Implementation and Analysis of Observer Studies in Medical Physics

W.F. Sensakovic, PhD, DABR, MRSC

Relevant Conflicts of Interest

Attendees/trainees should not construe any of the discussion or content of the session as insider information about the American Board of Radiology or its examinations.

- **Task is complex**
  - **Outline subtle tumor**
- **Unquantifiable human element**
  - **Clinical decision or Human visual system**
- **Human response is goal**
  - **Does widget “A” make it easier *for the observer* to detect the microcalcification?**

Bunch of observers look at Bunch of subject images to create data that is then analyzed

CC 3.0 Aaron Dodson, from The Noun Project

CC 3.0: Zzyzx11

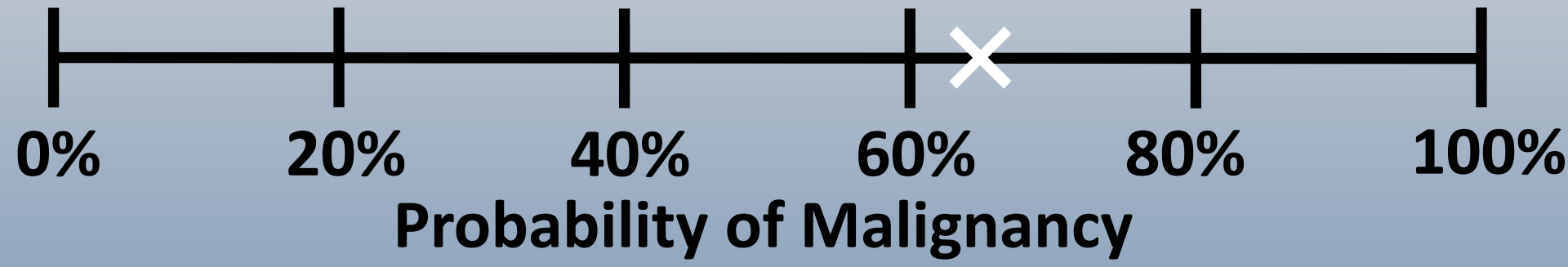
Diagnosis: What is it?

Bunch of radiologists look at bunch of CT scans (FBP or Iterative Recon) to record probability of malignancy for each. ROC analysis determines if resolution impacts diagnosis.

Widely Used Scale
Definitely or almost definitely malignant
Probably malignant
Possibly malignant
Probably benign
Definitely or almost definitely benign
Based on: Swets JA, et al. Assessment of Diagnostic Technologies. Science 205(4408):753 (1979)

Including clinically relevance
Malignant—diagnosis apparent—warrants appropriate clinical management
Malignant—diagnosis uncertain—warrants further diagnostic study/biopsy
I'm not certain—warrants further diagnostic study
Benign—no follow-up necessary
Based on: Potchen EJ. Measuring Observer Performance in Chest Radiology: Some Experiences. J Am Coll Radiol 3:423 (2006)

- **Typically 5-7 categories**
- **Validated scale if available and appropriate**
- **Include clinical relevance if possible**



- **Continuous vs. Categorical difference biggest for single reader studies**
  - Wagner RF et al. Continuous versus Categorical Data for ROC Analysis Some Quantitative Considerations. Acad Rad 8(4): 328 (2001).
- **No practical difference between discrete and continuous scales for ratings**
  - Rockette HE, et al. The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging techniques. Invest Radiol 27(2):169 (1992).

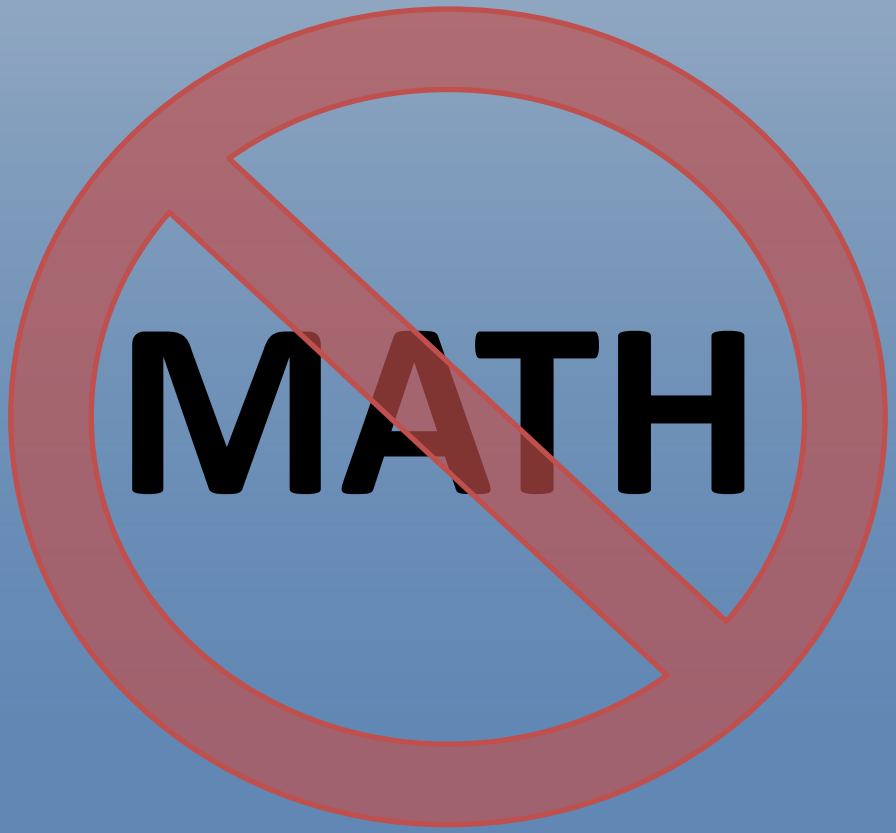
Truth

- **Best**
  - **Abnormal: Biopsy or other gold standard**
  - **Normal: Follow-up (e.g., 1-year) post imaging**
- **Combined reads (expert panel)**
  - **In a 3 system comparison the "best" system depended on whether majority vote, consensus opinion, expert judgment, feedback review, or clinical/pathologic proof was used for truth**
    - Revesz G et al. The effect of verification on the assessment of imaging techniques. Invest Radiol 18:194 (1983).
  - **Report variability in consensus**
    - Bankier AA et al. Consensus Interpretation in Imaging Research: Is There a Better Way? Radiology 257:14 (2010).

ROC Analysis Study Design

- **Traditional, Fully-Crossed, Paired-Case Paired-Reader, Full Factorial**
  - **Every observer, reads every case, in every modality**
  - **Data correlations all us to get the highest power and lowest sample requirements**

- **Software (free or not) does it for you**
  - **MRMC ROC Software and listed later**
    - **Some unsupported and not functional on modern computers, but may still run on an emulator such as dosbox (*https://www.dosbox.com*)**

