Digital Mammography,
Computer-aided Diagnosis,
Multiple-reader studies –

and the Holy Grail of
Imaging Physics

RF Wagner, SV Beiden (OST, CDRH, FDA)
Gregory Campbell (OSB, CDRH, FDA)
RM Gagne, RJ Jennings (OST, CDRH, FDA)
CE Metz, Y Jiang (Radiology Research, U of Chicago)
H-P Chan (Radiology Research, U of Michigan)
OUTLINE

I)  A Preamble: What we mean by the metaphor of the Holy Grail of Imaging Physics

II) Some Fundamentals of the Imaging Physics

III) Some Fundamentals of Scorekeeping for Diagnostic Imaging in Clinical Studies
     -- Why ROC & Multiple-Reader ROC Studies

     (from Public Panel Meeting on Digital Mammo)

IV) The Next Step: How (Some) Systems for Computer-Aided Diagnosis may help us toward our goals
PREAMBLE

Two Cultures

A:

Physical laboratory-elaborate physical measurements on performance of diagnostic imaging systems

-- and models of observers & lesion detection

B:

The Clinical setting – measurements of diagnostic performance of imaging systems on patients

i.e., measurements on real observers & real patients

The "Holy Grail"

(as the metaphor is used in these communities)

To predict the clinical ranking of imaging systems from the ranking obtained in physical laboratory

To use the info from A to reduce the burden of B.
World’s shortest course on physical image quality

Image quality depends on

Contrast, Resolution, Noise and the Task

More photons or quanta allow for finer tasks –
Physical measures: generalized concept of “quanta”
Family of Required Physical Measurements

Physical composition of input background & "lesions"

X-ray source distribution

X-ray spectral content

Input exposure

Spectral contrast transfer

Gray-scale transfer

Scatter-rejection mechanism

Modulation Transfer Function (MTF)

Noise Power Spectrum
Summary of Signal Detection Theory & ICRU Report #54: The Assessment of Image Quality

Detectability SNR in additive Gaussian noise:
\[ \Delta = \text{separation of class means}; \]
\[ \Sigma = \text{covariance matrix of noise} \]
\[ d^2 = \Delta^t \Sigma^{-1} \Delta \]

Fourier domain measurements in medical imaging:
\[ \text{NEQ}(f) = \frac{G^2 \text{MTF}^2(f)}{\text{NPS}(f)} \]
\[ \text{DQE}(f) = \frac{\text{NEQ}(f)}{Q} \]

For a discrimination task characterized by \( \Delta S(f) \):
\[ d^2 = \int |\Delta S(f)|^2 \text{NEQ}(f) \, df \]

Contemporary studies (non-stationary statistics):
\[ d^2 = \Delta^t \Sigma^{-1} \Delta \]
\[ \ldots \text{in spatial domain} \]
"Contrast-detail" curves

Contrast vs diameter (required for detection)

Symbols = Observed
Lines = predicted from Laboratory measurements

Contrast vs exposure (conventional mammo)

Contrast vs exposure (optimized digital system)
A slightly longer course on
Clinical scorekeeping in diagnostic radiology:

I) The ROC Paradigm

II) Two Classical Papers on
Variability in Mammography

III) Multiple-reader studies –
what’s involved

IV) Contemporary example – digital mammography
1) The ROC Paradigm

A typical conventional ROC curve, showing three possible operating points.
2) Classic papers on Variability in Mammography

(a) Elmore et al. Study of 10 Radiologists (1974)

True Positive Fraction vs False Positive Fraction
Based on Recommendation for Biopsy
(123 noncancers / 27 proven cancers)

D'Orsi & Swets’ Plotted in ROC Space (1995)
<table>
<thead>
<tr>
<th>Radiologist</th>
<th>Patients with Cancer (N = 27)</th>
<th>Patients without Cancer (N = 123)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Any Immediate Workup*</td>
<td>Biopsy</td>
</tr>
<tr>
<td>A</td>
<td>96</td>
<td>82</td>
</tr>
<tr>
<td>B</td>
<td>96</td>
<td>70</td>
</tr>
<tr>
<td>C</td>
<td>93</td>
<td>63</td>
</tr>
<tr>
<td>D</td>
<td>93</td>
<td>78</td>
</tr>
<tr>
<td>E</td>
<td>89</td>
<td>70</td>
</tr>
<tr>
<td>F</td>
<td>85</td>
<td>48</td>
</tr>
<tr>
<td>G</td>
<td>85</td>
<td>67</td>
</tr>
<tr>
<td>H</td>
<td>81</td>
<td>52</td>
</tr>
<tr>
<td>I</td>
<td>78</td>
<td>48</td>
</tr>
<tr>
<td>J</td>
<td>74</td>
<td>33</td>
</tr>
</tbody>
</table>

†percentage of patients

*Any immediate workup was defined as a recommendation to obtain additional mammographic views, ultrasound studies, or a biopsy.

Elmore et al.
(b) Beam, Layde, Sullivan 1996
Study of 108 randomly selected Radiologists

**Sensitivity vs Specificity**
Based on Recommendation for Biopsy
(45 cancers, 19 normal, 15 benign)

Figure 2. Scatterplot of sensitivity vs specificity among the 108 US radiologists who participated in the study.

