Principles of Radiology Study Design

Chaya S. Moskowitz, PhD

With thanks to Nancy Obuchowski

Financial disclosures: None

1

Outline

- What is “study design”?
- Building blocks of radiology studies
- Strategies to improve study efficiency

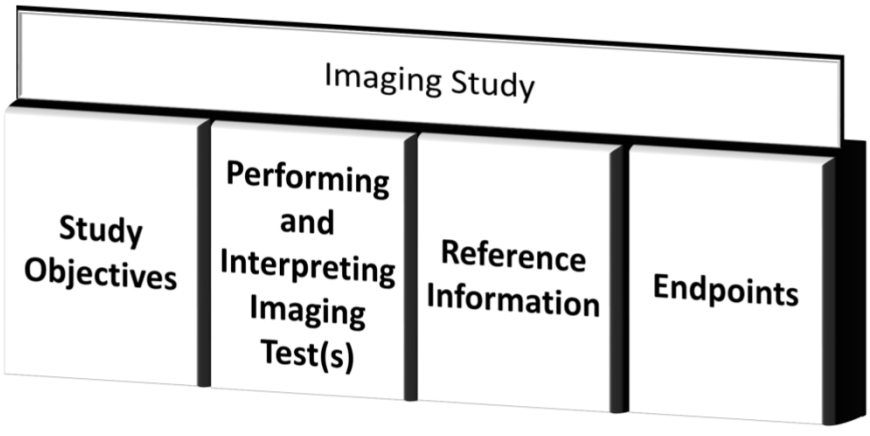
2

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; imaging test readings and endpoint collection

3

Building Blocks of Imaging Studies



4

Developmental stages

Phase I: Discovery
- Technical performance

Phase II : Introductory
- Diagnostic accuracy
- Impact on diagnostic thinking

Phase III: Mature
- Diagnostic accuracy
- Impact on diagnostic thinking
- Impact on therapeutic planning
- Effect on health outcomes

Phase IV: Disseminated
- Effect on health outcomes
- Costs & benefits to society

5

6

Study Objectives

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

7

Study Objectives

- 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

8

Example

Suppose you want to compare the accuracy of contrast-enhanced mammography (CEM) and abbreviated breast MRI (AB-MRI)…

9

Choice of Study Objective:

#1: To show that AB-MRI is better than CEM.

What aspect of performance is being evaluated?

10

Choice of Study Objective:

#2: To show the diagnostic accuracy of AB-MRI is better than CEM.

Be specific about the endpoint.

11

Choice of Study Objective:

#3: To show that breast-level sensitivity and specificity of AB-MRI are better than CEM.

What is the population(s) of interest?

12

Choice of Study Objective:

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

State the objective in a detached way.

13

Choice of Study Objective:

#5 To estimate and compare the breast-level sensitivity and specificity of AB-MRI and CEM in high-risk women when read by board-certified mammographers

14

15

Performing and Interpreting Imaging Tests

- Timing of the imaging tests
  - When in patients' history?
  - Sequential vs. cross-over? Wash-out period?
- Collecting data from the imaging tests

16

Design of Studies Involving Readers

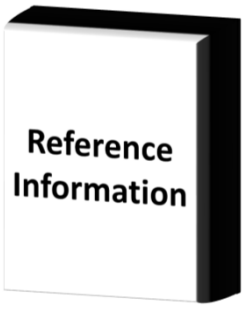
- Environment for reader interpretations
  - "In the field" vs. "test per se" (Begg and McNeil, 1988)
- Single reader vs. multi-reader
- Selection of readers
  - Target-reader population?
    - Early phase studies, narrow target-reader population
    - Late phase studies, broad target-reader population
  - Single institution, multi-institution, core-lab

17

Design Options for Accuracy Studies of Tests Requiring Reader Interpretation

Options	Convenience	Inter-reader variability	Generalizeable estimate of accuracy
Single reader	****		
Consensus reading	**		
Core-Lab reading	*	***	**
Multi-Reader study	*	****	****

18



19

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

20

Reference Standard

- Sometimes called “gold standard”
- Rarely is it “gold”

