Receiver Operating Characteristic (ROC) Methods in Diagnostic Imaging

Elizabeth A. Krupinski, PhD
Department Radiology & Imaging Sciences
Emory University

Bit of History

- Developed early 1950s based on principles SDT for eval radar operators detecting enemy aircraft & missiles
- Contributions from engineering, psychology & mathematics
- Lee Lusted introduced medicine 1960s with significant effort on gaining better understanding decision-making
- Result of radiology studies after WWII to determine which of 4 radiographic & fluoroscopic techniques better for TB screening
- Goal = single imaging technique outperform others
- Found intra & inter-observer variation so high impossible determine
- Necessary to build systems generate better images so radiologists' performance could improve (i.e., reduce observer variability) & develop methods evaluate these new systems & assess impact on observer performance

Basics

- ROC traditionally binary decision task – target/signal (e.g., lesion, disease, missile) present versus target/signal absent, or in case classification rather than detection target/signal belongs to class 1 (e.g., cancer, enemy) or class 2 (e.g., not cancer, friend)
- ROC analysis these two conditions must be mutually exclusive

2 x 2 Matrix

	Decision = Target Present	Decision = Target Absent
Truth = Target Present	True Positive (TP)	False Negative (FN)
Truth = Target Absent	False Positive (FP)	True Negative (TN)

Common Performance Metrics

- Sensitivity = TP/(TP + FN)
- Specificity = TN/(TN + FP)
- Accuracy = (TP + TN)/ (TP + FN + TN + FP)
- Positive Predictive Value (PPV) = TP/(TP + FP)
- Negative Predictive Value (NPV) = TN/(TN + FN)

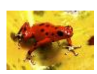
Trade-Offs

- Between sensitivity & specificity – as increase one decrease other
- If want detect more targets (high sensitivity) often occurs as result making more FPs (decreased specificity)
- Why would you want to use sensitivity/specificity versus PPV/NPV?

Prevalence

- Basically former are independent of prevalence of targets in case sample while latter are not
- Suppose have observer expert at visually detecting specific poisonous frog in jungle versus similar but non-poisonous frog
  - Her sensitivity is 95% & specificity 80%
  - In jungle #1 are 1000 frogs total with prevalence 50% poisonous (n = 500)
  - In jungle #2 are also 1000 frogs total but only 25% poisonous (n = 250)

- Jungle #1:TP = 475 FN = 25 FP = 100 TN = 400
- Accuracy = (475 + 400)/(475 + 25 + 100 + 400) = 0.88 or 88%
- PPV = 475/(475 + 100) = 0.83 or 83%
- NPV = 400/(400 + 25) = 0.94 or 94%



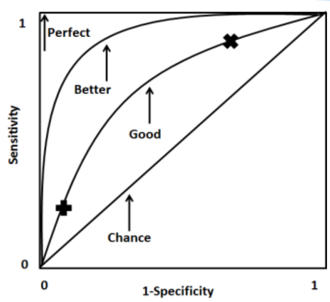
- Jungle #2: TP = 238 FN = 12 FP = 150 TN = 600
- Accuracy = (238 + 600)/(238 + 12 + 150 + 600) = 0.84 or 84%
- PPV = 238/(238 + 150) = 0.61 or 61%
- NPV = 600/(600 + 12) = 0.98 or 98%

Why ROC Useful?

- Many cases sensitivity & specificity adequate measures performance but becomes complicated when test sets contain cases with range difficulty
- Observer's decision threshold for reporting can change as function many things, including but not limited to nature target, target prevalence, background complexity within which the target is embedded, number & type targets, observer experience or expertise

ROC Curve

- Captures relationship sensitivity & specificity plus range decision thresholds every observer
- Curve = representation relationship sensitivity (TP fraction) vs 1-specificity (FP fraction or 1 – TN/(TN + FP) = FP/(FP + TN)) for all decision thresholds
- Axes go 0 to 1 & diagonal line = chance or guessing
- Curves indicate better performance as move to upper left corner = perfect performance



+ = conservative; x = liberal

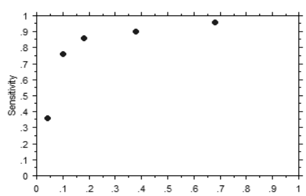
Example of distribution of confidence scores for a subject in an observer performance study with a 6-point confidence scale & images with target present or absent (truth)

Truth	1	2	3	4	5	6
Present	2	3	2	5	20	18
Absent	16	15	10	4	3	2

Sensitivity, specificity & FP fraction can then be determined at each threshold or cutoff point

Result positive is ≥	Sensitivity	Specificity	FP fraction
2 - probably absent	0.96 (48/50)	0.32 (16/50)	0.68
3 - possibly absent	0.90 (45/50)	0.62 (31/50)	0.38
4 - possibly present	0.86 (43/50)	0.82 (41/50)	0.18
5 - probably present	0.76 (38/50)	0.90 (45/50)	0.10
6 - definitely present	0.36 (18/50)	0.96 (48/50)	0.04

ROC curve generated from the data in Table 2



Fitting the Curve

- "Connecting dots" empirically based version but creates stepped or jagged plot
- Smooth curve reflecting theoretical "true" curve much more desirable
  - Non-parametric has no assumptions structure underlying distribution & essentially smooths histograms of output data for 2 classes
  - Parametric relies on validity underlying distribution assumptions
  - Most researchers prefer parametric

