Critical reading
of diagnostic imaging studies

Constantine Gatsonis
Center for Statistical Sciences
Brown University

ACRIN Fall 2010 Annual Meeting

Lecture  Goals

1. Review diagnostic imaging evaluation goals and endpoints.
2. Identify building blocks of study design.
3. Describe common study designs, including studies of accuracy in diagnosis and prediction and studies of outcomes.
4. Discuss structured reporting of studies

Topics

1. Diagnostic imaging evaluation:
   1. Overview
   2. Measures of performance
   3. Reporting checklists
2. Elements of study design
   1. Studies of accuracy in detection
   2. Studies of accuracy in prediction
   3. Studies of outcomes
3. Generalizability and bias considerations
4. Reporting revisited
Topics
1. Diagnostic imaging evaluation:
   1. Overview
   2. Reporting checklists
   3. Measures of performance
2. Elements of study design
   1. Studies of accuracy in diagnosis
   2. Studies of accuracy in prediction
   3. Studies of response
3. Generalizability and bias considerations
4. Reporting revisited

Imaging integral to all key aspects of health care

The Cancer Paradigm

Diagnosis vs therapy
- The dominant paradigm for evaluating medical care is provided by the evaluation of therapy.
- Diagnostic tests generate information, which is only one of the inputs into the decision making process.
- Most effects of the diagnostic information are mediated by therapeutic decisions.
- Diagnostic technology evolves very rapidly (moving target problem)
Endpoints for diagnostic test evaluation

- Diagnostic performance:
  - measures of accuracy
  - measures of predictive value

- Intermediate process of care:
  - Diagnostic thinking/decision making
  - Therapeutic thinking/decision making

- Patient outcomes:
  - Quality of life, satisfaction, cost, mortality, morbidity

STARD checklist

Sections
- Title/Abstract
- Introduction
- Methods
  - Participants
  - Test methods
  - Stat Methods
- Results
  - Participants
  - Test results
  - Estimates
- Discussion

Measures of accuracy

- Test results can be binary (e.g. yes/no for presence of target condition), ordinal categorical (e.g. degree of suspicion on a 5-point scale), or continuous (e.g. SUV for PET).
- Analysis of diagnostic performance typically assume binary truth (e.g. “disease present” vs “disease absent”).
- The well known 2x2 table for binary test and truth:

<table>
<thead>
<tr>
<th>Target condition</th>
<th>Test -</th>
<th>Test +</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absent</td>
<td>TN</td>
<td>FP</td>
<td>N_i</td>
</tr>
<tr>
<td>Present</td>
<td>FN</td>
<td>TP</td>
<td>N_i</td>
</tr>
<tr>
<td>Total</td>
<td>TN+FN</td>
<td>FP+TP</td>
<td>N</td>
</tr>
</tbody>
</table>

ACRIN Fall 2010
Constantine Gatsonis, PhD

gatsonis@stat.brown.edu

Measures of accuracy for binary tests

- Starting with disease status:
  - Sensitivity: Probability that test result will be positive given that the target condition is present.
  - Sens = TPF = Pr(T+ | D+), 1 - Sens = FNF = Pr(T- | D+)
  - Specificity: Probability that test result will be negative given that target condition is absent.
  - Spec = TNF = Pr(T- | D-), 1 - Spec = FPF = Pr(T+ | D-)
- Starting with test result:
  - Positive Predictive Value: PPV = Pr(D+|T+)
  - Negative Predictive Value: NPV = Pr(D-|T-)
- Positive likelihood ratio:
  - LR+ = Pr(T+|D+) / Pr(T+|D-)=Sens/(1-Spec)
- Post-test odds = pretest odds x LR

Fundamental conceptualization: Threshold for test positivity
ROC curves

- ROC curve: plot of all pairs of (1-Spec., Sens.) as positivity threshold varies

Topics

1. Diagnostic imaging evaluation:
   1. Overview
   2. Measures of performance
   3. Reporting guidelines

2. Elements of study design
   1. Studies of accuracy in detection
   2. Studies of accuracy in prediction
   3. Studies of outcomes
   4. Generalizability and bias considerations
   5. Reporting modified

Fundamental blocks of clinical study design

- Imaging intervention:
  - Technique: Technical characteristics of modalities to be studied
  - Test interpretation: selection of reader population, viewing experience, training, learning potential

- Measures of endpoints

- Reference information (“gold standard”)
Design template for prospective study of accuracy

Enroll predefined participant population
> Perform and interpret imaging studies

Evaluate study endpoints
> Collect information on reference standard (e.g. biopsy, follow-up)

