Implementation and Analysis of Observer Studies in Medical Physics

W.F. Sensakovic, PhD, DABR, MRSC

Attendees/trainees should not construe any of the discussion or content of the session as insider information about the American Board of Radiology or its examinations.

Relevant Conflicts of Interest

Public Domain

Public Domain
• Task is complex
  – Outline subtle tumor
• Unquantifiable human element
  – Clinical decision or Human visual system
• Human response is goal
  – Does widget “A” make it easier for the observer to detect the microcalcification?

Bunch of observers look at Bunch of subject images to create data that is then analyzed.
Bunch of radiologists look at bunch of CT scans (FBP or Iterative Recon) to record probability of malignancy for each. ROC analysis determines if iterative reconstruction impacts diagnosis.
• Typically 5-7 categories
• Validated scale if available and appropriate

- Widely Used Scale

<table>
<thead>
<tr>
<th>Definitely or almost definitely malignant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Probably malignant</td>
</tr>
<tr>
<td>Possibly malignant</td>
</tr>
<tr>
<td>Probably benign</td>
</tr>
<tr>
<td>Definitely or almost definitely benign</td>
</tr>
</tbody>
</table>


- Including clinically relevance

<table>
<thead>
<tr>
<th>Malignant—diagnosis apparent—warrants appropriate clinical management</th>
</tr>
</thead>
<tbody>
<tr>
<td>Malignant—diagnosis uncertain—warrants further diagnostic study/biopsy</td>
</tr>
<tr>
<td>I’m not certain—warrants further diagnostic study</td>
</tr>
<tr>
<td>Benign—no follow-up necessary</td>
</tr>
</tbody>
</table>


- Continuous vs. Categorical difference biggest for single reader studies


- No practical difference between discrete and continuous scales for ratings

Truth

• Best
  – Abnormal: Biopsy or other gold standard
  – Normal: Follow-up (e.g., 1-year) post imaging

• Combined reads (expert panel)
  – In a 3 system comparison, the “best” system depended on method used for truth
  – Report variability in consensus

ROC Analysis Study Design

• Task is binary (e.g., Malignant vs. Benign)
• Multi-Reader, Multi-Case (MRMC)
• Multiple treatments (e.g., IR vs FBP)
• Traditional, Fully-Crossed, Paired-Case Paired-Reader, Full Factorial
  – Every observer, reads every case, in every modality
  – Data correlations all us to get the highest power and lowest sample requirements
• Software (free or not) does it for you
  – ROC Software listed later
    • Some unsupported or not functional on modern computers, but may still run on an emulator such as dosbox ([https://www.dosbox.com](https://www.dosbox.com))

---

![Image](image.png)

True Positive (TP)
  – Sensitivity

False Positive (FP)
  – 1-Specificity

True Negative (TN)

False Negative (FN)
Does iterative reconstruction impact diagnosis of malignancy in lung lesions?

<table>
<thead>
<tr>
<th>Case #/Truth</th>
<th>Obs. 1</th>
<th>Obs. 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/Malignant</td>
<td>10.0</td>
<td>8.9</td>
</tr>
<tr>
<td>2/Benign</td>
<td>4.4</td>
<td>6.3</td>
</tr>
<tr>
<td>3/Benign</td>
<td>3.4</td>
<td>2.7</td>
</tr>
<tr>
<td>5/Malignant</td>
<td>5.6</td>
<td>5.2</td>
</tr>
<tr>
<td>6/Malignant</td>
<td>7.7</td>
<td>7.0</td>
</tr>
<tr>
<td>7/Malignant</td>
<td>9.2</td>
<td>8.1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

With IR

- True Positive Fraction (Sensitivity)
- False Positive Fraction (1-Specificity)

W/O IR

- True Positive Fraction (Sensitivity)
- False Positive Fraction (1-Specificity)
Does iterative reconstruction impact diagnosis of malignancy in lung lesions?

- Yes, it improves diagnosis
- By how much?

- AUC = 0.8
Does iterative reconstruction impact diagnosis of malignancy in lung lesions?

