How low can CT dose go?

Are the dose reduction claims justified?

Frederic Noo, PhD
Department of Radiology and Imaging Sciences
University of Utah

Scope

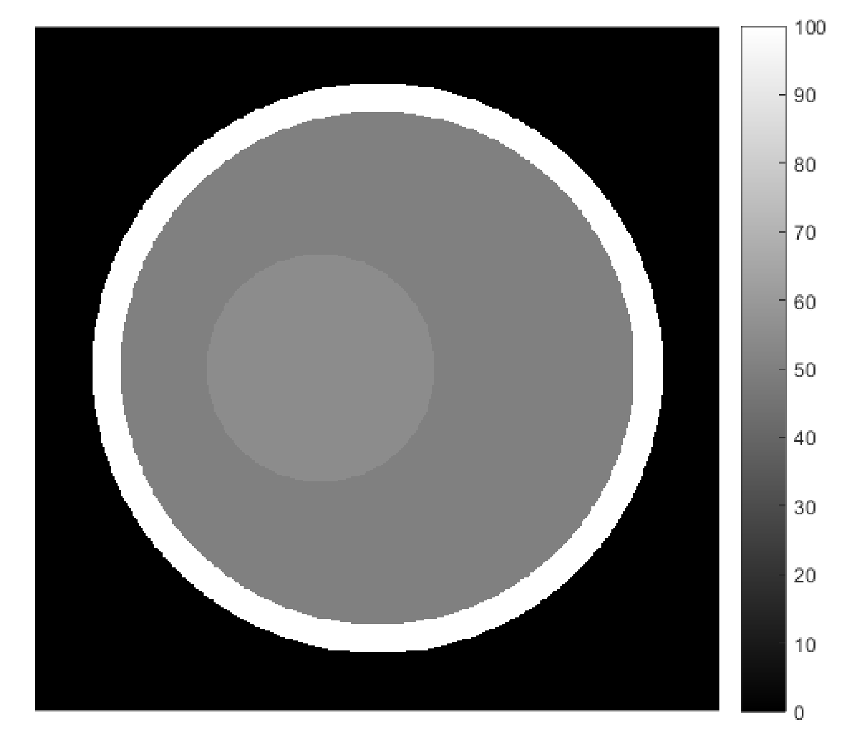
- A dose reduction claim is only of interest if the image quality is maintained
- What is the meaning of image quality?
- How should we assess image quality?
- What are the strengths and limitations of various definitions?

Subjective approach

- Most “crass” technique: show two images and claim that one is better than the other one for whatever reason

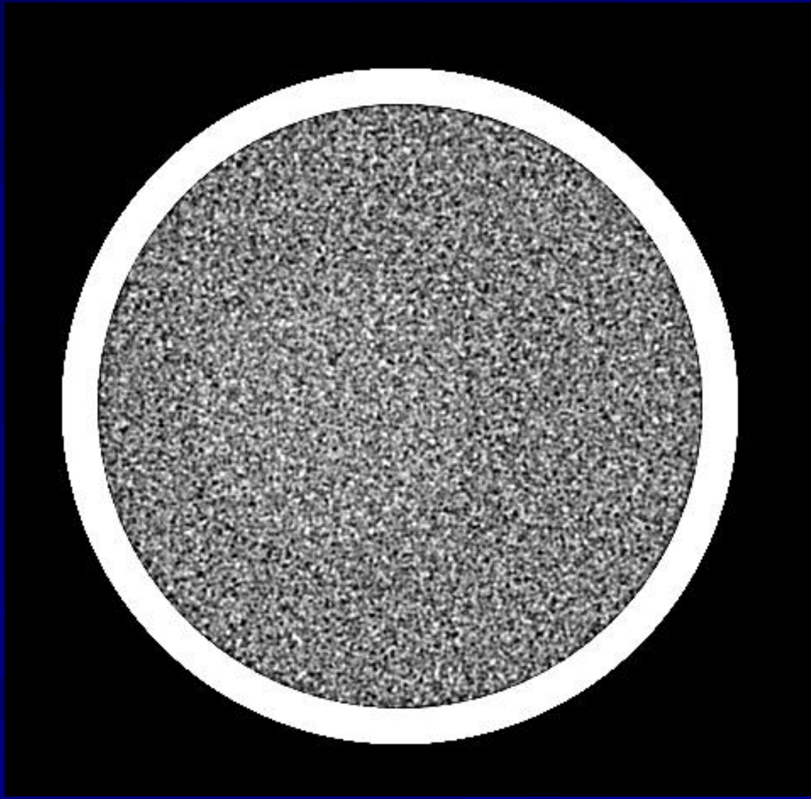
Subjective approach

- Most “crass” technique: show two images and claim that one is better than the other one for whatever reason

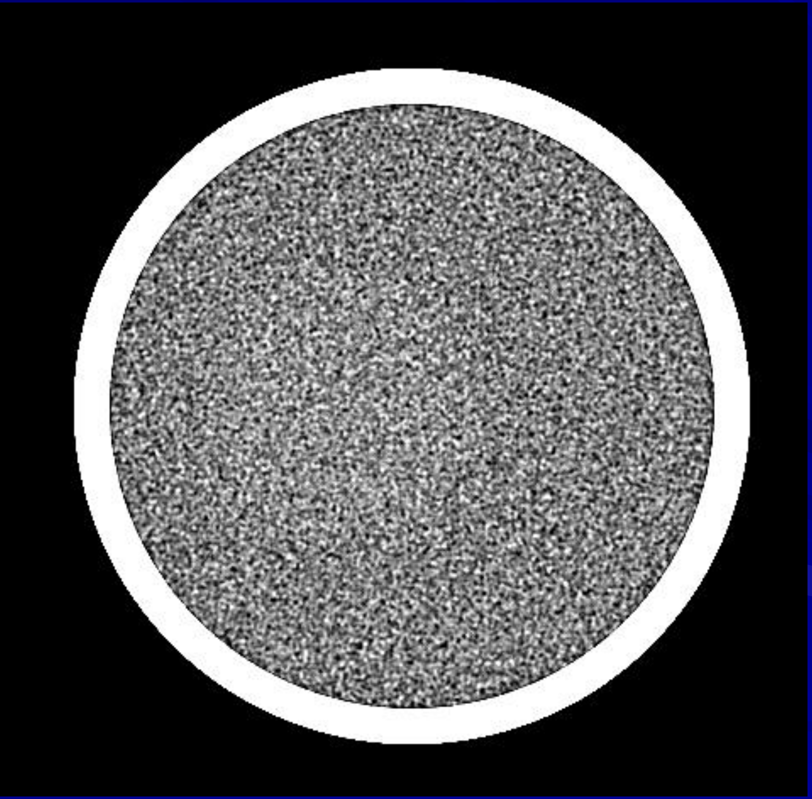


Subjective approach

- Most “crass” technique: show two images and claim that one is better than the other one for whatever reason