CT and MRI in staging cervical cancer. (ACRIN 6651)

208 women with documented cervical cancer, scheduled to undergo surgery

Participants undergo CT and MR to determine cancer stage prior to surgery. Studies interpreted locally and centrally

Diagnostic accuracy measures: predictive value, sensitivity, specificity, ROC

Participants undergo surgery. Stage determined by surgical and pathology data

Hricak, Gatsonis et al, JCO, 2005

Predictive value of CT and MRI for higher stage in cervical Cancer

Results from ACRIN 6651, JCO 2005

Graph showing predictive value (PV) vs. reference standard (RS) with CT and MR data points.
Sensitivity and specificity of CT and MRI for detecting advanced stage in cervical cancer. (ACRIN 6651, JCO 2005)

Digital vs Film Mammography. (ACRIN 6652, DMIST)

- Participants undergo both digital and plain film mammography (paired design)
- Primary reading at participating institutions. Secondary readings (of subsets) by selected readers
- Work-up of positive mammograms and 1-year follow-up information

(Pisano, Gatsonis et al, NEJM, 2005)

Comparison of ROC curves for Digital and Film Mammo. DMIST results NEJM, Oct 2005.

- Age <50
- Age 50+
- Full cohort
National CT Colonography study, ACRIN 6664

2600 participants scheduled to undergo routine colonoscopy

Participants undergo CT Colonography, interpreted at participating institutions.

Estimate sensitivity, specificity, PPV, NPV of CTC with colonoscopy as the reference standard.

Participants undergo colonoscopy.

ACRIN Fall 2010 Constantine Gatsonis, PhD gatsonis@stat.brown.edu

Performance Measure

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.90 (0.84-0.96)</td>
<td>0.00 (0.04-0.06)</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.86 (0.83-0.89)</td>
<td>0.14 (0.12-0.16)</td>
</tr>
<tr>
<td>Positive predictive value</td>
<td>0.23 (0.20-0.27)</td>
<td>0.00 (0.01-0.05)</td>
</tr>
<tr>
<td>Negative predictive value</td>
<td>0.99 (0.97-0.99)</td>
<td>0.99 (0.97-1.00)</td>
</tr>
</tbody>
</table>

ACRIN Fall 2010 Constantine Gatsonis, PhD gatsonis@stat.brown.edu

Predictive value of imaging

- Imaging as a biomarker:
  - Predicting the course of disease with or without therapy
  - Serving as surrogate marker for therapy study endpoints

Johnson et al, NEJM 2008

ACRIN Fall 2010 Constantine Gatsonis, PhD gatsonis@stat.brown.edu
Types of biomarkers

I: Biomarkers for disease onset or presence

II: Biomarkers for course of disease

- **Prognostic** biomarkers: used to predict disease outcome at time of diagnosis – usually without reference to potential therapy (e.g. tumor size or stage)
- **Predictive** biomarkers: used to predict the outcome of particular therapy (e.g. PET SUV change from baseline)
- **Monitoring** biomarkers: used to measure response to treatment and to achieve early detection of disease progression or relapse (e.g. FDG PET or DCE MRI)

MRI as predictor of response to therapy for breast cancer (ACRIN 6657, ISPY-1)

- 237 women receiving neoadjuvant breast cancer therapy.
- Participants undergo breast MRI at pre-specified time points pre, during, and after therapy.
- Estimate the ability of MRI to predict response and long term outcome.
- Clinical information on response to therapy and survival is obtained.
- Ongoing study
FDG-PET as predictor of survival after therapy for lung cancer (ACRIN 6668).

NSCLC patients undergoing chemoradiotherapy.

Participants undergo FDG PET and baseline and after therapy.

Estimate the ability of FDG-PET to predict survival and response.

Clinical information on response to therapy and survival is obtained.

Research questions and statistical formulation

• Example: Can SUV change from baseline to week 2 of therapy predict survival?

• Two approaches:
  – A. Compare survival curves among patient groups defined by a fixed cutoff value of marker.
  – B. Compute sens/spec/ROC curve for ref. standard defined by occurrence of future event.

• Note: A and B do not answer same question!

• A: do metabolic responders live longer?

• B: Can PET detect early those patients who will respond or survive past certain times?

A. Comparison of survival curves

• Divide patients into two groups, using specific threshold of marker (e.g. median value or other value obtained from previous studies).

• Compare survival curves.

• Or use regression analysis

Potential problems:

• Choice of threshold?

