Training Readers to Use Multi-Level ROC Scores

Brandon D. Gallas, Qi Gong, and Kyle J. Myers

Division of Imaging, Diagnostics, and Software Reliability, OSEL/CDRH/FDA, Silver Spring, MD, USA

Outline

- Introduction to VIPER study
- Examples of typical and bad ROC data
- Designing ROC data
- Reader training
  - Study aims $\neq$ clinical aims
  - Concept of ranking
  - Heuristic link between
    numerical score and clinical decisions
- VIPER eCRF
- VIPER score distributions
- Summary

Context: VIPER:

Validation of imaging premarket evaluation and regulation

- Aim 1:
  - Significant (AUC) difference in large prospective clinical trial
    can be achieved in a
  - Small, controlled, cancer-enriched lab study.
- Aim 2:
  - Sensitivity-specificity in a large prospective trial
    compared to
  - Small controlled lab study.
- Secondary aims:
  - Variability with focus on the reader

- Application Area: Breast cancer detection
  - SFM vs. FFDM
  - DMIST reported for women with dense breasts
    AUC(FFDM) = 0.78 > 0.68 = AUC(SFM)

Context: VIPER:

Validation of imaging premarket evaluation and regulation

- Study populations
  - Low prevalence (10%) screening study
  - Moderate prevalence (31%) screening study
  - Moderate prevalence (31%) challenge study
  - Moderate prevalence (50%) screening study
  - Moderate prevalence (50%) challenge study

DMIST:

- Prevalence = 0.7%

- Prevalence achieved by cancer enrichment
- Screening study:
  - Normal population: match screening
- Challenge study:
  - Normal population: only BI-RADS 0

DMIST:

- BI-RADS 1-2 = 91%
- BI-RADS 0 = 9%

Extrapolating Data:

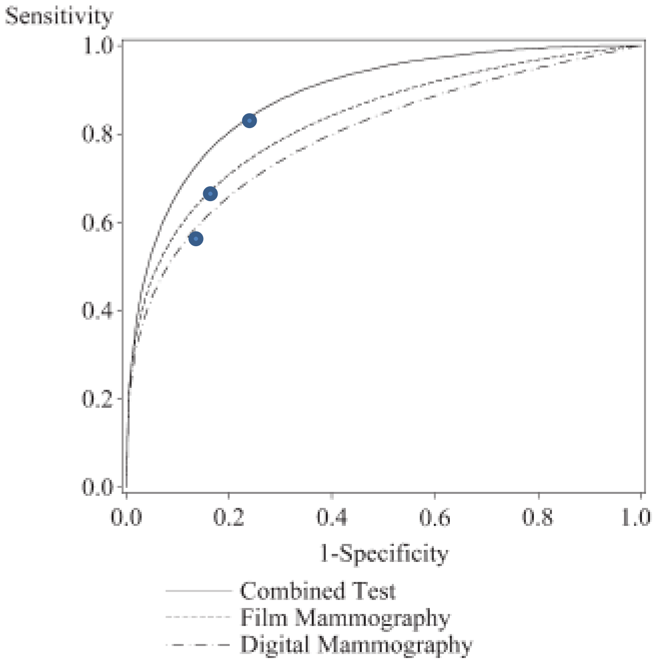
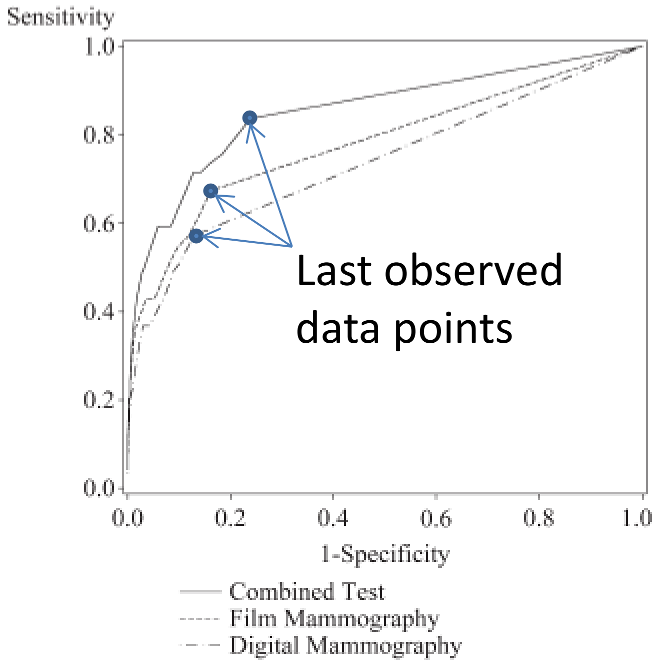


Figure 1. ROC curves for film mammography results, digital mammography results, and the combined test results in the parametric analysis.