- Yes, it improves diagnosis
- By how much?
  - $\text{AUC} = 0.8$
  - $\text{AUC} = 0.7$

Average percent correct if observers shown random malignant and benign and asked to choose the malignant
• ROC Software will (generally):
  – Calculate ROC and AUC for each observer
  – Calculate combined ROC and AUC with dispersion
  – Perform hypothesis test to determine if AUC’s from 2 treatments significantly differ

---

**ROC Software Selection**

• Non-parametric ROC gives bias underestimates with a small number of rating categories

• Parametric (semi-parametric) may perform poorly if there are too few samples or if ratings are confined to a narrow range
  – Metz CE. Practical Aspects of CAD Research Assessment Methodologies for CAD. Presented at the AAPM annual meeting.

• Only generalizable to population of all observers if observer is treated as a random effect instead of fixed effect
  – Similarly, for cases
Case Selection

• Comparisons should be on same cases
  – Sensitivity 25%-100% depending on case selection

• The normal case subtlety must be considered to ensure sufficient number of false-positive responses

• Study disease prevalence does not need to match disease population prevalence
  – ROC AUC stable between 2%-28% study prevalence, but small increases in observer ratings are seen with low prevalence
We need to know:

- Minimum effect size of interest
  - Smaller needs more cases for testing
  - Appendix C of ICRU 79: ΔSe (at Sp) → ΔAUC

  - How much the difference varies
    - More variation needs more cases for testing

Sample size software (see references)

- Run a small pilot
- Program uses pilot data and resampling/Monte Carlo simulation to estimate variance for various model components (reader, case, etc.)

- Typical power 0.8 and α of 0.05
- Typical numbers are 3-5 observers and 100 case pairs (near equal for normal/abnormal)

ICRU Report 79
Pilot Data

H0: \( \text{AUC}_A - \text{AUC}_B = 0.00 \), two-sided alternative, 95% significance, 5 Readers, 50 Normal cases, 50 Disease cases.
\( \text{AUC}_A = 0.757, \text{AUC}_B = 0.716, \text{AUC}_A - \text{AUC}_B = 0.041, \text{S.E}(\text{total}) = 4.561 \times 10^{-2} \)

- 50 observers, 530 cases each . . . Probably pass

Design Considerations

- Observer training
  - Non-clinical task, specialized software, new modality
- Data/truth verification
  - 45% of truth cases contained errors
- Display and acquisition
  - Clinical conditions and equipment
Design Considerations

- **Bias from re-reading**
  - A few weeks rule-of-thumb (unless case is unusual)
  - Block study design (see refs. below)

- **Observer Experience**
  - **Sp 0.9:**
    - Se - 0.76 (high volume mammographers)
    - Se - 0.65 (low volume mammographers)

Instructions to Observers

- **According to ICRU Report 79**
  - Study description mindful of blinding
  - Types of relevant abnormalities and their precise study definition
  - How to perform task and record data
  - Unique conditions observers should or should not consider
• ROC is costly (time and or money)
• Best used when looking for small to moderate, but important differences
  – ~5% (ICRU Report 79)
  – Bigger difference could be seen with easier testing methodology
  – Smaller differences might be too costly or clinically insignificant

Detection: Anything there?

1. No localization
Bunch of radiologists look at bunch of chest radiographs (CR and DR) to determine if pneumonia is present. ROC determines if the modalities are equivalent.
• Rating scales, sample size, and truth essentially the same as in diagnosis observer study . . .
• . . . but the tasks are very different!

Clinical relevance reduces variability and thus sample size requirements
2. Localization

Bunch of radiologists look at bunch of radiographs with and without CAD system to, mark centroid of nodules if present, and give confidence ratings. FROC determines if CAD helps.

- Mark lesion centroid
- Determine how close mark must be for “hit”
  - 50% ROI overlap
  - Radius based on size of largest lesion

Delineation: How big? Exactly where?

• Bunch of dosimetrists outline the brainstem on CT scans displayed two different window/level settings. “Distance” between outlines is calculated. ANOVA is used to test if outlines are impacted by window/level settings.

Truth

• Phantom
  – Know exact size
  – Clinically relevant?

