INTRODUCTION TO ROC ANALYSIS

Andriy I. Bandos
Department of Biostatistics
University of Pittsburgh

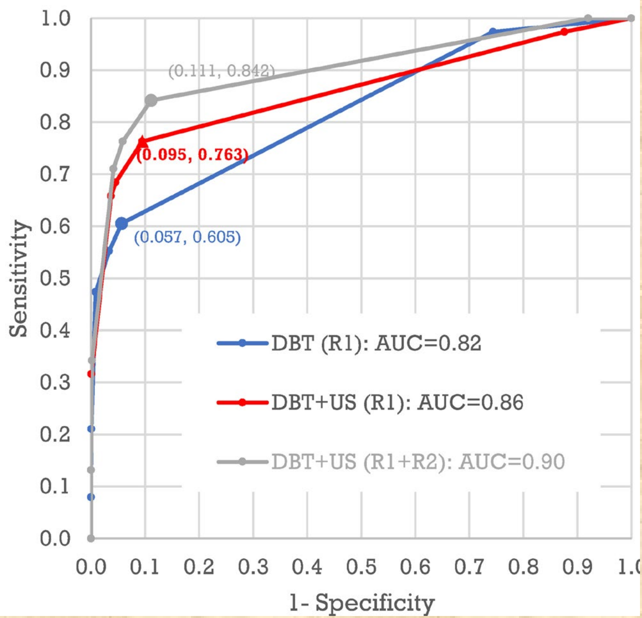
With thanks to
Drs. Sam Wieand, Nancy Obuchowski, and Todd Alonzo

Outline

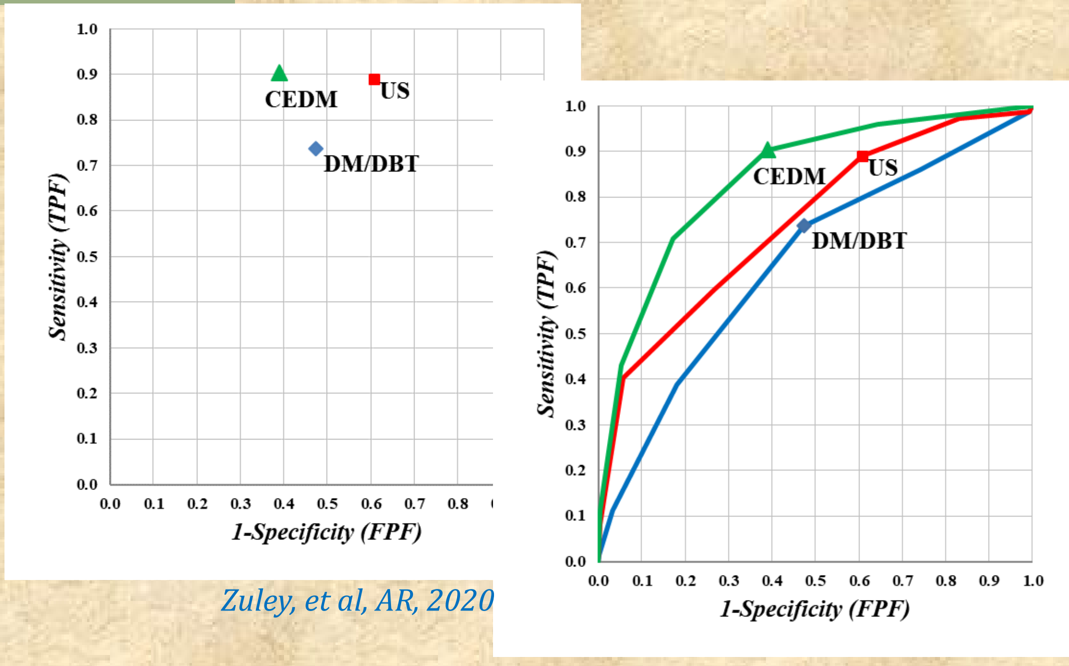
1. Basics of diagnostic accuracy evaluation
2. Why do we need ROC analysis?
3. How to construct and use the ROC curve

** Focus on the structure and interpretation of ROC tools*
(aside from the very important analysis of statistical uncertainty)