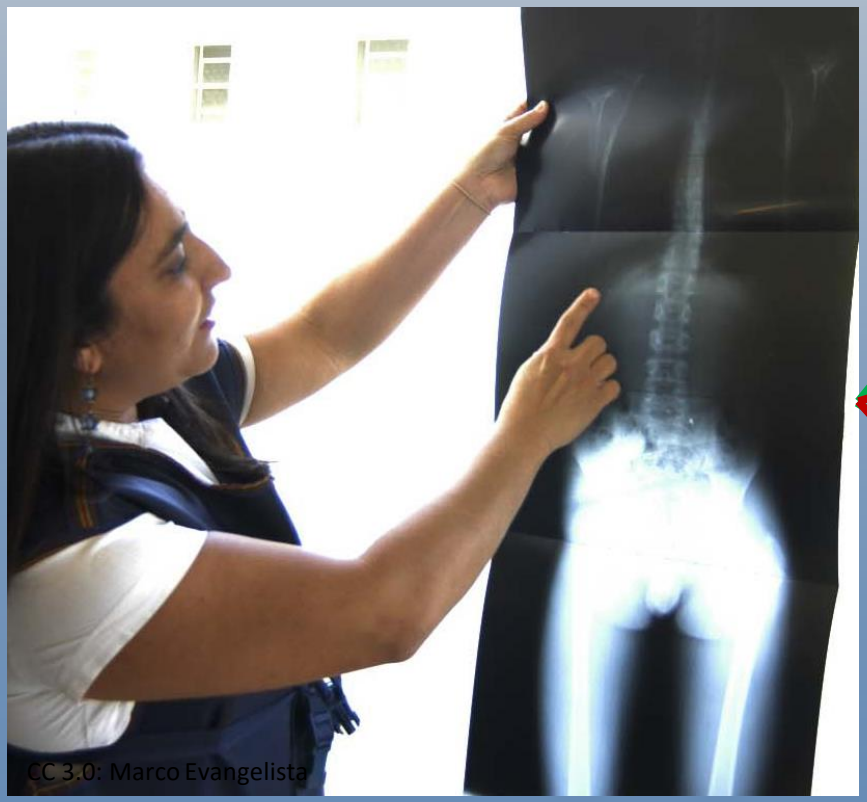


Does iterative reconstruction impact diagnosis of malignancy in lung lesions?

With IR

Case #/Truth	Obs. 1
1/Malignant	10.0
2/Benign	4.4
3/Benign	3.4
5/Malignant	5.6
6/Malignant	7.7
7/Malignant	9.2
…	…

- 0.0: Definitely Benign
- 2.0: Probably Benign
- 5.0: Indeterminate
- 8.0: Probably Malignant
- 10.0: Definitely Malignant



True Positive (TP)

- **Sensitivity**

False Positive (FP)

- **1-Specificity**

True Negative (TN)

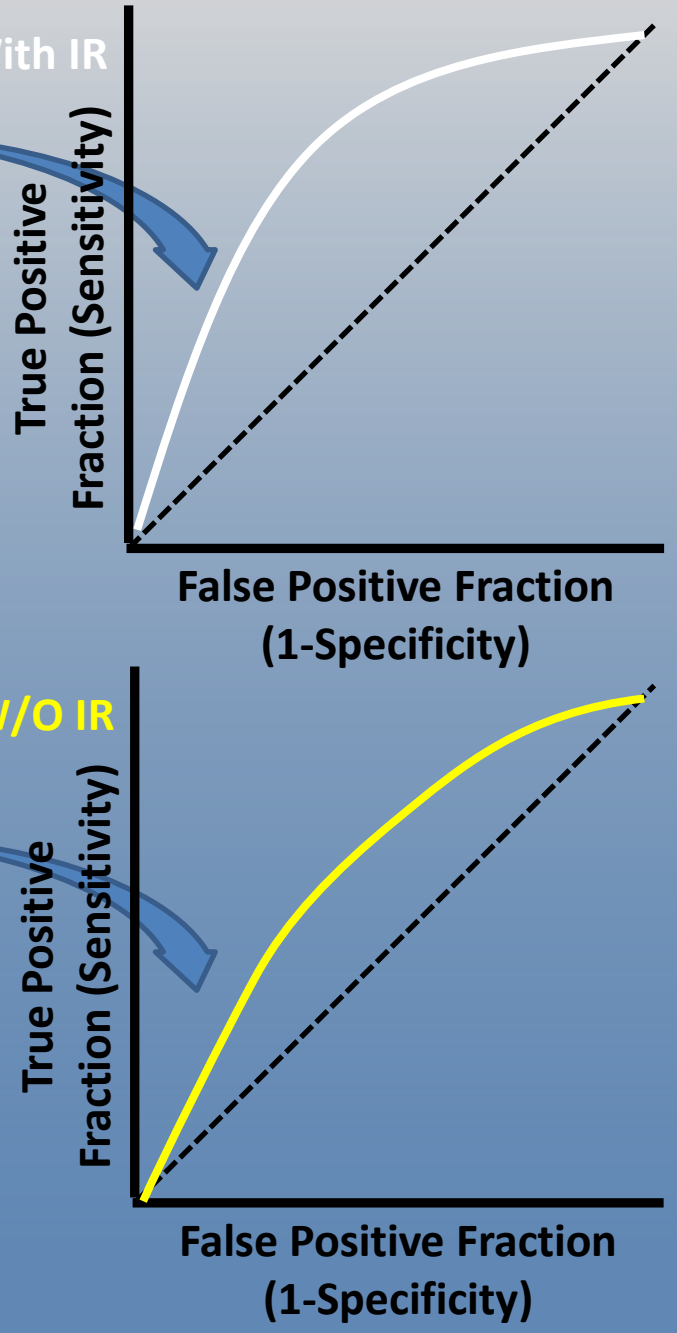
False Negative (FN)

With IR

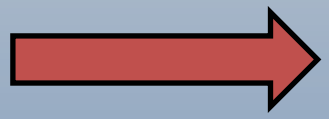
Case #/Truth	Rating
1/Malignant	10.0
2/Benign	4.4
3/Benign	3.4
5/Malignant	5.6
6/Malignant	7.7
7/Malignant	9.2
…	…

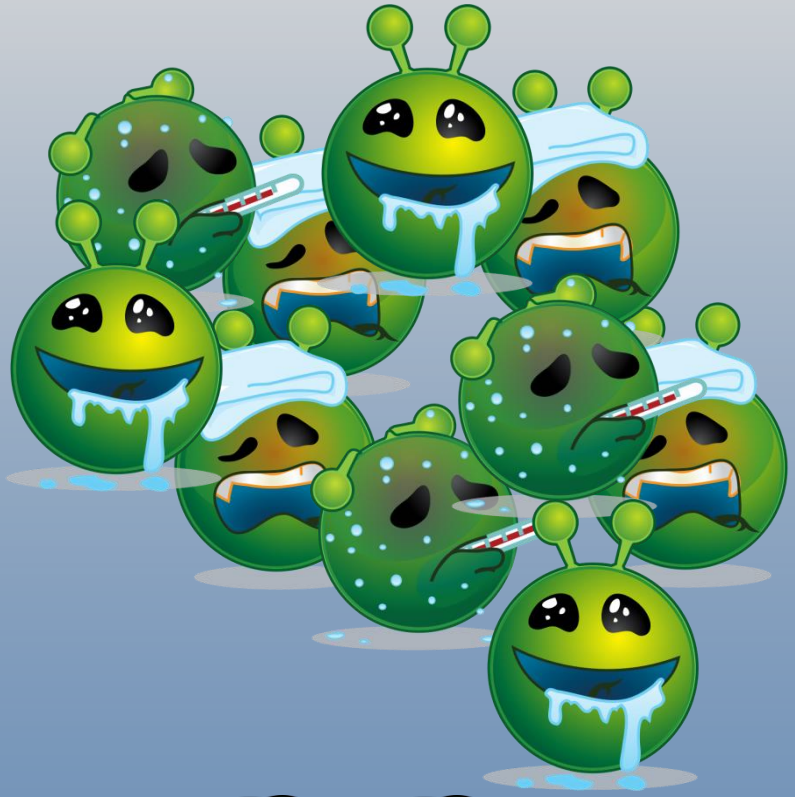
W/O IR

Case #/Truth	Rating
8/Malignant	9.1
9/Benign	4.3
10/Benign	4.4
11/Malignant	4.6
12/Malignant	7.0
13/Malignant	8.8
…	…

