Multi-Reader, Multi-Modality Studies

RSNA Clinical Trials Workshop

Jeffrey D. Blume
School of Data Science
University of Virginia

1

Outline

- Variability: observers / readers / sites
- Agreement
- Accuracy
- ROC curves in Multi-reader studies
- Issues with sample size projections

2

Variability differs

- Identify sources of variability
- Variation in radiologist performance is reflected in
  - ROC, AUC, Sens, Spec, Agreement
- Training, routine, experience, all contribute
  - Random (due to random variation)
  - Explained (due to knowable factors, i.e. experience)
- Possible to adjust for explained variation (simplest approach is to stratify)
- Good studies identify sources of variability

3

Agreement

- Agreement does not imply accuracy or truth
- Measures of Agreement/Reliability
  - Kappa, weighted kappa
  - Multireader kappa
  - Intraclass correlation coefficient
- Graphical assessment: Bland-Altman plot
  - Two observers only
  - Plots difference of scores versus average
  - Look for lack of patterns

4

Combined Screening With Ultrasound and Mammography vs Mammography Alone in Women at Elevated Risk of Breast Cancer

Berg, Blume, et. al. JAMA, 2008

2809 Women enrolled and randomized to order of imaging studies

		Integrated mammography plus ultrasound test results[a] +	Integrated mammography plus ultrasound test results[a] −	Total
Mammography test results	+	110	31	141
	−	216	2355	2571
Total		326	2386	2712

K = 0.43 (0.38,0.469)

2637 Had reference standard (analysis set)

Positive reference standard

		Integrated test results +	Integrated test results −	Total
Mammography test results	+	19	1	20
	−	12	8	20
Total		31	9	40

K = 0.35 (0.04,0.45)

Negative reference standard

		Integrated test results +	Integrated test results −	Total
Mammography test results	+	88	28	116
	−	187	2294	2481
Total		275	2322	2597

K = 0.413 (0.36,0.46)

*Actual analysis incorporated reference standard

5

Data/Response	Agreement Measure	What it measures	Extensions
Dichotomous or Categorical	Kappa	Percent agreement corrected for chance ('The diagonal')	Multireader Kappa
Ordered Categorical	Weighted Kappa	Percent agreement corrected for chance, but partial credit is given for being 'close' ('The diagonal' + partial credit for close 'off diagonals')	Weighted Multi-reader Kappa
Continuous	Intraclass Correlation Coefficient (ICC)	Proportion of total response variance due to readers	Analysis and reporting of variance components

6

Data/Response	Agreement Measure	Scale
Dichotomous	-1 < Kappa <= 1	> 0.75 is excellent > 0.40 & <0.75 is fair to good < 0.40 is poor
Categorical	Weighted Kappa	Same as Kappa
Continuous	0 <= ICC <= 1	> 0.85 nearly perfect reliability > 0.75 & <0.85 excellent reliability > 0.60 & <0.75 good reliability > 0.40 & <0.60 fair reliability < 0.40 poor reliability

7

Inference for agreement

- Most statistical packages test the hypothesis that the agreement statistic, such as kappa, is zero
- This is effectively useless
- Avoid this problem by reporting the confidence interval for kappa or ICC
- Sample size projections based on these statistics are complex

8

Radiology

Early Invasive Cervical Cancer: CT and MR Imaging in Preoperative Evaluation—ACRIN/GOG Comparative Study of Diagnostic Performance and Interobserver Variability[1]

Hedvig Hricak, MD, PhD
Constantine Gatsonis, PhD
Fergus V. Coakley, MD
Bradley Snyder, MS
Caroline Reinhold, MD
Lawrence H. Schwartz, MD
Paula J. Woodward, MD
Harpreet K. Panna, MD
Marco Amendola, MD

Purpose: To retrospectively compare diagnostic performance and interobserver variability for computed tomography (CT) and magnetic resonance (MR) imaging in the pretreatment evaluation of early invasive cervical cancer, with surgical pathologic findings as the reference standard.

Materials and Methods: This HIPAA-compliant study had institutional review board approval and informed consent for evaluation of preoperative CT (n = 146) and/or MR imaging (n = 152) studies in 150 women (median age, 43 years; range, 22–81 years) from

CT: 4 readers, 146 cases
MR: 4 readers, 152 cases

Radiology 2007; 245:491–498

9

Table 1

Reader Agreement in Retrospective Interpretation of CT and MR Imaging Studies

Parameter	Multirater κ Value*		P Value†	
	CT	MR Imaging	CT	MR Imaging
Tumor visualization	0.16 (0.12 to 0.29)	0.32 (0.22 to 0.41)	<.001	<.001
Invasion of right parametrium	−0.04 (−0.02 to 0.13)	0.10 (0.06 to 0.27)	.961	<.001
Invasion of left parametrium	−0.05 (−0.01 to 0.11)	0.12 (0.05 to 0.29)	.981	<.001
Overall parametrial invasion‡	−0.04 (−0.02 to 0.13)	0.11 (0.05 to 0.29)	. . .	. . .
Staging§	0.26 (0.23 to 0.34)	0.44 (0.34 to 0.56)	<.001	<.001

* Data in parentheses are ranges of pairwise κ values. A pairwise κ value of greater than 0.00 but less than 0.40 was considered to represent poor agreement; a value of 0.40–0.75, fair to good agreement; and a value greater than 0.75, excellent agreement (8).

† For testing whether multirater κ values were significantly greater than zero.

‡ Average of multirater κ values in left and right parametrium. (Data in parentheses are ranges of pairwise values over both left and right parametrium.)

§ For staging of tumors as IIA or lower versus IIB or higher.

10

Eur Radiol
DOI 10.1007/s00330-008-1191-7

NEURO

Birgit B. Ertl-Wagner
Jeffrey D. Blume
Donald Peck
Jayaram K. Udupa
Benjamin Herman
Anthony Levering
Ilona M. Schmalfuss
The members of the ACRIN 6662 study group

Reliability of tumor volume estimation from MR images in patients with malignant glioma. Results from the American College of Radiology Imaging Network (ACRIN) 6662 Trial

- **Retrospective reader study, designed to assess the value of two semi-automated systems for calculating volumes of brain tumors on MR images, in patients with new, postoperative, and recurrent malignant gliomas.**
- **16 readers evaluated 24 cases on each platform.**

11

Table 4 95% prediction intervals, sum of variances and ICCs depending on the level of professional expertise of the reader (*CI* confidence interval)

	3DVIEWNIX-TV		Eigentool		Manual	
	Staff/Fellows	Technologists	Staff/Fellows	Technologists	Staff/Fellows	Technologists
95% prediction interval (95% CI)						
Gd	51 (40, 61)	53 (42, 61)	86 (73, 98)	46 (39, 51)	211 (185, 235)	153 (134, 171)
FLAIR	181 (111, 231)	179 (115, 226)	104 (86,119)	96 (80, 109)	250 (202, 290)	364 (294, 422)
Sum of variances (95% CI)						
Gd	172 (105, 238)	181 (117, 245)	485 (349, 622)	135 (102, 169)	2,915 (2,223, 3,607)	1,534 (1,161, 1,908)
FLAIR	2,132 (798, 3,467)	2,092 (854, 3,328)	697 (478,917)	597 (418, 775)	4,064 (2,656, 5,473)	8,617 (5,620, 11615)
ICC (95% CI)						
Gd	0.53 (0.42, 0.65)	0.465 (0.31, 0.62)	0.29 (0.10, 0.48)	0.187 (0.02, 0.35)	0.145 (0.00, 0.30)	0.170 (0.07, 0.27)
FLAIR	0.98 (0.97, 0.99)	0.920 (0.88, 0.96)	0.37 (0.25, 0.50)	0.335 (0.17, 0.50)	0.445 (0.32, 0.58)	0.45 (0.31, 0.59)

12

Table 1 Summary of the variance components contributing to the total variance when estimating volume differences in the FLAIR hyperintensity over time with the respective methods (*LCB* lower confidence bound, *UCB* upper confidence bound)