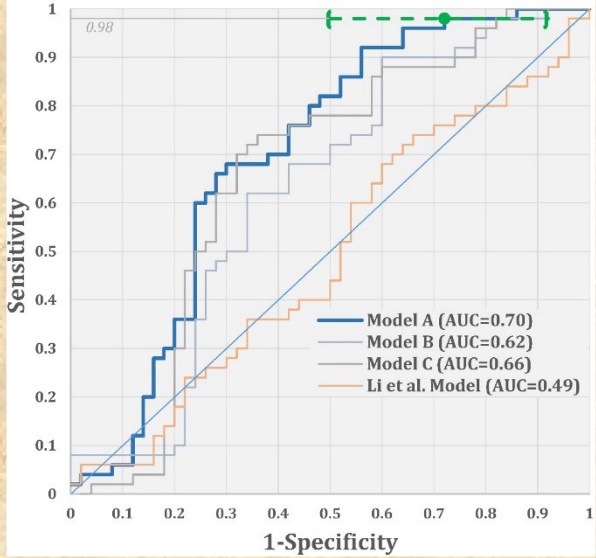
Examples



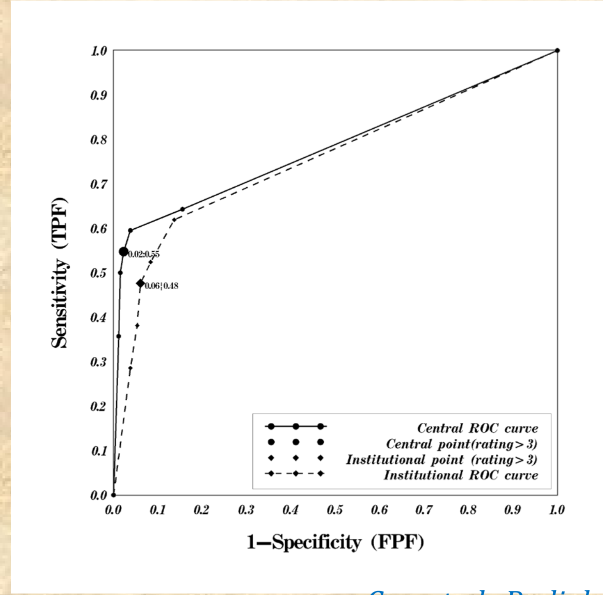
Berg et al, JCO, 2022



Zuley, et al, AR, 2020



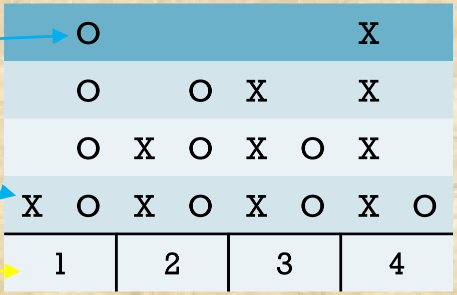
Pu et al, European Radiology, 2020



Gee, et al., Radiology, 2017

Basic set-up for accuracy evaluation

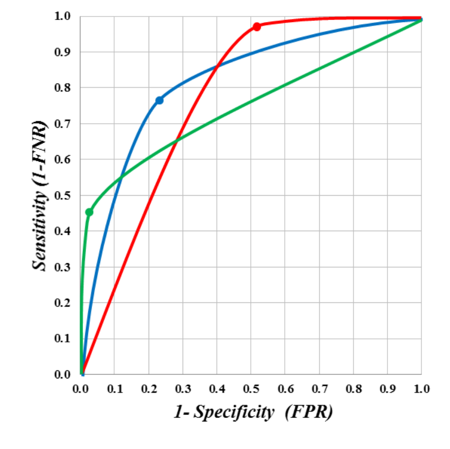
- A sample of subjects, for every subject ("diagnostic unit"):
  - the presence/absence of the *condition of interest*, or "**true status**" ("***normal*"/"*abnormal***") as determined by the "**Gold** (Reference) **Standard**"
  - the diagnostic **test result** (***score 1-4*, "*positive*"/ "*negative*",** etc.) as obtained (*from biomarker, radiologist, prediction model, etc.*)



	O					X	
	O		O	X		X	
	O	X	O	X	O	X	
X	O	X	O	X	O	X	O
1		2		3		4	

- **Diagnostic accuracy** is a vague term ≈ "agreement" between the *test result* and *true status*

- "Good" performance scenarios
  - accurate in determining <u>both</u> levels of the true status (high agreement overall)
  - accurate in determine <u>*either*</u> normal or abnormal true status (high agreement for only some results)

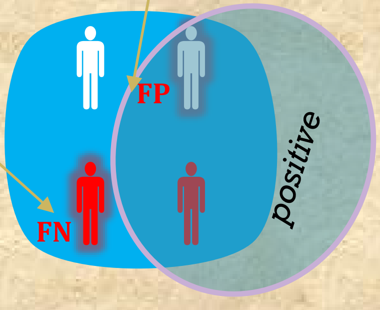


Illustrative Examples

- **Studies:**
  - *Ultrasound after Tomosynthesis in dense breast* (Berg et al, JCO'23)
  - *FDG PET-CT for distant metastatic disease* (Gee et al., Radiology 2017)
  - *CEDM to reduce breast biopsies* (Zuley et al., AR 2019)
  - Image marker for COVID-19 (Pu et al, European Radiology, 2020)
- **Conditions of Interest:**
  - *presence/absence of breast cancer, distant metastatic disease, active COVID-19, malignant/benign nature of the index lesion, ….(future events)…*
- **Reference (Gold) Standard:**
  - *pathology and follow-up radiology reports, repeated PCR tests …*
- **Test result:**
  - *BI-RADs ("positive" biopsy recommendation = "≥ 4A") ,*
    scores 1-6 ("positive"= ">3") for presence of distant metastases,…

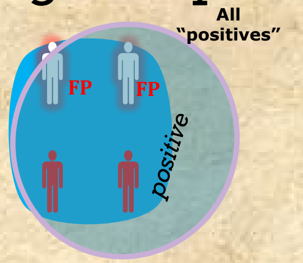
Errors in testing/decisions

- □ Two basic errors: "**false positive**" FP (positive for "normal") and "**false negative**", FN (negative for "abnormal")

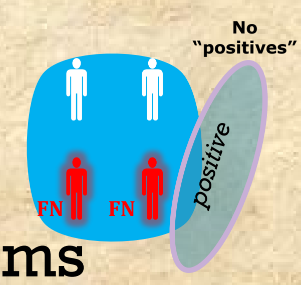


- □ Different errors ↔ different consequences
  - **FN** → higher grade of disease, spread of infection, …
  - **FP** → unnecessary procedures, surgeries, quarantine….
- □ Both decision errors must be considered **simultaneously**
  - errors can always be exchanged, by changing the "positivity" criteria

trivial cases: "*all positive*" ⇒ *only FP errors:*



"*no positive*" ⇒ *only FN errors*)



- □ Need to quantify errors in absolute and relative terms
  - **how few is few enough?**

