Designing Studies of Diagnostic Imaging

Chaya S. Moskowitz, PhD

With thanks to Nancy Obuchowski
Outline

• What is “study design”?  
• Building blocks of imaging studies  
• Strategies to improve study efficiency
What is “study design”? 

• **Where are you going?**
  • Research goals, study objectives

• **How will you get there?**
  • Available resources (Examples: patients, funding, expertise, time)

• **What will you do?**
  • Examples: Prospective vs. retrospective; paired vs. unpaired; randomization; study population; diagnostic test readings and endpoint collection
Building Blocks of Imaging Studies

- Study Objectives
- Performing and Interpreting Imaging Test(s)
- Reference Information
- Endpoints
Study Objectives
Study Objectives

• Where are you going?
• Active statement about the specific steps to answer the research question.
Developmental stages

Phase I: Discovery

Phase II: Introductory

Phase III: Mature

Phase IV: Disseminated
Study Objective

• Where are you going?
• Active statement about the specific steps to answer the research question.

• Should **not** re-state the research hypothesis
• Should **not** state investigators’ hope for a statistically significant finding
Example

Suppose you want to compare the accuracy of mammography and breast MRI...
Choice of Study Objective:

#1: To show that breast MRI is better than mammography.

What aspect of performance is being evaluated?
Choice of Study Objective:

#2: To show the diagnostic accuracy of breast MRI is better than mammography.

Be specific about the endpoint.
#3: To show the breast-level sensitivity and specificity of MRI are better than mammography.

What is the population(s) of interest?
Choice of Study Objective:

#4: To show that board-certified mammographers interpreting MRI of high-risk women have better breast-level sensitivity and specificity than when interpreting mammograms.

State the objective in a detached way.
Choice of Study Objective:

#5 To estimate and compare the breast-level sensitivity and specificity of board-certified mammographers interpreting MRI and mammograms of high-risk women
Performing and Interpreting Imaging Test(s)
Performing and Interpreting Imaging

• Timing of the imaging tests
  – When in patients’ history?
  – Sequential vs. cross-over? Wash-out period?

• Collecting data from the imaging tests
Design of Studies Involving Readers

• Environment for reader interpretations
  – “In the field” vs. “test per se” (Begg and McNeil, 1988)

• Single reader vs. multi-reader

• Selection of readers
  – Target-reader population?
    • Early phase studies, narrow target-reader population
    • Late phase studies, broad target-reader population
  – Single institution, multi-institution, core-lab
Design Options for accuracy studies of tests requiring reader interpretation

<table>
<thead>
<tr>
<th>Options</th>
<th>Convenience</th>
<th>Inter-reader variability</th>
<th>Generalizeable estimate of accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single reader</td>
<td>****</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consensus reading</td>
<td>**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Core-Lab reading</td>
<td>*</td>
<td>***</td>
<td>**</td>
</tr>
<tr>
<td>Multi-Reader study</td>
<td>*</td>
<td>****</td>
<td>****</td>
</tr>
</tbody>
</table>
Reference Information

• Source of information, completely different from the test or tests under evaluation, which tells us the true condition status of the patient

Zhou et al 2011
Reference Standard

- Sometimes called “gold standard”
- Rarely is it “gold”
- How nearly “gold” the reference standard needs to be depends on the phase of the test assessment
Examples of Reference Standards

- Pathology
- Surgery
- Reference standard expert panel
- Another imaging test
- Follow-up

You can use multiple, but equally good, reference standards in the same study – Composite reference standard
Diagnosis of PE was based on any one of the following:

1. Positive contrast enhanced spiral CT angiogram showing PE in a main or lobar pulmonary artery irrespective of clinical assessment.
2. Positive contrast-enhanced spiral CT angiogram or venous phase imaging of the lower extremities in a patient with a high or intermediate clinical probability by the Wells criteria.
3. High probability ventilation-perfusion lung scan in a patient with no previous PE and high or intermediate probability clinical assessment by the Wells criteria.
4. Positive venous ultrasound in a patient with no previous deep venous thrombosis at the same site in a patient with a nondiagnostic V/Q scan and high or intermediate clinical probability by the Wells criteria.
5. Positive conventional digital subtraction pulmonary angiogram.
What if there is no reference standard?

• Some types of studies do not require reference standards

• Examples:
  – Correlation study: New test results correlate with standard test
  – Agreement study: New test results agree with standard test

• For an accuracy study, it’s very important to have a reference standard
Endpoints

• Measurements required by study objectives

• The success of a study, of any phase, depends critically on the choice of a primary endpoint.