3DVIEWNIX-TV

	Variance component	95% LCB	95% UCB	% total (ICC)	95% LCB	95% UCB
Cases	2,009.08	1,895.21	2,128.23	0.95	0.92	0.98
Readers	0.00	0.00	19.28	0.00	0.00	0.01
Error	102.82	38.95	156.53	0.05	0.02	0.07
Total	2111.89	–	–	–	–	–

Eigentool

	Variance component	95% LCB	95% UCB	% total (ICC)	95% LCB	95% UCB
Cases	226.53	175.53	344.45	0.35	0.26	0.57
Readers	1.66	0.00	79.44	0.00	0.00	0.10
Error	422.40	206.33	564.02	0.65	0.42	0.69
Total	650.60	–	–	–	–	–

Manual

	Variance component	95% LCB	95% UCB	% total (ICC)	95% LCB	95% UCB
Cases	2,782.05	1,655.03	4,790.47	0.44	0.33	0.61
Readers	0.00	0.00	425.42	0.00	0.00	0.06
Error	3,564.82	2,125.08	4,455.45	0.56	0.37	0.66
Total	6,346.87	–	–	–	–	–

13

Table 2 Summary of the variance components contributing to the total variance when estimating volume differences of the Gd-enhancing lesion over time with the respective methods

3DVIEWNIX-TV

	Variance component	95% LCB	95% UCB	% total (ICC)	95% LCB	95% UCB
Cases	89.08	74.76	118.47	0.51	0.40	0.71
Readers	0.00	0.00	10.22	0.00	0.00	0.06
Error	87.05	38.58	129.27	0.49	0.28	0.59
Total	176.13	–	–	–	–	–

Eigentool

	Variance component	95% LCB	95% UCB	% total (ICC)	95% LCB	95% UCB
Cases	76.84	26.59	196.72	0.25	0.12	0.53
Readers	0.00	0.00	43.45	0.00	0.00	0.09
Error	232.64	73.63	379.43	0.75	0.44	0.85
Total	309.47	–	–	–	–	–

Manual

	Variance component	95% LCB	95% UCB	% total (ICC)	95% LCB	95% UCB
Cases	397.98	283.37	1,118.21	0.18	0.14	0.39
Readers	0.00	0.00	383.66	0.00	0.00	0.10
Error	1,825.26	617.52	3,225.44	0.82	0.58	0.84
Total	2,223.23	–	–	–	–	–

14

Pulmonary Nodules: Interscan Variability of Semiautomated Volume Measurements with Multisection CT— Influence of Inspiration Level, Nodule Size, and Segmentation Performance

Gietema et al, Radiology 2007

20 patients, scanned twice with low dose CT

Conclusion**:** ***Variation of semiautomated volume measurements of pulmonary nodules can be substantial. Segmentation represents the most important factor contributing to measurement variability. Change in inspiration level has only a weak effect for completely segmented nodules.***

15

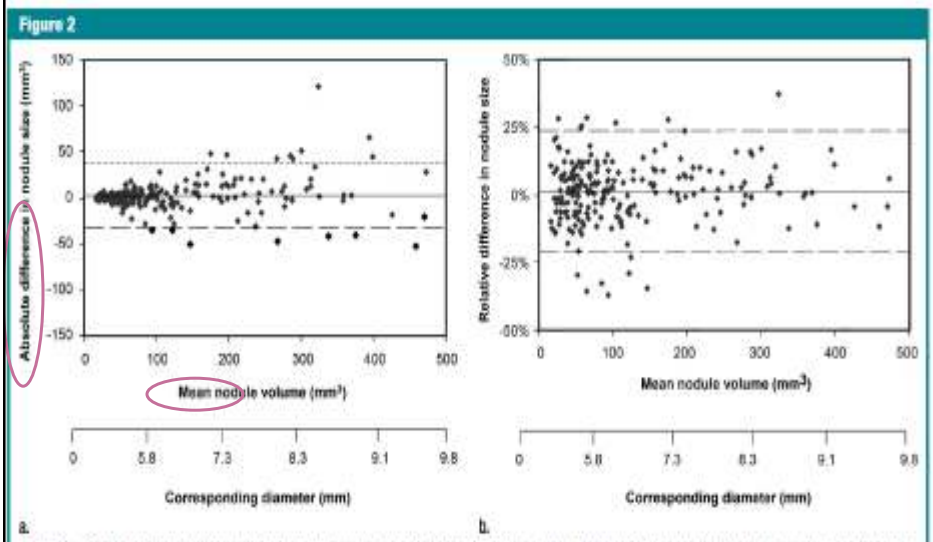


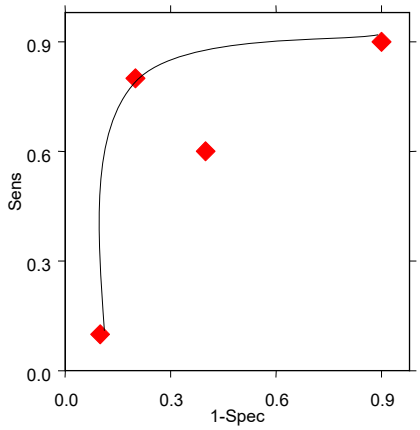
Figure 2: Graphs show interscan agreement of volume measurements of all nodules. (a) Absolute and (b) relative differences between both measurements are plotted against mean nodule volume. Mean difference is shown by continuous line; upper and lower limits of agreement are shown by dashed lines.

16

Need to account for threshold

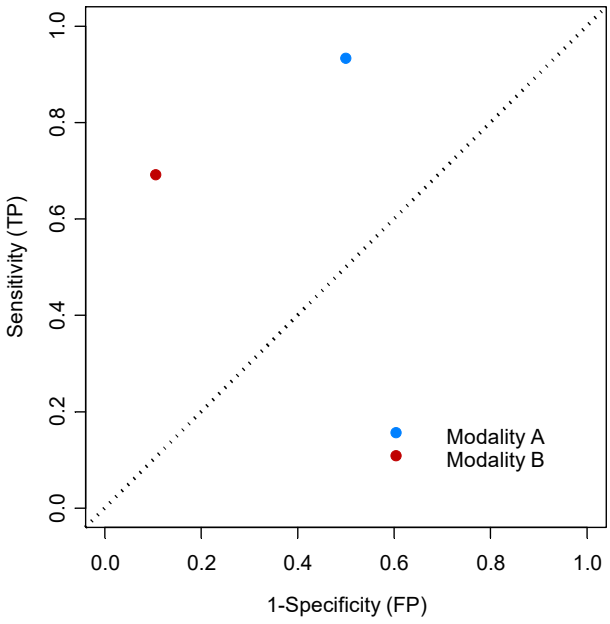
Averaging sensitivities and specificities can be misleading