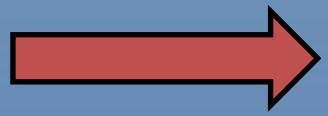


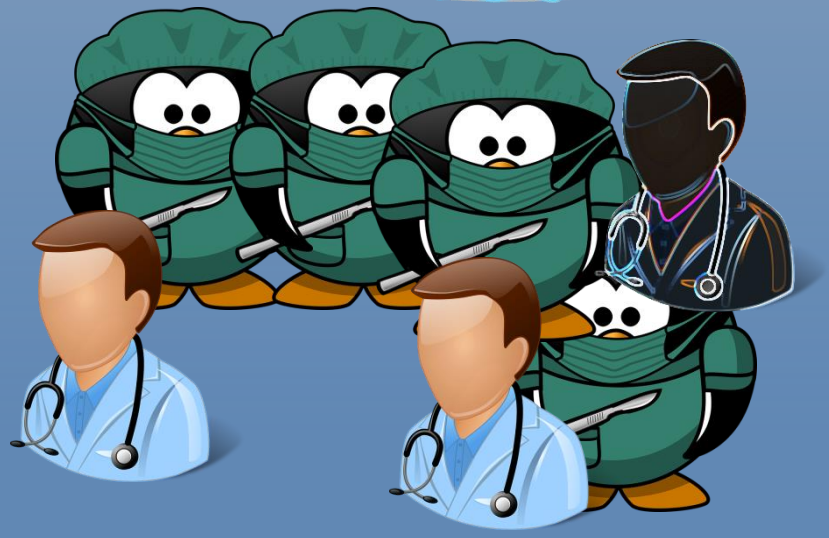

Public Domain



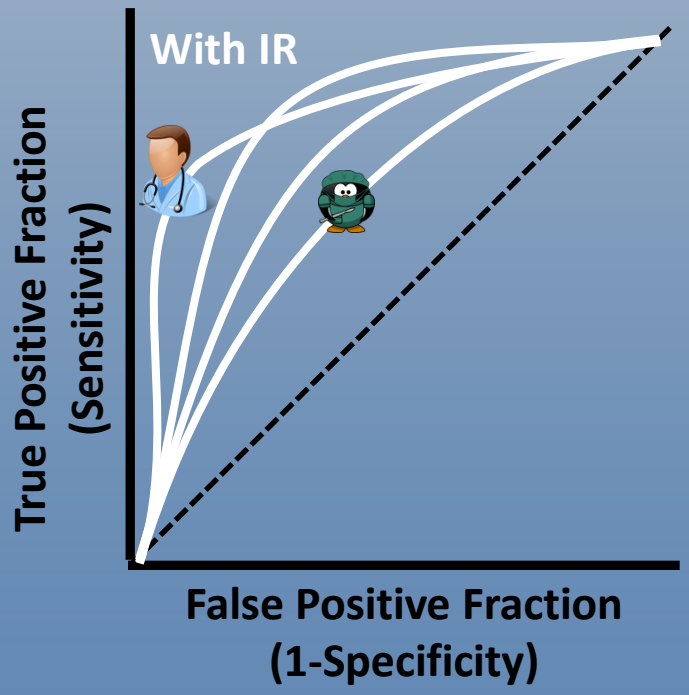


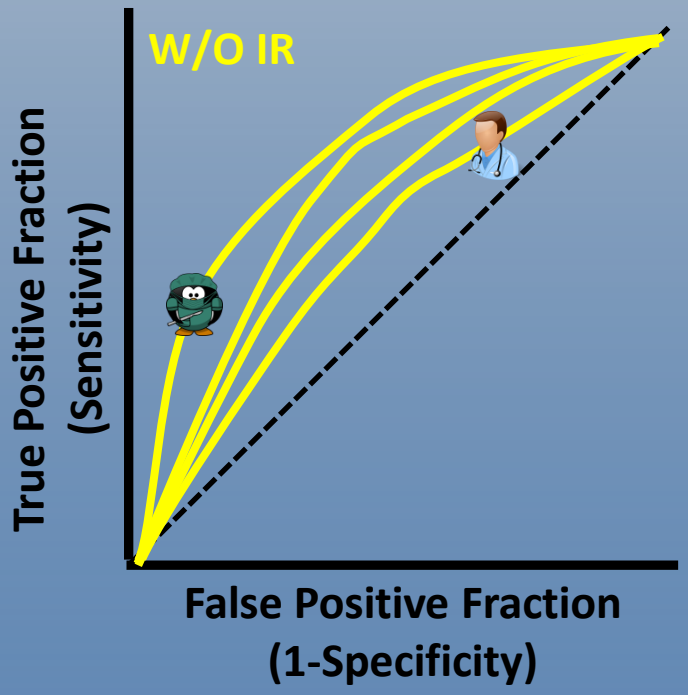

Public Domain





Multi-Reader, Multi-Case (MRMC)