• Endpoint must:
  • Correspond to the study objectives
  • Be sensitive to the effect you are measuring
  • Be reliably measured
  • Be clinically relevant
## Choice of endpoints

- Should be appropriate for development phase

<table>
<thead>
<tr>
<th>Technical parameter</th>
<th>Accuracy</th>
<th>Effect on patient care decisions</th>
<th>Patient outcome</th>
<th>Effect on society</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery</td>
<td>yes</td>
<td>yes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Introduction</td>
<td></td>
<td>yes</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>Mature</td>
<td></td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>Disseminated</td>
<td></td>
<td></td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>
Details make a difference!

• Consider a study of the accuracy of CT for detecting colon polyps.
• Suppose you want to compare accuracy without computer-aided detection (CAD) vs. with CAD.
Without CAD, reader finds a polyp in descending colon
Computer-Aided Detection – marks 4 suspicious areas
With CAD, reader finds a second polyp.
Did CAD increase the reader’s accuracy?
Second example:
Without CAD, reader finds a FP in the ascending colon but misses the actual polyps
With CAD, reader finds the polyps. Did CAD increase the reader’s accuracy?
How do you define a True Positive?

- If a patient has disease, is it sufficient to find *anything* (even a FP)?
- Does the reader need to correctly locate the lesion?
- Does the reader need to find all lesions?

You need to state these details in study protocol.
Order of building blocks is not fixed

• Prospective
• Retrospective
Retrospective Study Flow

1. Study Objectives
   - Identify patients

2. Performing and Interpreting Imaging Test(s)

3. Reference Information
   - Based on reference

4. Endpoints
   - (reread images)

Done in the past
Two Examples of Diagnostic Accuracy Studies

1) Prospective design
2) Retrospective design
Prospective study example

Study Objective:
- Compare mammography and MRI
- Enroll 400 patients at high-risk for breast cancer

Imaging Tests:
- Perform MRI and mammography annually for three years
Breast Imaging Example

• Paired design: All patients have both MRI and mammogram

• Paired (vs. unpaired) designs
  – Control for case-difficulty, eliminate confounding
  – Statistically more efficient $\rightarrow$ smaller sample size
  – Pairing *ensures* that two groups are same
    • (Randomization makes two groups *similar.*)
Breast Imaging Example

- Blinded design:
  Technicians perform MRI without knowledge of mammogram (and vice versa), AND images read by radiologists who haven’t seen competing images.

- Images read by scheduled radiologist at time of scan, but design calls for a multi-reader interpretation at a later date.
Prospective study example

Study Objective:
Compare mammography and MRI
Enroll 400 patients at high-risk for breast cancer

Imaging Tests:
Perform MRI and mammography annually for three years

Reference Information:
Perform biopsy, surgery (if indicated); follow all patients
Breast Imaging Example
Reference Standard

• Composite reference standard
• Combination of biopsy results within 365 days of the imaging tests and clinical follow-up at 1 year
  – Includes interview with participant and medical record review.
Prospective study example

**Study Objective:**
Compare mammography and MRI
Enroll 400 patients at high-risk for breast cancer

**Imaging Tests:**
Perform MRI and mammography annually for three years

**Reference Information:**
Perform biopsy, surgery (if indicated); follow all patients

**Endpoint:**
Breast-level sensitivity and specificity
Breast Imaging Example
Reference Standards - Endpoints

• Of 400 high-risk patients, only 5 developed cancer over the three-year study.

Unable to reliably estimate sensitivity!

• This is one reason why retrospective designs are used often for accuracy studies
Retrospective study example

- **Imaging Tests:**
  - CT Colonography performed *in the past*

- **Reference Information:**
  - Optical Colonoscopy performed *in the past*

- **Study objective:**
  - Estimate and compare accuracy of CTC with and without CAD
  - (Identify 50 patients with and 50 without polyps on OC)

(Dachman et al 2010)
Colon CAD CT Study Example

Study Objectives

• **Primary objective:**
  Estimate and compare the **segment-level** area under the ROC curves for board-certified radiologists with and without CAD

• **Secondary objective:**
  Estimate and compare the **patient-level** area under the ROC curves for board-certified radiologists with and without CAD
Retrospective study example

