Methodologies for Evaluation of Standalone CAD System Performance

Berkman Sahiner, PhD
USFDA/CDRH/OSEL/DIAM
AAPM CAD Subcommittee in Diagnostic Imaging

INTRODUCTION

- CAD: CADe and CADx
  - CADe: Identify portions of an image to reveal abnormalities during interpretation by reader
  - CADx: Provide assessment of disease; specify disease severity, type, or stage to the reader
- Standalone assessment
  - Assessment of the performance of device alone
    - Assessment of the effect of CAD on the reader is next talk

CAD SYSTEM ASSESSMENT

- Measure the performance of your system
  - Inform users, regulators, scientific community, and yourself
  - Establish its effectiveness for use
  - Compare with other systems with a similar intended use
- If you can't assess it, you will not know how to improve it

STANDALONE VERSUS WITH READERS

- The effect of CAD on the radiologists' performance is the ultimate test
  - Currently, CAD devices in radiology are intended for use by radiologists
  - Not standalone or triage use
- The effect of CAD on the readers' performance may be more burdensome to assess than standalone

STANDALONE VERSUS WITH READERS

- Merits of standalone assessment
  - Potential impact at early stage of development, prior to testing with readers
  - Potentially large datasets, more amenable to subset analysis
  - Reader variability is eliminated

COMPONENTS OF CAD ASSESSMENT

- Dataset
- Reference standard
- Mark-labeling
- Assessment metric

DATASETS

- Training
  - In theory, known properties of abnormals and normals may suffice for CAD
  - In practice, many parameters are determined using a training data set
- Test
  - Used for performance assessment
- Mixing training and test sets introduces optimistic bias to CAD assessment

DATASETS

- Images and data components used as inputs to the CAD system
- Other images necessary for reference standard
- Other data to provide context and perform sub-group analysis
  - Age, demographics, disease type, lesion size, concomitant diseases

TRAINING DATASET

- Ideally, covers the spectrum of intended task
- May not need to be representative
  - Sub-group may be over-represented if thought to be more difficult or more important
- May include
  - Phantom images
  - Electronically altered images

TEST DATASET

- Independent of the training data set used at any stage of development
- Should include the range of abnormalities for the target population
- Image acquisition and patient preparation parameters should be consistent with those in the target population
- Should be large enough for adequate statistical power to demonstrate study objectives

ENRICHMENT

- Low prevalence disease
  - Enhance with cases containing disease
    - Will not affect sensitivity, specificity, area under the ROC curve
    - In an observer study, may affect the reader's behavior

SPECTRUM OF DIFFICULTY

- Spectrum of difficulty for test cases versus spectrum of difficulty for intended population:
  - If different, test results may be biased
- Bias may be acceptable if
  - Comparing two modalities

  and
  - both modalities are affected similarly by spectrum bias

STRESS TESTING

- Study differences between competing modalities using cases selected to challenge those differences*
- Example in CADe
  - Excluding obvious cases because they will be detected both with and without CAD

RF Wagner et al, "Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review," Acad Radiol 14, 723-748 (2007).

TEST DATASET REUSE

- Can I keep using the same test dataset while trying to improve my CAD system?
  - Starting over with a completely new dataset
    - Burdensome
    - Does not promote enlarging the dataset, i.e., reducing uncertainty in performance estimates
- Danger: Tuning the CAD system explicitly or implicitly to "test" dataset

TEST DATASET REUSE

- Risks / benefits need to be weighed depending on
  - The stage of CAD algorithm design
    - e.g., an early-stage CAD design for a new modality
      - Should acknowledge data set reuse
  - How dataset reuse occurred
    - e.g., were detailed results reported back to algorithm design?

COMMON SEQUESTERED DATASET

- Some public datasets available, but not sequestered
- Sequestered dataset for independent testing
- Must ensure
  - CAD systems are not tuned to sequestered dataset
  - Dataset evolves over time, does not become obsolete

DATASET SUMMARY

- Very critical in both design and assessment
- For assessment purposes, training does not need to be “optimal”
  - Training dataset may not have to follow the distribution of intended population
- Independent test dataset essential
- Prevalence enrichment often necessary

REFERENCE STANDARD

- Disease status
  - Ideally, independent of the modality that CAD is designed for
- Location and extent of disease
  - Ideally, additional data or images are used to complement the images targeted by CAD

REFERENCE STANDARD: DISEASE STATUS

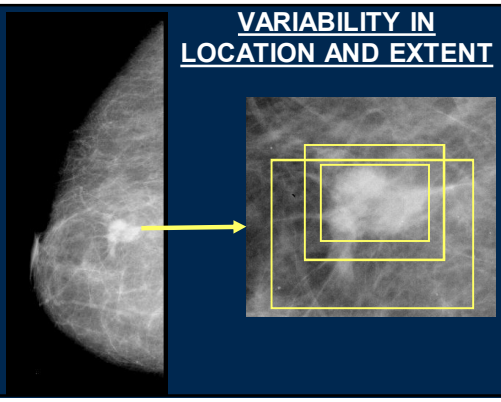
- Disease status often known by biopsy, follow-up, or other method with very high accuracy
  - Mammography
    - However*,
      - 11-gauge vacuum-assisted biopsy: 0.8–1.7% rate of discordance
      - 14-gauge vacuum-assisted biopsy: 3.3–6.2% rate of discordance
- If long-term follow-up is missing, negative cases may have uncertainty
- In other situations, the imaging modality for CAD may be standard of care
  - CT for pulmonary embolism