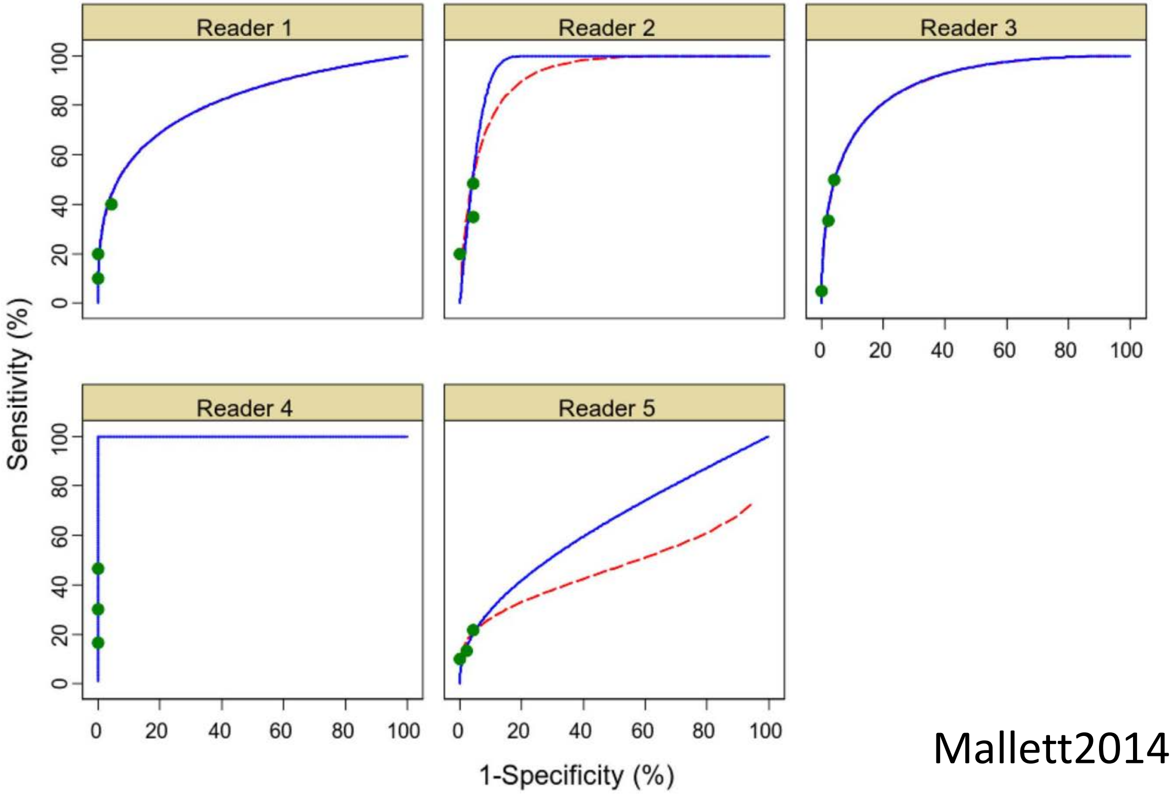
Last observed data points

Figure 2. Graph of AUC lines for ROC nonparametric analysis.

Glueck2007_Acad-Radiol_v14p670

Polyp-based Scores

- Figure 4. Different curve fitting methods. ROC plots each for an individual reader using CT colonography without CAD. Green dots indicate real data points underlying curve fitting. ROC curve are shown extrapolated from these data using DBM MRMC (red dotted line) and PROPROC software (blue solid line). Five readers are shown in plots labelled 1 to 5.