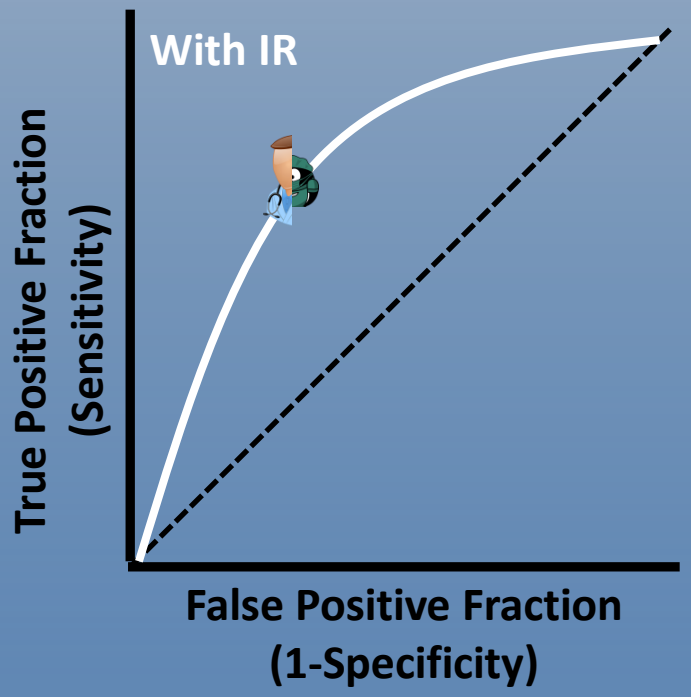


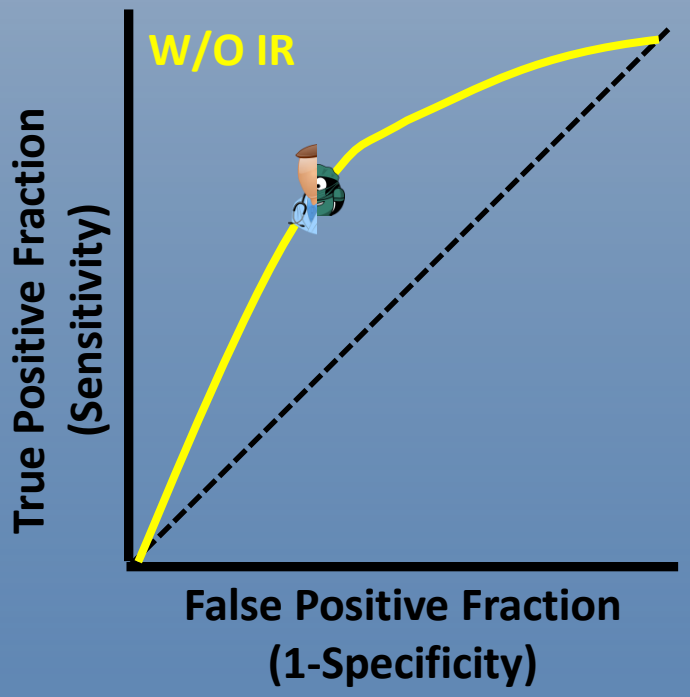


Ensemble Curve and AUC

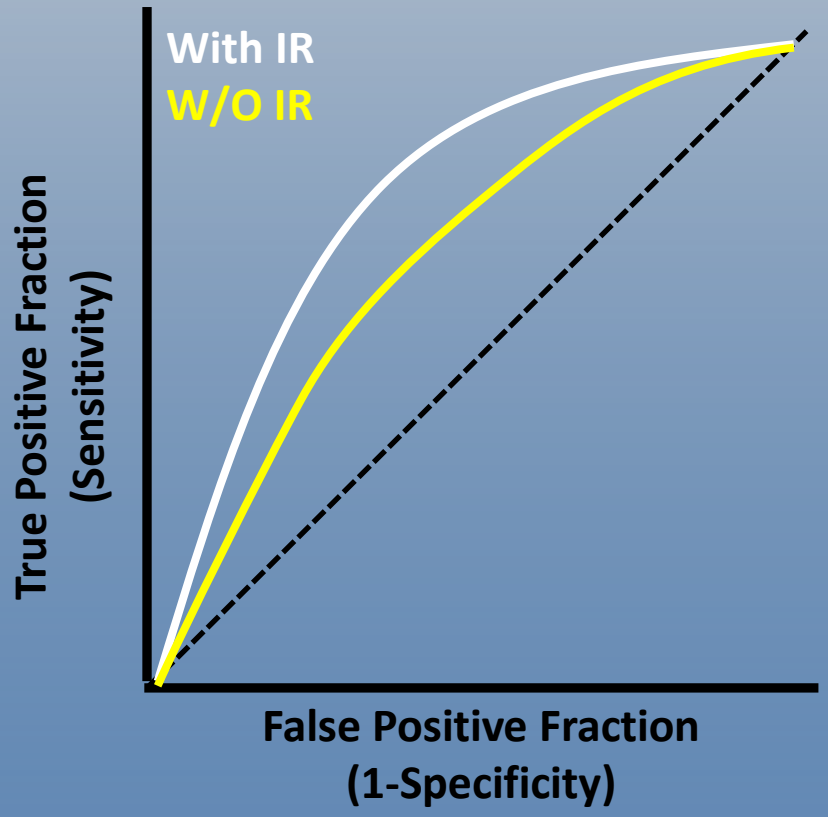
ANOVA or Other Analysis

Resampling (Jackknife or Bootstrap)



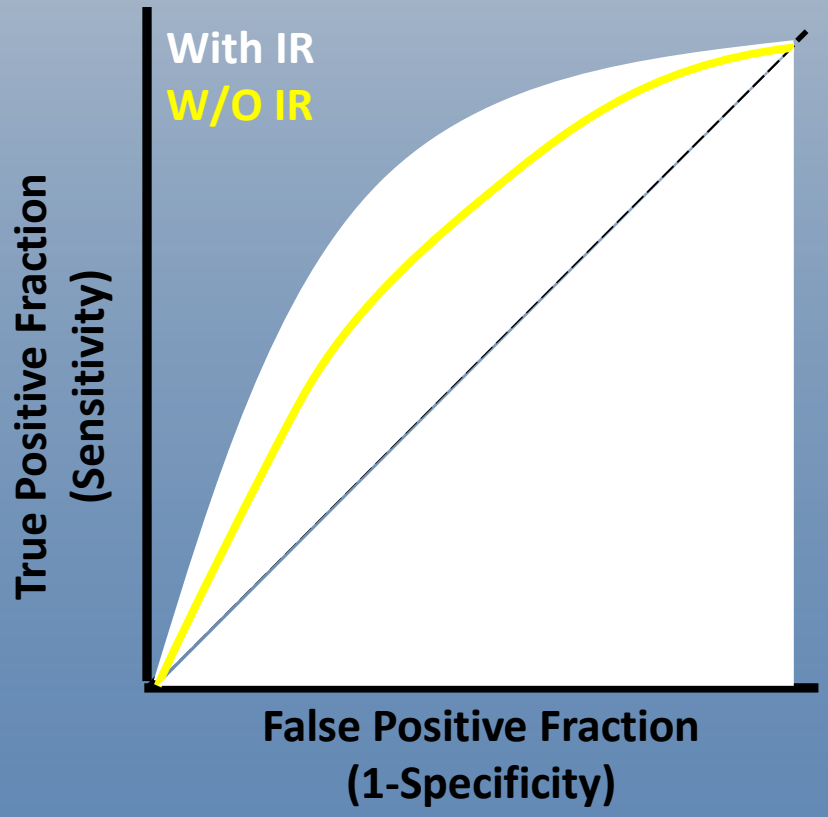


Does iterative reconstruction impact diagnosis of malignancy in lung lesions?



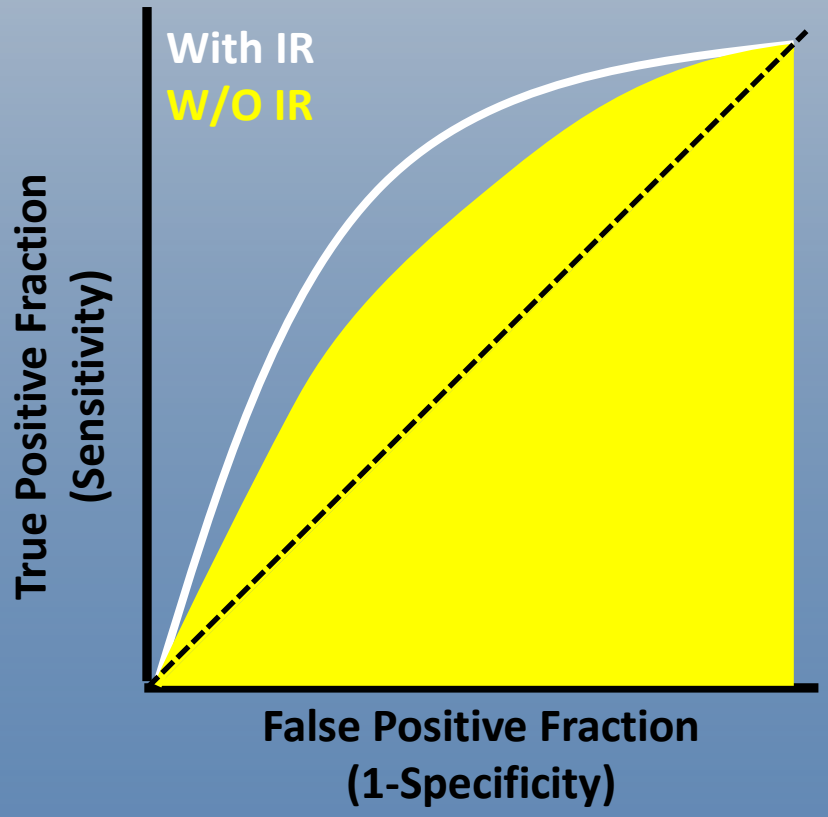
- Yes, it improves diagnosis
- By how much?

Does iterative reconstruction impact diagnosis of malignancy in lung lesions?



