Bernoulli trials
- PC is a binomial random variable

$$\mathrm{var}\left(\widehat{\mathrm{PC}}\right) = v_{\mathrm{PC}} = \frac{1}{N}\mathrm{PC}(1-\mathrm{PC})$$

$$\hat{v}_{\mathrm{PC}} = \frac{1}{N}\left(\sum_{i=1}^{N}\frac{s_i s_i}{N} - \sum_{i=1}^{N}\sum_{i\neq i}^{N}\frac{s_i s_{i'}}{N(N-1)}\right)$$

$$\hat{v}_{\mathrm{PC}} = \frac{1}{N-1}\widehat{\mathrm{PC}}\left(1-\widehat{\mathrm{PC}}\right)$$



## Variance: Single Reader AUC

- Nonparametric AUC
  - Successes are correlated Bernoulli trials
  - U-statistics
  - method of moments
- Maximum Likelihood AUC
  - part of the solution machinery
  - output of software
  - Fisher Information Matrix

# Variance: Single Reader AUC

- Nonparametric AUC

$$v_{\mathrm{AUC}} = \mathrm{var}\left(\widehat{\mathrm{AUC}}\right)$$

$$\hat{v}_{\mathrm{AUC}} = \sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\frac{s_{ij}s_{ij}}{(N_0N_1)^2} + \sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{i'\neq i}^{N_0}\frac{s_{ij}s_{i'j}}{(N_0N_1)^2}$$

$$+\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{j'\neq j}^{N_1}\frac{s_{ij}s_{ij'}}{(N_0N_1)^2} - \frac{(N_0+N_1-1)}{(N_0-1)(N_1-1)}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{i'\neq i}^{N_0}\sum_{j'\neq j}^{N_1}\frac{s_{ij}s_{i'j'}}{(N_0N_1)^2}$$



# MRMC ROC experiment

- MRMC
  - Multi-Reader
  - Multi-Case
- Everything here has a 2AFC/PC analog
- Study Designs
  - Fully crossed study design
  - Doctor-patient
  - Hybrid
- Compare two modalities

# Sample cases

| | |
|---|---|
| $N_0$ sig-abs images | $N_1$ sig-pres images |



# Sample Readers

$R$ readers

# Collect Scores

| | $N_0$ sig-abs images | $N_1$ sig-pres images |
|---|---|---|
| $R$ readers | Fully Crossed | |



# For the $r^{th}$ reader

| $N_0$ sig-abs images | $N_1$ sig-pres images |
|---|---|
| $\cdots\ t_{0ir}\ \cdots$ | $\cdots\ t_{1jr}\ \cdots$ |

# For the $r^{th}$ reader

| $N_0$ sig-abs images | $N_1$ sig-pres images |
|---|---|

$\mathbf{t}_r \rightarrow a(\mathbf{t}_r)$

$$a(\mathbf{t}_r) = \begin{array}{l}\text{Wilcoxon Statistic} \\ \text{Percent Correct}\end{array}$$



# Figure of Merit

| | $N_0$ sig-abs images | $N_1$ sig-pres images |
|---|---|---|
| $R$ readers | $\vdots$ $\mathbf{t}_r$ $\vdots$ | |