Mallett2014_PLoS-One_v9pe107633

Design ROC Data

- Involve readers
  - Understand clinical task, workflow, and language
  - Identify and isolate binary task
  - Develop task for reader study
  - Develop training for reader study
- Two-step process
  - Step 1: Ask clinical question first
  - Step 2: Then get more information

Instructions for Reporting the *"Numeric Score"*

- Establish study aim
  - IS to evaluate imaging technology or perception!
  - IS NOT to evaluate clinical performance

- "Numeric Score" IS NOT
  - Part of standard clinical report
  - Previously defined
  - Previously trained

- "Numeric Score" IS
  - More Information
  - More Quantitative
  - For evaluating the imaging technology

Instructions for Reporting the *"Numeric Score"*

- Establish the concept of ranking

- The "Numeric Score" arises from:
  - *Given two cases, which one is more suspicious for cancer?*

- Given two cases you would recall:
  - *Could you decide which one is more suspicious for cancer?*
- Given two cases you would not recall:
  - *Could you rank one as more suspicious than the other?*
- Given a stack of cases:
  - *Could you sort them from the least suspicious to the most suspicious (allowing for some ties)?*

Instructions for Reporting the *"Numeric Score"*

- Establish heuristic link between
  - "Numeric Score"

  And
  - Clinical task, workflow, and language

- Step 1: Recall and No Recall Decision
- Step 2: Score
  - Treat Recall and No Recall separately
  - Provide language and anchors that illustrate and help being quantitative
  - *Scoring is challenging and unfamiliar; just do your best*

Recall scores easier to link to
Level of confidence,
Likelihood,
Probability
No recall scores harder
No confidence,
No likelihood,
No probability

Data Collection
Case Report Form

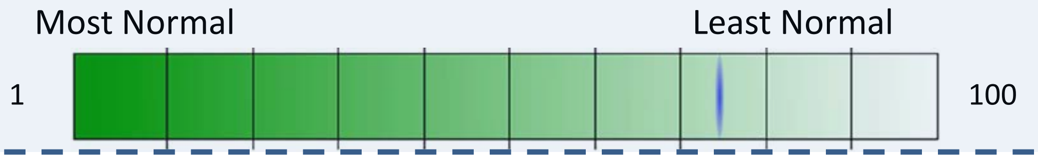
- Force consistency
- Summarize key instructions
- Allow digital and analog scales

Would you recall the 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?