	Sens	Spec
	.10	.90
	.80	.80
	.90	.10
Mean	.60	.60

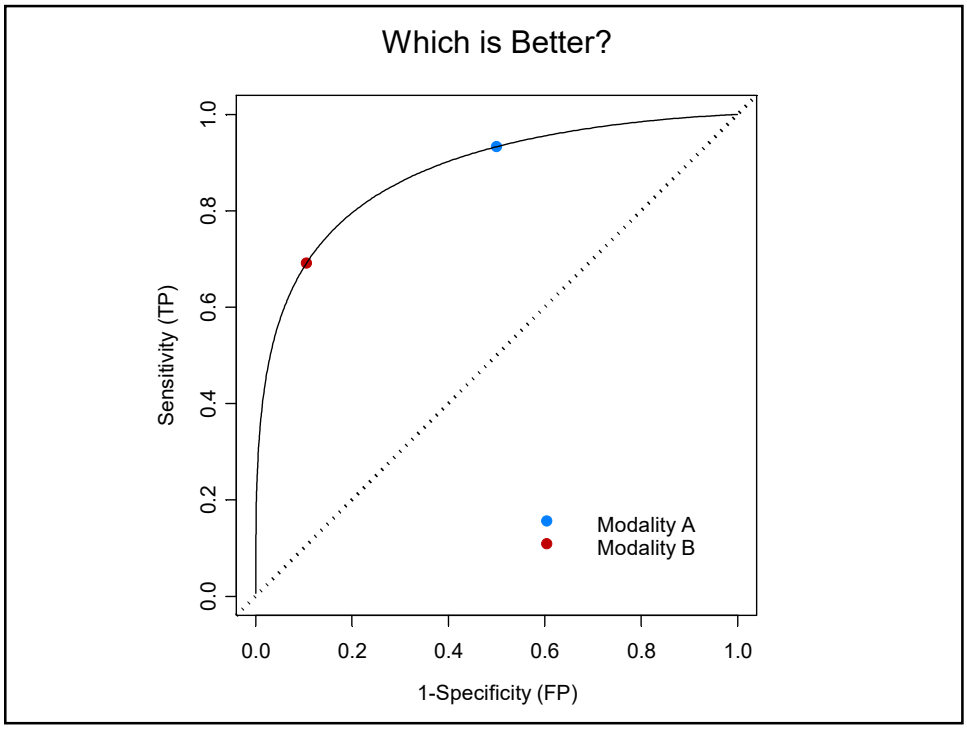


17

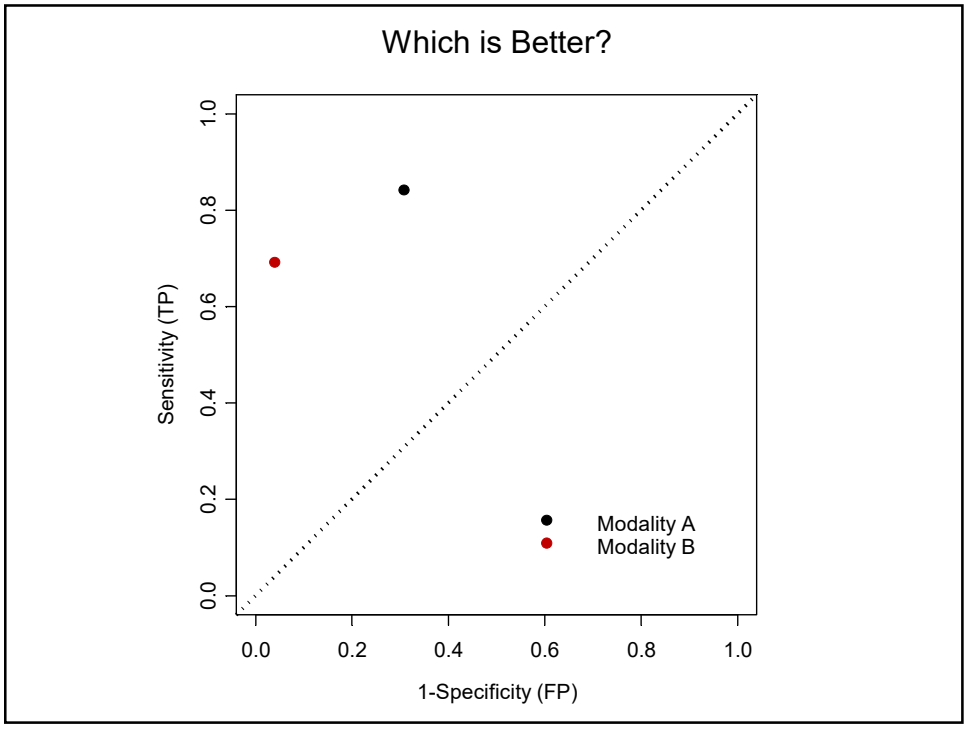
Which is Better?



18

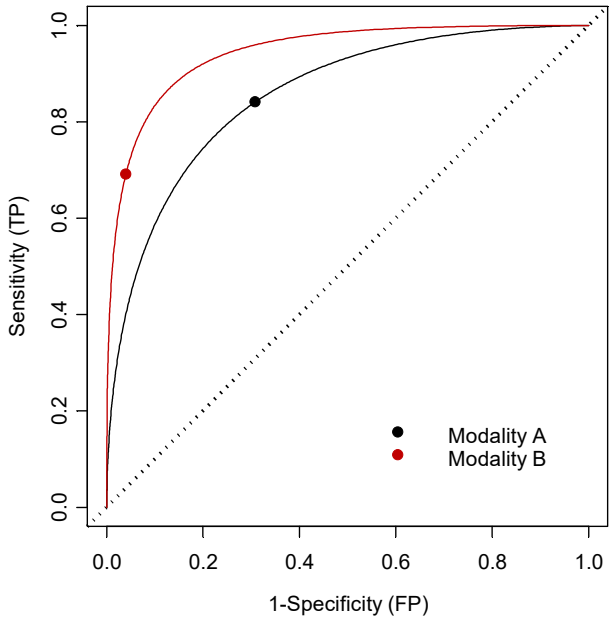

Which is Better?
Sensitivity (TP)
1-Specificity (FP)
0.0
0.2
0.4
0.6
0.8
1.0
Modality A
Modality B

19


Which is Better?
Sensitivity (TP)
1-Specificity (FP)
0.0
0.2
0.4
0.6
0.8
1.0
Modality A
Modality B

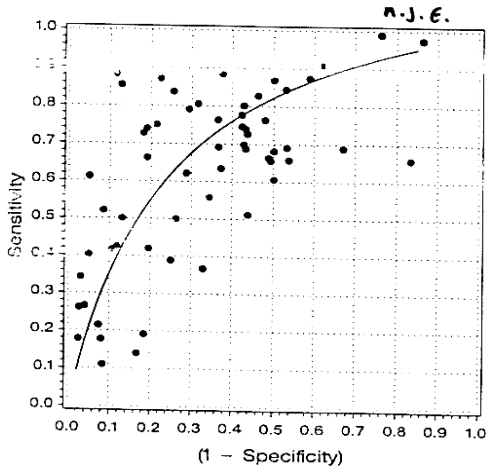
20

Which is Better?



21

Accuracy of Pap test, (Fahey et al, Am J Epi, 1996)



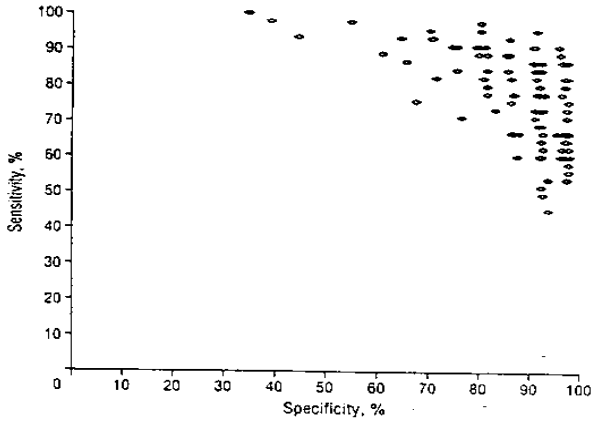
22

Variability is the focus

- Multi-reader studies: involve multiple readers interpreting common sets of imaging studies derived by one or more diagnostic modalities
- Uniform imaging protocol
- Studies can not assess/estimate reader variability unless they use same set of cases
- Variety of available analytic approaches

23

Performance of mammographers interpreting a common set of images (Beam, Arch Int Med, 1996)



24

CT and MRI for cervical cancer

Table 4

Detection of Advanced Stage (≥IIB) Cancer by Retrospective Readers of CT and MR Imaging Studies

Parameter	CT*	MR Imaging*	*P* Value
Mean sensitivity	0.28 (0.14–0.38)	0.47 (0.40–0.57)	.104
Mean specificity	0.90 (0.84–1.00)	0.79 (0.77–0.80)	.099
Mean positive predictive value	0.55 (0.38–1.00)	0.36 (0.32–0.39)	.001
Mean negative predictive value	0.83 (0.81–0.84)	0.85 (0.83–0.87)	.305

* Data in parentheses are ranges over the readers.