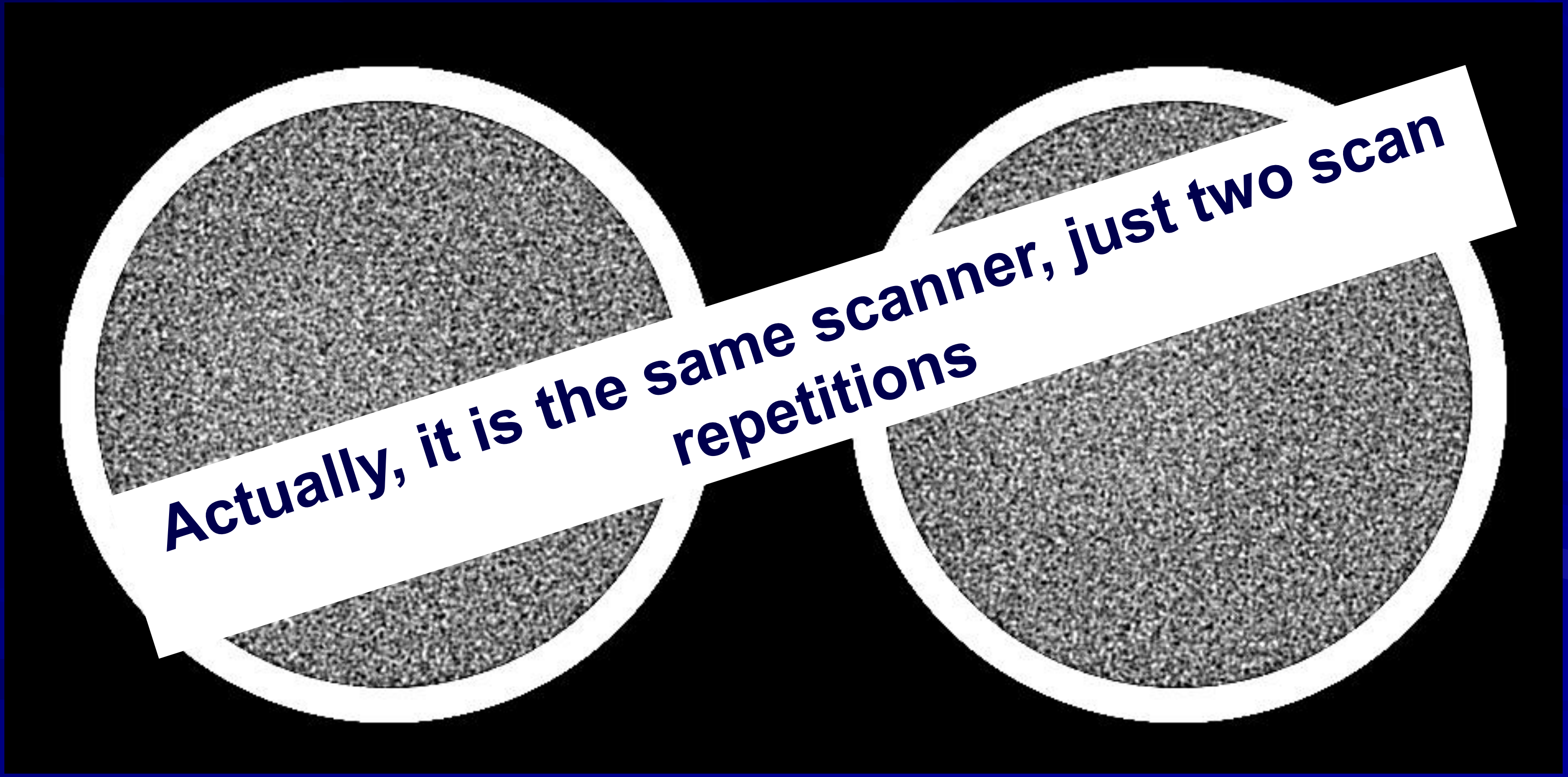
Scanner A



Scanner B

Subjective approach

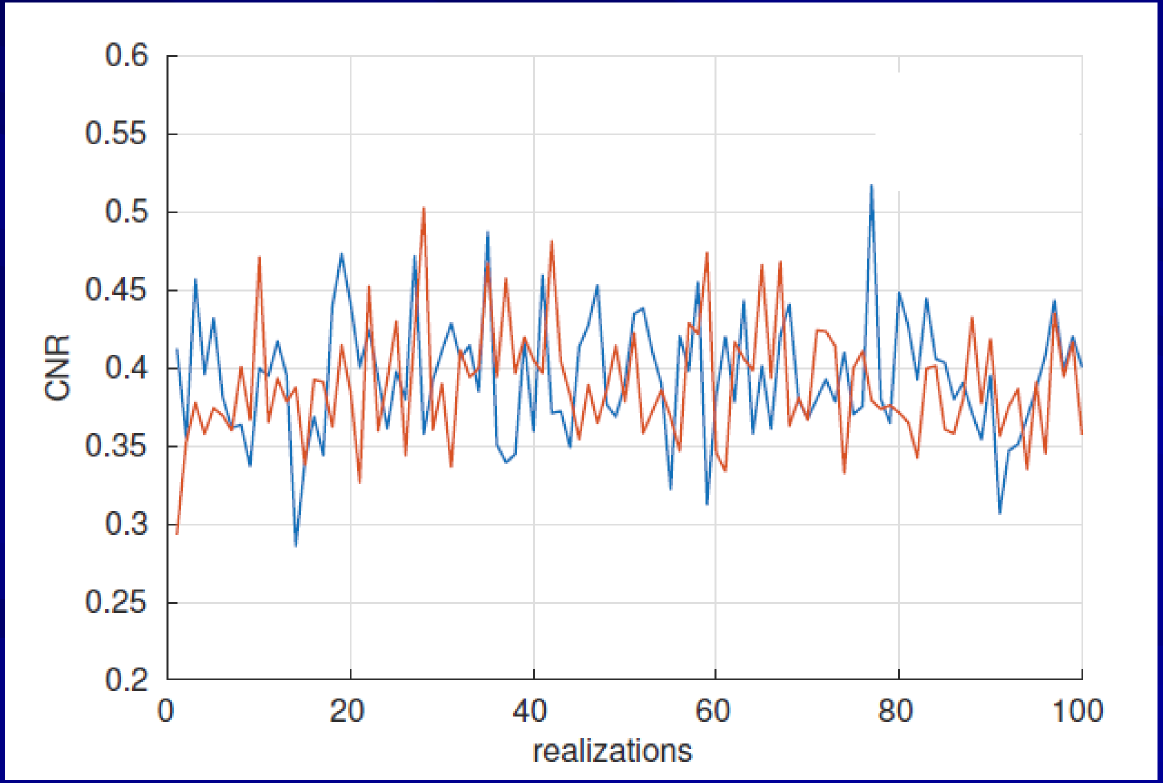
- Most “crass” technique: show two images and claim that one is better than the other one for whatever reason



Actually, it is the same scanner, just two scan repetitions

Subjective approach: what happened …

- A CT image is a multivariate random deviate
- Hardware changes decorrelate image comparisons
- Also possible for weakly correlated image comparisons



Absolute objective metrics

- Image sharpness/resolution: (MTF, SSP)

- Image noise (pixel variance, noise power spectrum)

- Image artifacts

Unfortunately, modern imaging solutions have complicated their use:

  - Resolution can depend on background and contrast
  - Ensemble-averaged resolution not the same as resolution measured from one image
  - Noise linked to background uniformity
  - Noise over a uniform ROI poorly predict image quality
  - Non-linear post-processing or deep learning techniques can hide artifacts

Relative objective metric

Mean squared error, peak signal-to-noise-ratio

$$\mathrm{MSE}(x,y) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(x_{ij}-y_{ij})^2, \quad \mathrm{PSNR} = 10\log_{10}\left(\frac{1}{\mathrm{MSE}}\right)$$

Features:

- Quantitative comparison between two images (e.g., full dose, low dose), but …
- Not related to visual perception
- Single pair of images
- Sensitive to image distortions (rigid and not)

A newly-popular relative objective metric

Structure Similarity Index Metric (SSIM). The closer to unity the better.

$$\mathrm{SSIM}(x,y) = [l(x,y)]^{\alpha}\ [c(x,y)]^{\beta}\ [s(x,y)]^{\gamma}$$

$$l(x,y) = \frac{2\mu_x\mu_y + b_1}{\mu_x^2 + \mu_y^2 + b_1} \quad c(x,y) = \frac{2\sigma_x\sigma_y + b_2}{\sigma_x^2 + \sigma_y^2 + b_2} \quad s(x,y) = \frac{2\sigma_{xy} + b_3}{\sigma_x\sigma_y + b_3}$$