$\rightarrow a(\mathbf{t}_0)$

$\vdots$

$\rightarrow a(\mathbf{t}_r)$

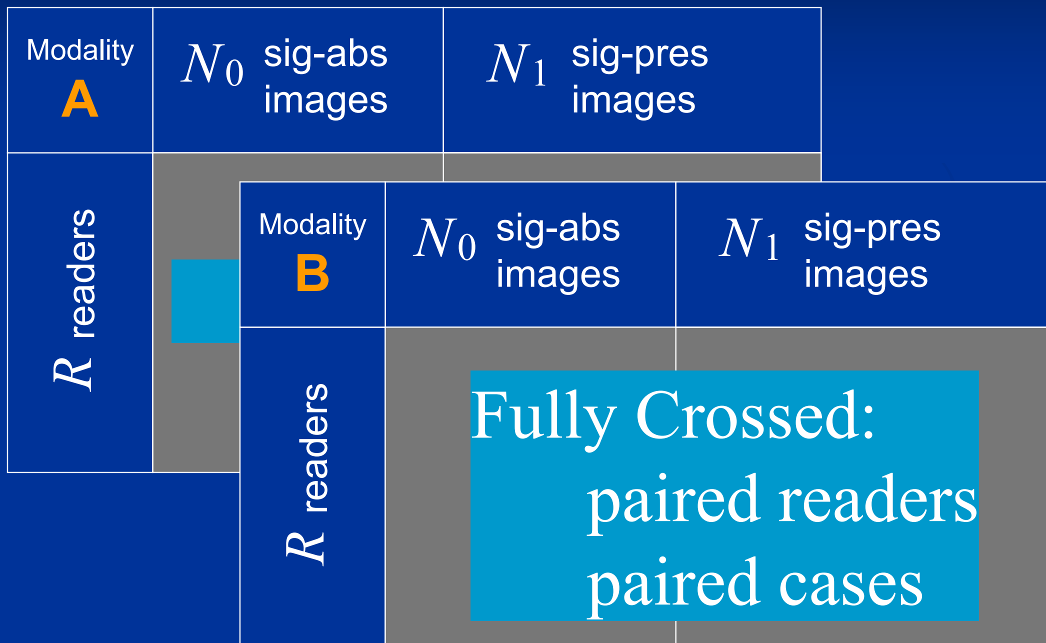
$\vdots$

$\downarrow$

$A(\mathbf{T})$

- Want a variance that generalizes to
  - new readers
  - new cases

# MRMC comparing modalities




# MRMC Variance

| Existing Methods | Parametric | Resampling |
|---|---|---|
| The jackknife/ANOVA ...................... Dorfman, Berbaum and Metz | Yes | Yes |
| ANOVA and correlation model ........ Obuchowski | Yes | No |
| Ordinal regression ........................... Toledano and Gatsonis | Yes | No |
| The bootstrap .................................. Beiden, Wagner, and Campbell | Yes & No | Yes |
| The one-shot .................................. Gallas (based on theory by Barrett, Clarkson, & Kupinski) | No | No |

# MRMC Variance Nonparametric AUC

- Single reader PC had 2 terms:
  - pairs of cases

$$i' = i, i' \neq i$$

- Single reader variance had $2^2$=4 terms:
  - signal-present, signal-absent cases

$$i' = i, i' \neq i \quad \times \quad j' = j, j' \neq j$$

- MRMC AUC variance has $2^3$=8 terms:
  - signal-present cases, signal-absent cases, readers

$$i' = i, i' \neq i \quad \times \quad j' = j, j' \neq j \quad \times \quad r' = r, r' \neq r$$



# One-shot estimate

$$\mathrm{var}(A(\mathbf{T}_{G\Gamma}))$$
$$= \frac{1}{R}\left[c_1\widehat{M}_1 + c_2\widehat{M}_2 + c_3\widehat{M}_3 + c_4\widehat{M}_4\right]$$
$$+ \frac{R-1}{R}\left[c_1\widehat{M}_5 + c_2\widehat{M}_6 + c_3\widehat{M}_7 + c_4\widehat{M}_8\right]$$
$$- \widehat{M}_8$$

# One-shot estimate

$$\widehat{M}_1 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\frac{s(t_{1jr}-t_{0ir})^2}{RN_0N_1}$$

$$\widehat{M}_2 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{i'\neq i}^{N_0}\frac{s(t_{1jr}-t_{0ir})s(t_{1jr}-t_{0i'r})}{RN_0N_1(N_0-1)}$$

$$\widehat{M}_3 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{j'\neq j}^{N_1}\frac{s(t_{1jr}-t_{0ir})s\left(t_{1j'r}-t_{0ir}\right)}{RN_0N_1(N_1-1)}$$

$$\widehat{M}_4 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{i'\neq i}^{N_0}\sum_{j'\neq j}^{N_1}\frac{s(t_{1jr}-t_{0ir})s\left(t_{1j'r}-t_{0i'r}\right)}{RN_0N_1(N_0-1)(N_1-1)}$$



# One-shot estimate

$$\widehat{M}_5 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{r'\neq r}^{R}\frac{s(t_{1jr}-t_{0ir})s\left(t_{1jr'}-t_{0ir'}\right)}{RN_0N_1(R-1)}$$

$$\widehat{M}_6 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{r'\neq r}^{R}\sum_{i'\neq i}^{N_0}\frac{s(t_{1jr}-t_{0ir})s\left(t_{1jr'}-t_{0i'r'}\right)}{RN_0N_1(R-1)(N_0-1)}$$

$$\widehat{M}_7 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{r'\neq r}^{R}\sum_{j'\neq j}^{N_1}\frac{s(t_{1jr}-t_{0ir})s\left(t_{1j'r'}-t_{0ir'}\right)}{RN_0N_1(R-1)(N_1-1)}$$

$$\widehat{M}_8 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{r'\neq r}^{R}\sum_{i'\neq i}^{N_0}\sum_{j'\neq j}^{N_1}\frac{s(t_{1jr}-t_{0ir})s\left(t_{1j'r'}-t_{0i'r'}\right)}{RN_0N_1(R-1)(N_0-1)(N_1-1)}$$