Interpreting ROC Curve

- Most common AUC or Az
- Diagonal = chance = AUC 0.5
- Top left = perfect & encompasses all area = 1.0
- Curve between chance & perfect = 0.5 - 1.0
- As with generation curve there are variety methods calculate AUC & most programs use one of these methods
- Less common = d´, d_e´, Dm, B and Z_k

EMORY UNIVERSITY SCHOOL OF MEDICINE

Partial AUC

- AUC often not appropriate as not all decision thresholds equally important - real life observers may not actually operate at se threshold
- Diagnostic test with low specificity may not be clinically acceptable so select “acceptable” FP rate & determine its associated sensitivity (TP rate) then calculate AUC only up to operating point (i.e., capturing part total AUC)
- Very common in development of CAD algorithms

EMORY UNIVERSITY SCHOOL OF MEDICINE

Comparing Curves

- Visually not always possible tell difference is significant
- Especially true if curves cross (usually upper right)
- Methods developed = parametric & non-parametric options
- Most common comparing multiple observers & multiple cases = Multi-Reader Multi-Case method developed by Dorfman, Berbaum, Metz

EMORY UNIVERSITY SCHOOL OF MEDICINE

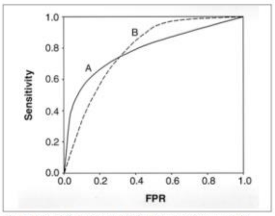


Fig. 3. Two ROC curves (A and B) with equal area under the ROC curve. However, these two ROC curves are not identical. In the high false positive rate range (or high sensitivity range) test B is better than test A, whereas in the low false positive rate range (or low sensitivity range) test A is better than test B.

```
=============================================================================
*****                           Estimates                              *****
=============================================================================

TREATMENT x READER AUC ESTIMATES

                  TREATMENT
          -----------------------
READER         1            2
------    ----------   ----------
    1     0.86872166   0.83994471
    2     0.81920700   0.77936491
    3     0.87579463   0.88957602
    4     0.81715215   0.82460854
    5     0.89398208   0.84914684
    6     0.81649540   0.76688132

TREATMENT AUC MEANS (averaged across readers)
---------------------------------------------
      1      0.84855882
      2      0.82492039


TREATMENT AUC MEAN DIFFERENCES
------------------------------
   1 - 2     0.02363843


=============================================================================
*****             ANOVA Tables (OR analysis of reader AUCs)            *****
=============================================================================

TREATMENT X READER ANOVA of AUCs
(Used for global test of equal treatment AUCs and for treatment differences
 confidence intervals in parts (a) and (b) of the analyses)

Source             SS               DF             MS
------   --------------------     ------   ------------------
     T             0.00167633          1           0.00167633
     R             0.01463444          5           0.00292689
   T*R             0.00189007          5           0.00037801


READER ANOVAs of AUCs for each treatment
(Used for single treatment confidence intervals in part (c) of the analyses)

                        Mean Squares
Source     df   Treatment 1   Treatment 2
------    ---   -----------   -----------
     R      5    0.00121751    0.00208739
```

Other ROCs

- In real life images contain multiple targets & FPs can occur in both target present & absent stimuli
- Traditional ROC analysis typically does not ask or require observer to locate target once detected
- Always some question whether actually detected true target (TP) or called something else in image (FP) thereby actually missing true target (FN)

LROC

- LROC (Location ROC) - observer reports somewhere in image is target & marks location most suspicious region
- Only allows single target
- Hard to generate curve & AUC

FROC

- Free Response ROC - observers mark different locations & provide confidence each mark
- Curve plots lesion localization fraction (LLF) on y-axis & non-LL (NLF) fraction on x-axis (denominators = total # targets & total # images respectively)
- Plot: y-axis goes from 0 to 1 but x-axis goes from 0 to some number depending on number FP making calculation of AUC difficult

AFROC

- Alternative FROC method developed to address
- Creating plot that has both axes going from 0 to
- Jackknife AFROC (JAFROC) method was then developed to allow for generalization to population of readers and cases in same way that MRMC ROC does

Other Considerations

- Truth or gold standard
- Lesions: how subtle, mix, locations, sizes, background, prevalence
- How long to display, zoom/pan. window/level
- Sample size - # images & observers
  - metric under consideration (Az) & design (e.g., repeated measures with same observers viewing same images 2+ conditions or different readers viewing images in different conditions

There is Software!

- University of Iowa http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/tabid/120/Default.aspx
- University of Chicago http://metz-roc.uchicago.edu/
- FROC http://perception.radiology.uiowa.edu/Software/ReceiverOperatingCharacteristicROC/tabid/120/Default.aspx

More Software!

- MedCalc Statistical Software https://www.medcalc.org/manual/roc-curves.php
- Analyse-It http://analyse-it.com/docs/220/method_evaluation/roc_curve_plot.htm
- NCSS Statistical Software http://www.ncss.com/software/ncss/procedures/
- SPSS Statistics http://www-03.ibm.com/software/products/en/spss-statistics
- STATA Data Analysis and Statistical Software http://www.stata.com/features/overview/receiver-operating-characteristic/