21

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 –

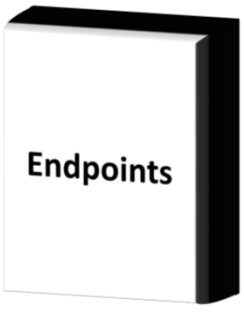
Composite reference standard

22

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

23



24

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

25

Choice of Endpoints

- Should be appropriate for development phase

	Technical parameter	**Accuracy**	**Effect on patient care decisions**	**Patient outcome**	**Effect on society**
Discovery	yes	yes			
Introduction		yes	yes		
Mature		yes	yes	yes	
Disseminated		yes		yes	yes

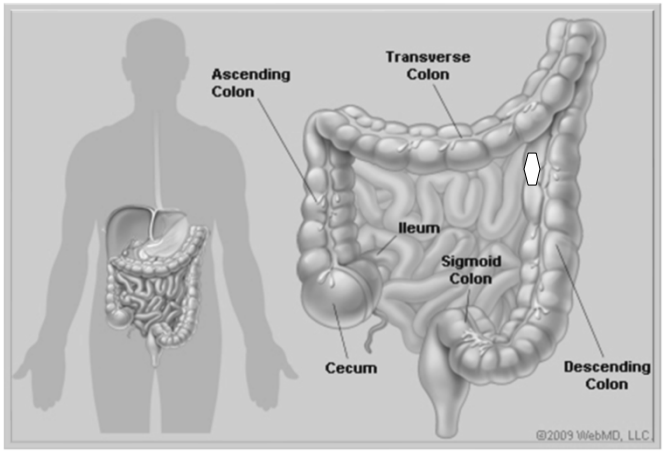
26

Details Make a Difference!

- Consider a study of the accuracy of CT for detecting colon polyps.
- Suppose you want to compare accuracy with vs. without knowledge of artificial intelligence-assisted polyp detection (AI)

27

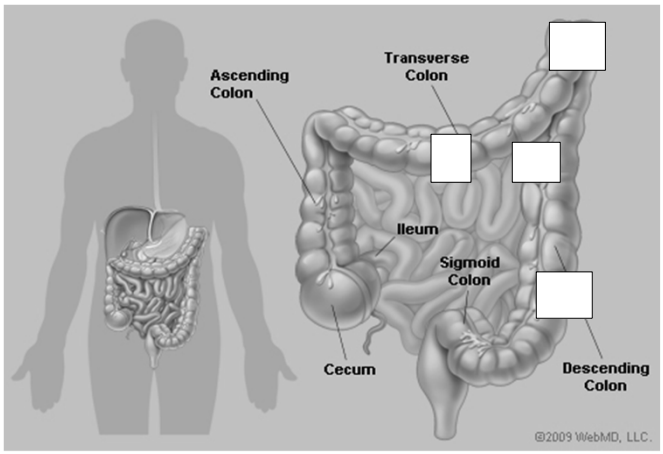
Without AI, reader finds a polyp in descending colon