Quantifying Errors/Correct classifications

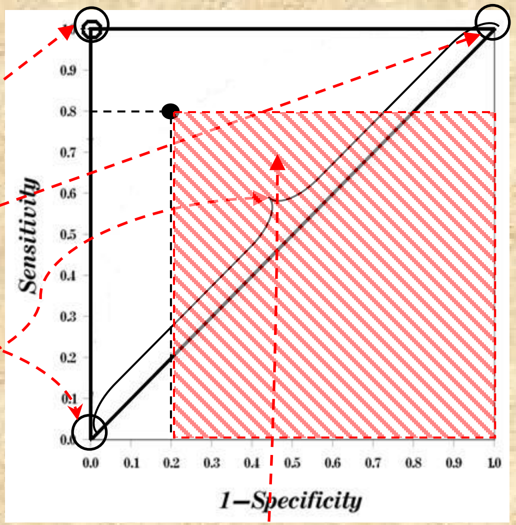
TRUE STATUS	**TEST RESULT**		
	Negative (-)	***Positive*** (+)	
Normal	*# True Negatives=* ***353***	*#False Positives=5*	#"Normal"=**358**
Abnormal	*#False Negatives=17*	*#True Positives=****31***	#"Abnormal"=**48**
	#Negatives=**370**	#Positives=**36**	Total=406

- How frequent are the correct classifications ?
  - *31 True Positives*
    - 31 out of 406≈0.08 → Probability of True Positives (***Detection Rate***)
    - 31 out of 48 ≈ 0.65 → ***Sensitivity, Sens,*** (or *True Positive Fraction,* ***TPF***) (*complement of False Negative Fraction,* ***FNF***)
    - 31 out of 36 ≈ 0.86 → *Positive Predictive Value* (***PPV***)
  - *353 True Negatives*
    - 353 out of 406 ≈ 0.87 → Probability of True Negatives (*complement of* ***False Recall Rate***)
    - 353 out of 50 ≈ 0.99 → ***Specificity*** (*complement of False Positive Fraction,* ***FPF***)
    - 353 out of 20 ≈ 0.95 → *Negative Predictive Value* (***NPV***)

- *Sens* and *Spec* are usually preferred, because they are
  - robust (e.g., do not depend on prevalence)
  - have fixed benchmarks of what is large (1) and what is small (0)

Graphical representation: ROC space

- ROC coordinates: ***Sens*** (or ***TPF***) as a vertical and ***1-Spec*** (or ***FPF***) as a horizontal axes
- Characteristics of benchmark tests:
  - perfect (no errors): *Spec=1, Spec=1*
  - most liberal (all "positive"): *Spec=0, Sens=1*
  - most strict (all "negative"): *Spec=1, Sens=0*
  - guess (flip of a coin): *1-Spec=Sens*

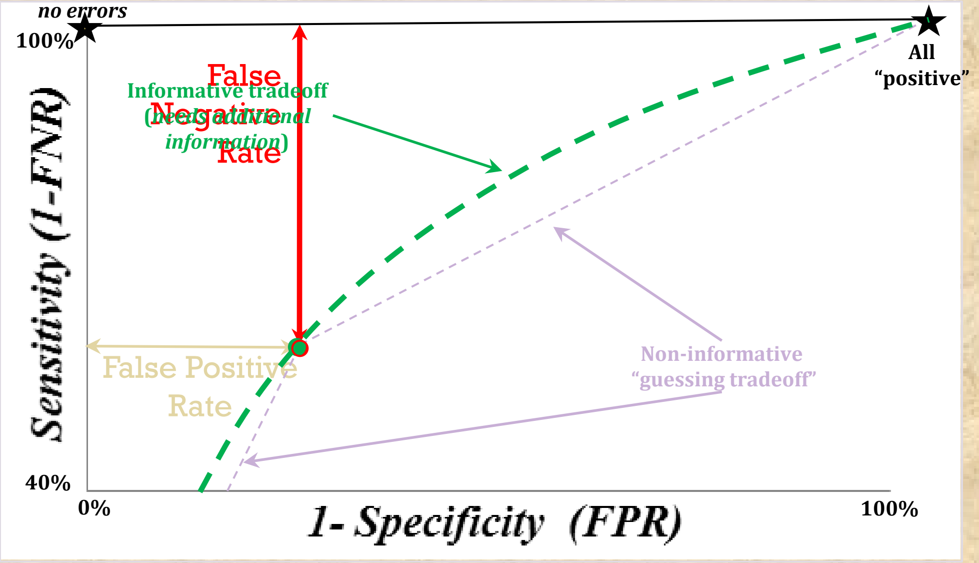


- There is always a test with better Sens, or better Spec
  - ⇒**MUST consider both *Sens* and *Spec***
- Simultaneous Interpretation of values of *Sens* and *Spec*
  - "Bad" – comparable to performance of a guess (*1-Sp≈Se, or close to the diagonal*)
  - "Good" – close to the perfect (Sens*≈1, Spec≈1; or FPF≈0*)
- A tool with worse *Sens* and *Spec* is objectively worse

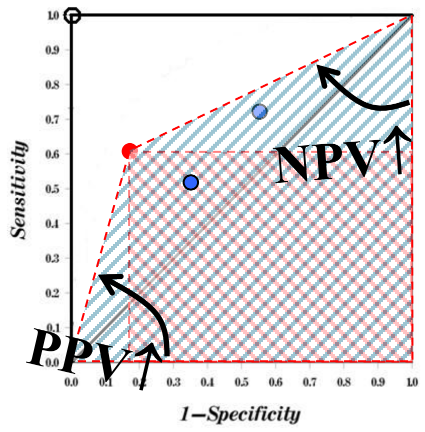
Importance of the join assessment

- The exchange of errors is always possible
  - E.g., by randomly reclassifying the given results