- Yes, it improves diagnosis
- By how much?
  - AUC = 0.8

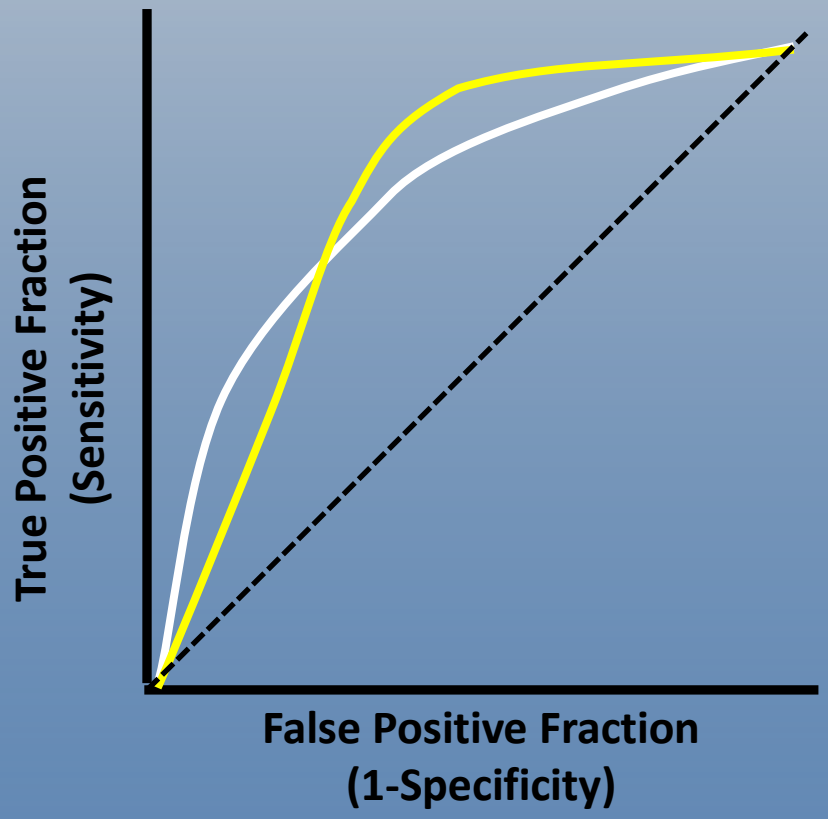
Does iterative reconstruction impact diagnosis of malignancy in lung lesions?



- Yes, it improves diagnosis
- By how much?
  - AUC = 0.8
  - AUC = 0.7

- **AUC probability a randomly selected malignant case is rated higher than a randomly selected benign case**
- **Average TFP over all FPF**
- **Average percent correct if observers shown random malignant and benign and asked to choose the malignant**
  - **2-alternative forced choice**

AUC comparison not appropriate if ROC curves cross each other



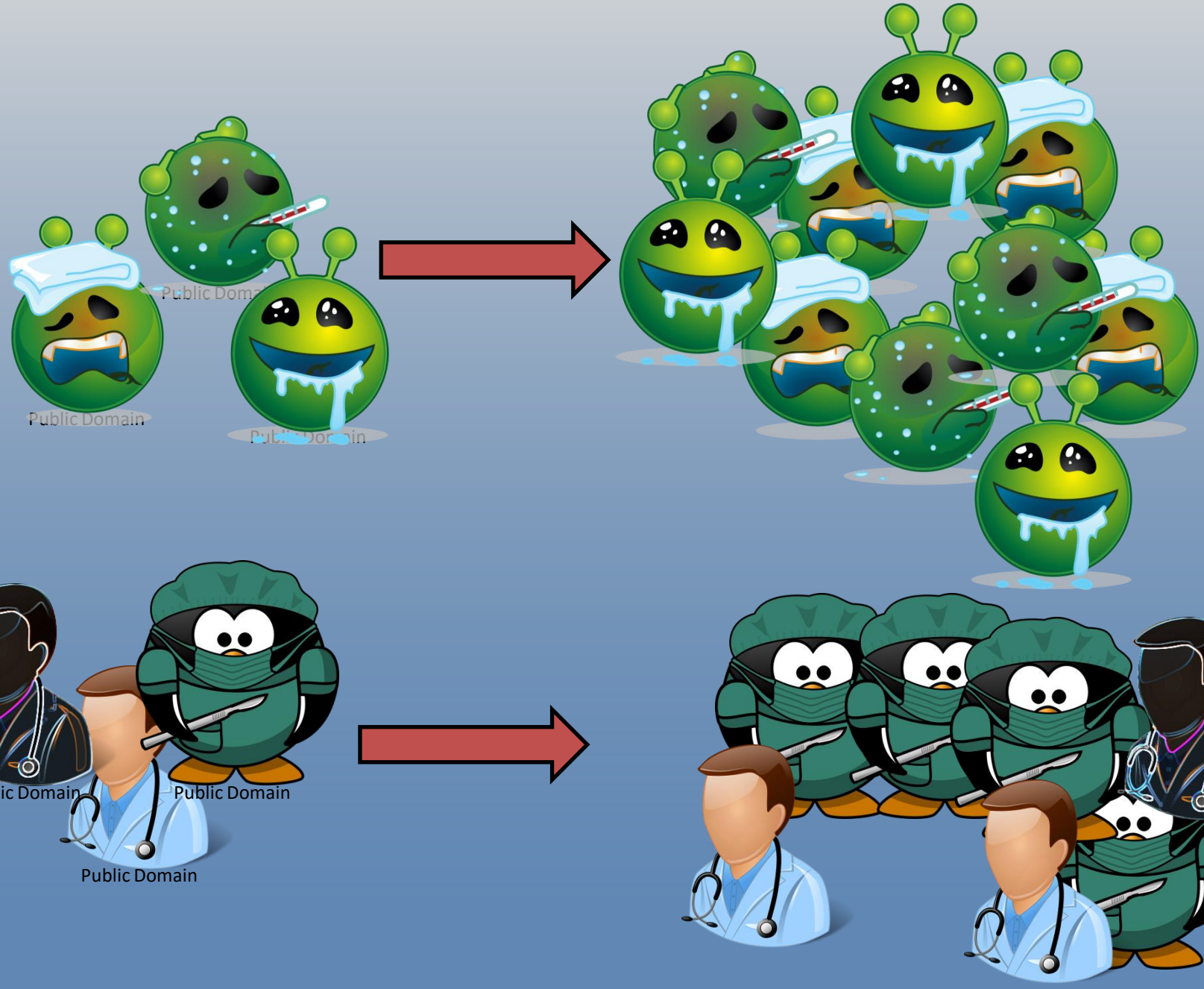
- Better for Screening
- Maybe better for Diagnostic
- Partial AUC
  - McClish DK. Analyzing a Portion of the ROC Curve. Medical Decision Making 9 (3): 190 (1989)

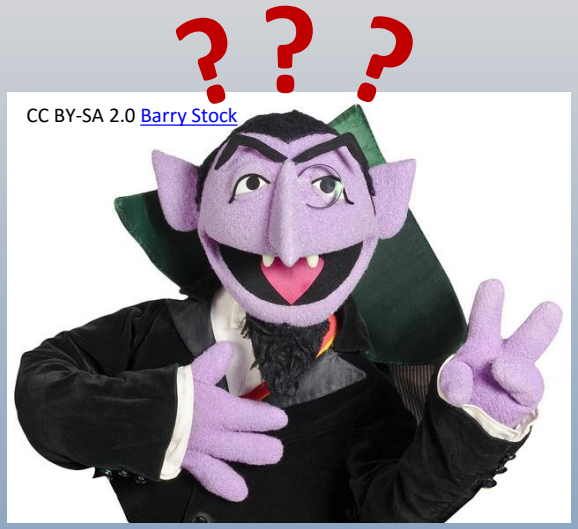
MRMC ROC Program Selection

- **Non-parametric ROC gives bias underestimates with a small number of rating categories**
  - Zweig MH, Campbell G. Receiver operating characteristic plots: a fundamental evaluation tool in clinical medicine. Clin Chem 39:561 (1993).
- **Parametric (semi-parametric) may perform poorly if there are too few samples or if ratings are confined to a narrow range**
  - Metz CE. Practical Aspects of CAD Research Assessment Methodologies for CAD. Presented at the AAPM annual meeting.
- **Only generalizable to population of all observers if observer is treated as a random effect instead of fixed effect**
  - **Similarly, for cases**