28

28

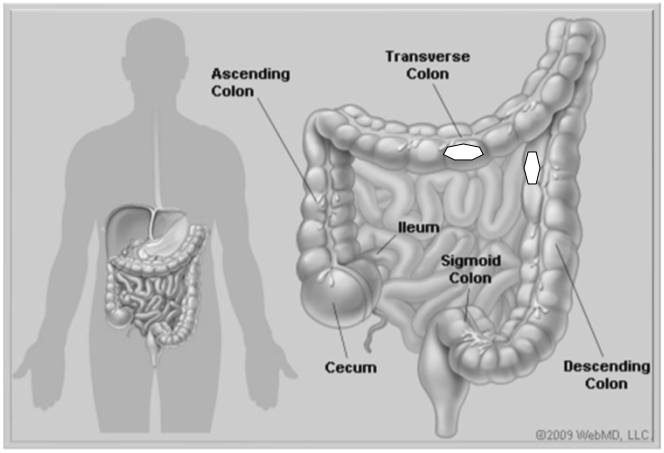
AI marks 4 suspicious areas



29

29

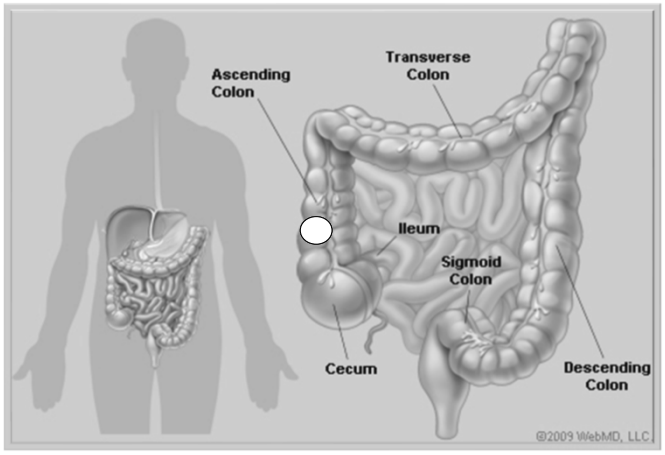
AI, reader finds a second polyp.
Did AI increase the reader's accuracy?



30

30

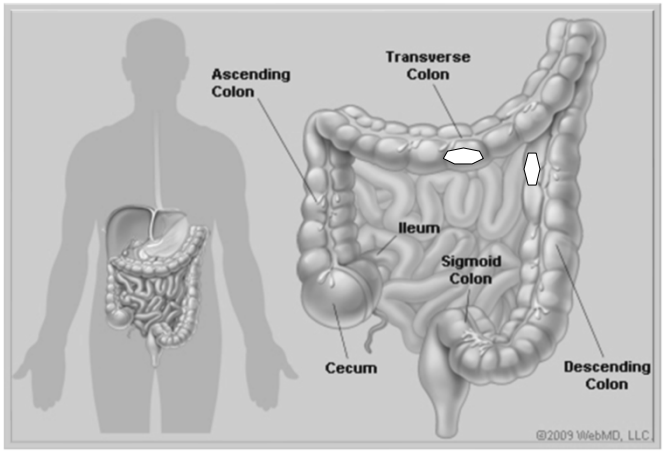
Second example:
Without AI, reader finds a FP in the ascending colon but misses the actual polyps



31

31

With AI, reader finds the polyps.
Did AI increase the reader's accuracy?



32

32

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.