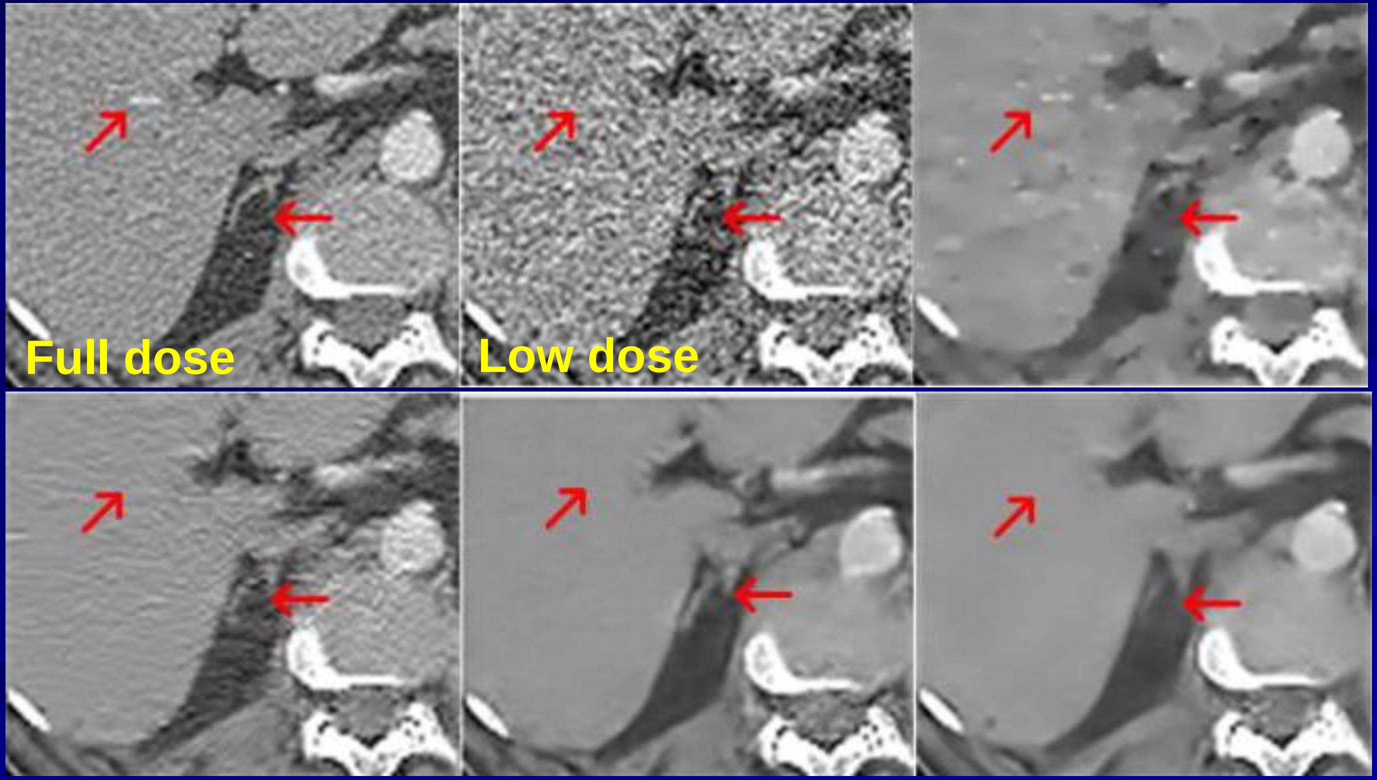
Features:

- Related to visual perception, but …
- Single pair of images
- Sensitive to image distortions (rigid and not)

Implementation:

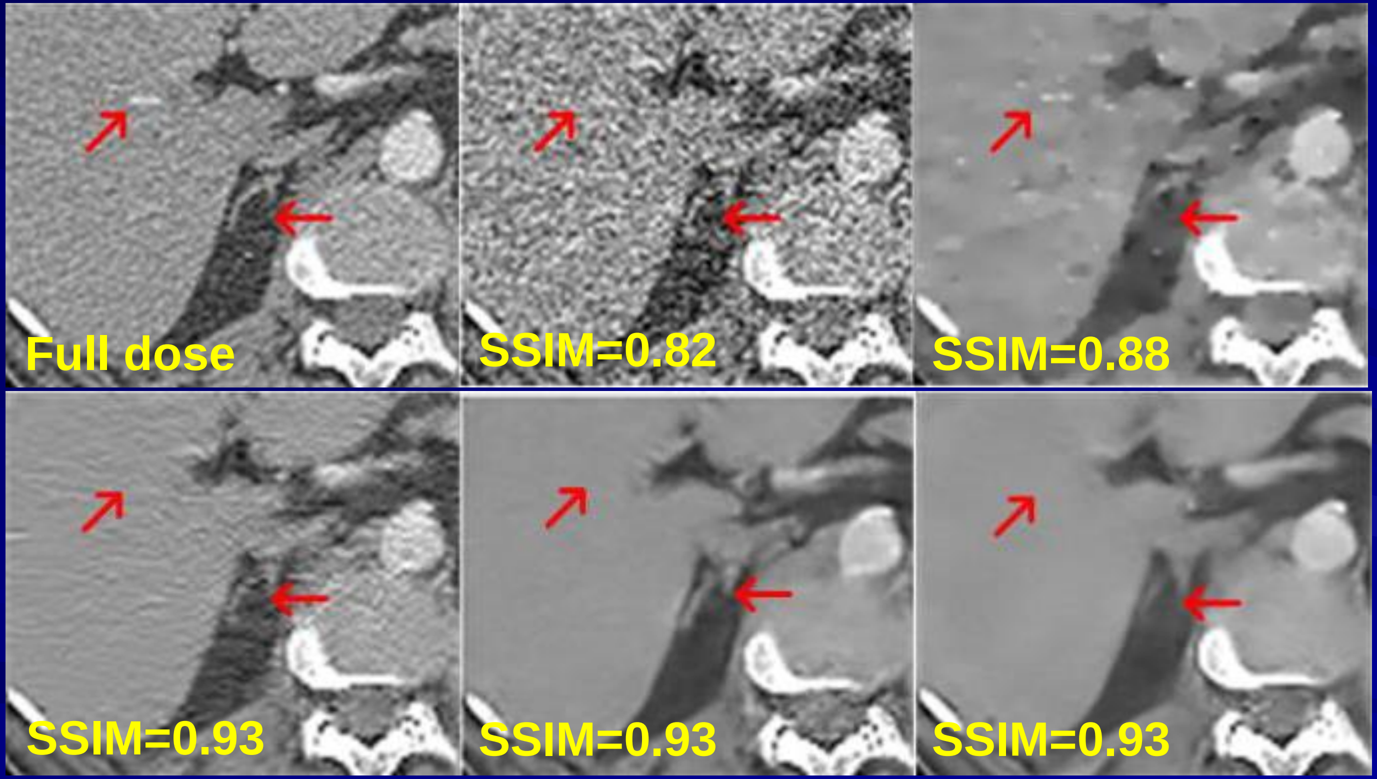
- Rescale to [0, 1.0] or [0, 255] after clipping to grayscale window

SSIM in action …



SSIM in action …

SSIM in agreement with general visual perception, but does not capture subtle changes



More elaborated: reader preference study

- Show images to radiologists and ask them to provide scores (Likert scale) for various features (noise level, artifacts, clarity of vessels).
- Strengths:
    - Relies on a population of cases (unlike CNR, MSE, SSIM)
    - Involves an observer
- Weaknesses:
    - Not predictive of diagnostic value
    - Essentially still a “beauty” or “art” contest

➡ Images should be evaluated relative to the task(s) they are built for

Essential elements of objective task-based assessment of image quality (HH Barrett, K Myers)