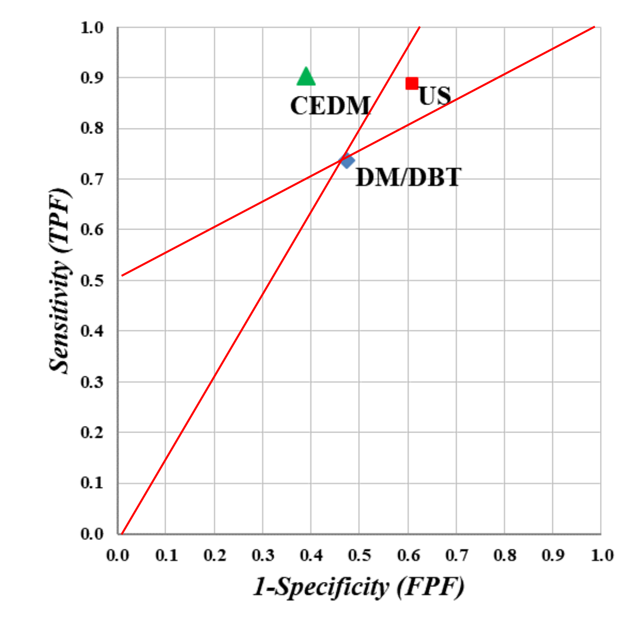
- *Informative exchange of errors → ROC curve*
  - By changing the threshold on underlying score/result

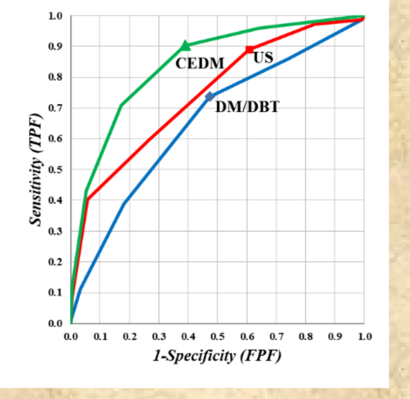


More on Comparison of Diagnostic Tests



- □ Both *Sens* and *Spec* are higher ⇒better test
  - ■ DBT+US+CEDM vs DBT
- □ Both *PPV* and *NPV* are higher (in the same population) ⇒ better test (⇐ *recall relabeling*)
- □ Higher *PPV*, but lower NPV ⇒ ?
  - ■ DBT versus DBT+US
- □ This problem can be objectively solved by constructing ROC curves:





Zuley et al, AR, 2019

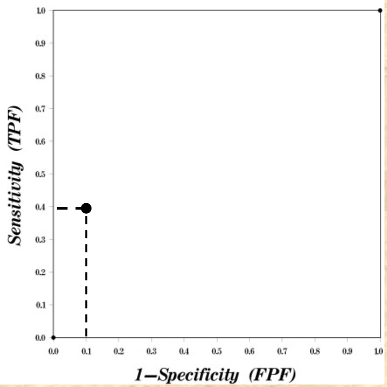
ROC curve construction (make-up example)

	O					X	
	O		O	X		X	
	O	X	O	X	O	X	
X	O	X	O	X	O	X	O
1		2		3		4	

“positivity” threshold

TRUTH	TEST “-”	“+”	
Normal	9	1	10
Abnormal	6	4	10

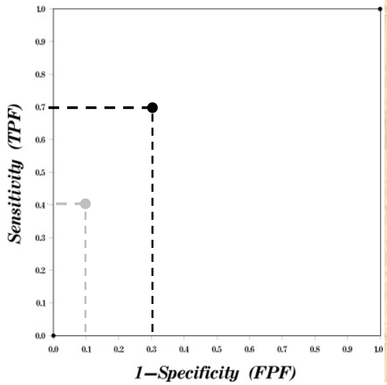
Sens(c)=0.4
1-Spec(c)=0.1



	O					X	
	O		O	X		X	
	O	X	O	X	O	X	
X	O	X	O	X	O	X	O
1		2		3		4	

TRUTH	TEST “-”	“+”	
Normal	7	3	10
Abnormal	3	7	10

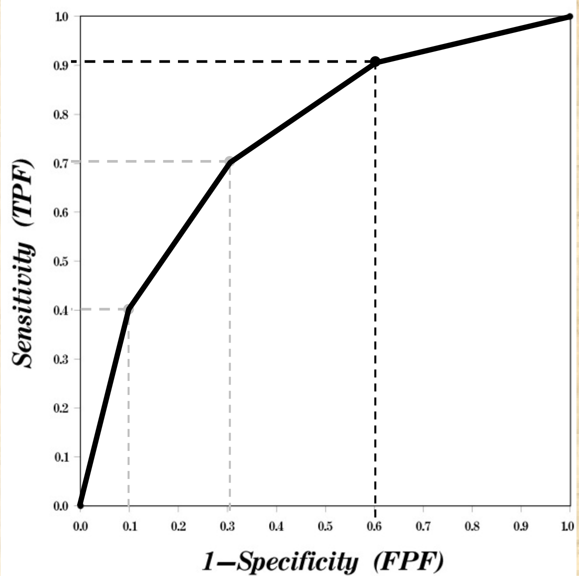
Sens(c)=0.7
1-Spec(c)=0.3



	O					X	
	O		O	X		X	
	O	X	O	X	O	X	
X	O	X	O	X	O	X	O
1		2		3		4	

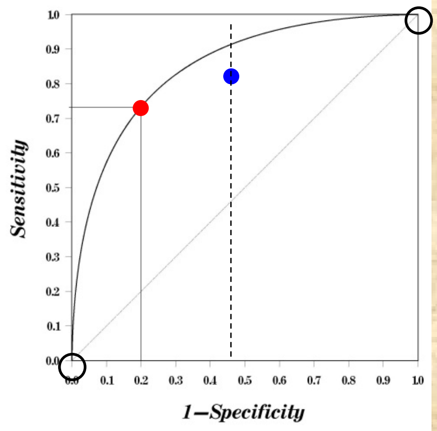
TRUTH	TEST “-”	“+”	
Normal	4	6	10
Abnormal	1	9	10