# One-shot Covariance

- Form of covariance is identical to variance
- Instead of success outcomes from one modality, use two
- For example,

$$\hat{M}_5 = \sum_{r=1}^{R}\sum_{i=1}^{N_0}\sum_{j=1}^{N_1}\sum_{r'\neq r}^{R} \frac{\overbrace{s(t_{1jr}-t_{0ir})}^{\text{Outcomes Modality 1}}\,\overbrace{s\left(t_{1jr'}-t_{0ir'}\right)}^{\text{Outcomes Modality 2}}}{RN_0N_1(R-1)}$$



# Power

## (statistical efficiency)

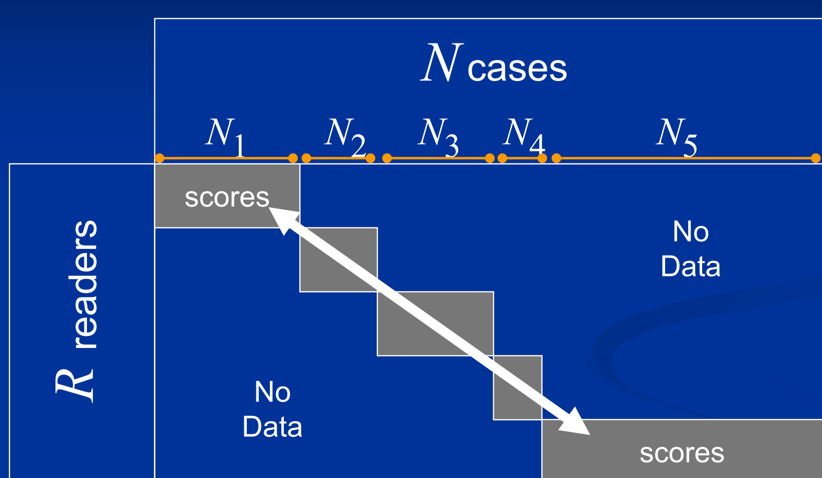
- Fully-Crossed, paired reader, paired cases
  - readers reading same cases
  - readers same in both modalities
  - cases same in both modalities
- Correlations increase power to detect difference

# Doctor-Patient Study design

- Each reader reads their own cases (necessary for in vivo diagnostics)
- Readers may read different numbers of cases
- More variety in averaging performance?