*ES Burnside et al., "A probabilistic expert system that provides automated mammographic–histologic correlation: Initial experience." AJR 182. 481-488 (2004)

REFERENCE STANDARD: LOCATION AND EXTENT

- Required in CADe if truth location is part of the assessment
  - Generally the case for standalone CADe assessment
- Other imaging data often available to locate disease
  - Breast cancer: Images acquired during biopsy
  - Colon cancer: Optical colonoscopy
- In other situations, additional imaging data may not be available
  - CT for pulmonary embolism

VARIABILITY IN LOCATION AND EXTENT



LACK OF GOLD STANDARD

- Expert panel
  - Combine expert readers' interpretations into a reference standard
  - Example:
    - Each reader first reads independently
    - Interpretations are merged using an adjucation method
      - Majority, independent arbiter
  - Uncertainty in truth

REFERENCE STANDARD - SUMMARY

- In practice, a perfect reference standard may be difficult to establish for many CAD applications
  - Practical scenario: Use as much information as possible, but recognize that the reference standard may not be perfect
- Expert panels
  - May be beneficial or may be the only option in some applications
  - Additional uncertainty in truth

MARK-LABELING

- Rules for declaring a mark as a TP or FP
  - Applies to CADe only

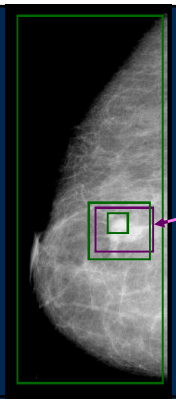
MARK-LABELING

- By a human:
  - A human may be a good judge for deciding whether a mark points to a FP
  - May be subjective
    - Labeler should not have a stake in the outcome of assessment to reduce bias
  - May be burdensome if repeated mark-labeling is desired

MARK-LABELING

- Automated:
  - Compare computer mark to reference standard mark using an automated rule
    - Overlap of computer and reference standard marks
    - Centers of computer and reference standard marks
    - Distance of centroids
- Some methods better at the task than others

MARK-LABELING



MARK-LABELING

- Most studies do not report the mark-labeling protocol
  - Randomly-selected publications on CADe
  - Nodule detection on CT
    - 47/58 (81%) did not report mark-labeling protocol
  - Polyp detection in CT colonography
    - 9/21 (43%) did not report mark-labeling protocol

MARK-LABELING SUMMARY

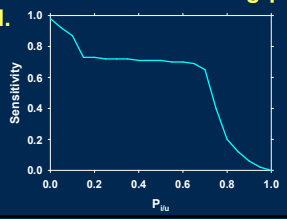
- It is important to specify the mark-labeling method in a study
  - It can have a major effect on the reported performance of the CADe system*
- Methods that have the potential to label clearly unhelpful marks as TPs should be avoided

*M Kallergi et al., "Evaluating the performance of detection algorithms in digital mammography," Med Phys 26, 267-275 (1999)

MARK-LABELING SUMMARY

- If a parameter is used in mark labeling
  - e.g., Area(intersection) / Area(union) > $P_{i/u}$

  it is helpful to study how performance is affected when the mark-labeling parameter is modified.



PERFORMANCE MEASURES: BINARY OUTPUT

- Many CAD systems internally produce continuous (or multi-level) scores
  - If so, assume a threshold has been used
- CADx system with binary output
  - Positive
  - Negative
- CADe system that marks potential lesions
  - Mark
  - No mark

CADx: TRUE AND FALSE-POSITIVE FRACTIONS

$$TPF = \frac{\text{Number of units (images) correctly called positive}}{\text{Total number of positive units (images)}}$$

$$FPF = \frac{\text{Number of units (images) incorrectly called positive}}{\text{Total number of negative units (images)}}$$

Unit: 2D or 3D image, region-of-interest, case

CADe: LESION AND NON-LESION LOCALIZATION FRACTIONS

- Lesion localization fraction (LLF) ~ Sensitivity
- Non-lesion localization fraction (NLF) ~ Number of FPs per unit

$$LLF = \frac{\text{Number of correctly marked locations}}{\text{Total number of abnormalities}}$$

$$NLF = \frac{\text{Number of incorrectly marked locations}}{\text{Total number of negative units (images)}}$$

(TPF, FPF) AND (LLF, NLF) PAIRS

- Always in pairs
- Should always be accompanied with uncertainty estimates or confidence intervals
  - TPF, FPF, LLF: Binomial
    - Normal approximation, Wald interval
    - More accurate: Agresti-Coull*, or Jeffreys** interval
  - NLF: Poisson
    - Normal approximation, Wald interval
    - More accurate: Jeffreys** interval