- Task: estimation, characterization, detection

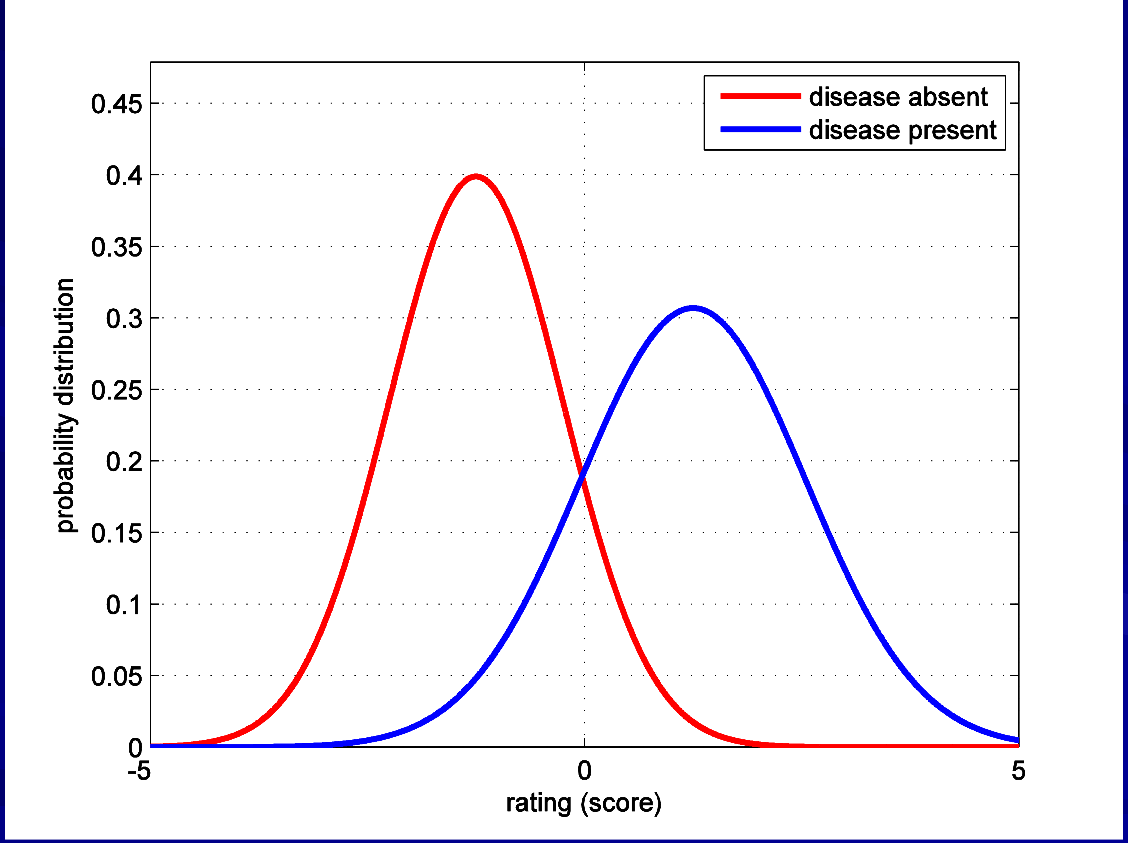
- Observer: human or computerized (model observer)

- Images (cases): population based

- Figure-of-merit for task performance:
  - Estimation: continuous variable (e.g., cardiac ejection fraction, lesion diameter)
  - Characterization: ROC curve for binary classification
  - Detection: ROC curve with localization

ROC curve for a diagnostic test (e.g., PSA)

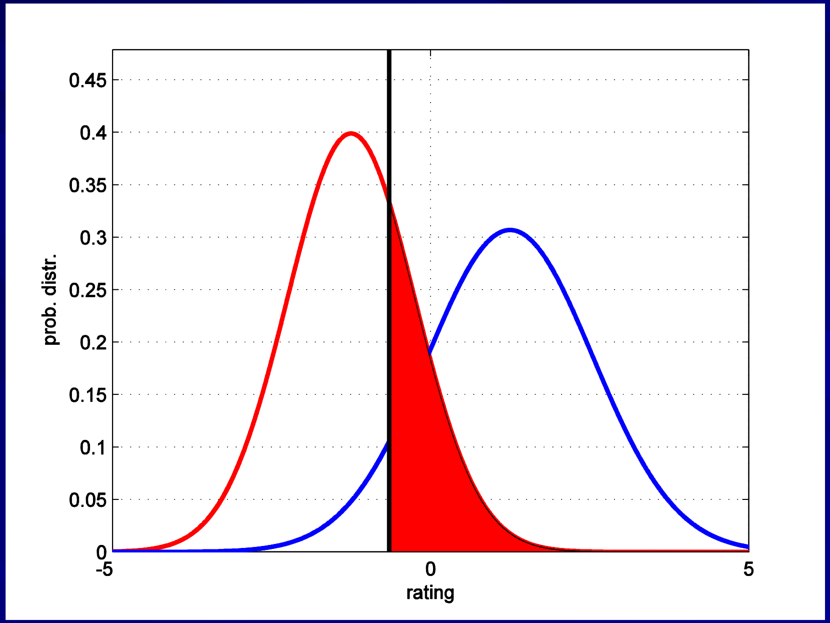
1. Class-based scores



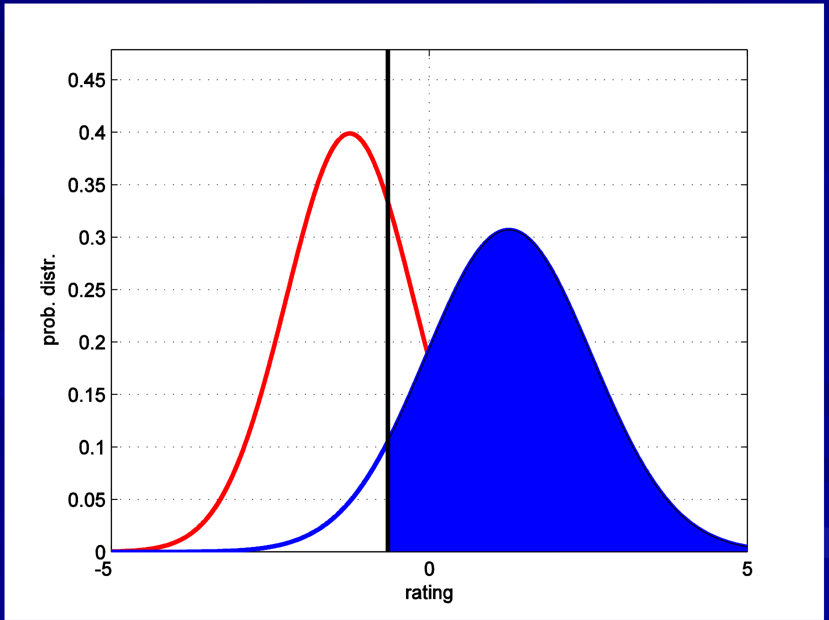
*Scale unimportant up to any monotonic transformation
**Convention: higher scores for disease present

ROC curve for a diagnostic test (cont'd)

2. Decision based on a threshold



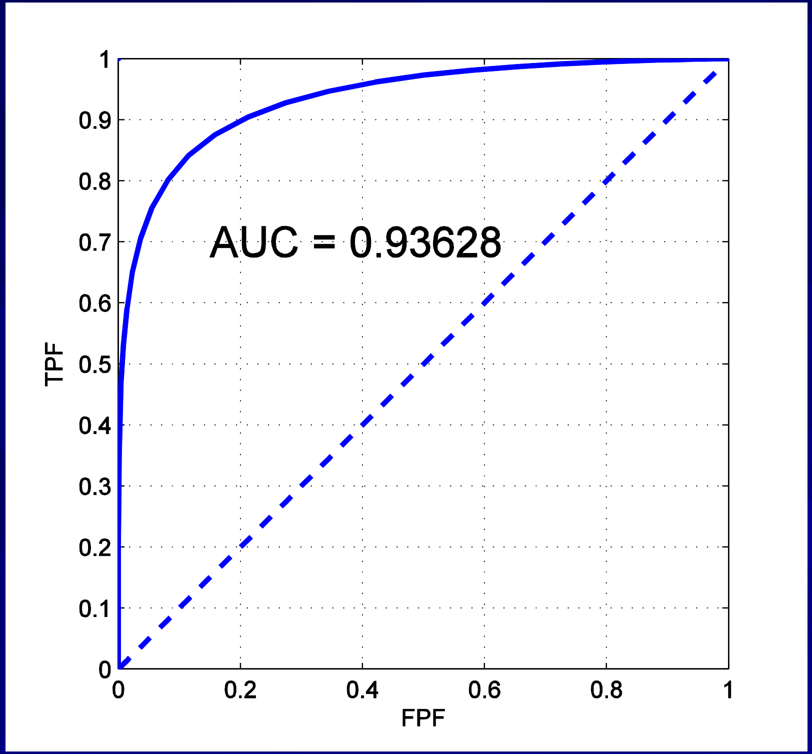
False positive fraction (FPF)