25

Accuracy of CT Colonography for Detection of Large Adenomas & Cancers

- ACRIN; prospective; 2600 asymptomatic participants; 15 radiologists (Johnson et al. NJEM, Sept 2008)
- Sensitivity for the detection of adenomas or cancers measuring 10 mm or more in diameter (based on the identification of all lesions measuring 5 mm or more)
- Graph: radiologists are ordered according to the total number of cases read; the size of each square (point estimate) is proportional to the square root of the total number of cases read. The number of positive cases (at least one adenoma or cancer 10 mm) is shown below each confidence interval.

26

Variability among readers in NCTC study

Prospective design,

2600 asymptomatic participants

15 radiologists

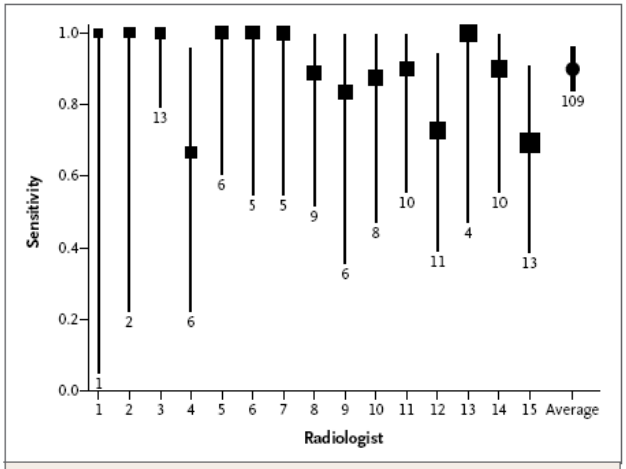


Figure 1. Individual Estimates of the Sensitivity of CT Colonography for the Detection of Adenomas or Cancers.

Johnson et al. NEJM, 2008

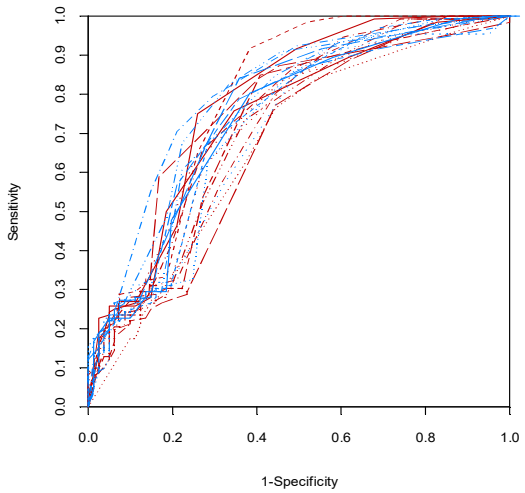
27

Reader Study: Breast MR w/ & w/o CAD

- *Accuracy and Efficiency of Computer Aided Diagnostics Among Novice and Expert Breast MRI Readers.* AJR. Lehman CD, Blume JD, et al.
- Aim: To compare the diagnostic accuracy of breast MR imaging interpretation with and without a computer-aided diagnostic (CAD) system in novice and expert readers.
- 20 readers reading with CAD and without CAD.
- 9 experts and 11 novices
- 70 cases, 27 were benign and 43 were malignant
- Test result scale: Probability of Malignancy scale (5 cats)

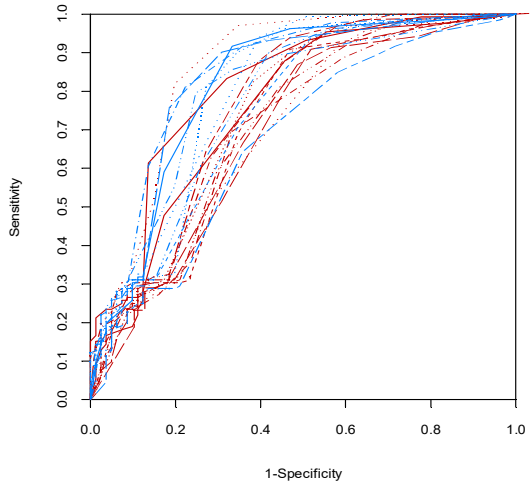
28

ROC **without** CAD (expert in blue and novice in red)



29

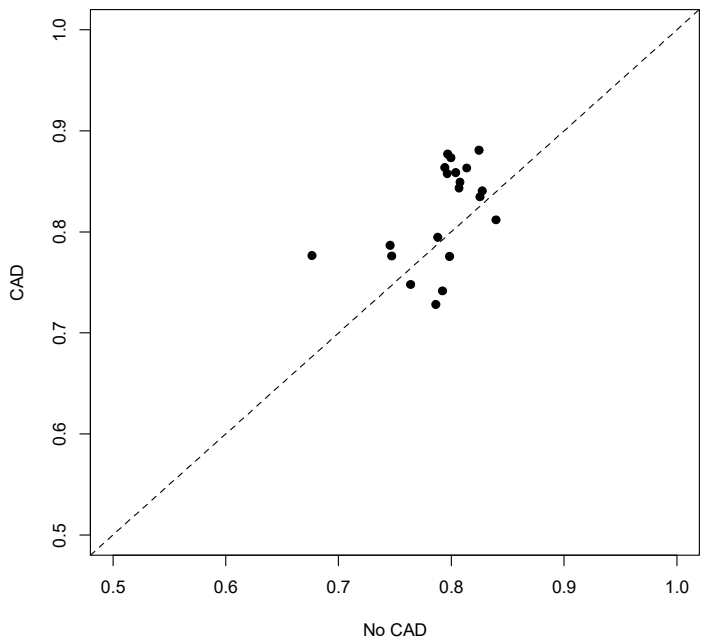
ROC **with** CAD (expert in blue and novice in red)