Sens(c)=0.9
1-Spec(c)=0.6



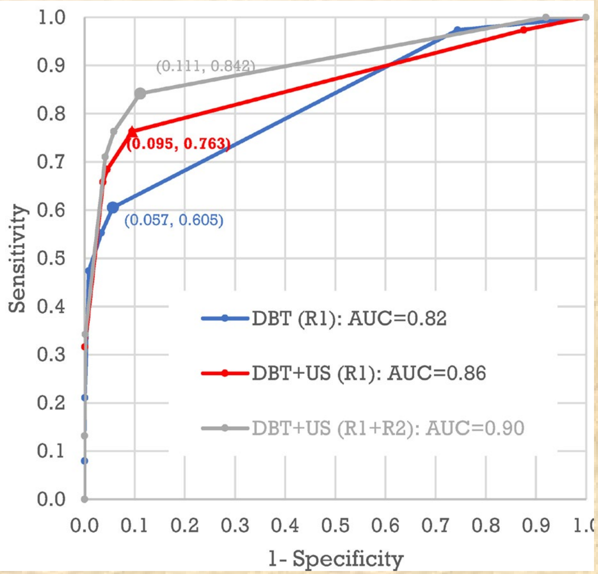
ROC curve: comparing performance levels

- ROC describes all Sens-Spec values that we can obtain by changing the threshold

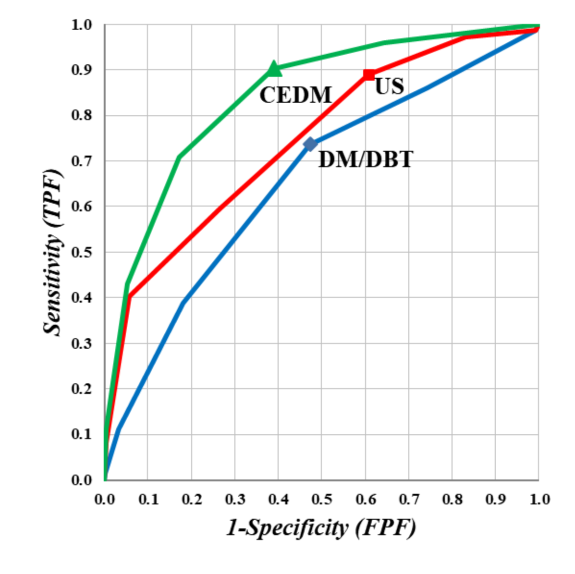


- A classic application of the ROC curve:
  - Can one test be tuned to achieve higher Sens and Spec than another ?

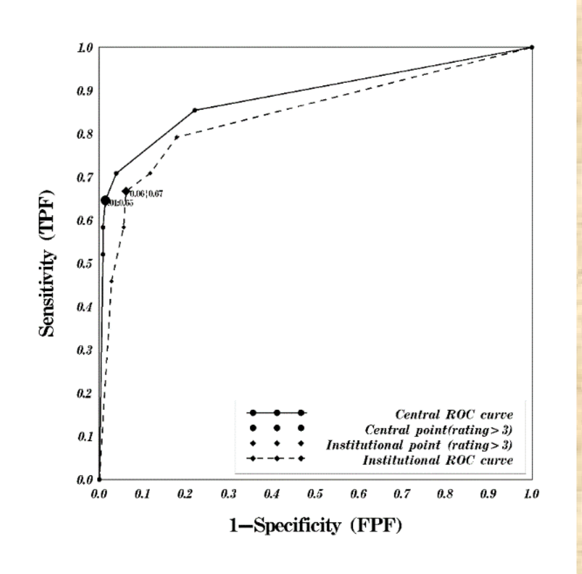
- ROC curve helps determine if higher sensitivity is justified:



Berg et al, JCO, 2022



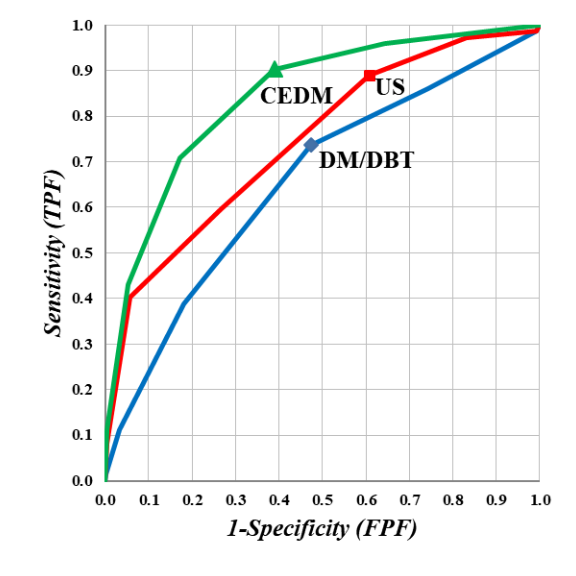
Zuley et al., Academic Radiology, 2019



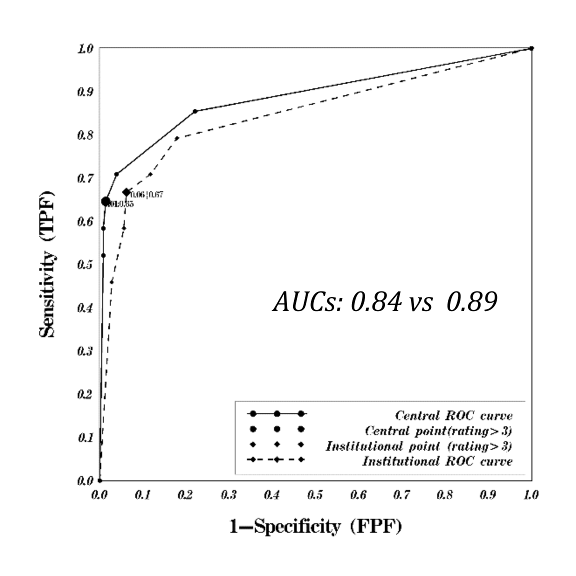
Gee et al., Radiology, 2017

ROC curves: overall comparison

- □ An overall better test has uniformly higher ROC curve

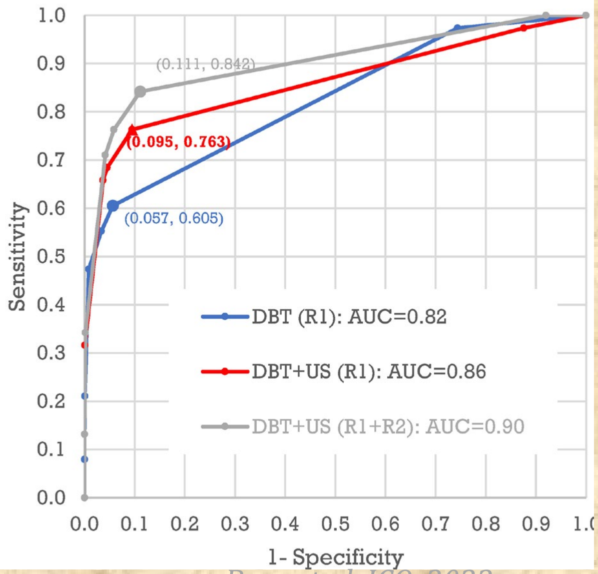


Zuley et al., Academic Radiology, 2019



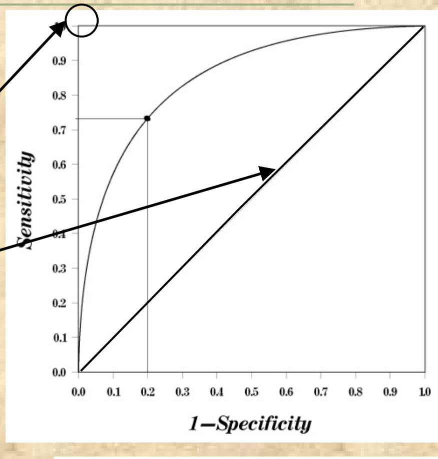
Gee et al., Radiology 2017

- □ Sometimes, one ROC curve is higher only in some ranges
  - ⇒ *practical purpose should drive considerations*

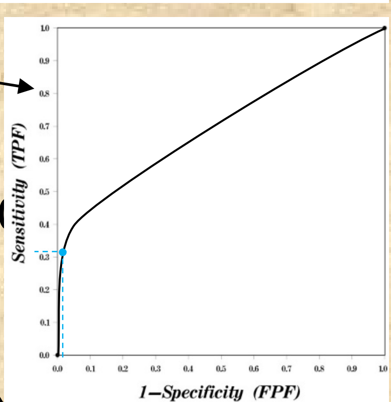


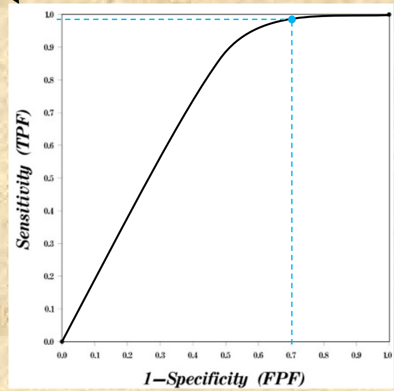
Berg et al, JCO, 2022

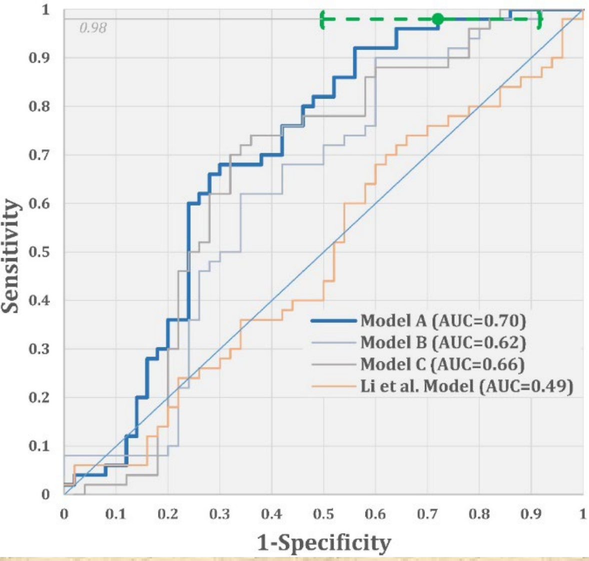
ROC curve: types of “good” for the purpose



- Recall: ROC landmarks/benchmarks:
  - *Perfect:* two segments connecting at *(Sens=1, Spec=1)*
  - *Guessing:* diagonal (random choice) *1-Spec=Sens*
- But, tests with relatively low ROC curves could still be very useful for targeted decisions, e.g.
  - for identifying a subset of “diseased”: Sens>>0, Spec ≈1 (*e.g., screening task*)
  - for identifying a subset of “disease-free”: Sens≈1, Spec>>0 (*e.g., triaging task*)