• Combined outlines on patient images
  – Union/Intersection
  – P-Map

  – STAPLE
• Jaccard Similarity Coefficient

\[ J \left( A, B \right) = \frac{|A \cap B|}{|A \cup B|} \]

- Count pixels in intersection
- Count pixels in union
- Divide intersection by union

• Dice
  - \( D = \frac{2J}{1+J} \)
- **Average Euclidean Distance**
  - Easy to understand
  - Meaningful units
• **Average Euclidean Distance**
  – Find shortest absolute distance from each boundary point of A to each the boundary point of B
  – Repeat for B to A
  – Summary stats

• **Fail to capture difference**
  – Dice/Jaccard
    • ~0.9
  – Average distance
    • <1mm
• Hausdorff distance
  – Take a point in A and find the shortest distance to B
  – Repeat for all points of A
  – Take the maximum of shortest distances
    • \( h(A,B) \)

• Hausdorff distance
  – Take a point in A and find the shortest distance to B
  – Repeat for all points of A
  – Take the maximum of shortest distances
    • \( h(A,B) \)
  – Repeat for \( h(B,A) \)
  – Max of \( h(A,B) \) and \( h(B,A) \)
Design Considerations

• Observers tend to agree with whatever is already drawn
  – 48% increase in Jaccard when previous outline used

• Different boundary definitions can alter measurements by 20%

• Summary statistics, regression, hypothesis testing
  – See “Practical Statistics for Medical Physicists” from the last two AAPM annual meetings
Conclusions

• Many intricacies to running an observer study
• Most respected studies in radiology and medicine in general

Further Reading

• Review (ROC and some FROC)
  – ICRU Report 79

• Comparing ROC Methods

• Study Design

• Power and Sample Size
  – See also some of the software listed
Further Reading

- FROC and JAFROC

Software (Usually Free)

- http://www.lerner.ccf.org/qhs/software/
- http://metz-roc.uchicago.edu/MetzROC/software
- http://didsr.github.io/iMRMC/
- Websearch your favorite software package and ROC
Cite

- 2015 (Virtual Library)
  - Phases, Levels, Controls, and All That: An Informal Session On Clinical Trials
    - http://www.aapm.org/education/VL/vl.asp?id=4986
  - Use and Abuse of Common Statistics in Radiological Physics
    - http://www.aapm.org/education/VL/vl.asp?id=4985
  - Uncertainty and Issues in Biological Modeling for the Modern Medical Physicist
    - http://www.aapm.org/education/VL/vl.asp?id=4687
- 2016 (Handouts Available)
  - Clinical trials and the medical physicist: design, analysis, and our role
  - Implementation and Analysis of Observer Studies in Medical Physics
  - Analysis of Dependent Variables: Correlation and Simple Regression
  - Hypothesis or Hypotheses, That is the Question

Image Copyright Requirements

- https://creativecommons.org/licenses/
Correlation: Review of Terminology

• Dependent vs. Independent Variables
  • Standard plot: X is Independent
    Y is Dependent

• Linear vs. Monotonic
  • Linear: increase in X leads to proportional increase in Y
  • Monotonic: increase in X leads to some increase in Y
Correlation: Review of Terminology

• Variable Type
  • Continuous
    • Example: Ionization chamber charge collected vs. Dose delivered
  • Discrete
    • Example: Number of patients seen vs. Calendar year
  • Ordinal
    • Example: Severity of normal tissue toxicity vs. Prescription Level
  • Categorical
    • Example: RECIST response classification vs. Radiologist Observer

Correlation: Metrics of Interest

• Four big categories of data
  • Continuous
  • Discrete
  • Ordinal
  • Categorical
Correlation: Metrics of Interest

- Four big categories of data
  - Continuous
  - Discrete
  - Ordinal
  - Categorical

Three major correlation metrics
- Pearson’s $r$
- Spearman’s $\rho$
- Fleiss’ $\kappa$
Correlation: Pearson’s $r$

- “Linear” or “Product-Moment” correlation
- Applies only to continuous data
- Parametric correlation
  - Tendency of dependent variable to increase linearly with the independent variable
- Key Point:
  - There is an assumed form to the relationship
  - Linear, and therefore also monotonic
Correlation: Pearson’s $r$