Numeric Score

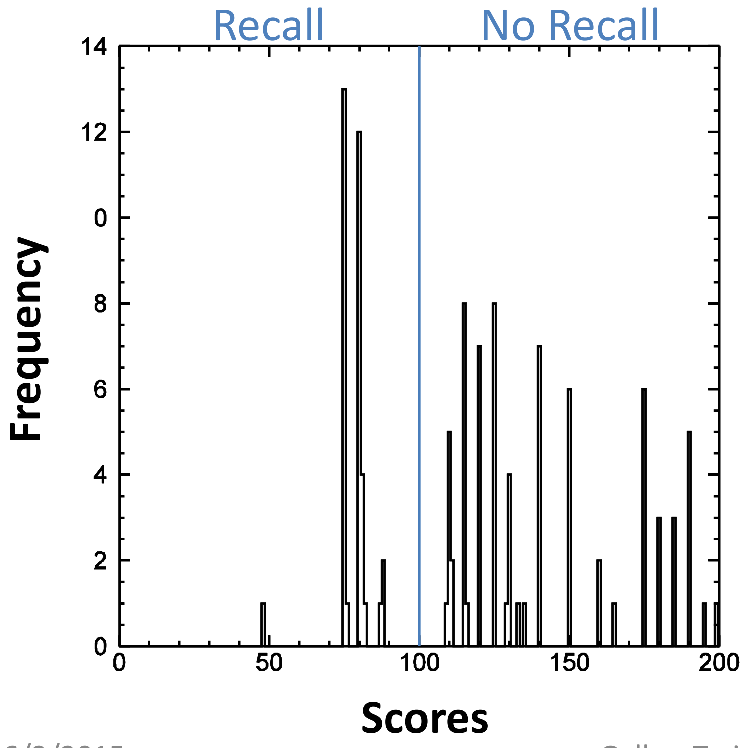
□□□

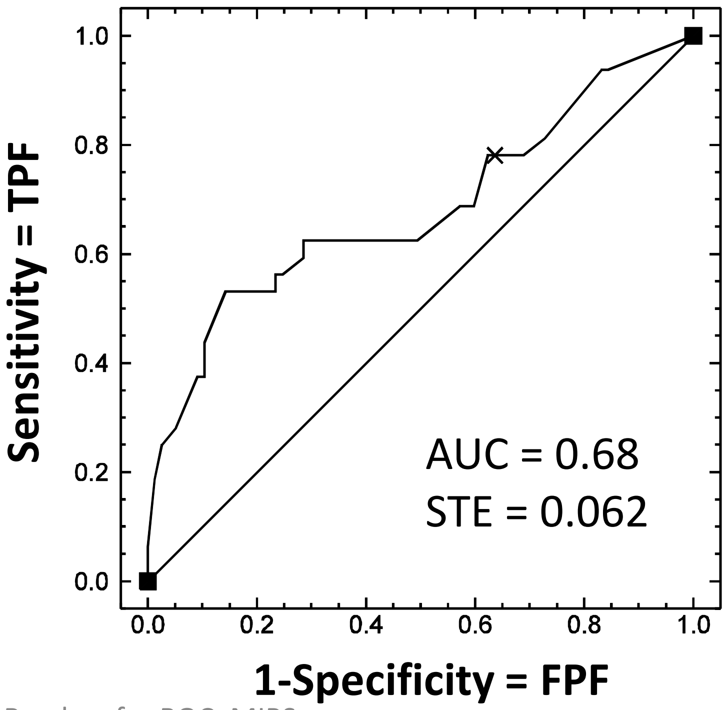
Results Reader 01, FFDM

Challenge study (prevalence = 0.30)