• Choice of specific form of regression model?

“Predictive value” type of analysis
B. Sens/Spec/ ROC analysis

- Research question viewed as detection of a future event.
- For a future time \( t \):
  - Test result: SUV change
  - Gold standard: dead or alive at time \( t \)
- Using a threshold on SUV change, sensitivity and specificity are estimated.
- Using continuous test result, ROC curve can be derived (time-dependent ROC).
- Technical issue: Censored observations present technical challenge

![ROC curves for two types of %S-phase measurements as predictors of survival after breast cancer diagnosis.](Heagerty et al, Biometrics, 2000;)

---

Reliability and standardization

- Reliability of measurements
  - e.g. are SUV measurements reliable for single patient?
  - e.g. are SUV measurements comparable across institutions?
  - Assessment of reliability is aim in current clinical trials, such as ACRIN 6678 (FDG-PET in advanced lung cancer patients)
  - Is there a required level of reliability for marker validation?
- Standardization of measurements across institutions, machine types, is challenging but absolutely essential.
Studies of patient outcomes

Patient presentation → Diagnostic workup → Treatment Decisions → Patient outcomes

National Lung Screening Trial (ACRIN 6654)

52,000 participants, at high risk for lung cancer. Randomized

CXR screening
Baseline, plus two annual screens

Helical CT screening
Baseline, plus two annual screens

Primary endpoint: Comparison of lung cancer mortality between randomization arms.

Participants followed up in subsequent years
ACRIN 4005
Low- to Intermediate- Risk Patients: People 30 and older who present to the ED with symptoms consistent with potential ACS.

RANDOMIZATION

GROUP A: TRADITIONAL "RULE-OUT" ARM

GROUP B: CT CORONARY ANGIOGRAPHY "RULE-OUT" ARM

POSITIVE TEST
NORMAL TEST

FOLLOW UP: 30 DAYS AND 1 YEAR

ACRIN Fall 2010
Constantine Gatsonis, PhD
gatsonis@stat.brown.edu

Another example of randomized study of outcomes

Patient with lower back pain

Baseline functional status, HQRL

Fast MRI
Plain film

• Process of care
• Patient outcomes
• Health care utilization


Topics

1. Diagnostic imaging evaluations:
   1. Overview
   2. Measures of performance
   3. Reporting checklist

2. Elements of study design
   1. Studies of accuracy in detection
   2. Studies of accuracy in prediction
   3. Studies of effectiveness

3. Generalizability and bias considerations

4. Reporting revised
Building blocks of study design

**Clinical question**
- start with defining clinical setting
- recruit representative sample ("consecutive clinical series" usually works but is difficult to achieve)
- ensure patients with all forms of disease in sample
- sample prevalence may influence interpretation because of limited spectrum or, even with representative spectrum, because of factors such as reader vigilance

**Patient population**

**Imaging intervention:**
- **Technique**: Technical characteristics of modalities to be studied
- **Test interpretation**: setting and reader population, training, experience, learning potential

**Measures of endpoints**

**Reference information** ("gold standard")

---

Clinical question
- **Clinical question /Patient population/ Case mix/spectrum:**
  - start with defining clinical setting
  - recruit representative sample ("consecutive clinical series" usually works but is difficult to achieve)
  - ensure patients with all forms of disease in sample
  - sample prevalence may influence interpretation because of limited spectrum or, even with representative spectrum, because of factors such as reader vigilance

**Patient population**

**Imaging intervention:**
- **Technique**
- **Test interpretation**
- **Imaging Intervention: Technical characteristics of the imaging process**
  - precise description of techniques
  - reproducible at other clinics
  - should reflect expected clinical practice, but this often varies across institutions – set minimum acceptable techniques, or allowable range
Imaging intervention:
- Technique
- Test interpretation

Imaging intervention: Reader population
- expert readers or professionals 'at large'
- variation across readers & institutions
- extent of reader experience
- want to generalize beyond sample on the study, but do not want to bias against new technology if readers have little experience

The “moving target” problem within studies
- Most changes occurring during study unlikely to affect endpoints appreciably ("hottest thing since sliced bread" syndrome).
- Need to anticipate changes and implement in controlled way during study.
- Statistical methods for adaptive study design and analysis, usually based on Bayesian approaches, can be used.

Reference information ("gold standard")
- Procedure for determining presence of target condition
  - usually (but not always) involves pathology (or cytology) examination.
  - May need to supplement pathology with follow-up for subjects who do not have biopsies
  - needs unambiguous definition
  - uniform assessment across centers
  - sometimes use “truth committee”.
Variability among pathologists often persists even when common protocols are in place.