*A Agresti and BA Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," American Statistician 52, 119-126 (1998)

** LD Brown, et al., "Interval estimation in exponential families," Statistica Sinica 13, 19-49 (2003)

COMPARISON OF TWO STANDALONE SYSTEMS A AND B

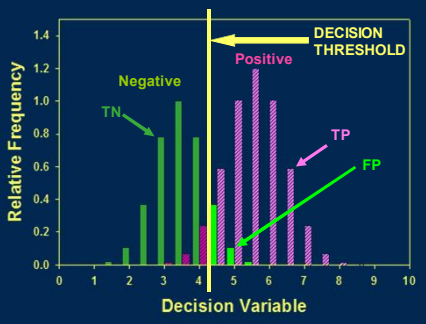
- System A is better if
  - TPF_A is significantly higher that TPF_B

  and
  - FPF_A is significantly lower than FPF_B
- In practice, a high bar to achieve

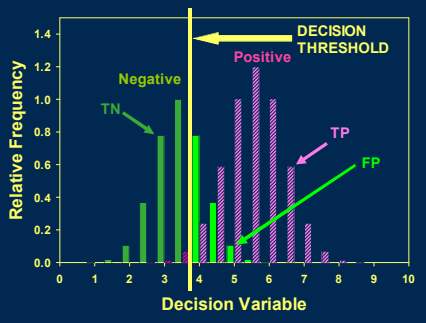
COMPARISON OF TWO CADx SYSTEMS

- Often, both members of the (TPF, FPF) pair are higher for one system compared to the other
  - Higher TPF but also higher FPF
  - Lower TPF but also lower FPF
- Instead of (TPF, FPF) at a fixed threshold, use the continuous scores for each unit (image)
  - Compare ROC curves

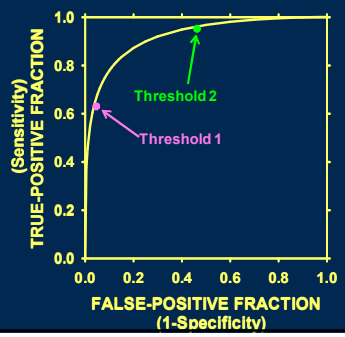
THRESHOLDING



THRESHOLDING



RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE



FIGURES OF MERIT

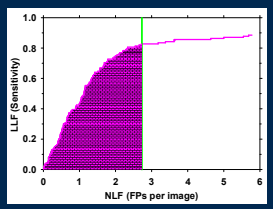
- Area under the curve (AUC)
- Partial area under the curve
  - Important to pre-specify which part of the ROC curve you are interested in *before* performing the comparison
- Point estimates should always be accompanied with confidence intervals

ROC ANALYSIS

- Numerous methods in the literature
- To fit the data and estimate uncertainties
  - Parametric
- To estimate FOMs and uncertainties
  - Both parametric and non-parametric
- To statistically compare FOMs of two systems
  - Both parametric and non-parametric

LOCATION-SPECIFIC ROC ANALYSIS

- ROC: Scores
- Location-specific ROC: (Mark, Score) pair
  - LROC, AFROC, FROC, EFROC



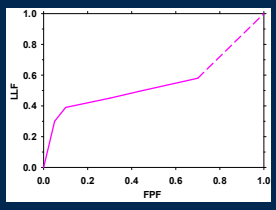
Area under FROC (FPPI threshold)

Bootstrapping*

* FW Samuelson and N Petrick, "Comparing image detection algorithms using resampling," IEEE Int Symp on Biomedical Imaging: 1-3, 1312-1315 (2006)

LOCATION-SPECIFIC ROC ANALYSIS

- ROC: Scores
- Location-specific ROC: (Mark, Score) pair
  - LROC, AFROC, FROC, EFROC



FROC Data →

AFROC Data →

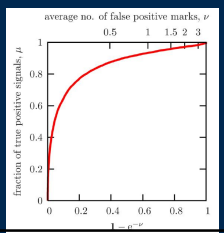
Area Under AFROC →

JAFROC*

*DP Chakraborty, "New Developments in Observer Performance Methodology in Medical Imaging," Semin Nucl Med, 41:401-418 (2011)

LOCATION-SPECIFIC ROC ANALYSIS

- ROC: Scores
- Location-specific ROC: (Mark, Score) pair
  - LROC, AFROC, FROC, EFROC



FROC data →

Exponential transform →

EFROC curve →

Area under EFROC*

*LM Popescu, "Nonparametric signal detectability evaluation using an exponential transformation of the FROC curve," Med Phys 38, 5690-5702 (2011)

PERFORMANCE MEASURES - SUMMARY

- (TPF, FPF) or (LLF, NLF) pairs are good starting points
- If you have continuous scores, you can do more
  - ROC
  - FROC, AFROC, EFROC
- Point estimates should always be accompanied with confidence intervals or measures of variability

SUMMARY

- Standalone CAD assessment has its own merits compared to assessment of CAD systems' effect on users
- Important components in CAD assessment:
  - Dataset, reference standard, mark-labeling procedure, assessment metric