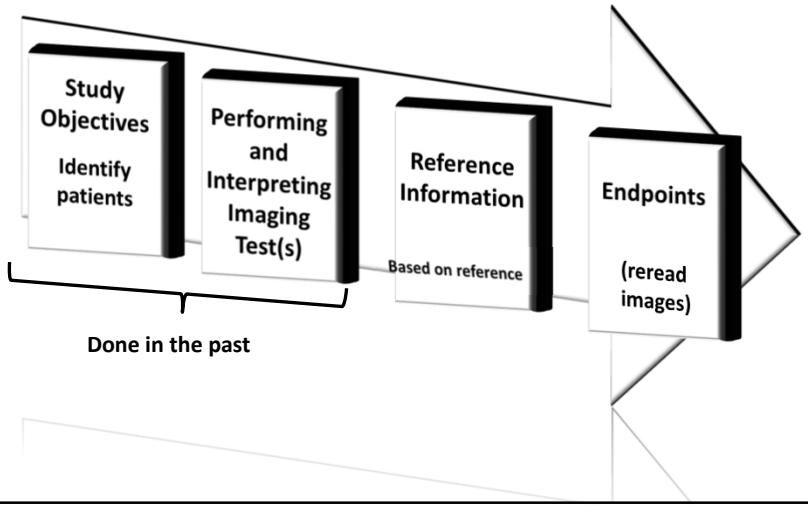
33

Order of Building Blocks is Not Fixed

- Prospective
- Retrospective

34

Retrospective Study Flow



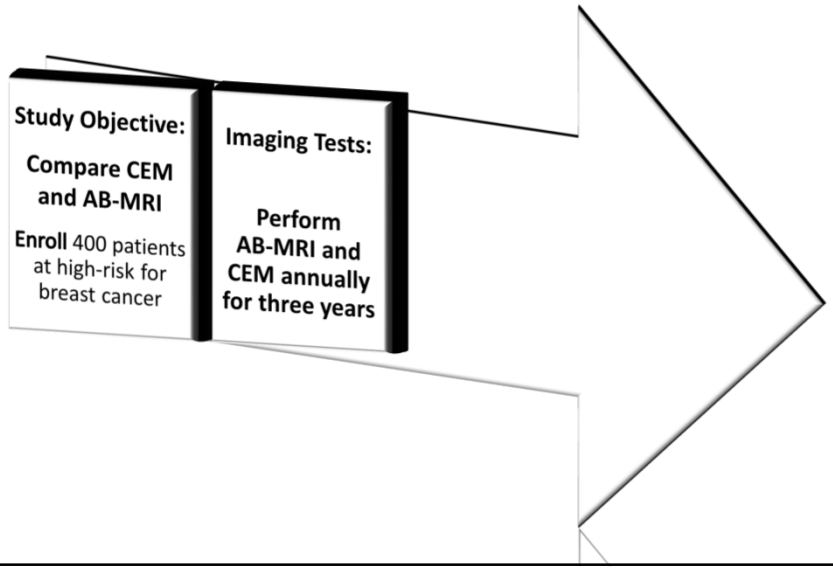
35

Two Examples of Diagnostic Accuracy Studies

1) Prospective design
2) Retrospective design

36

Prospective study example



37

Breast Imaging Example

- Paired design: All patients have both AB-MRI and CEM
- Paired (vs. unpaired) designs
  - Control for case-difficulty, eliminate confounding
  - Statistically more efficient → smaller sample size
  - Pairing *ensures* that two groups are same
    - (Randomization makes two groups *similar.*)

38

Breast Imaging Example

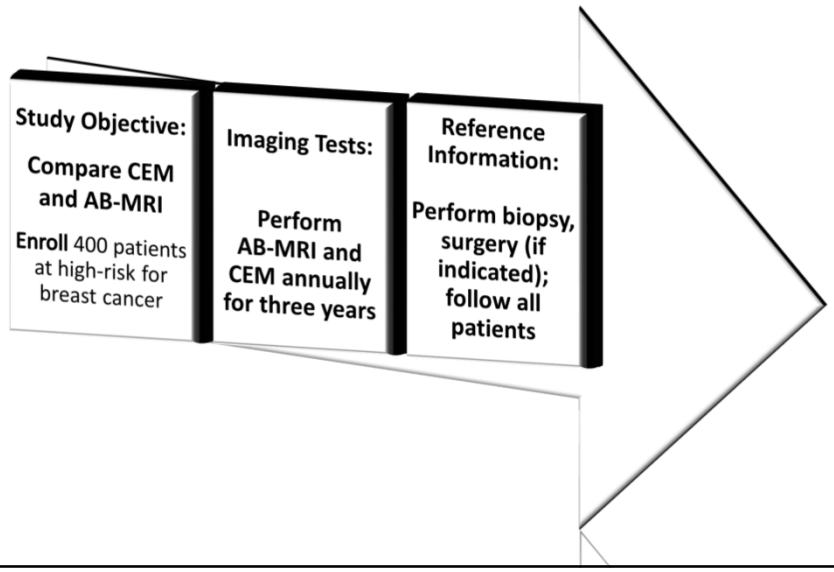
- Blinded design:

  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.

39

Prospective study example



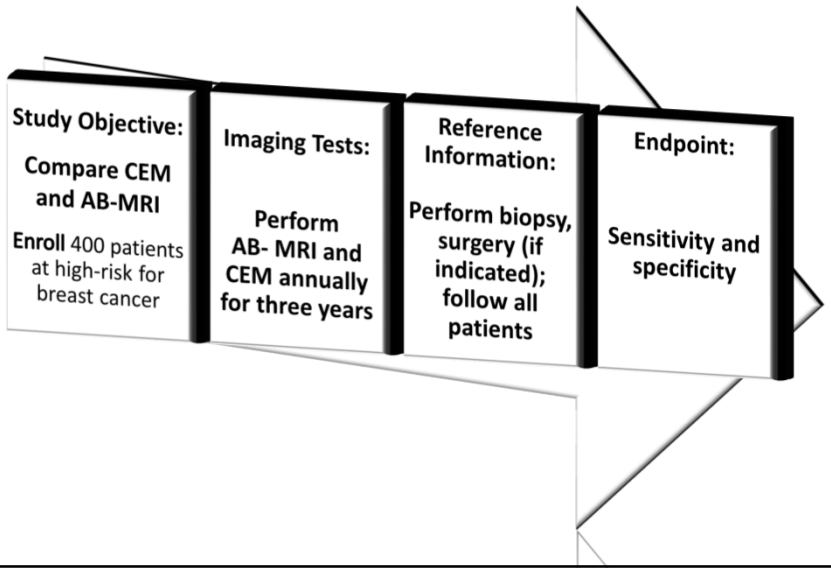
40

Breast Imaging Example
Reference Standard

- Breast cancer defined by 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.