- **Imaging Tests:** CT Colonography performed in the past
- **Reference Information:** Optical Colonoscopy performed in the past
- **Study objective:** Estimate and compare accuracy of CTC with and without CAD (Identify 50 patients with and 50 without polyps on OC)
- **Endpoints:** Segment- and Patient-level AUCs (19 readers interpret images)

(Dachman et al 2010)
Colon CAD CT Study Example
Re-Reading Images for Endpoints

Cross-Over Design

19 readers interpreted images in 4 reading sessions:

• Session 1: 50 cases without CAD
• Session 2: 50 cases with CAD
• One month wash-out
• Session 3: first 50 cases with CAD
• Session 4: second 50 cases without CAD
Colon CAD CT Study Example
Endpoint Definitions

Primary Analysis: (segment-level)
True Positive: reader must correctly locate at least one polyp in segment

Secondary Analysis: (patient-level)
True Positive: reader must correctly locate at least one polyp in patient
Results of Colon CT CAD Study

Primary Analysis:

Without CAD: AUC=0.737
With CAD: AUC=0.758 p=0.015

Secondary Analysis:

Without CAD: AUC=0.711
With CAD: AUC=0.727 p=0.071
Augmenting Retrospective Studies

- Augment sample with diseased patients
  Increase prevalence of diseased cases.

- *Increasing prevalence doesn’t cause bias!*
Enriching Studies

• **Enrich sample with challenging cases.**
  – Ensures we have a good idea of how the modality works in hard cases.
  – In study comparing two modalities, many cases will be obvious, thus easily diagnosed by both modalities. It doesn’t help the study power to have these cases in study.

• Enriching the sample with difficult cases improves study power, *but estimates of accuracy will be lower* (so need to understand/discuss this in paper).
Disadvantages of Retrospective Readings:

Reader interpretations may suffer from lab-effect:

- Reading conditions may be substantially different from typical clinical reading
- Behavior of readers may be affected by knowledge that decisions won’t impact patient, and that they are being watched
What about Randomization?
Appropriate when there is *Equipoise*

- Unpaired design

- Genuine uncertainty as to whether one intervention or one imaging test will be more beneficial than another
Strategies for Randomization

- Simple Randomization:
  - Random assignment of subjects to different groups

- Block Randomization:
  - Randomization is performed in small blocks, so that at the end of each block there is balance between the arms of the study.

Disadvantage for simple and block randomization: in small studies, there can be imbalances in subject covariates
Strategies for Randomization: Stratified Randomization

• Generate separate strata for each combination of covariates

• Potential advantages
  – Ensures balance for known covariates
  – May protect against Type I and Type II errors
  – May increase efficiency (sufficient power with a smaller sample size)
  – May help with subgroup analyses and interim analyses

Kernan et al. J Clin Epidemiol 1999
Strategies for Randomization: Stratified Randomization

- Rarely harmful, but there may be disadvantages:
  - Can be administratively more difficult
  - Over-stratification (too many strata incompletely filled) can lead to imbalance between the arms

- How and when to use:
  - Best when stratification factors have a large effect on patient outcome / primary endpoint
  - Most important for smaller trials (< 100)
  - Keep number of strata used to a minimum
  - Adjust for stratification factors in the analysis

Kernan et al. J Clin Epidemiol 1999
Example: Patient Outcome Study

• Randomized Clinical Trial (RCT) of lower back pain patients randomized to either rapid MRI or x-ray

• 380 patients recruited from 4 diverse sites
  – University outpatient clinic
  – Private, nonprofit teaching hospital
  – Private, for-profit multispecialty clinic with onsite radiology
  – Private, for-profit, free-standing imaging center

Jarvik et al JAMA 2003
A Randomized Study Comparing Test Outcomes

Patient with lower back pain

Baseline functional status

Fast MRI

Plain film

• Primary endpoint:
  – Modified Roland back pain score at 12 months after randomization
• Secondary endpoints:
  – Patient outcomes
  – Health care utilization, costs

Jarvik et al JAMA 2003
Example of 4-strata Randomization

<table>
<thead>
<tr>
<th></th>
<th>Acute pain</th>
<th>Chronic pain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age &lt;50 years</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age ≥ 50 years</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Stratified Randomization in the Trial

<table>
<thead>
<tr>
<th>Recruitment site</th>
<th>Radiograph (n=190)</th>
<th>Rapid MRI (n=190)</th>
<th>Total (n=380)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>73</td>
<td>75</td>
<td>148</td>
</tr>
<tr>
<td>2</td>
<td>30</td>
<td>30</td>
<td>60</td>
</tr>
<tr>
<td>3</td>
<td>46</td>
<td>46</td>
<td>92</td>
</tr>
<tr>
<td>4</td>
<td>41</td>
<td>39</td>
<td>80</td>
</tr>
</tbody>
</table>

- Block randomization within strata with varying block sizes
Why did the investigators recruit patients from 4 diverse sites?
Why did the investigators recruit patients from 4 diverse sites?