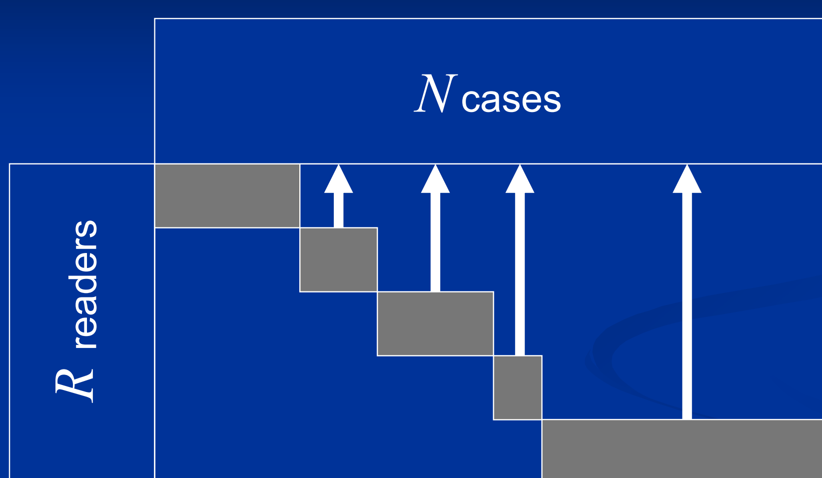
# Doctor-Patient Study design



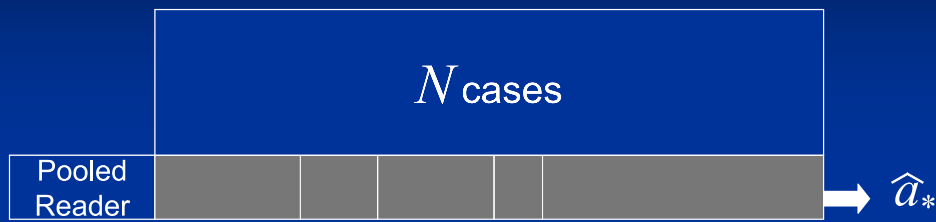
# Reader-averaged Performance




# Pooled-reader Performance



# Pooled-Reader Performance






# Statistical Properties

## Reader-averaged PC vs. Pooled-Reader PC

- Expected values are the same
- Can estimate MRMC variance of both statistics!
- Variances are different
  - depends on distribution of cases among readers
  - depends on reader variance, case variance, and interaction
- There is an optimal statistic
  - optimal = minimum variance

# Statistical Properties

**Reader-averaged AUC vs. Pooled-Reader AUC**

- Expected Values are different!
  - Pooling is dangerous because readers use scales differently
- Can estimate MRMC variance of both statistics!
- Variances
  - Definitely Doable for Nonparametric AUC
  - Probably Doable for MLE AUC

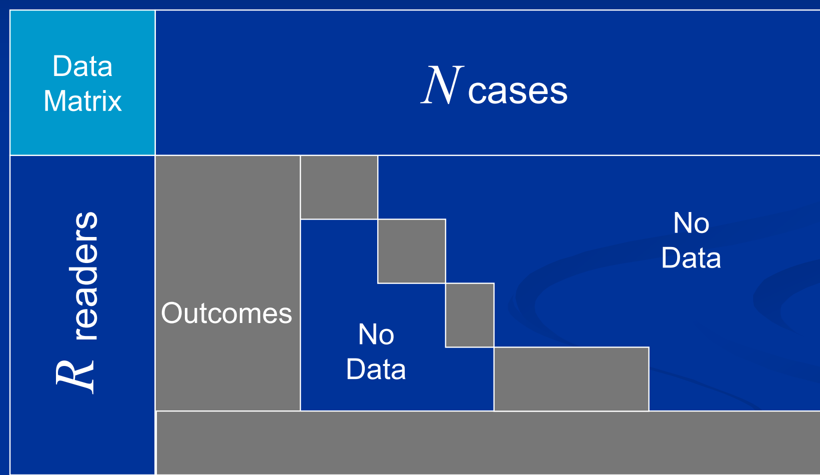


# Power

**(Statistical Efficiency)**

- Doctor-Patient:
  paired reader, paired cases
  - ~~readers reading same cases~~
  - readers same in both modalities
  - cases same in both modalities
- Readers don't need to be paired
- Cases don't need to be paired

# Hybrid Study design



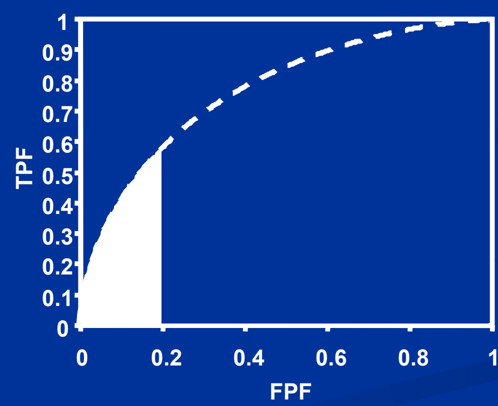



# Extending ROC Methods

- Partial-area ROC
- Location: partition image
  - One choice per case: MAFC
  - Multiple scores per case: ROI analysis
- Location: continuous specification
  - One score per case: LROC
  - Multiple scores per case: FROC

# Partial Area

- Interested in high specificity decisions (Specificity > 0.8)






# Agreement Statistic

- Compare two readers
- Compare model observer to human
  - Typically compare AUCs
  - How about comparing rankings!
  - Prediction probability

# Agreement Statistic

## Prediction Probability

- Similar to Kendall's tau
- Generalization of Wilcoxon AUC

| Human / Model | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| 5 | x | x | x | x | x |
| 4 | x | x | x | x | x |
| 3 | x | x | x | x | x |
| 2 | x | x | x | x | x |
| 1 | x | x | x | x | x |

AUC
AUC
AUC
AUC