nCases = 109, nBinsUsed = 29

- This reader used the fewest bins



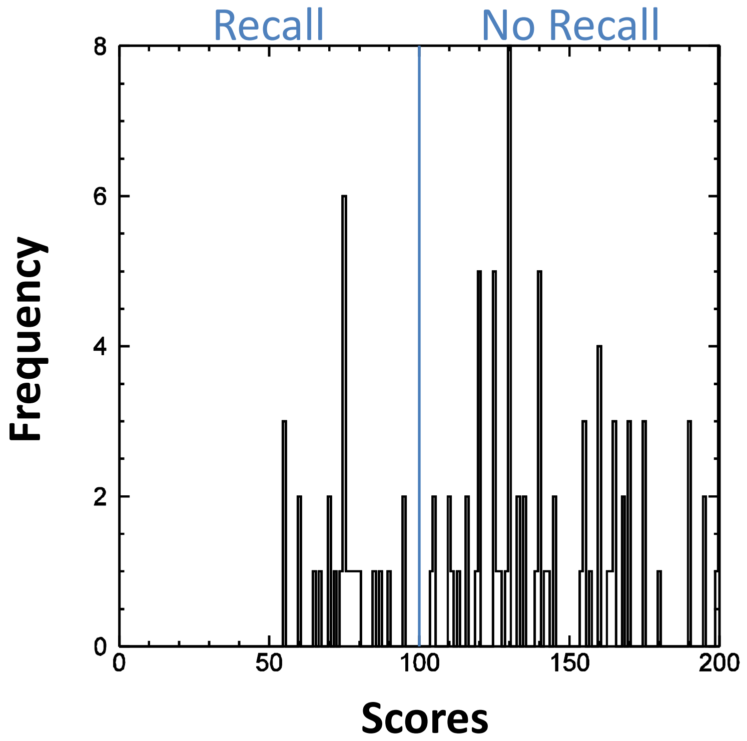


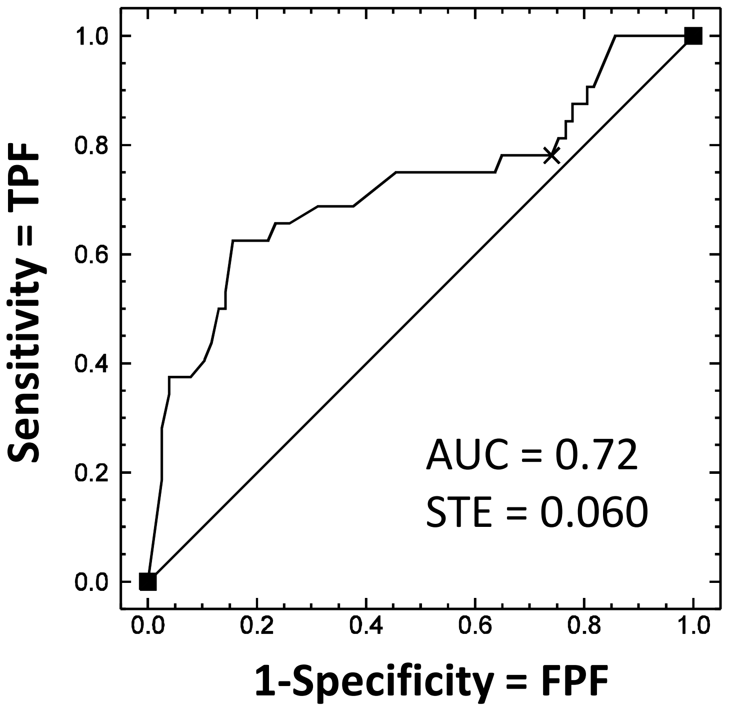
Results Reader 02, FFDM

Challenge study (prevalence = 0.30)

nCases = 109, nBinsUsed = 52

- This reader used a moderate number of bins



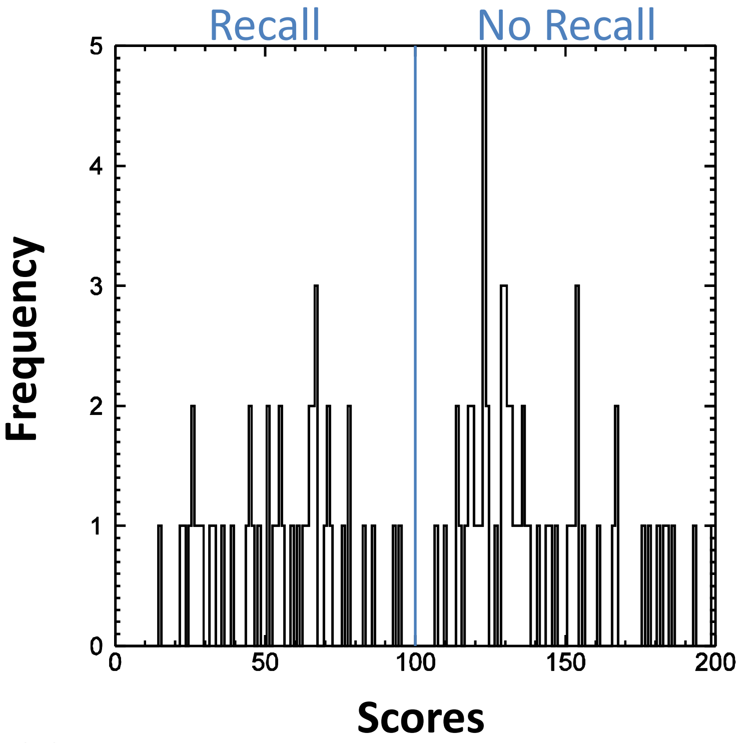


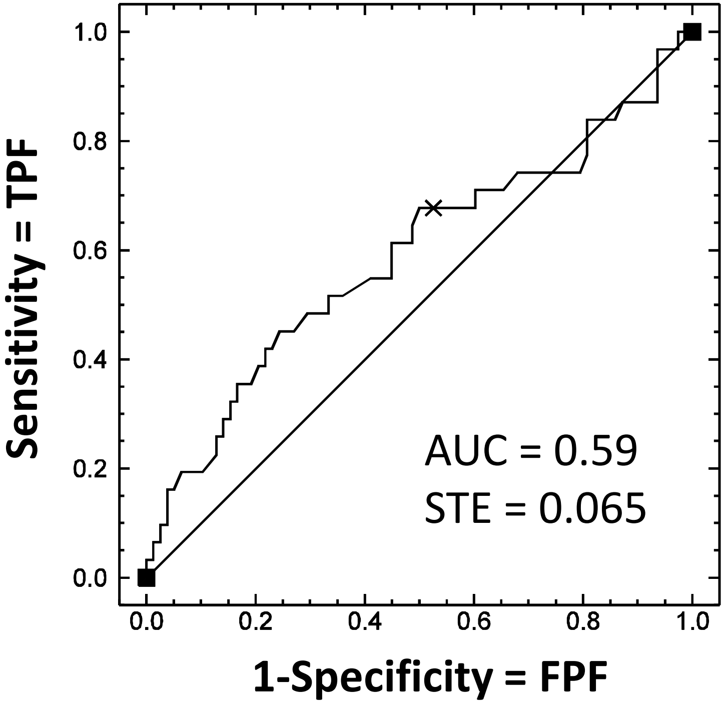
Results Reader 13, FFDM

Challenge study (prevalence = 0.30)