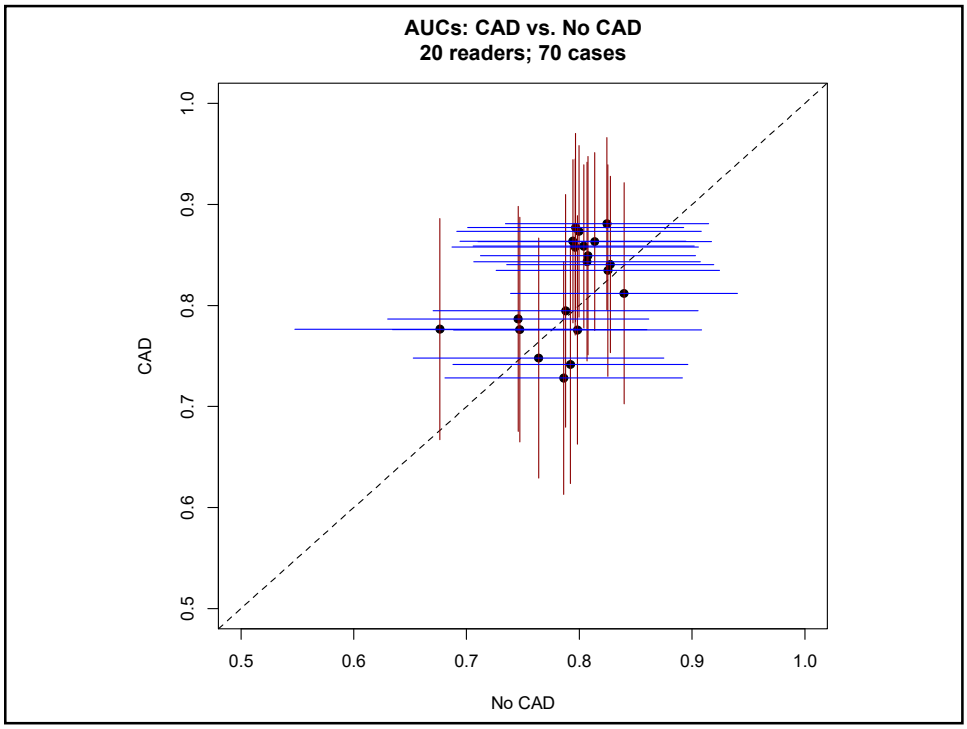
30

	CAD		NoCAD		
Reader	**AUC**	**SE**	**AUC**	**SE**	**p-value**
1	0.8586	0.0411	0.8039	0.0500	0.2062
2	0.8771	0.0475	0.7967	0.0489	0.1231
3	0.8405	0.0445	0.8274	0.0468	0.7929
4	0.7761	0.0567	0.7471	0.0575	0.6399
5	0.8577	0.0444	0.7963	0.0557	0.2588
6	0.7765	0.0559	0.6763	0.0656	0.1230
7	0.8119	0.0558	0.8396	0.0513	0.4504
8	0.7757	0.0577	0.7984	0.0561	0.6596
9	0.7281	0.0588	0.7862	0.0537	0.2734
10	0.7866	0.0568	0.7458	0.0591	0.4677
11	0.7479	0.0606	0.7639	0.0567	0.8310
12	0.8493	0.0501	0.8077	0.0486	0.2715
13	0.7416	0.0601	0.7921	0.0532	0.4056
14	0.8809	0.0434	0.8245	0.0460	0.2128
15	0.8434	0.0502	0.8068	0.0512	0.5048
16	0.8636	0.0412	0.7942	0.0511	0.1472
17	0.8346	0.0534	0.8253	0.0505	0.8443
18	0.8733	0.0433	0.7997	0.0553	0.0446
19	0.8632	0.0449	0.8136	0.0529	0.0969
20	0.7946	0.0587	0.7879	0.0599	0.8289
Overall	0.8191		0.7917		0.0865

31



32



33

Some results

- Overall test averages AUCs using random-reader effects model. The p-value is 0.0865.
- The 95% CIs for the difference is [-.0043, 0.0591]
- 95% CIs for Mean Accuracy of Each Modality
- For CAD : [0.7367, 0.9014]
- For No CAD : [0.7108, 0.8726]

34

		Readers	CAD Average AUC	NoCAD Average AUC	p-value
BIRADS	All	20	0.8091	0.7841	0.0890
	Expert	9	0.8266	0.7972	0.1308
	Novice	11	0.7949	0.7734	0.2665
Prob. of Malignancy Scale	All	20	0.8191	0.7917	0.0865
	Expert	9	0.8383	0.8058	0.0752
	Novice	11	0.8033	0.7801	0.2390
% Prob. of Malignancy (PPM)	All	20	0.8238	0.8036	0.2529
	Expert	9	0.8431	0.8231	0.1941
	Novice	11	0.8080	0.7876	0.3881

35

How to Combine ROC data?

- Must compare ROC curves (threshold=confounder)
- Must distinguish between variability of
  - Operating point (Threshold)
  - Accuracy (Roc Curve/Area)
- What is an average ROC curve for a population of readers?
  - Several possible ways to define such a curve, each with its own advantages and disadvantages. For example, area under average curve is not equal to the average area.

36

There are many ways to 'average'

- Many ways to average (which is why you need statistical help!)
  - Compute estimate, then average estimates.
  - Pool data, then compute estimate
  - Pool data under model that addresses reader variation, then compute estimate
- Regression models for outcomes typically do #3
- 'Average' ROC curve *vs.* Average area
  - Average over binormal parameters
  - Average over ordinal regression parameters

37

An example of a study with low variability across readers

Prostate Cancer: Sextant Localization at MR Imaging and MR Spectroscopic Imaging before Prostatectomy—Results of ACRIN Prospective Multi-institutional Clinicopathologic Study[1]

Jeffrey C. Weinreb, MD
Jeffrey D. Blume, PhD

Purpose: To determine the incremental benefit of combined endorec-

Radiology 2009; 251:122–133

38

ACRIN 6659

- MR Spectroscopy of the Prostate
- Assess the performance of MRS compared to MR alone in localizing and staging prostate cancer in sextants of the prostate.
- Reader study, 134 patients across seven sites (one reader per site plus PI, all cases distributed to each site)
- Consensus panels to determine pathology imaging matching.

39

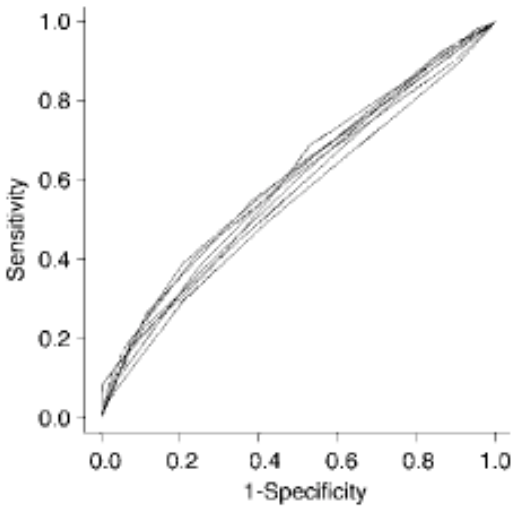


Figure 3: Receiver operating characteristic curves of combined MR imaging–MR spectroscopic imaging values for individual readers.

40

AUCs for Sextant PZ Prostate Cancer Localization with MR Imaging Alone and Combined MR Imaging–MR Spectroscopic Imaging

Reader	MR Imaging	Combined MR Imaging–MR Spectroscopic Imaging	*P* Value
1	0.6028	0.5856	.2614
2	0.5723	0.5425	.2399
3	0.6264	0.5907	.0475
4	0.6054	0.5933	.5311
5	0.6163	0.6050	.6271
6	0.6092	0.5741	.0268
7	0.5863	0.5930	.7359
8	0.6099	0.6029	.6338
All Readers	0.6007	0.5844	.0892

41

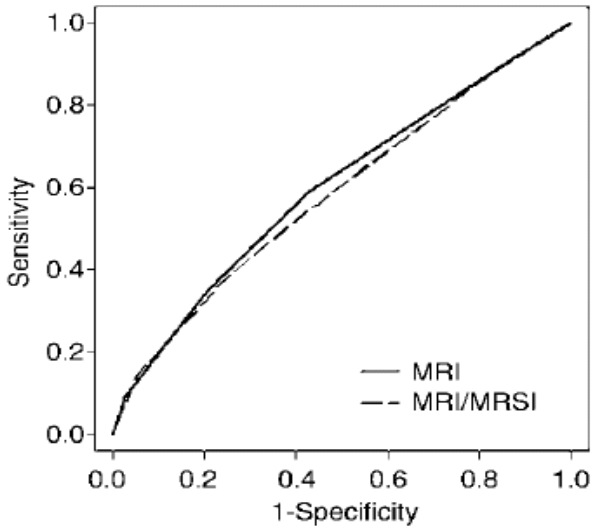


Figure 2: Receiver operating characteristic curves of MR imaging values versus combined MR imaging–MR spectroscopic imaging *(MRSI)* values for all readers.

42

Commentary

- Extensive variability in accuracy exists among test interpreters in both screening & diagnostic contexts.

- Magnitude of variability is of interest
- Ideally, the analysis of accuracy assesses both
  - an average value of diagnostic performance , and
  - the variation across test interpreters

43

Analysis techniques

- Multi-center or Multi-reader studies
  - Report results by reader
  - Ignore the site / reader and pool the data (not recommended – often leads to attenuation)
  - Model response score and/or ROC
  - Combine ROCs with an (weighted) average or (fancy) model
  - Use regression models, fancy models, bootstrap techniques on AUCs to tease out different sources of variability

44

Generalizability of results

- Reader population
  - expert readers vs. professionals 'at large'
  - variation across readers & institutions
  - extent of reader experience
  - want to generalize beyond sample on the study, but do not want to bias against new technology if readers have little experience

45

Generalizability of results

- Case mix (spectrum)
  - representative sample
  - all forms of disease in sample
  - sample prevalence may influence interpretation because of limited spectrum or, even with representative spectrum, because of factors such as reader vigilance

46

Generalizability of results

- Technical characteristics of the imaging process
  - precise description of techniques
  - reproducible at other clinics
  - should reflect expected clinical practice, but this often varies across institutions – set minimum acceptable techniques, or allowable range

47

Multi-reader, Multi-modality studies

- Commonly used design for studies comparing accuracy of modalities.
- Typically, a set of J readers interprets N scans on the same cases by two or more modalities ("fully crossed" design)
- Several other variants of the design exist, to be used in special circumstances when fully crossed design is not practically feasible.
- Goal is to compare average measure of performance between modalities, while accounting for correlations in the data and for reader variability.
- Rarely each reading is repeated K times

48

Multi-reader, multi-modality designs

- The variance of the difference in AUC estimates is (Zhou, Obuchowski, McClish, 2002)

$$\mathbf{Var}[\hat{\theta}_{1\bullet\bullet} - \hat{\theta}_{2\bullet\bullet}] = \frac{2}{J}\left[\sigma_b^2(1-r_b) + \frac{\sigma_w^2}{K} + \sigma_c^2(1 - r_1 + (J-1)(r_2 - r_3))\right]$$

where

- σ^2_b is the between-reader variability
- σ^2_w is the within-reader variability
- σ^2_c is the case variability
- Note that *K* is almost always *1*.