Case Selection

- **Comparisons should be on same cases**
  - **Sensitivity 25%-100% depending on case selection**
    - Nishikawa RM, et al. Effect of case selection on the performance of computer-aided detection schemes. Med Phys 21, 265 (1994)
- **The normal case subtlety must be considered to ensure sufficient number of false-positive responses**
  - Rockette, et al. Selection of subtle cases for observer-performance studies: The importance of knowing the true diagnosis (1998).
- **Study disease prevalence does not need to match disease population prevalence**
  - **ROC AUC stable between 2%-28% study prevalence, but small increases in observer ratings are seen with low prevalence**
    - Gur D, et al. Prevalence effect in a laboratory environment. Radiology 228:10 (2003).
    - Gur D, et al. The Prevalence Effect in a Laboratory Environment: Changing the Confidence Ratings. Acad Radiol 14:49 (2007).
  - **Important to ensure a database represents range of disease presentations (e.g., nodules from 3mm-3cm)**


Public Domain
Public Domain
Public Domain
Public Domain
Public Domain
Public Domain



$$\mathbf{CR} = \frac{\mathbf{AUC_1} - \mathbf{AUC_2}}{\boldsymbol{\sigma}(\mathbf{AUC_1} - \mathbf{AUC_2})}$$

- **We need to know:**
  - **Minimum effect size of interest**
    - **Smaller needs more cases for testing**
    - **Appendix C of ICRU 79: ΔSe (at Sp) → ΔAUC**
  - **How much the difference varies**
    - **More variation needs more cases for testing**

- **Sample Size Programs (see references)**
  - **Run a small pilot**
  - **Program uses pilot data and resampling/Monte Carlo simulation to estimate variance for various model componenets (reader, case, etc.)**
- **Typical power 0.8 and α of 0.05**
- **Typical numbers are 3-5 observers and 100 case pairs (near equal for normal/abnormal)**
  - ICRU Report 79

Pilot Data

H0: AUC_A - AUC_B = 0.00, two-sided alternative, 95% significance, 5 Readers, 50 Normal cases, 50 Disease cases.
AUC_A = 0.757, AUC_B = 0.716, AUC_A - AUC_B = 0.041, S.E(total) = 4.561E-2

Significance level 0.05 Effect Size 0.05 #Reader 5 #Normal 50 #Diseased 50
Sizing Analysis: S.E=4.561E-2
Large Sample Approx(Normal) , Power= 0.19

Significance level 0.05 Effect Size 0.05 #Reader 50 #Normal 50 #Diseased 50
Sizing Analysis: S.E=3.954E-2
Large Sample Approx(Normal) , Power= 0.24

Significance level 0.05 Effect Size 0.05 #Reader 50 #Normal 265 #Diseased 265
Sizing Analysis: S.E=1.784E-2
Large Sample Approx(Normal) , Power= 0.80

- **50 observers, 530 cases each . . . Probably pass**

Design Considerations

- Observer training
  - Non-clinical task, specialized software, new modality
- Data/truth verification
  - 45% of truth cases contained errors
    - Armato SG, et al. The Lung Image Database Consortium (LIDC): Ensuring the Integrity of Expert-Defined "Truth" Acad Radiol 14:1455 (2007)
- Display and acquisition
  - Clinical conditions and equipment

General Guidance

- **Bias from re-reading**
  - **A few weeks rule-of-thumb (unless case is unusual)**
  - **Half of readers read subset A first and B second, the other half read B first and A second**
    - Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 24:234 (1989).
    - Metz CE. Fundamental ROC analysis. In: Beutel J, et al. Handbook of medical imaging. Vol 1. Bellingham, WA: SPIE Press, 2000.
- **Observer Experience**
  - **Sensitivity at Sp 0.9 was 0.76 (high volume mammographers) and 0.65 (low volume mammographers)**
    - Esserman L, et al. Improving the accuracy of mammography: volume and outcome relationships. J Natl Cancer Inst 6;94(5):369 (2002)

Instructions to Observers

- According to ICRU Report 79
  - Study description mindful of blinding
  - Types of relevant abnormalities and their precise study definition
  - How to perform task and record data
  - Unique conditions observers should or should not consider
- FROC studies should <u>not</u> indicate how many lesions may be present in a given study

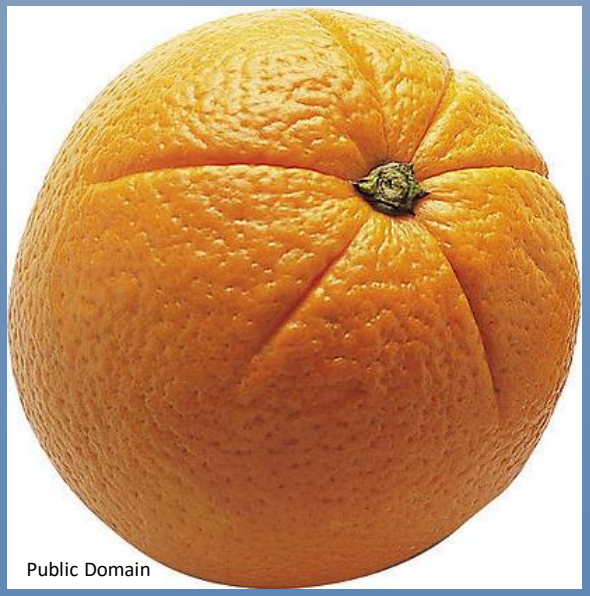
- ROC is costly (time and or money)
- Best used when looking for small to moderate, but important differences
  - ~5% (ICRU Report 79)
  - Bigger difference could be seen with easier testing methodology
  - Smaller differences might be too costly or clinically insignificant

Detection: Anything there?

1. No localization

Bunch of radiologists look at bunch of chest radiographs (CR and DR) to determine if pneumonia is present. ROC determines if the modalities are equivalent.

- **Rating scales and Truth essentially the same as in diagnosis observer study . . .**
- **. . . but the tasks are very different!**


Public Domain


CC 2.0: Abhijit Tembhekar

Widely Used Scale
Definitely or almost definitely abnormal
Probably abnormal
Possibly abnormal
Probably normal
Definitely or almost definitely normal
Swets JA, et al. Assessment of Diagnostic Technologies. Science 205(4408):753 (1979)

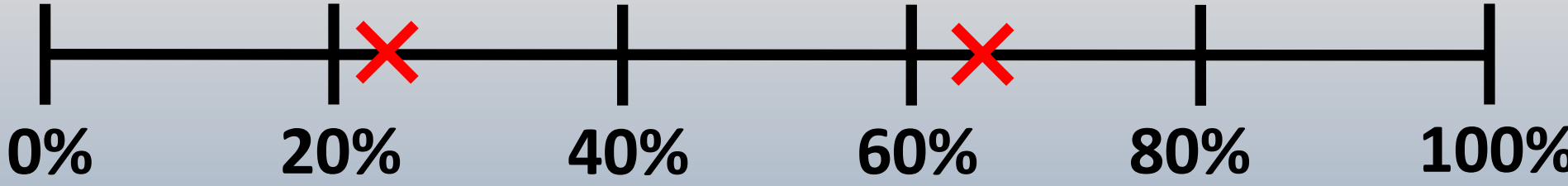
Reduced Observer Variability 20%→7%
Abnormal—diagnosis apparent—warrants appropriate clinical management
Abnormal—diagnosis uncertain—warrants further diagnostic study
I'm not certain—warrants further diagnostic study
Abnormal—but not clinically significant
Normal
Potchen EJ. Measuring Observer Performance in Chest Radiology: Some Experiences. J Am Coll Radiol 3:423 (2006)