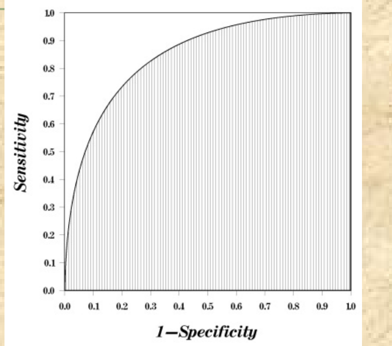




Pu et al, European Radiology, 2020

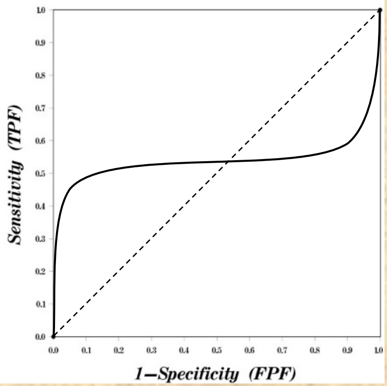
Most typical ROC summary index

- Area Under the ROC curve (AUC)
  - Single-value summary index of the entire ROC curve *perfect* ROC⇒AUC=1; *guessing* ROC ⇒AUC=0.5
  - difference between distributions of test results for "normal" and "abnormal" (i.e., P(X<Y))



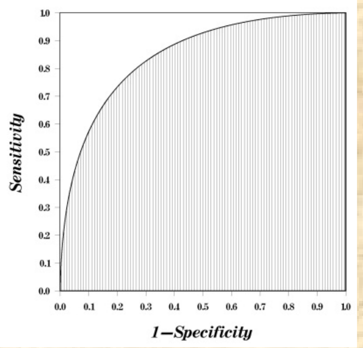
- Technical advantages: well-known, objective, easy to use

- Limitations
  - not very practically relevant
  - summarizes over the operating points outside of practical interest (e.g. Sp< 0.5)
  - can be misleading (*as any scalar index for the entire curve*), e.g., non-guessing ROC with AUC=0.5

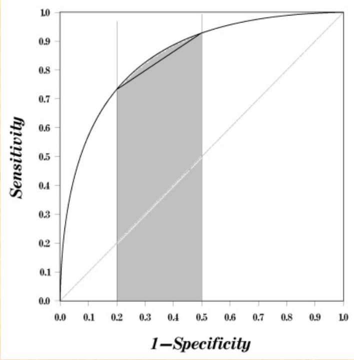


Summary Indices for the ROC curve

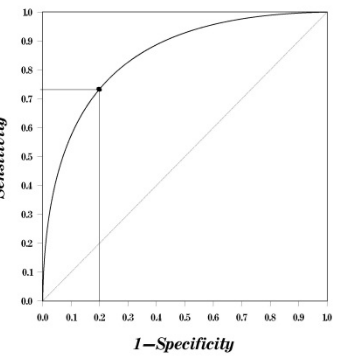
- Area under the ROC curve (AUC)
  - "+" does not require subjective conjectures
  - "-" summarizes over the many useless operating points
  - "+" one of more precise summary indices



- Partial AUC (pAUC), e.g., for s_1<Spec<s_2
  - "-" requires specification of the range of interest
  - "+" focuses on multiple points of potential interest
  - "-" often requires larger samples than AUC



- Sens corresponding to a given Spec (Sens|spec=0.9) (*or vice versa Spec|sens=0.9*)
  - "-" requires specification of the range of interest
  - "+" focuses on practically relevant operating point
  - "-" often requires larger sample than pAUC

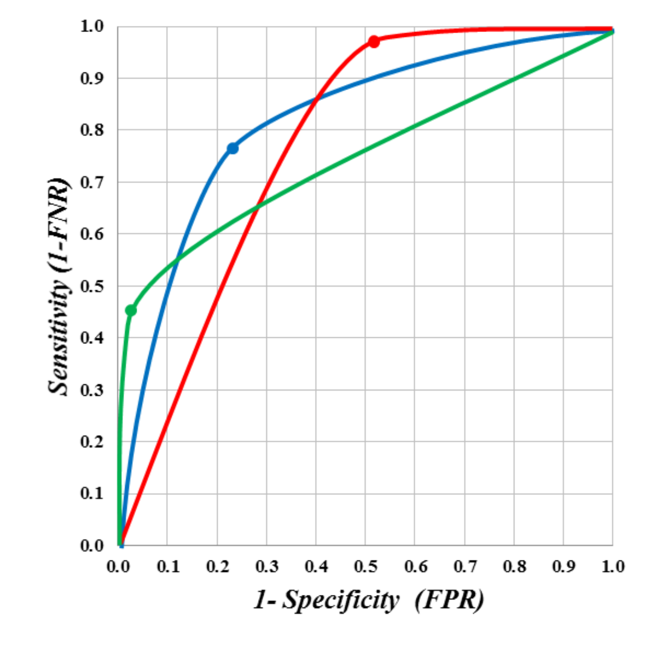


Problems with ROC indices

- □ Index always loses some information about the ROC curve

  ⇒ different indices could contradict to each other, e.g.:

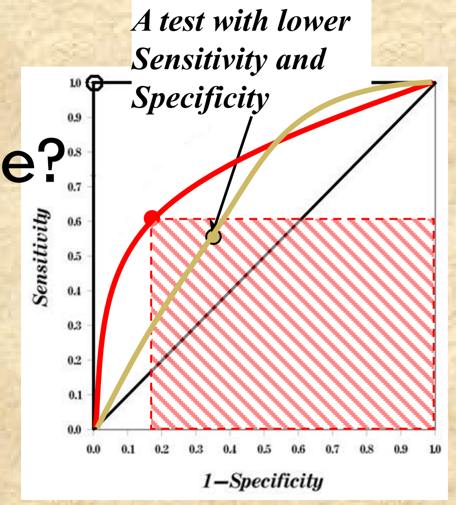
  - curves with the same AUCs can be different at almost all points
  - curve with higher AUC can be lower in the region of interest