49

Multi-reader, multi-modality designs

$$\mathbf{Var}[\hat{\theta}_{1\bullet\bullet} - \hat{\theta}_{2\bullet\bullet}] = \frac{2}{J}\left[\sigma_b^2(1-r_b) + \frac{\sigma_w^2}{K} + \sigma_c^2(1 - r_1 + (J-1)(r_2 - r_3))\right]$$

and $$L = 2Z_{\alpha/2}\sqrt{\mathbf{Var}[\hat{\theta}_{1\bullet\bullet} - \hat{\theta}_{2\bullet\bullet}]}$$

- r_1 = corr. b/w area est., same reader, diff. modality
- r_2 = corr. b/w area est., diff. readers, same modality
- r_3 = corr. b/w area est., diff. reader, diff. modality
- r_b = corr. b/w area est., set of readers, diff. modality

50

Multi-reader, multi-modality designs

- Consider power as well.
- Values for parameters need to be assessed on the basis of previous studies. Design is sensitive to parameter values.
- Examples of values used:
  - $\sigma^2_b = 0.000625$, $\sigma^2_w = 0.0001$
  - $\sigma_c^2 = (\mathrm{Var}[\hat{\theta}_1] + \mathrm{Var}[\hat{\theta}_2])/2$ (exponential assumption)
  - $r_b = 0.82$
  - $(r_1, r_2, r_3) = (0.44, 0.33, 0.29)$ or $(0.3, 0.1, 0.05)$

51

Multi-reader studies

- Multi-reader study designs typically introduce a trade-off between required cases and readers and can thus lead to studies with fewer required cases.
- Computations for design and analysis are complex.
- Power to compare average AUC of two modalities, if difference in areas in 0.10, and average AUC is 0.85.

	4 readers		6 readers		8 readers	
Design	**n=50**	**n=100**	**n=50**	**n=100**	**n=50**	**n=100**
Fully paired	**0.5**	**0.61**	**0.77**	**0.88**	**0.89**	**0.97**
Unpaired case, paired reader	**0.29**	**0.42**	**0.44**	**0.64**	**0.53**	**0.75**
Paired case, unpaired reader	**0.36**	**0.44**	**0.6**	**0.71**	**0.76**	**0.86**

Zhou , Obuchowski & McClish, 2002

52

Table 9.9 The Estimated Power for Various Study Designs and Sample Sizes

Study Design	$J = 4$, $m = 50$	$J = 4$, $m = 100$	$J = 6$, $m = 50$	$J = 6$, $m = 100$	$J = 8$, $m = 50$	$J = 8$, $m = 100$
Paired-patient, paired-reader	$\hat{\lambda} = 8.16$ Power = 0.50	$\hat{\lambda} = 11.02$ Power = 0.61	$\hat{\lambda} = 11.33$ Power = 0.77	$\hat{\lambda} = 15.69$ Power = 0.88	$\hat{\lambda} = 14.07$ Power = 0.89	$\hat{\lambda} = 19.90$ Power = 0.97
Unpaired-patient, paired-reader	$\hat{\lambda} = 4.08$ Power = 0.29	$\hat{\lambda} = 6.58$ Power = 0.42	$\hat{\lambda} = 4.94$ Power = 0.44	$\hat{\lambda} = 8.27$ Power = 0.64	$\hat{\lambda} = 5.51$ Power = 0.53	$\hat{\lambda} = 9.49$ Power = 0.75
Paired-patient, unpaired-reader	$\hat{\lambda} = 5.31$ Power = 0.36	$\hat{\lambda} = 6.82$ Power = 0.44	$\hat{\lambda} = 7.57$ Power = 0.60	$\hat{\lambda} = 9.90$ Power = 0.71	$\hat{\lambda} = 9.62$ Power = 0.76	$\hat{\lambda} = 12.78$ Power = 0.86
Unpaired-patient, unpaired-reader	$\hat{\lambda} = 3.43$ Power = 0.26	$\hat{\lambda} = 5.05$ Power = 0.35	$\hat{\lambda} = 4.29$ Power = 0.39	$\hat{\lambda} = 6.59$ Power = 0.54	$\hat{\lambda} = 4.89$ Power = 0.48	$\hat{\lambda} = 7.79$ Power = 0.67
Paired-patient-per-reader, paired-reader	$\hat{\lambda} = 9.27$ Power = 0.54	$\hat{\lambda} = 11.99$ Power = 0.64	$\hat{\lambda} = 13.90$ Power = 0.84	$\hat{\lambda} = 17.99$ Power = 0.92	$\hat{\lambda} = 18.54$ Power = 0.96	$\hat{\lambda} = 23.99$ Power = 0.99

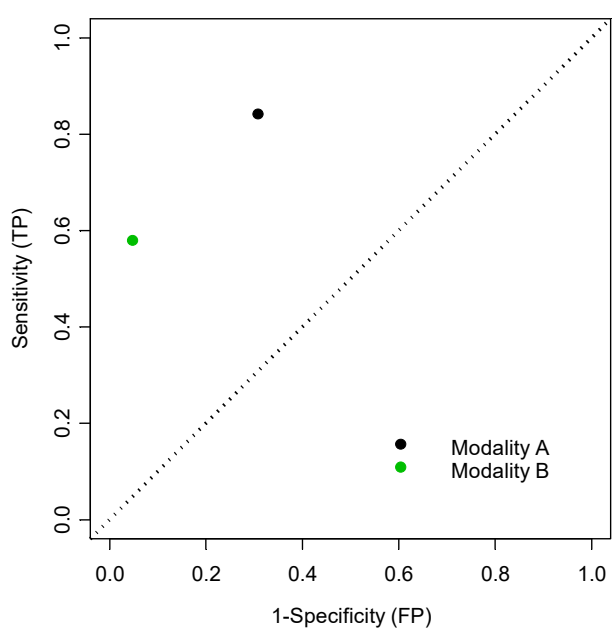
Note: For paired-patient study designs, m is the total number of patients with malignant lesions needed for the study; for unpaired-patient designs, m is the total number of patients with malignant lesions needed per diagnostic test; for paired-patient-per-reader designs, a total of m patients with malignant lesions are needed for each of the J reader.