• **External validity** – results are generalizable to other institutions
• **Internal validity** – results are free of bias (i.e. there is no systematic error that skewed the results)
Trial Results

• No statistically significant difference in the primary endpoint:
  – Mean Roland score: 8.75 (x-ray) vs. 9.34 (MRI)
  – 95% CI for difference: -1.69 to 0.87

• Is this a type II error?
When your results are not statistically significant (p-value > 0.05), there are two possibilities:

1. There really is no difference between MRI and x-ray and your study correctly reflects that.
2. There really is a difference between MRI and x-ray and your study missed it.
Interpreting Negative Results

When your results are not statistically significant (p-value>0.05), there are two possibilities:

1. There really is no difference between MRI and x-ray and your study correctly reflects that.
2. There really is a difference between MRI and x-ray and your study missed it.

*This can happen because there is a bias, your sample is too small, or just by chance.*
Clinical vs. Statistical Significance

• Need to define *clinically relevant difference*
  – Is the difference big enough to matter to patients or their physicians?

• Statistical significance is related to whether or not a statistical test meets a criterion
  – Not the same thing as clinical significance

• Could have a statistically significant difference that is not clinically relevant

• Important part of the design is the sample size calculation
  – Match the sample size to the minimal clinically relevant difference
Clinical vs. Statistical Significance Example

- Jarvik et al: 2-unit difference in Roland score
  - Need 372 patients to detect this difference with 80% power (so 20% risk of type II error)
  - Study used N=380

- The observed difference and corresponding confidence interval were < 2
  - Mean difference -0.59, 95% CI (-1.69, 0.87)

- Authors concluded that there is no difference between MRI and x-ray.
Interim Analyses

• Analyze data mid-way through data collection

• Used often in clinical trials to identify:
  – if there are safety issues
  – if the results are unimpressive ("futility")
  – if the results are impressive ("benefit")

• Interim analyses are planned during the design phase of the study
Stopping Rules

• Formal statistical rules
  – Control trials’ operating characteristics
• In design phase, set up stopping rules that control “multiplicity problem” (Type I error).
“Multiplicility Problem”

- Statisticians calculate a test statistic to test hypotheses.
- At the start of a study, the test statistic = 0
- If there is no benefit, test statistic randomly fluctuates near zero.
- If there is benefit, test statistic randomly fluctuates and gradually moves away from zero.
“Multiplicity Problem”

• If there is no benefit, test statistic randomly fluctuates near zero.
• If you calculate test statistic often, you will find instances when it is far from zero just by chance.
Stopping Rules

• Formal statistical rules
  – Control trials’ operating characteristics
• In design phase, set up stopping rules that control “multiplicity problem” (Type I error).
• There are lots of methods to do this.
Statistical Methods for Stopping Rules

• Two example of methods
  – Pocock
    • Same alpha level (critical value) at each interim analysis
    • More aggressive at early analyses
  – O’Brien-Fleming
    • Smaller alpha levels at earlier looks
    • More conservative at early analyses

• Example of a trial with 2 equally-spaced interim looks:

<table>
<thead>
<tr>
<th></th>
<th>1st Interim Look</th>
<th>2nd Interim Look</th>
<th>End of study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pocock Method</td>
<td>0.022</td>
<td>0.022</td>
<td>0.022</td>
</tr>
<tr>
<td>O’Brien-Fleming Method</td>
<td>0.0005</td>
<td>0.014</td>
<td>0.045</td>
</tr>
</tbody>
</table>
Stopping Rules

• Formal statistical rules
  – Control trials’ operating characteristics
• In design phase, set up stopping rules that control “multiplicity problem” (type I error).
• There are lots of methods to do this.
• Should be written into the protocol:
  – Which method you will use
  – How many interim looks and when they will take place
Conclusion

- Studies of diagnostic tests are important, nationally recognized.
- Many possible study designs for imaging studies
- Details of the study design determine its worth

- This week take time to carefully consider details of your study’s design
- Listen to other students deliberate their designs