True positive fraction (TPF)

ROC curve for a diagnostic test (cont'd)

The ROC curve is a plot of TPF vs FPF



$$
\begin{aligned}
X &= \{\text{score given patient in class 1}\} \\
Y &= \{\text{score given patient in class 2}\} \\
\\
\mathrm{TPF} &= \mathrm{Prob}(Y > t_c) \\
\mathrm{FPF} &= \mathrm{Prob}(X > t_c) \\
\\
\mathrm{AUC} &= \mathrm{Prob}(Y > X)
\end{aligned}
$$

TPF : sensitivity
1-FPF : specificity

- Figure-of-merit: area under the curve (AUC), also called probability correct. The higher AUC, the better the test.
- AUC values are used to compare tests.

ROC for a characterization task in medical imaging

Class 1: images with disease absent (or benign lesion)
Class 2: images with disease present (or malignant lesion)

Score: reader confidence that disease is present or not (latent decision variable)

Figure-of-merit: reader-average AUC

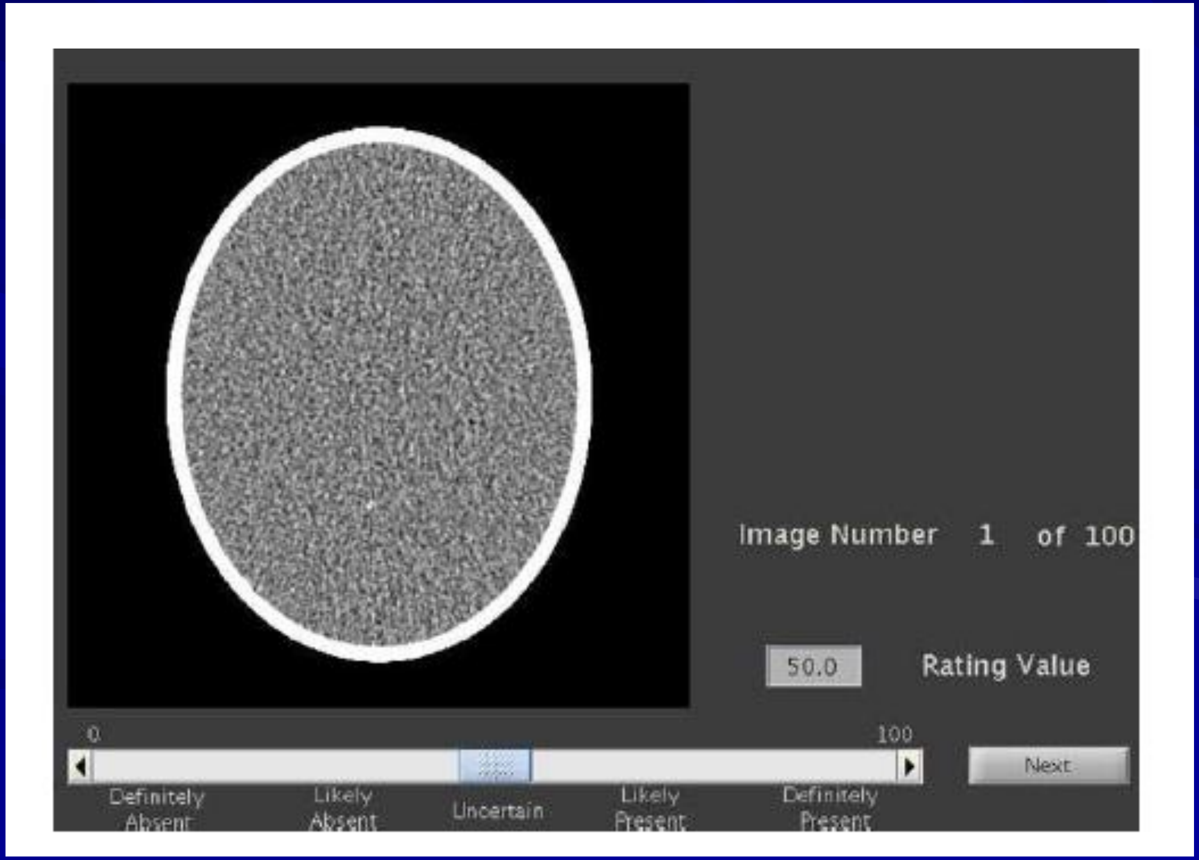


Image Number 1 of 100

50.0 Rating Value

ROC for a characterization task in medical imaging (cont'd)

Important implementation aspects

- Ground-truth needed
- Ordinal or continuous scoring scale can be used
- Thorough task explanation needed
- Typical experiment: training followed by testing
- Keep any reading session < 2 hours (reader fatigue)
- Use multiple reading sessions if needed
- Randomize order of compared methods, as well as cases
- Report results with proper statistical analysis (more on this later)

Lesion detection tasks

- ROC analysis is sub-optimal for lesion detection tasks: outcome of visual search not included

- In addition to deciding that disease is present, it is important to identify where the disease is

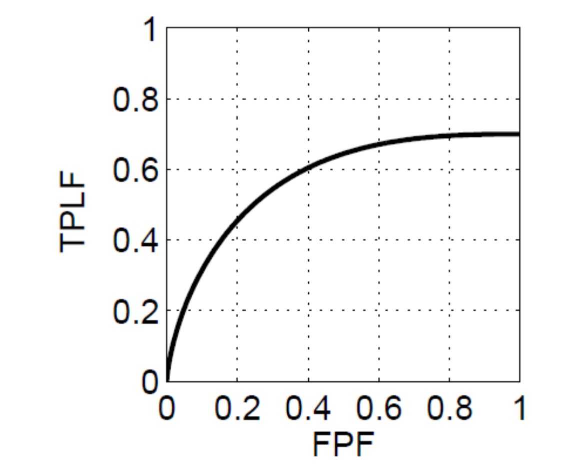
- A positive case should be deemed positive only if the lesion has been located

- A richer assessment paradigm is needed

→ Localization-ROC (LROC) analysis (Starr et al. 1975)

LROC analysis

- Class 1: no lesion. Class 2: one lesion
- Decide if the lesion is present or not, find its location and give a rating
- For each threshold replace TPF by TPLF: fraction of TPs that are correctly localized
- LROC curve: plot of TPLF vs. FPF