41

Prospective study example



42

Breast Imaging Example
Reference Standards - Endpoints

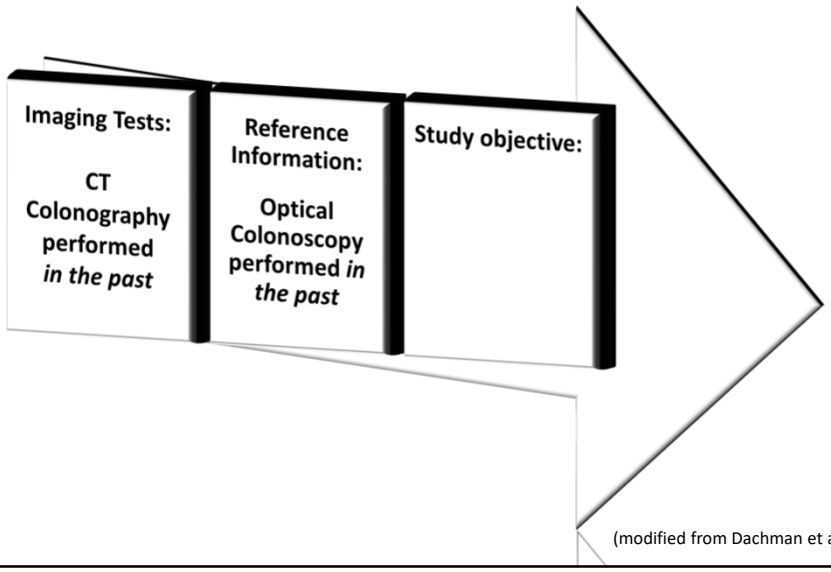
- Of 400 high-risk patients, only 5 developed cancer over the three-year study.

  Unable to reliably estimate sensitivity!

- This is one reason why retrospective designs are used often for accuracy studies

43

Retrospective study example



(modified from Dachman et al 2010)

44

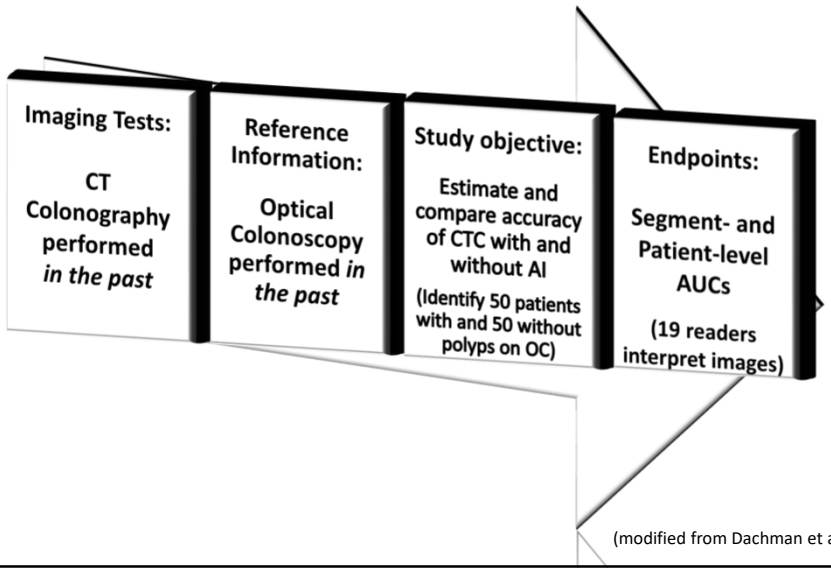
Colon CAD CT Study Example
Study Objectives