nCases = 109, nBinsUsed = 81

- This reader used the most bins



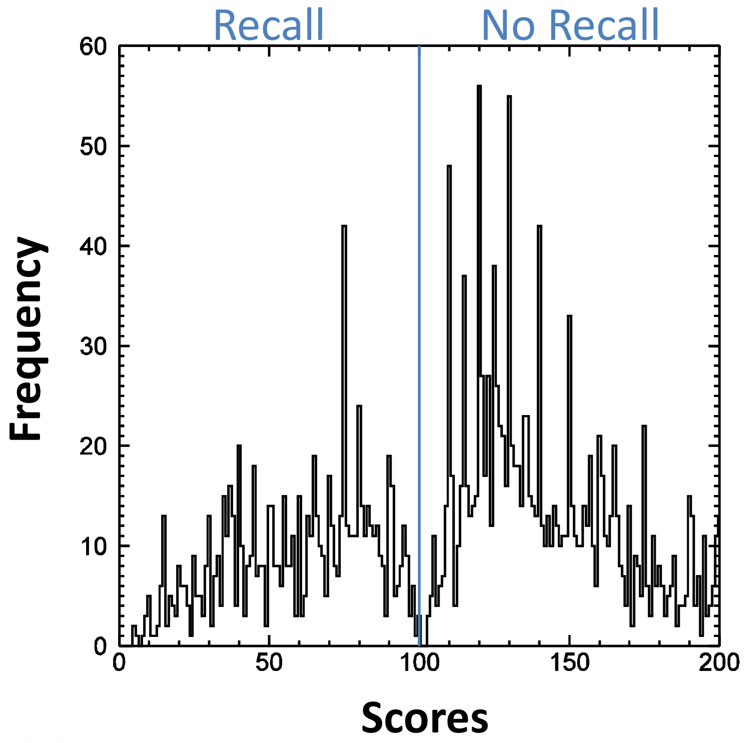


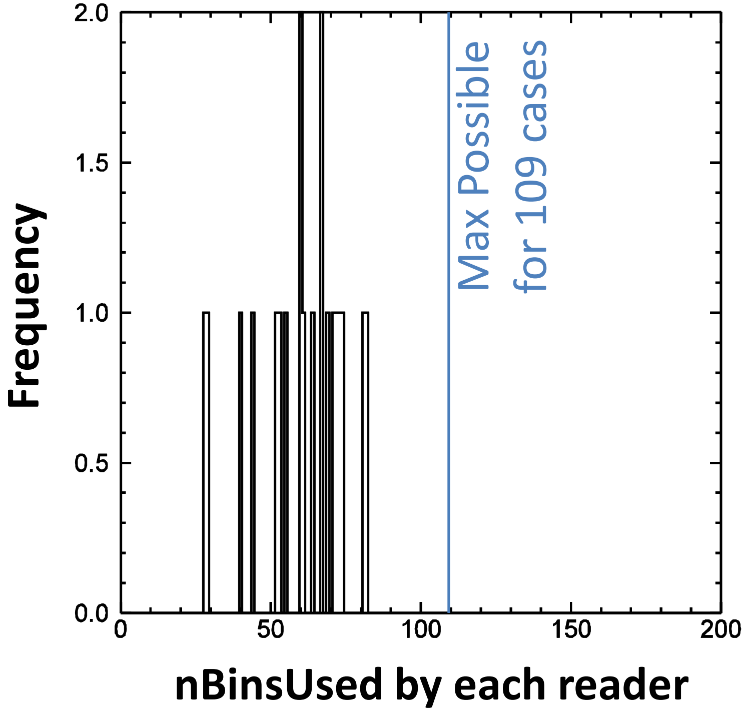
Summarize All Readers, FFDM

Challenge study (prevalence = 0.30)