303

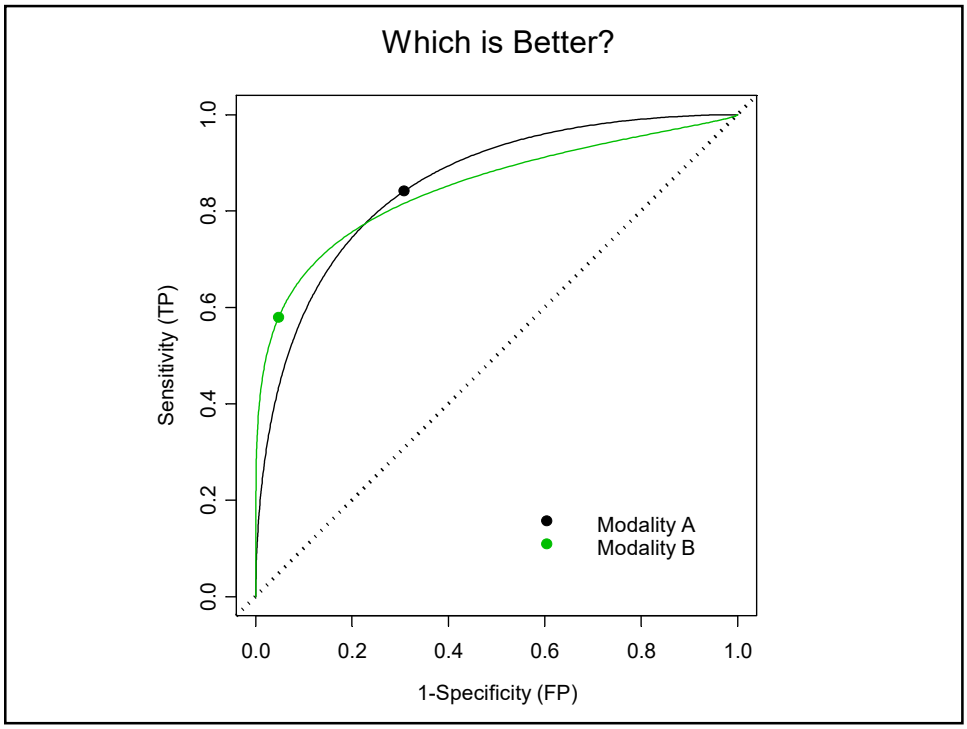
Zhou, Obuchowski, McClish. Statistical Methods in Diagnostic Medicine. 2002. page 303

53

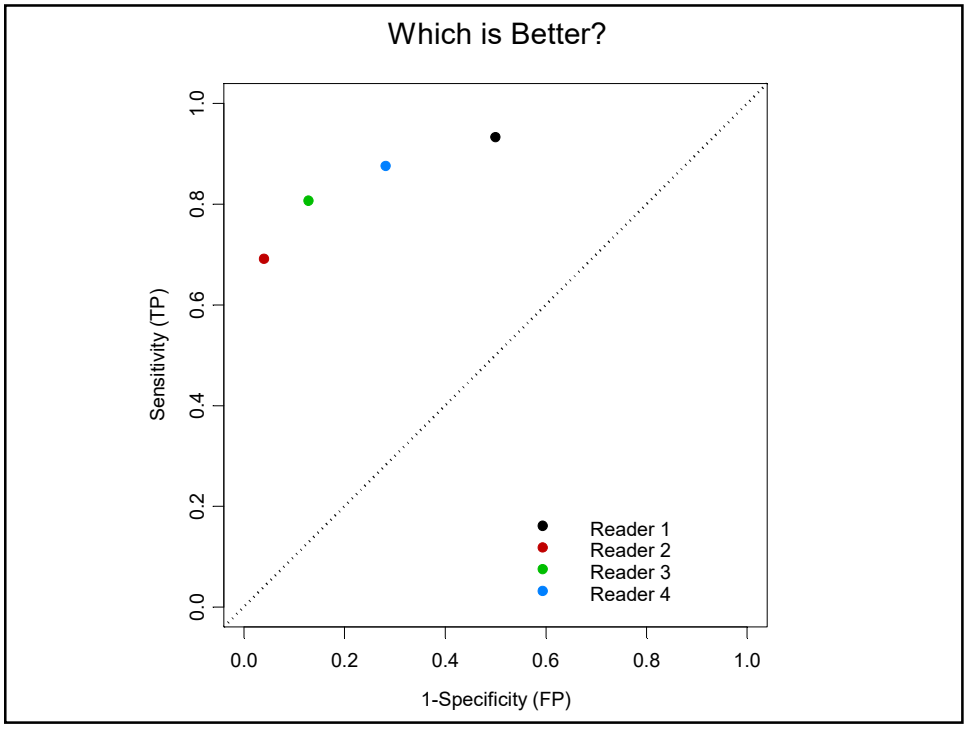
Which is Better?



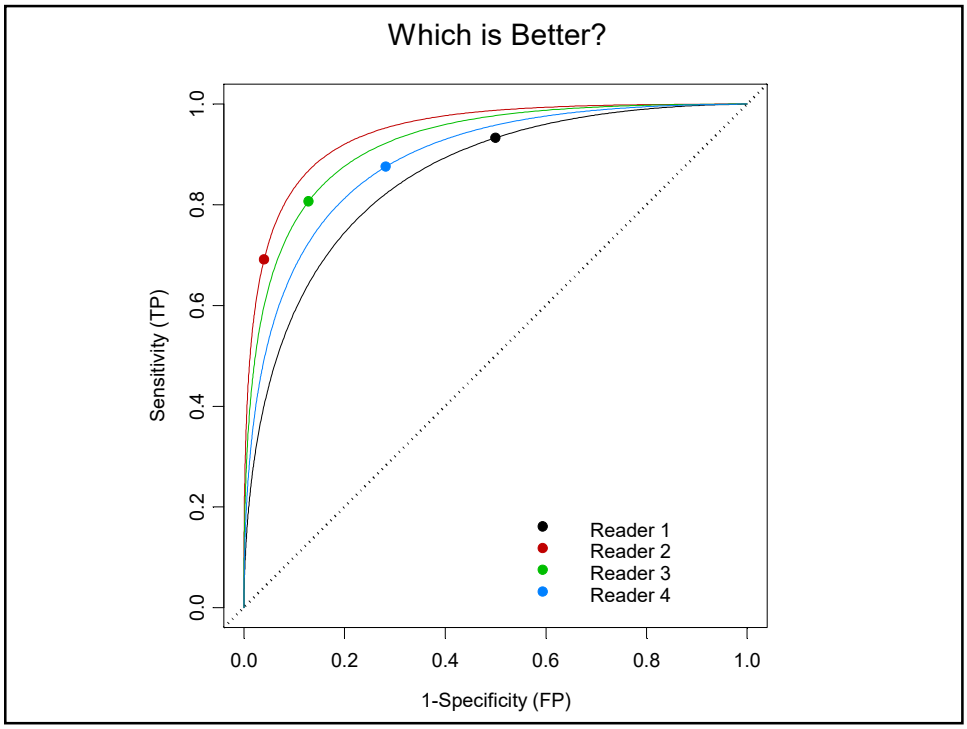
54


Which is Better?
Sensitivity (TP)
1-Specificity (FP)
0.0
0.2
0.4
0.6
0.8
1.0
Modality A
Modality B

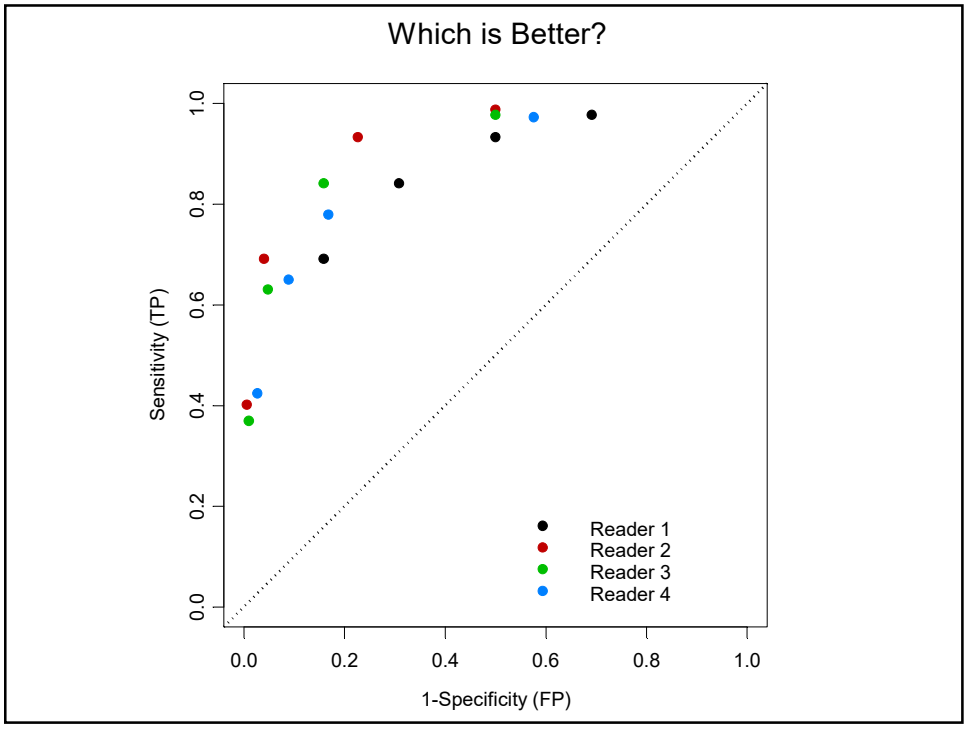
55


Which is Better?
Sensitivity (TP)
1-Specificity (FP)
0.0
0.2
0.4
0.6
0.8
1.0
Reader 1
Reader 2
Reader 3
Reader 4