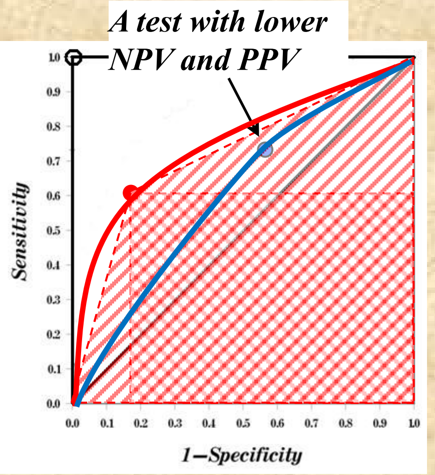
- □ **It is important to examine the ROC curve in addition to analyzing the summary indices**

ROCs and binary tests

- When binary test has better Sens and Spec does it also have a better underlying ROC curve?
  - Yes, at least for some operating points (as *ROC curve is non-decreasing*) but not necessarily for all thresholds


A test with lower Sensitivity and Specificity

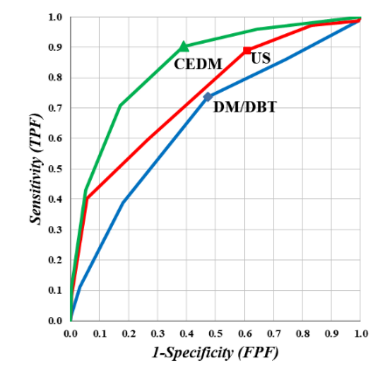
- When a binary test has better PPV and NPV (in the same sample) does it also have better ROC curve?
  - Yes, if a test is reasonable/optimized (*as then the ROC curve is bulging up, or "concave"*)


A test with lower NPV and PPV

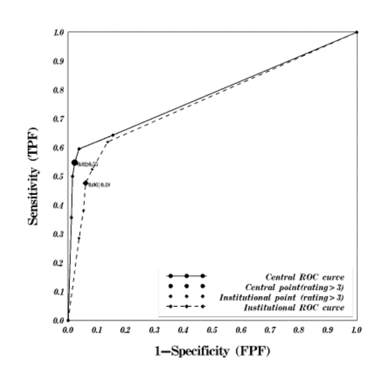
Limitations of the ROC curve

- Typical limitations
  - an entire curve can be difficult to interpret
  - a single-number summary of the ROC curve can be misleading
  - ROC curves for human observers might be difficult to interpret (*potential versus actually achievable performance*)

- ROC curves are not always needed
  - in some cases, a single point (FPF, TPF) (*a pair of Sens and Spec*) can provide sufficient information

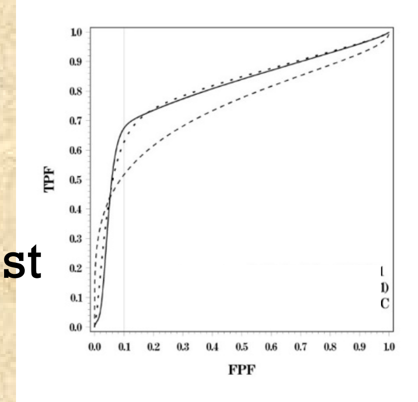


Zuley, et al., AR, 2019



Gee, et al., Rad., 2016

- ROC curves are not always definitive
  - the ROC curves can cross in the region of interest
  - improvement immediately outside the region of interest



Some recommendations

- If a reliable estimate of the ROC curve is available
  - use ROC curve to visually evaluate or compare diagnostic systems
  - quantify the results with appropriate summary indices (*AUC, partial AUC, Sens*$|_{sp=0.9}$, …)
- If only the binary results ("positive"/"negative") are known:
  - intrinsic characteristics (e.g., Sens, Spec) are preferable
  - prevalence-dependent characteristics (e.g., PPV, NPV) require careful handling (*due to dependence on the prevalence in the sample*)
  - single summary index (*odds ratio, Youden's, Sens|spec=0.9,..*) usually needs additional justification
    - no scalar summary index is better than others under all circumstances
    - value requires specific interpretations and is often application specific *e.g.: odds ratio of 3 could correspond to very poor classification tool (Pepe, 2004)*
- To ensure reliability and robustness of the conclusions
  - summary measures, and other design aspects, must be set a priori
  - statistical uncertainty must be properly quantified (*see STARD*)

ROC analysis as a toolbox

- Useful in various tasks of classification, predictions, etc.
- More sophisticated methods and extensions e.g.,
  - Advanced methods
    - parametric (e.g., *binormal ROC*), non-parametric (empirical), semi-parametric
    - adjusting for **covariates**: *modeling ROC curve or its indices* (e.g., ROC-GLM)
    - accounting for **multiple readers** (MRMC)
    - …
  - Extensions
    - time-to-event data – ***time-dependent*** *ROC*
    - more than two classes of truth: ***multi-class*** *ROC analysis*
    - multiple targets(lesions) per subject (**detection and localization problem**): *free-response ROC (FROC), regions of interest approach (ROI)*
    - ……..
- A couple of great textbooks on ROC analysis and related topics
  - Zhou, X.H., **Obuchowski, N.A.**, McClish D.K. (2011). Statistical methods in diagnostic medicine. 2nd edition. New York: Wiley & Sons Inc.
  - Pepe, M.S. (2003). The statistical evaluation of medical test for classification and prediction. Oxford: Oxford University Press.

Enjoy the Workshop!