- Primary objective:
  Estimate and compare the segment-level area under the ROC curves for board-certified radiologists evaluating CT Colonography images for polyps with and without AI

- Secondary objective:
  Estimate and compare the patient-level area under the ROC curves for board-certified radiologists evaluating CT Colonography images for polyps with and without AI

45

Retrospective study example



(modified from Dachman et al 2010)

46

Colon CT Study Example
Re-Reading Images for Endpoints

Cross-Over Design

19 readers interpreted images in 4 reading sessions:

- Session 1: 50 cases without AI
- Session 2: 50 cases with AI

- **One month wash-out**

- Session 3: first 50 cases with AI
- Session 4: second 50 cases without AI

47

Colon 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

48

Results of Colon CT Study

Primary Analysis:

Without AI:	AUC=0.74	(95% CI: 0.73-0.75)	
With AI:	AUC=0.76	(95% CI: 0.75-0.77)	p=0.015

Secondary Analysis:

Without AI:	AUC=0.71	(95% CI: 0.69-0.72)	
With AI:	AUC=0.73	(95% CI: 0.72-0.75)	p=0.071

49

Augmenting Retrospective Studies

- Augment sample with diseased patients
  Increase prevalence of diseased cases.
- *Increasing prevalence doesn't cause bias in estimates of sensitivity, specificity, and ROC curves!*

50

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).

51

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

52

What about Randomization?

53

Appropriate when there is *Equipoise*

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

54

Strategies for Randomization

- Basics: Simple, Block, Stratified
- Stratified Randomization
  - –Generate separate strata for each combination of covariates
  - –Potential advantages:
    - Ensures balance for known covariates
    - May increase efficiency
  - –Rarely harmful, but careful of over-stratification!
  - –Best when stratification factors have a large effect on patient outcome / primary endpoint
  - –Adjust for stratification factors in the analysis

Kernan *et al. J Clin Epidemiol* 1999

55

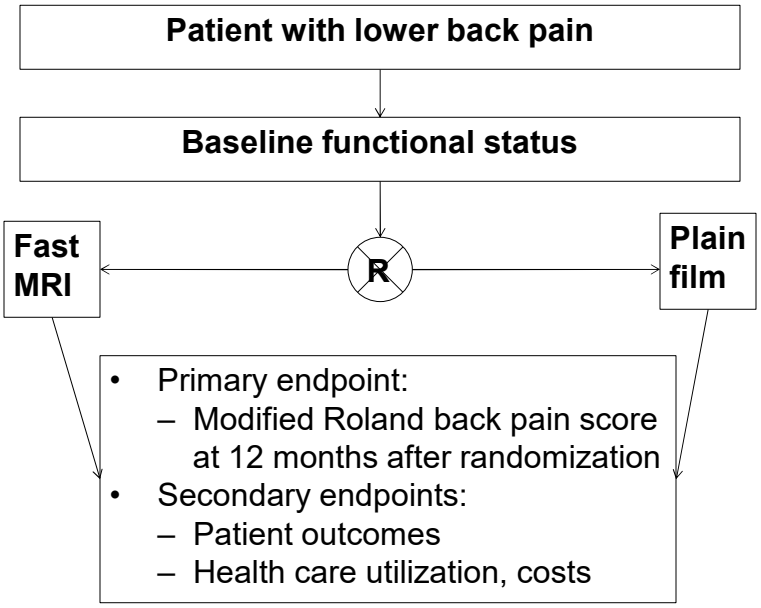
Example: Patient Outcome Study

- 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

56

A Randomized Study Comparing Test Outcomes



Jarvik *et al JAMA* 2003

57

Stratified Randomization in the Trial

Recruitment site	Radiograph (n=190)	Rapid MRI (n=190)	Total (n=380)
1	73	75	148
2	30	30	60
3	46	46	92
4	41	39	80

- Block randomization within strata with varying block sizes

58

Why did the investigators recruit patients from 4 diverse sites?

59

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)

60

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?

61

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.

62

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.

63

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 tests 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

64

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.

65

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

66

Stopping Rules

- Formal statistical rules
  - Control trials' operating characteristics
- In design phase, set up stopping rules that control "multiplicity problem" (Type I error).

67

"Multiplicity Problem"

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

68

"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.

69

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

70

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

71