56

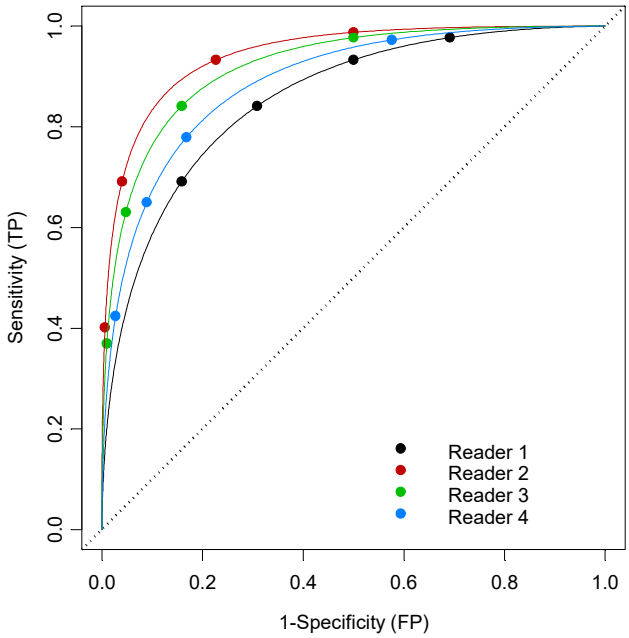

Which is Better?
Sensitivity (TP)
1-Specificity (FP)
0.0
0.2
0.4
0.6
0.8
1.0
Reader 1
Reader 2
Reader 3
Reader 4

57


Which is Better?
Sensitivity (TP)
1-Specificity (FP)
0.0
0.2
0.4
0.6
0.8
1.0
Reader 1
Reader 2
Reader 3
Reader 4

58

Which is Better?

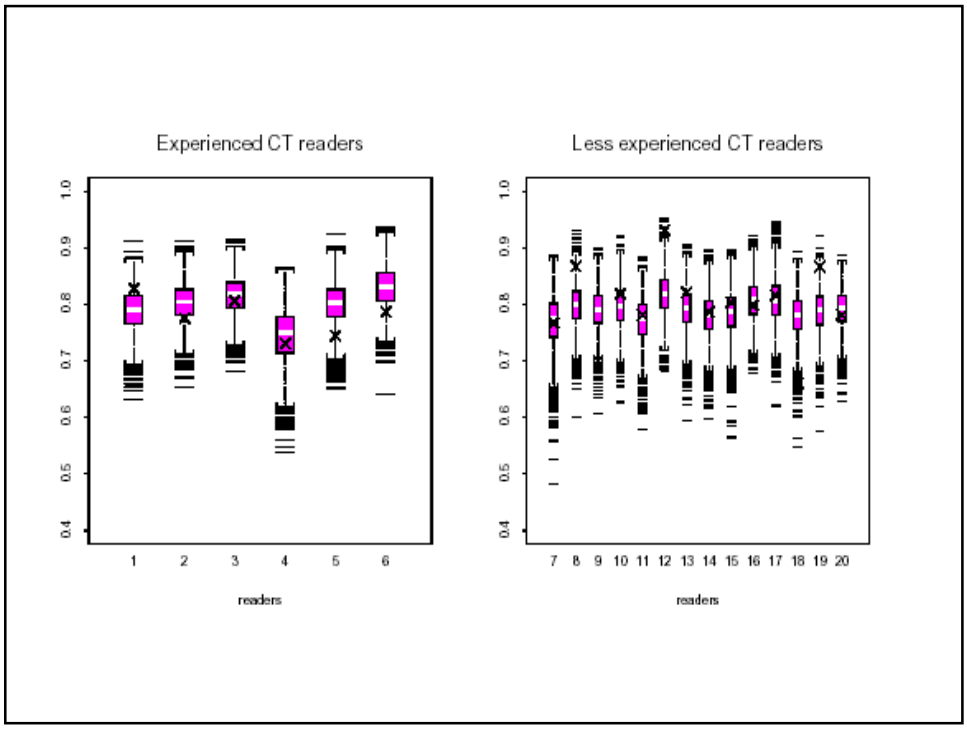


59

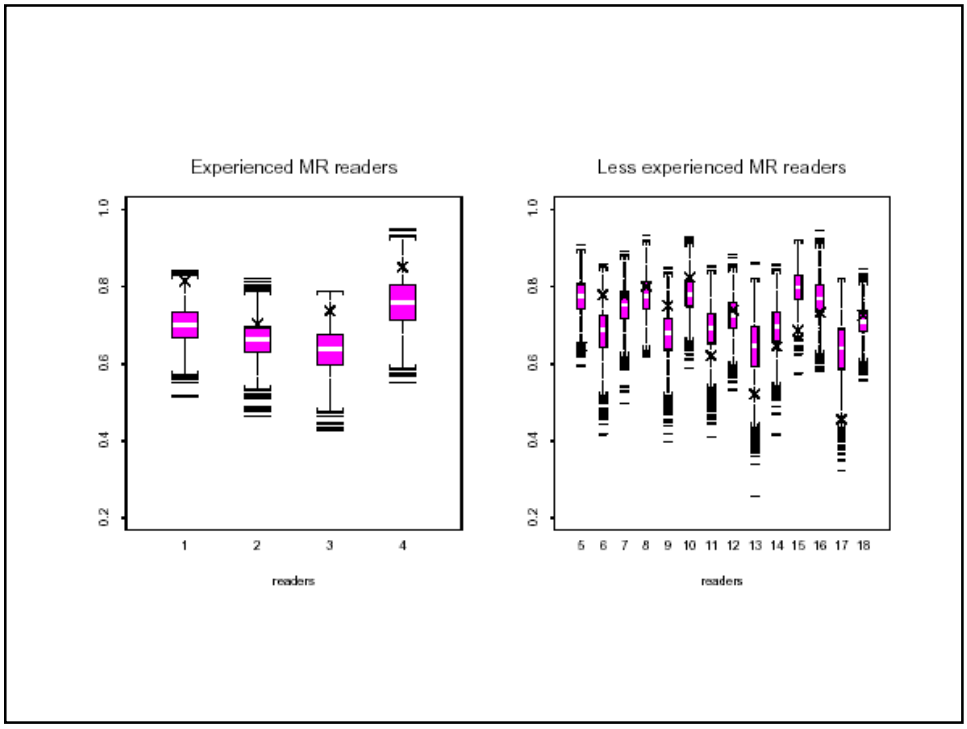
Example of Variability: Analysis of head and neck cancer data

- 38 radiologists interpreted CT and MRI scans on head and neck cancer patients. Each case was interpreted by 3 readers in each modality. Total of 20 CT readers, 18 MRI readers.
- Degree of suspicion about metastasis recorded on 5 point ordinal categorical scale.
- A fancy model accounts for correlations due to cases and readers (see Ishwaran and Gatsonis, 2000)
- Fancy (Posterior) estimates of AUCs presented in next two graphs.

60


Experienced CT readers
Less experienced CT readers
readers
readers

61


Experienced MR readers
Less experienced MR readers
readers
readers

62