$$X = \{\text{score for class 1}\}$$
$$Y = \{\text{score for class 2}\}$$

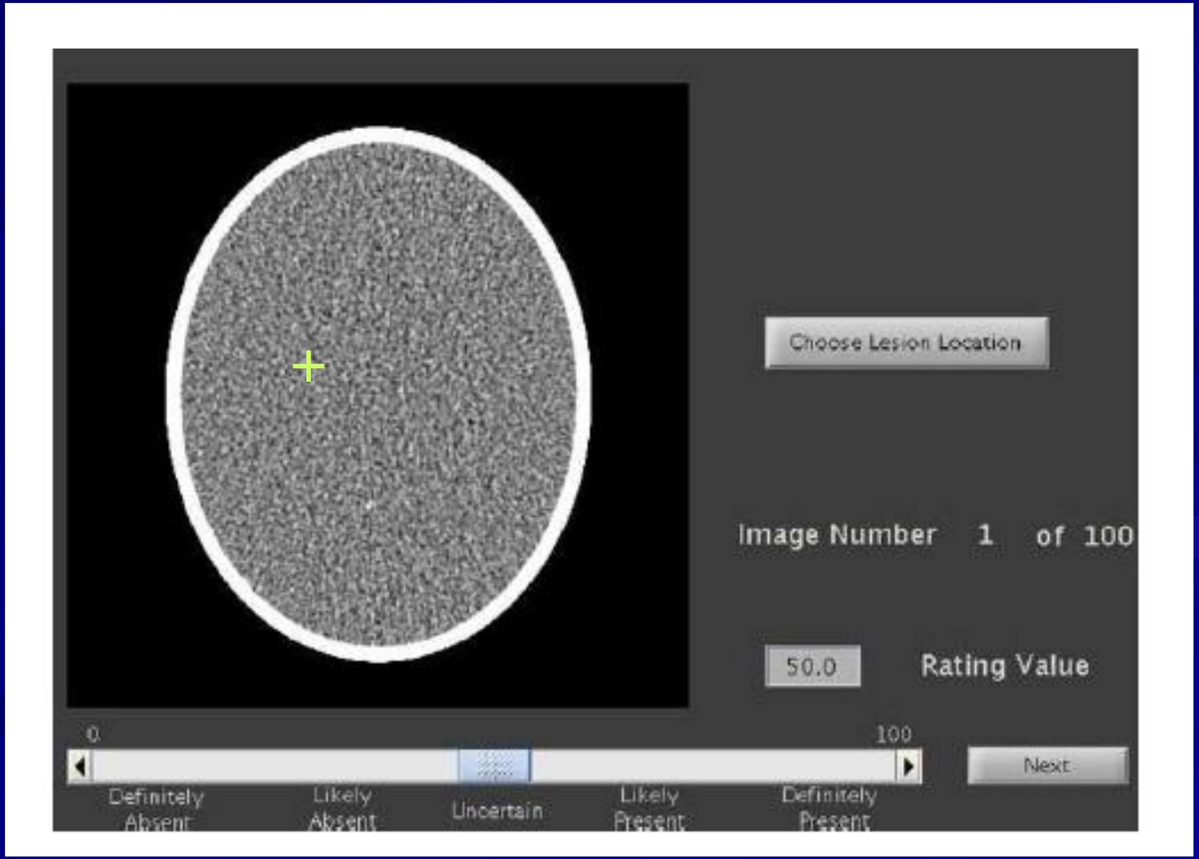
$$\text{TPLF} = \text{Prob}(Y > t_c \text{ and } \mathcal{L})$$
$$\text{FPF} = \text{Prob}(X > t_c)$$

$$\text{AUC} = \text{Prob}(Y > X \text{ and } \mathcal{L})$$

- Summary measure: AUC, which is again the probability of correct decision

LROC experiment

- Reader told that image contains one or no lesion
- Reader task: select a lesion location and provide a confidence rating



Beyond the single lesion model

- Free-ROC extends the concept to arbitrary number of lesions
- Many ways to build a curve and a summary measure

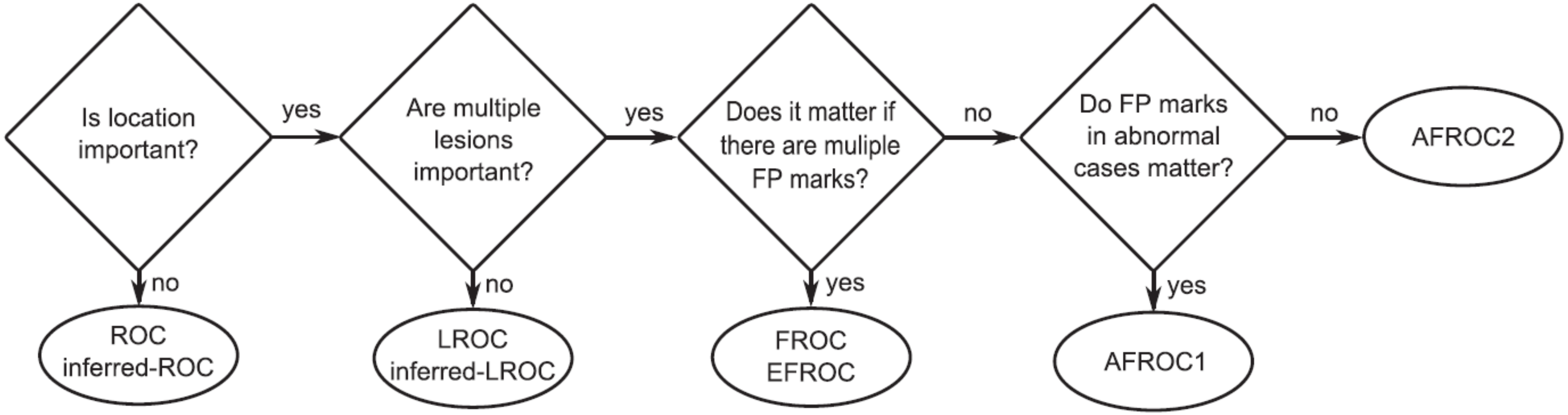

Is location important?
yes
no
ROC inferred-ROC
Are multiple lesions important?
yes
no
LROC inferred-LROC
Does it matter if there are muliple FP marks?
no
yes
FROC EFROC
Do FP marks in abnormal cases matter?
no
yes
AFROC1
AFROC2

FIG. 2. Flowchart to aid in the choice of a summary curve for observer performance on a lesion detection task.

*A Wunderlich and C K Abbey. Utility as a rationale for observer performance assessment, Med. Phys. 2013

Statistical analysis

- Results of LROC (and ROC) experiments are deviates of random variables
- Two sources of statistical variability: cases and readers
  → MRMC paradigm
- Cases can be seen as fixed or random effects
- Readers can be seen as fixed or random effects

Statistical analysis: result reporting

- Two options: hypothesis testing or confidence intervals

- Confidence intervals recommended

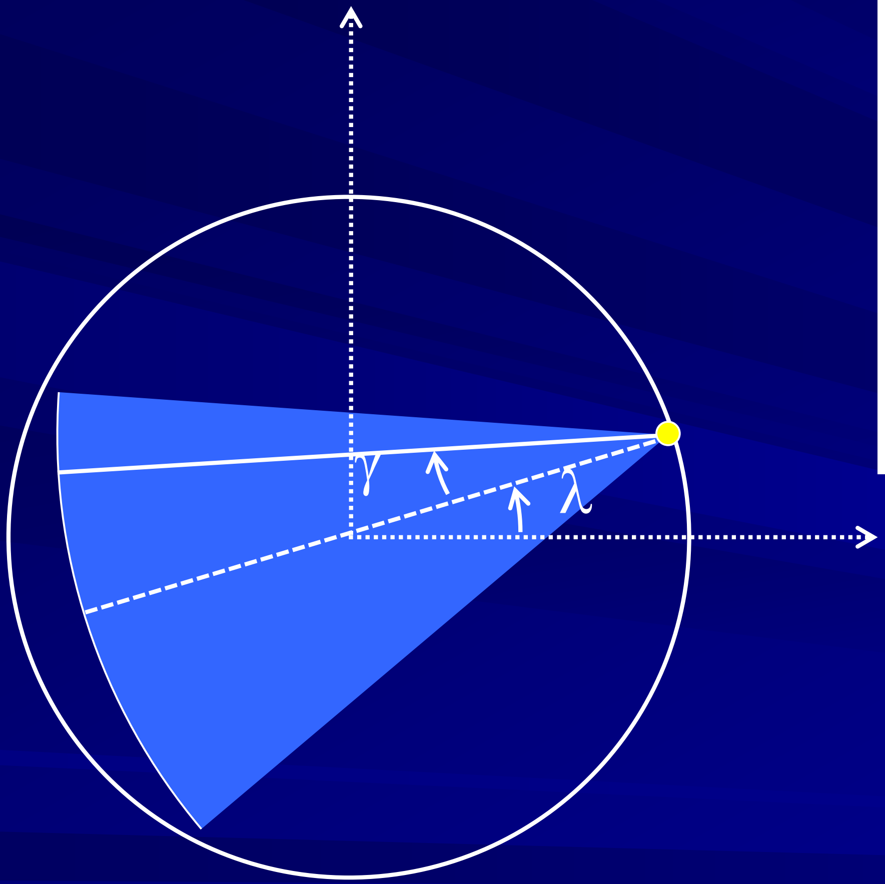
    - Confidence intervals allow direct assessment of effect size and statistical accuracy
    - CT studies are most often about “accepting the null hypothesis”
    - Journals increasingly demanding for CIs

- Multiple comparison adjustments needed (Bonferroni inequality, family-wise error correction)

Example using LROC analysis

- Comparison of 3 fan-beam CT reconstruction methods
- Task: detection and localization of one lesion only (LROC) within a uniform head phantom
- Random lesion position and contrast (25-to-35 HU), fixed size (5mm diameter)
- 4 observers
- Two sessions: 40 training and 125 testing images for each method (about 90 minutes effort per session)
- Partial pairing: pairing across algorithms but not across readers → maximize statistical power
- Final figure-of-merit: reader-averaged probability of correct decision for each reconstruction algorithm.