- Do all studies need central pathology interpretation.

- Reference information (“Gold standard”) may be in error.

Bias: sources and control

What is bias?

A systematic deviation of study-based inferences (e.g. estimates of accuracy) from the true values of parameters describing the phenomenon under study (e.g. true value of accuracy of the modality).
Biases in studies of diagnostic tests

Although some statistical methods for bias correction are available, the potential for bias is best addressed at the design stage.

Common types of bias:
• Verification Bias (workup bias)
• Interpretation Bias
• Uninterpretable Tests
• Selection/referral bias
• Temporal effects (“moving target” problem, “learning effects”)

Verification bias: Example

Suppose the full results of a study would be:

<table>
<thead>
<tr>
<th>Test result</th>
<th>Disease Status</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>160</td>
<td>30</td>
<td>190</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>120</td>
<td>510</td>
<td>630</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>280</td>
<td>540</td>
<td>820</td>
<td></td>
</tr>
</tbody>
</table>

Then we would estimate:
sensitivity = 160/280 = 57%
specificity = 510/540 = 94%

Verification bias example (cont.)

However, the investigators had reference standard only on a subset of cases:

<table>
<thead>
<tr>
<th>Test result</th>
<th>Disease Status</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Positive</td>
<td>80</td>
<td>15</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>Negative</td>
<td>20</td>
<td>85</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
<td>200</td>
<td></td>
</tr>
</tbody>
</table>

New estimates: sensitivity = 80% & specificity = 85%
• Note: 50% of cases testing positive were verified compared to 17% of cases testing negative
**Verification bias**

- May occur if definitive information is selectively available on a subset of cases
  - especially when selection for disease verification depends on test result.
- Assuming all participants without verified diseases are disease free leads to bias.
- Avoid by verifying all cases. May suffice to verify a randomly selected subset.
- Statistical correction is possible.

---

**Interpretation Bias**

Occurs when test interpreter knows results of
- other tests for the disease
- or (even worse) the reference standard

---

**Possible effect of interpretation bias**

<table>
<thead>
<tr>
<th>Test +</th>
<th>Test -</th>
</tr>
</thead>
<tbody>
<tr>
<td>Truth +</td>
<td>a b</td>
</tr>
<tr>
<td>Truth -</td>
<td>c d</td>
</tr>
</tbody>
</table>

If test is interpreted with knowledge of reference standard, both sensitivity and specificity could be overestimated.
Interpretation Bias

- Blinding to reference standard is absolutely essential.
- However, the broader question is:
  - What information should be available to reader?
- The answer depends on what is actually being assessed.
- Some questions:
  - Is total blinding to other information necessary?
  - Will such blinding enhance validity and generalizability of study?

What information should be available to reader?

- The evaluation of a diagnostic test needs to take into account the context in which the test will be used.
- Study design needs to specify the types and amount of information available to the reader.
- Information available to test interpreter should correspond to the context to which study results would generalize.
- Example: In studies of contrast agents, if reading of contrast will be done clinically always with access to baseline (non-contrast enhanced) scan,
  - Will the results be biased if contrast enhanced images were interpreted with access to baseline images?

Uninterpretable test results

- Common problem.
- What is meant by “uninterpretable result”?
  - technically unacceptable?
  - equivocal?
- Equivocal/intermediate results are not uninterpretable.
- Excluding uninterpretable test outcomes may result in bias; often leads to over-estimates of test accuracy.
- If at all possible, participants should be followed to determine reference standard information.
Investigating uninterpretable results

If test was to be repeated, could the problem be resolved?

- No: Need new test result category
- Yes: May repeat test but still need to report frequency

Investigating uninterpretable results (cont.)

Are uninterpretable results more common in diseased cases?

- If yes, then uninterpretable result may be of value in diagnosis/prediction

Topics

1. Diagnostic imaging evaluation:
   1. Overview
   2. Measures of performance
   3. Reporting checklist
2. Elements of study design
   1. Studies of accuracy in detection
   2. Studies of accuracy in prediction
   3. Studies of outcome
3. Generalizability and bias considerations
4. Reporting revisited
Review of Topics

1. Diagnostic imaging evaluation:
   1. Overview
   2. Measures of performance
   3. Reporting checklists

2. Elements of study design
   1. Studies of accuracy in detection
   2. Studies of accuracy in prediction
   3. Studies of outcomes

3. Generalizability and bias considerations

4. Reporting revisited