Beam later showed that this is not consistent with a single ROC curve and finite sampling statistics.
Two Correlated Diagnostic Modalities

probability of malignancy / modality 2

probability of malignancy / modality 1

Ca

NonCa
3) The Multiple-Reader, Multiple-Case (MRMC) ROC Paradigm:

Every reader reads every case (and where practical) in both modalities

Can then model and account for the following multivariate ROC Accuracy Parameters:

Variance due to case sampling

Variance due to reader sampling

Correlation of case variation across modalities

Correlation of reader variation across modalities

Within reader variability, or reader "jitter"

(Most models involve additional parameters)

Using Multiple Readers, Multiple Cases & Elaborate MRMC Software (refs. at end)
We are interested in assessing mean performance and the uncertainty (confidence intervals, C.I.) in estimates of the difference in performance over Multiple (#) Readers & Multiple (#) Cases.

The uncertainty of a diff. in average ROC parameters (between two modalities) =

\[
\frac{\text{Var (C)} \times (1 - \rho_c)}{\text{(# Cases)}} + \frac{\text{Var (Btw R)} \times (1 - \rho_r)}{\text{(# Readers)}} + \frac{\varepsilon}{\text{(# Cases x # Readers)}}
\]

(i.e.)

\[
\frac{\text{(Uncorrelated portion of Case Var)}}{\text{(# Cases)}} + \frac{\text{(Uncorrelated portion of Reader Var)}}{\text{(# Readers)}} + \frac{\text{(Within reader jitter)}}{\text{(# Cases x # Readers)}}
\]
4) Sponsor's Multiple-Reader, Multiple-Case Study

Cases:
625 women (one or both breasts)
997 breasts
44 breasts with cancer (no known bilateral Ca)

Readers:
5 MQSA-qualified radiologists

Modalities:
All cases imaged with both modalities

All 5 readers read all images from both modalities:
Readings of Digital & Analog separated by 30 days - Balanced Reading (half of cases: Digital first; & v. v.)

<table>
<thead>
<tr>
<th>Individual Reader Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (over 5 readers) ROC Area SFM (Analog): 0.77</td>
</tr>
<tr>
<td>Mean (over 5 readers) ROC Area FFDM (Digital): 0.76</td>
</tr>
<tr>
<td>Mean 95% C.I. about the difference of 0.01: +/- 0.11</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Multiple-Reader Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Generalizable to a Pop. of Readers &amp; a Pop. of Cases)</td>
</tr>
<tr>
<td>Mean 95% C.I. about the difference of 0.01: +/- 0.064</td>
</tr>
</tbody>
</table>
Implications of the present Variance Structure

for a post-approval trial

Drawn from a screening population –
assumed to have same sampling properties as current

(a) with goal of 95% C.I. ~ +/- 0.05

44 cancers, 10 readers
50 cancers, 6 or 7 readers
59 cancers, 5 readers

(b) with goal of 95% C.I. ~ +/- 0.03

78 cancers, 100 readers
100 cancers, 20 readers

... and screening yields about 5 cancers/1000 screened
The Clinical Studies

Hendrick RE, Lewin JM, D’Orsi CJ, Kopans DM, Conant E, Cutter GR, Sitzler A.

Interim clinical evaluation of FFDM in a screening cohort: Comparison with screen-film mammography in 4,965 exams.

We gratefully acknowledge

R. Edward Hendrick, Radiology Department, NWU & General Electric Medical Systems

for permission to present these results prior to publication
Formal concepts underlying the previous analysis --

& required to take the next step:

Analysis of computer-aided diagnosis
Components of Variability

(most general linear model for the present problem)

Components correlated across modalities:

\( c \) - range of case difficulty & finite-sample effect
\( r \) - range of variability of reader skill
\( r \times c \) - dependence of case difficulty on reader (or vice versa: of reader skill on case)

Components uncorrelated across modalities:

\( m \times c \) - "interaction"
\( m \times r \) - ""
\( m \times (r \times c) \) - ""
Analysis of Clinical Study of
Jiang, Nishikawa, Schmidt, Metz, Giger, Doi
Discriminating 46 Ca vs 58 Bn μcalc clust/10 readers
(Acad Radio 6, 1999)

Model B

Study #1 (Jiang et al.)
Analysis of Clinical Study of
Jiang, Nishikawa, Schmidt, Metz, Giger, Doi
Discriminating 46 Ca vs 58 Bn μcalc clust/10 readers
(Acad Radio 6, 1999)

Model B

Study #1 (Jiang et al.)
Toward the Holy Grail

If mature systems for computer-aided diagnosis can greatly reduce reader components of variance

... this will leave only the case component ...

and this carries the physics/technology

One possible scenario at that point:

Limit clinical studies to special cohorts who will challenge/stress the physically determined differences between systems
"Field Test"

A study that attempts to sample the modality
under completely realistic conditions
- attempts to reduce any sources of bias

"Stress Test"

A study that samples the application of the modality
under conditions that challenge
one or another features of the system
When setting up a Stress Test of Modality A vs Modality B - How to generate an "enriched" image set?

(i.e., a set with challenging images)

Caveat:

Selecting images using Modality A will:

Favorably bias Sensitivity of Modality A;

Favorably bias Specificity of Modality B;

Possibly indeterminate bias on ROC curve itself

Alternative solutions to the enrichment problem must be found
Principal references for this talk


Other Key Sources


"Table of 10 Radiologists' recommendations for biopsy" from JG Elmore, CK Wells, CH Lee et al, "Variability in radiologists' interpretations of mammograms." New England Journal of Medicine 1994; 331: 1493-99. Copyright 1994, Massachusetts Medical Society. All rights reserved.