- **Clinical relevance reduces variability**

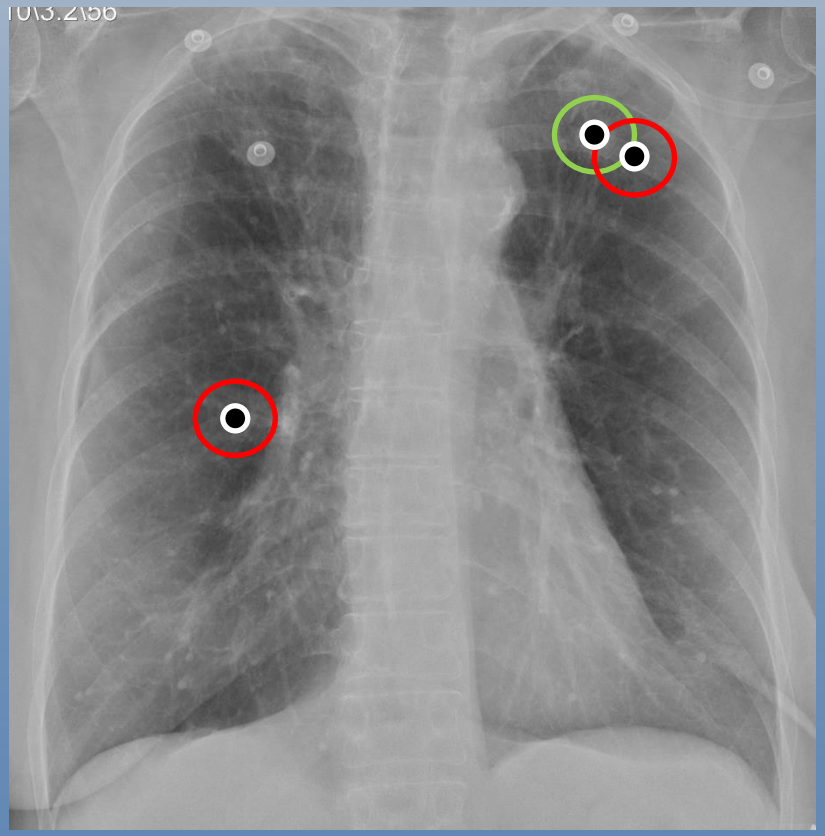
Detection: Anything there? Where?

2. Localization

Bunch of radiologists look at bunch of radiographs with and without CAD system to, mark centroid of nodules if present, and give confidence ratings. FROC determines if CAD helps.



Probability of Malignancy



- Mark lesion centroid
- Determine how close mark must be for “hit”
  - 50% ROI overlap
  - Radius based on size of largest lesion

- Haygood TM, et al. On the choice of acceptance radius in free-response observer performance studies. Br J Radiol 86 (2013)

Delineation: How big? Exactly where?

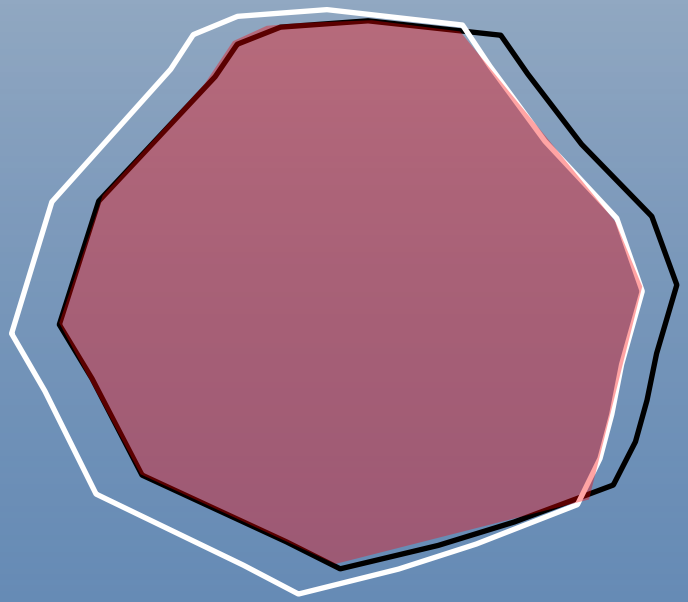
- Bunch of dosimetrists outline the brainstem on CT scans displayed two different window/level settings. “Distance” between outlines is calculated. ANOVA is used to test if outlines are impacted by window/level settings.

Truth

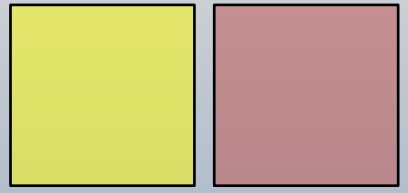
- Phantom
  - Know exact size
  - Clinically relevant?
    - Likely case size of 1 if physical phantom
- Combined outlines on patient images
  - Union/Intersection
  - P-Map
    - Meyer CR, et al. Evaluation of Lung MDCT Nodule Annotation Across Radiologists and Methods. Acad Radiol 13(10): 1254 (2006).
  - STAPLE
    - Warfield SK. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 23(7):903 (2004).

- **Jaccard Similarity Coefficient**

$$J = \frac{|A \cap B|}{|A \cup B|}$$

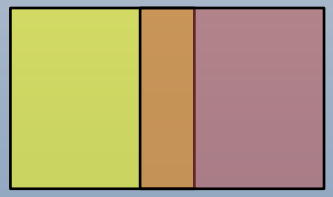


- **Jaccard**
  - **Count pixels in intersection**
  - **Count pixels in union**
  - **Divide intersection by union**
- **Dice**
  - **D = 2J/(1+J)**