- $r = 1.00$
- $r = 0.97$
- $r = 0.76$

Correlation: Spearman’s $\rho$

- “Rank” correlation
- Applies to continuous, discrete, or ordinal data
- Non-parametric correlation
  - Tendency of dependent variable to increase with the independent variable
- Key Point:
  - There is no assumed relationship, only monotonicity
- Math: Pearson’s $r$ of rank-transformed data
Correlation: Spearman’s $\rho$

Raw: (0,0)
Rank: (1,1)

(X,Y) pairs
Correlation: Spearman’s $\rho$

Raw: (0,0)
Rank: (1,1)

Raw: (0.05,0.0025)
Rank: (2,2)

$(X,Y)$ pairs

Pearson’s $r$ of rank-transformed data: 1.00
Correlation: Spearman’s $\rho$

$r = 1.00$
$\rho = 1.00$

$r = 0.97$
$\rho = 0.97$

$r = 0.76$
$\rho = 0.90$

$r = 0.96$
$\rho = 1.00$

$r = 0.96$
$\rho = 0.99$

$r = 0.75$
$\rho = 0.89$
Continuous variables; “When one goes up, does the other (reliably) go down?”

Answer: Spearman’s ρ
Correlation: Fleiss’ $\kappa$

- Categorical correlation
- Applies only to categorical data
  - Categorical data could be inherently ordinal
- Non-parametric correlation
  - How well do independent categories sort dependent categories?
- Math: number of dependent-independent pairs in agreement over the number expected by chance alone.

Example:
- 5 radiologists contour tumors in
- 31 patients
- Response classification from baseline to post-chemo CT scans

<table>
<thead>
<tr>
<th></th>
<th>Obs. 1</th>
<th>Obs. 2</th>
<th>Obs. 3</th>
<th>Obs. 4</th>
<th>Obs. 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Progression</td>
<td>6</td>
<td>11</td>
<td>7</td>
<td>11</td>
<td>14</td>
</tr>
<tr>
<td>Stable</td>
<td>17</td>
<td>10</td>
<td>19</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>Partial</td>
<td>7</td>
<td>10</td>
<td>5</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>Complete</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

$\kappa = 0.64$

Correlation vs. Agreement

• Quick tangent…

Important question:
Do you already know that the two variables will be correlated?

Example: Tumor volumes as assessed by Physician vs. Algorithm
Correlation vs. Agreement

• Absolute agreement vs. Relative agreement
  • Absolute: plot raw differences
  • Relative: plot log differences

\[ \ln \left( \frac{x}{y} \right) = \ln x - \ln y \]

• Get mean, SD of log-transformed data, then apply exponential to get relative agreement bounds

### Table: Observer Ratings

<table>
<thead>
<tr>
<th>Case #/Truth</th>
<th>Obs. 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>1/Malignant</td>
<td>10.0</td>
</tr>
<tr>
<td>2/Benign</td>
<td>4.4</td>
</tr>
<tr>
<td>3/Benign</td>
<td>3.4</td>
</tr>
<tr>
<td>5/Malignant</td>
<td>5.6</td>
</tr>
<tr>
<td>6/Malignant</td>
<td>7.7</td>
</tr>
<tr>
<td>7/Malignant</td>
<td>9.2</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

### Graphs

- **Graph 1**: Distribution of ratings for actually benign and malignant cases.
- **Graph 2**: ROC curve showing true positive fraction (Sensitivity) and false positive fraction (1-Specificity).
- **Graph 3**: Observer Malignancy Rating distribution with true negatives (TN), false negatives (FN), false positives (FP), and true positives (TP) highlighted.

**With IR**

**True Positive Fraction (Sensitivity)** vs **False Positive Fraction (1-Specificity)**
AUC comparison not appropriate if ROC curves cross each other

- Better for Screening
- Maybe better for Diagnostic
- Partial AUC


Sp and Se. . . Why bother with ROC?

- CR: Better Se, Worse Sp
- DR: Better Sp, Worse Se
- Which is better?

Based on example given by CE Metz during lectures at the University of Chicago, 2003
• JAFROC
  – Use localization information
  – More than one response per subject

• Practically, comparisons need whole ROC curves and not specific operating points (TP, FP) or (PPV, NPV) values
  – Reader ability and experience, societal norms, and even disease prevalence in the study can impact specific operating points and (PPV, NPV) values