nReaders = 20, nObservations = 2188, nBinsUsed = 193

- Most readers use between 50 and 80 bins

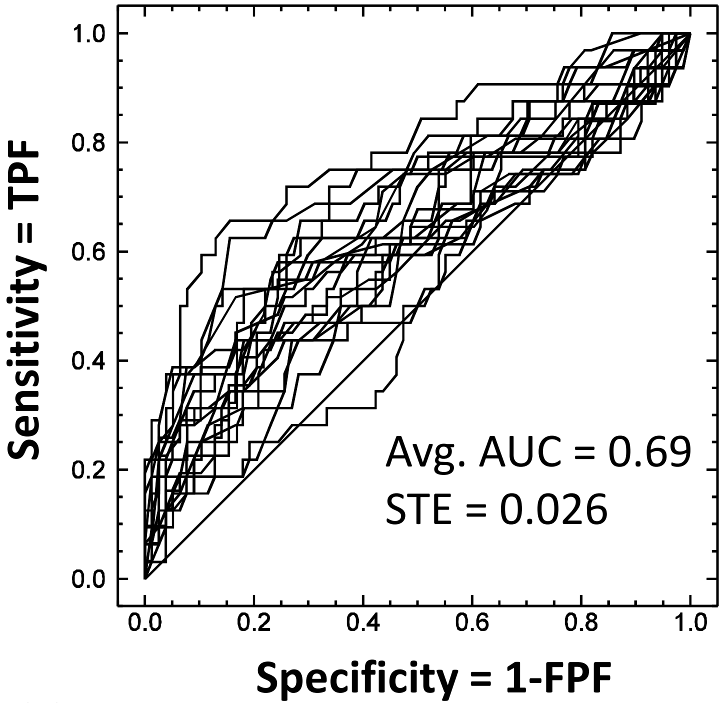


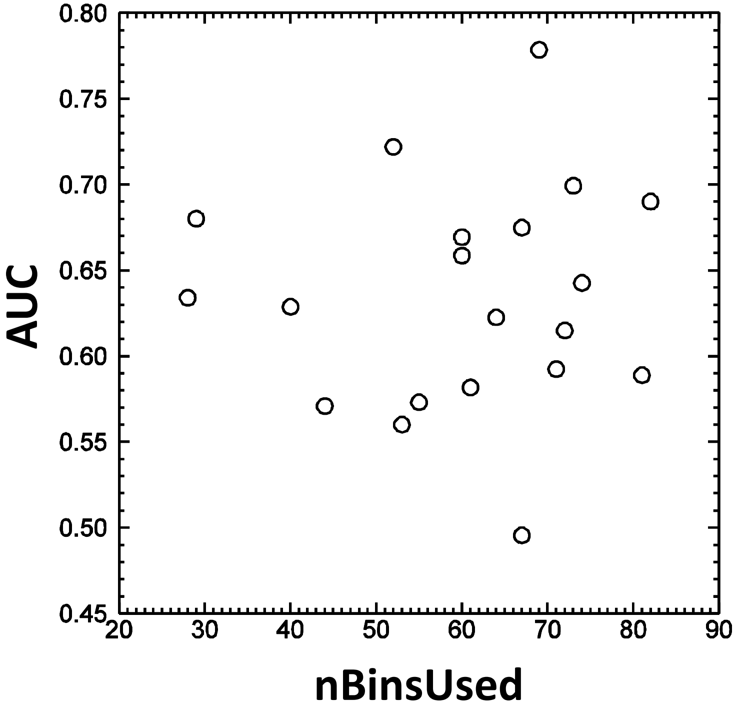


Summarize All Readers

Challenge study (prevalence = 0.30)
nReaders = 20

- nBinsUsed and AUC are uncorrelated





Summary

- Radiologists can be trained to use many scores
  - Involve readers in developing training
  - Explain: Not doing clinical task
  - Describe concept of ranking
  - Link to ROC score to clinical task
  - Provide strategies and anchors for being quantitative

- Training and data-collection strategy effectively sampled ROC space

- On the web:
  - https://code.google.com/p/imrmc/wiki/iMRMCGuide
  - Full scoring instructions
  - Mockup of Case Report Form
  - Data (eventually)

Google code is shutting down. We are likely moving to Github.