0.0

0.0



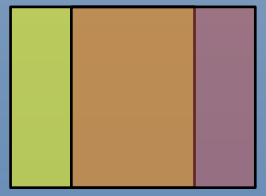
0.33

0.20

Dice

0.50

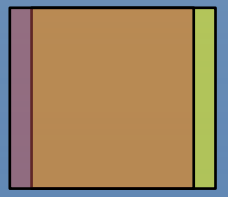
0.33



0.67

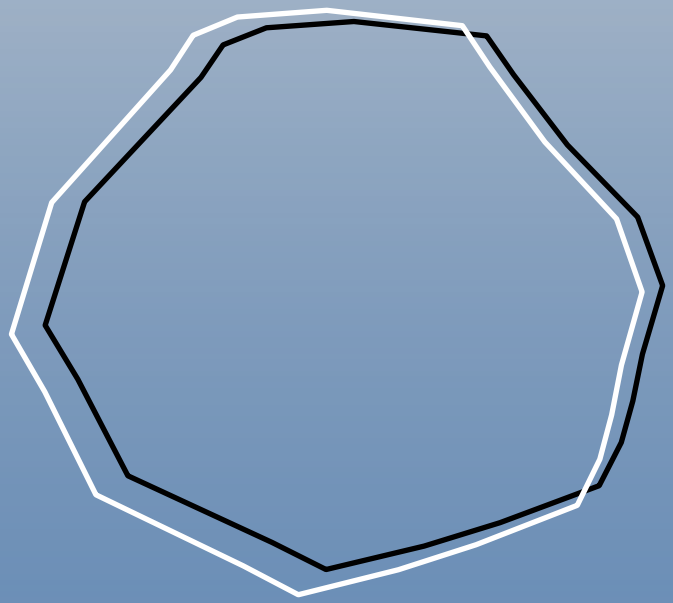
0.50

Jaccard

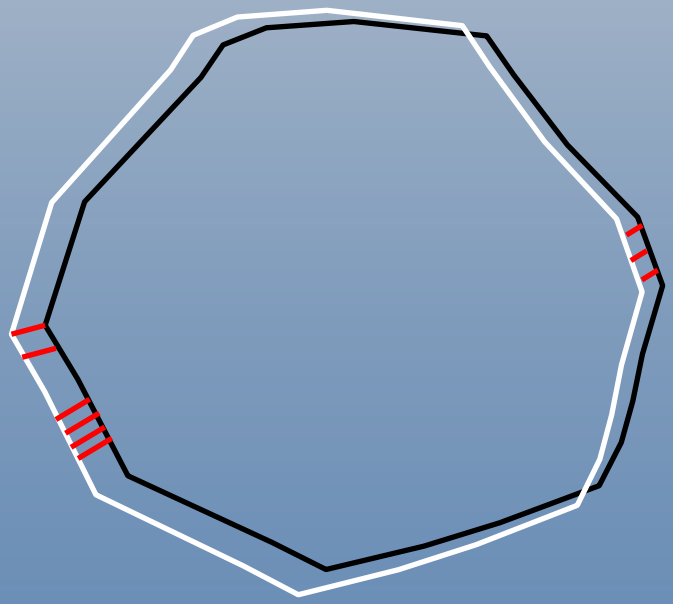


0.8

0.67

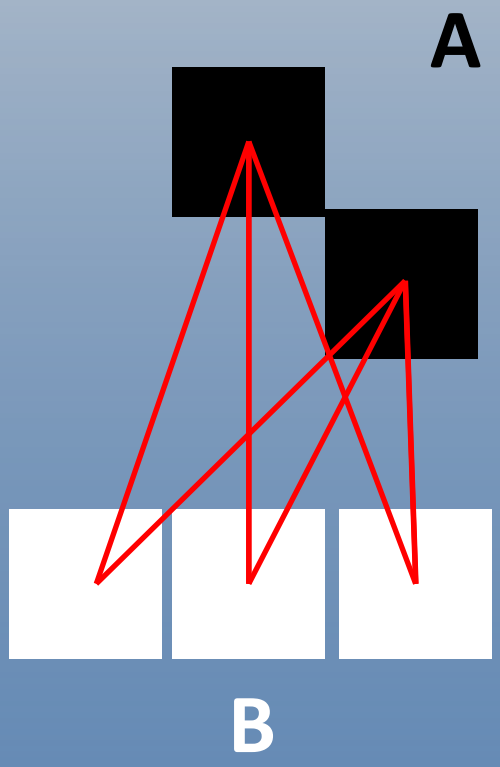


- **Average Distance**
  - **Easy to understand**
  - **Meaningful units**

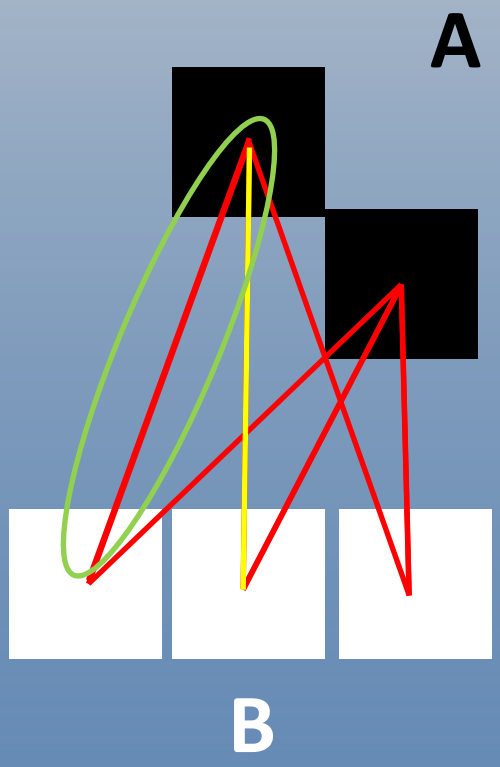


- **Average Distance**
  - **Find shortest absolute distance from each boundary point of A to each the boundary point of B**
  - **Repeat for B to A**
  - **Summary stats**

- **Fail to capture difference**
  - **Dice/Jaccard**
    - **~0.9**
  - **Average distance**
    - **<1mm**



- **Hausdorff distance**
  - **Take a point in A and find the shortest distance to B**
  - **Repeat for all points of A**
  - **Take the maximum of shortest distances**
    - **h(A,B)**



- **Hausdorff distance**
  - Take a point in A and find the shortest distance to B
  - Repeat for all points of A
  - Take the maximum of shortest distances
    - h(A,B)
  - Repeat for h(B,A)
  - Max of h(A,B) and h(B,A)

Design Considerations

- **Prior outlines strongly bias the decisions of expert observers**
  - Sensakovic et al. The influence of initial outlines on manual segmentation. Med Phys. 37(5):2153 (2010).
- **Define where the boundary is drawn and how size is calculated**
  - **???Sensakovic et al.**

- **Summary statistics, correlation, hypothesis testing**
  - **See other talks in session or previous year**

Conclusions

- Many intricacies to running an observer study and properly analyzing using ROC analysis
- Most respected studies in radiology and medicine in general
- Not time for FROC and other important variants . . . references at the end

Further Reading

- Review (ROC and some FROC)
  - ICRU Report 79
  - Wagner RF et al. Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review. Acad Radiol 14: 723 (2007)
  - Chakraborty DP. New Developments in Observer Performance Methodology in Medical Imaging. Semin Nucl Med 41(6): 401 (2011)
- Comparing ROC Methods
  - Obuchowski NA, Beiden SV, Berbaum KS, et al. Multi-reader, multicase ROC analysis: an empirical comparison of five methods. Acad Radiol 2004; 11:980 –995.
  - Toledano A. Three methods for analyzing correlated ROC curves: A comparison in real data sets from multi-reader, multi-case studies with a factorial design. Stat Med 2003; 22:2919 –2933
- Study Design
  - Obuchowski NA. Multireader receiver operating characteristic studies: a comparison of study designs. Acad Radiol 1995; 2:709 –716
  - Potchen EJ. Measuring Observer Performance in Chest Radiology: Some Experiences. J Am Coll Radiol 3:423 (2006)
- Power and Sample Size
  - Hillis et al. Power Estimation for the Dorfman-Berbaum-Metz Method. Acad Radiol 11:1260 (2004).

Further Reading

- **FROC and JAFROC**
  - **Chakraborty DP, et al. Observer studies involving detection and localization: Modeling, analysis, and validation. Medical Physics 31, 2313 (2004).**
  - **Chakraborty DP. New Developments in Observer Performance Methodology in Medical Imaging. Semin Nucl Med 41(6): 401 (2011).**
  - **Thompson JD, et al. Analysing data from observer studies in medical imaging research: An introductory guide to free-response techniques. Radiography 20: 295 (2014).**
  - **Thompson JD, et al. The Value of Observer Performance Studies in Dose Optimization: A Focus on Free-Response Receiver Operating Characteristic Methods. J Nucl Med Technol 41:57 (2013).**

Software (Usually Free)

- http://www.lerner.ccf.org/qhs/software/
- http://metz-roc.uchicago.edu/MetzROC/software
- http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/MRMCAnalysis/tabid/116/Default.aspx
- http://www.devchakraborty.com/index.php
- http://didsr.github.io/iMRMC/
- **Websearch your favorite software package and ROC**

Cite This Talk/Handout

- **Sensakovic, WF. MO-FG-206-02: Implementation and Analysis of Observer Studies in Medical Physics. Med Phys. 43, 3714 (2016); http://dx.doi.org/10.1118/1.4957320**

Image Copyright Requirements

- https://creativecommons.org/licenses/