Method A: direct fan-beam FBP reconstruction with equal ray weighting



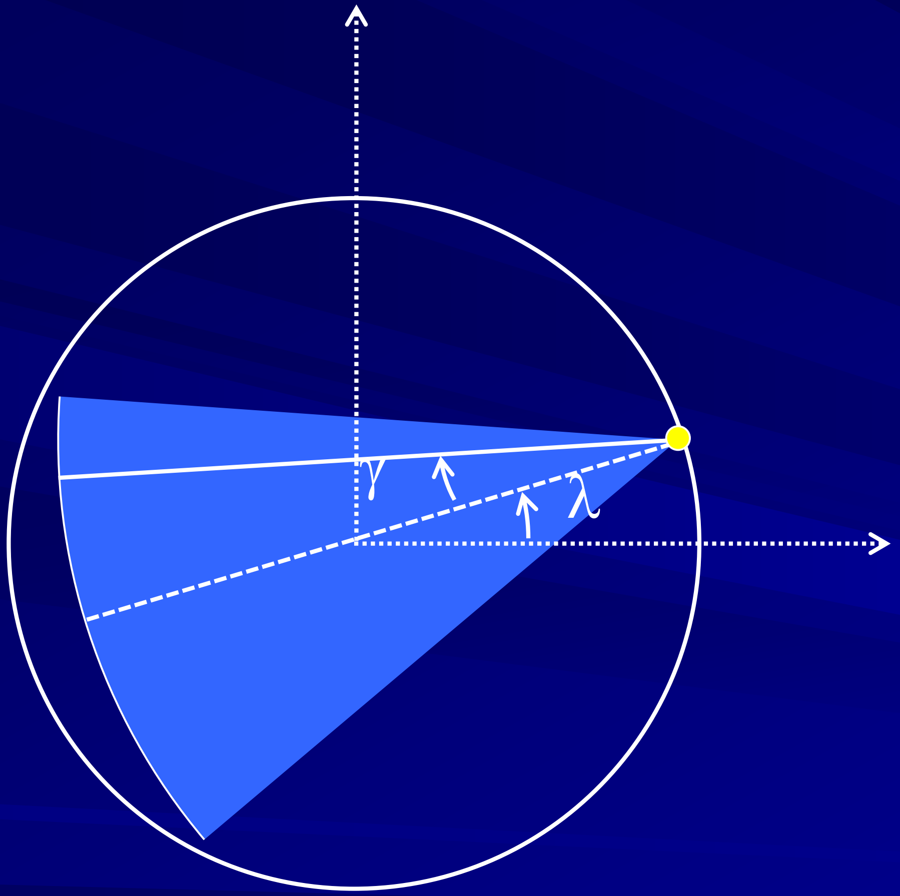
$$f(x,y) = \int_0^{2\pi} \frac{1}{\left(d(\lambda,x,y)\right)^2}\; g_F(\lambda,\gamma^*)\; d\lambda$$

$$g_F(\lambda,\gamma) = \int_{-\pi/2}^{\pi/2} \operatorname{sinc}^{-2}(\gamma-\gamma')\, h_{\text{ramp}}(\gamma-\gamma')\, w(\lambda,\gamma')\, g(\lambda,\gamma')\, d\gamma'$$

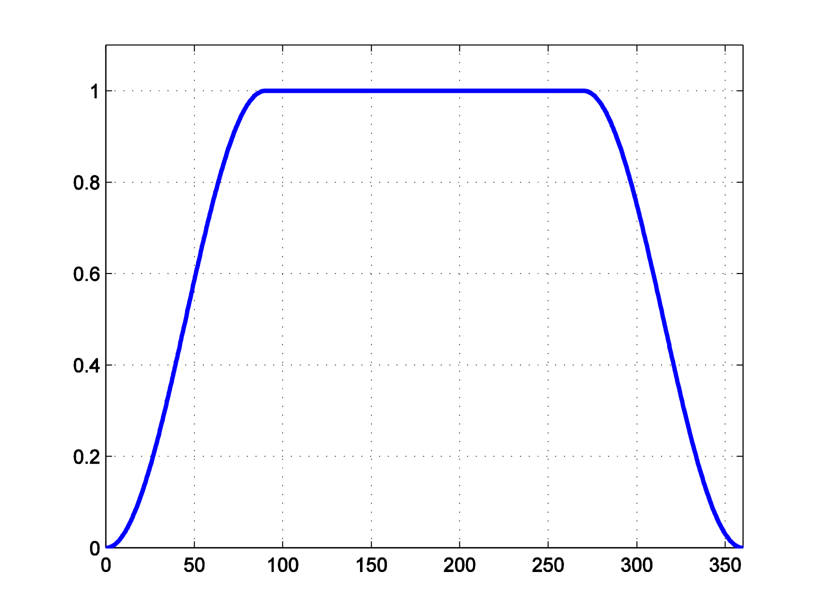
$$\gamma^* = \arctan\left(\frac{-x\sin\lambda + y\cos\lambda}{R - x\cos\lambda - y\sin\lambda}\right)$$

$$w(\lambda,\gamma) = 1/2$$

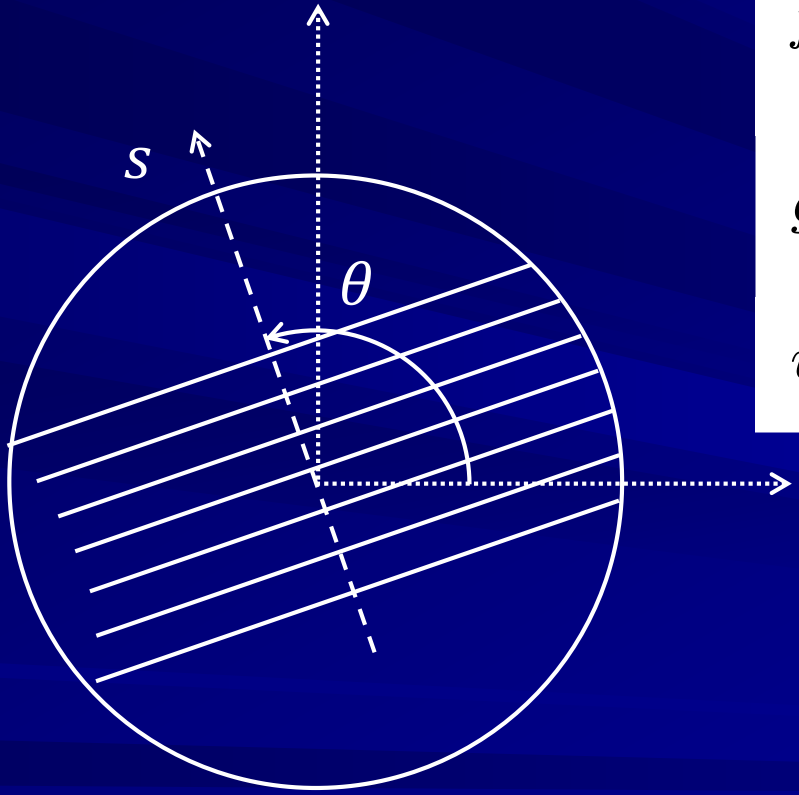
Method B: direct fan-beam FBP reconstruction with unequal ray weighting



$$w(\lambda,\gamma) = \frac{c(\lambda)}{c(\lambda) + c(\lambda + \pi - 2\gamma)}$$



Method C: indirect parallel-beam FBP reconstruction with equal ray weighting

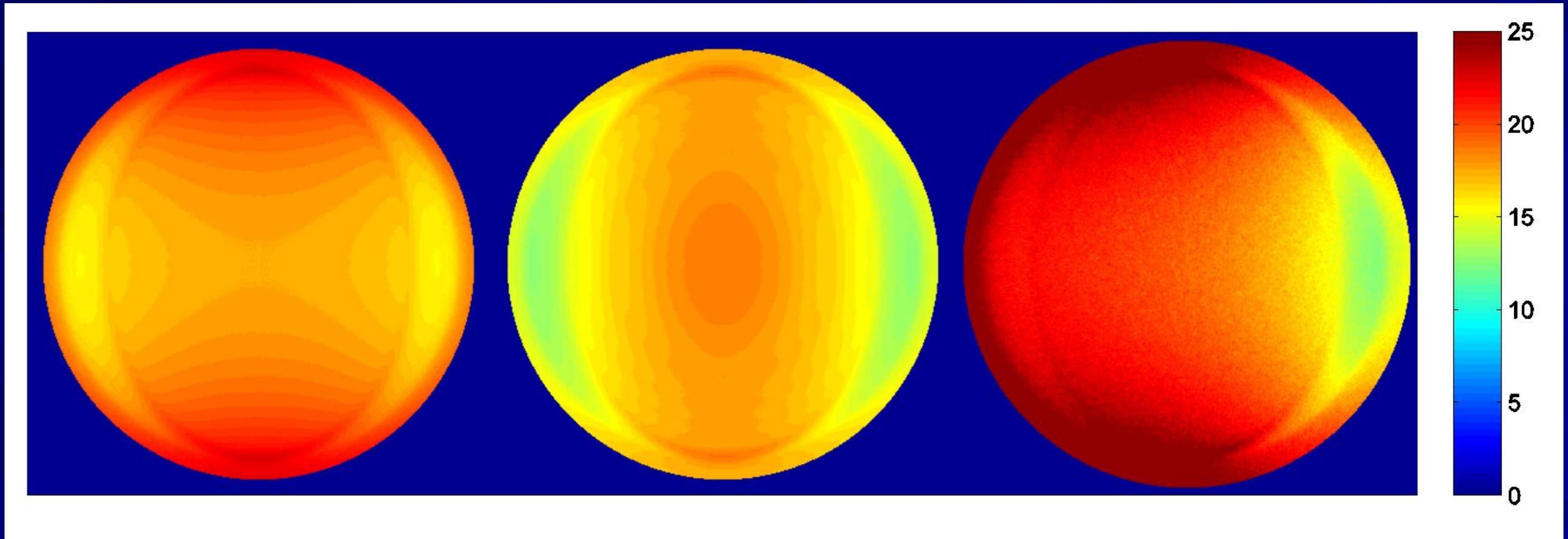


$$f(x, y) = \int_0^{2\pi} g_F(\theta, x\cos\theta + y\sin\theta)\, d\theta$$

$$g_F(\theta, s) = \int_{-\infty}^{\infty} h_{\text{ramp}}(s - s')\, w(\theta, s')\, g(\theta, s')\, ds'$$

$$w(\theta, s) = 1/2$$

Images of pixel noise (ensemble standard deviation):

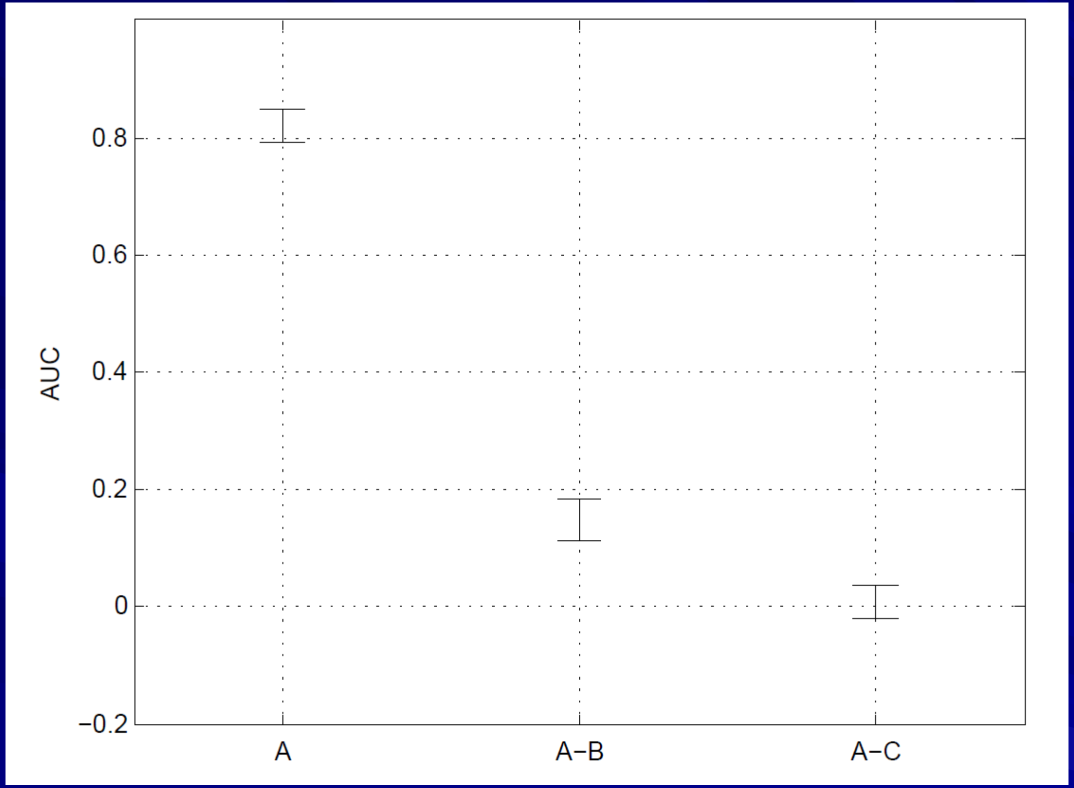


Method A:
direct FB

Method C:
indirect FB

Method B:
weighted direct FB

Human-observer study results (reader-averaged)



98.33% intervals, joint probability of 95% at least

Recall Bonferoni’s inequality:

$$P\left(\cap_{i=1}^{N} E_i\right) \geq \sum_{i=1}^{N} P(E_i) - (N-1)$$

Model (computerized) observers

- Cheaper
- Not subject to fatigue
- Can be chosen
  - to match human performance (can be challenging, channels and internal noise needed)
  - to provide an upper bound on image quality (ideal observer)
- Most often used with phantoms for early assessment of novel systems (e.g., MITA)

Important caveat

ROC analysis with model observers and phantom data may not be adequate to assess dose reduction claim related to deep learning based methods

- The network may perform poorly if phantom data are not part of the training
- Including phantom data within the training may result in dividing the network into two sub-networks, one for phantom data and one for human data

Statistical tools

https://github.com/DIDSR/IQmodelo

Parametric statistical methods for ROC performance analysis of linear model observers, including channelized Hotelling observers. Also includes functions for nonparametric ROC, LROC, EROC, and MAFC analysis with either fixed or random observers. All software is written for MATLAB.

https://github.com/DIDSR/

Conclusion

Are the dose reduction claims justified?

→ Is image quality maintained?
Was it appropriately and convincingly demonstrated?

Task
Observer
Images
Figure-of-merit

Did you like the selections made?