Recognizing and Controlling Bias

BY
NANCY OBUCHOWSKI, PHD

1

Disclosures

- Statistical consultant for IBM, Siemens, RSNA, Elucid, QURE, Takeda
- Contracts are with Dr. Obuchowski's institution

2

Learning Objectives

1. Identify common types of study design bias in imaging studies
2. Understand the effects of bias on your study
3. Develop strategies to minimize the bias

3

What is "BIAS" ?

4

BIAS:

ANY SYSTEMATIC DEVIATION OF OBSERVATION / STUDY RESULT FROM TRUE CLINICAL STATE

5

BIAS:

ANY SYSTEMATIC DEVIATION OF OBSERVATION / STUDY RESULT FROM TRUE CLINICAL STATE

Bias → Results → Conclusions

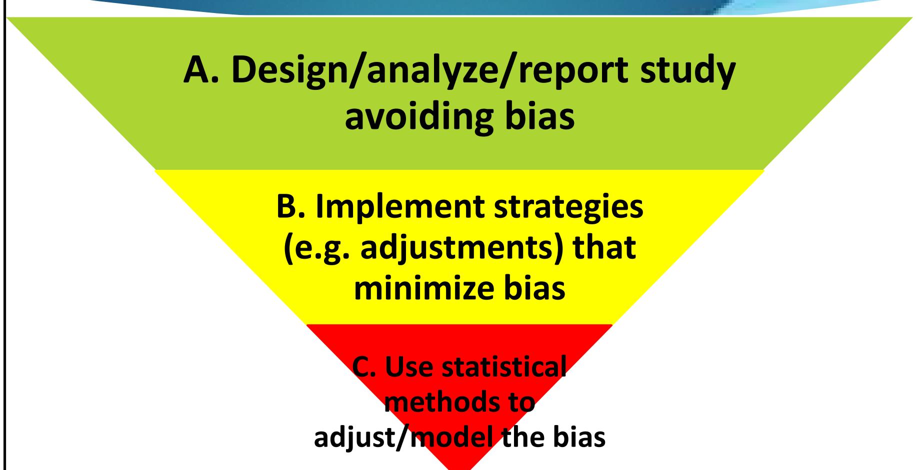
6

Outline

- Examples of bias
  - Study design (4)
  - Analysis (2)
  - Reporting (2)
- Controlling/Correcting for bias

7

Controlling/Correcting for bias



8

Study Design Biases

- Incorporation Bias
- Spectrum Bias
- Verification Bias
- Differential Reference Bias

9

Example 1: AI for detecting liver lesions

- New AI algorithm has been developed to help radiologists find liver lesions on CT
- We need to design a study to test whether radiologists' accuracy is improved with the AI

10

Example 1:
Proposed Reference Standard

All findings found by study readers while unaided + Findings found by the AI → 3+ Expert readers independently review consolidated findings and determine which ones are true lesions

Take majority decision of 3+ experts

11

Example 1:
Proposed Reference Standard

All findings found by study readers while unaided + Findings found by the AI → 3+ Expert readers independently review consolidated findings and determine which ones are true lesions

What's not to like here?

12

Incorporation Bias

- Results from diagnostic test under evaluation are incorporated, in full or part, into the evidence used to establish the ground truth
- It often favors the new modality.

13

Example 1: Reference Standard

All findings found by study readers while unaided + ~~Findings found by the AI~~ → 3+ Expert readers independently review consolidated findings and determine which ones are true lesions

"A" level correction

14

Example 1: Reference Standard

All findings found by study readers while unaided + ~~Findings found by the AI~~ → 3+ Expert readers independently review consolidated findings and determine which ones are true lesions

****Alternatively, just let the expert readers interpret the unmarked images and use the majority decision****

15

Example 2: New modality to detect intracranial hemorrhage (ICH)

- New rapid test to rule out hemorrhagic stroke
- We need to design a study to determine if new rapid test can replace standard CT imaging

16

Example 2: Proposed patient recruitment

200 consecutive stroke patients (NIHSS ≥ 7)

~170 ischemic

~30 non-ischemic

17

Example 2: Proposed patient recruitment

200 consecutive stroke patients (NIHSS ≥ 7)

~170 ischemic

~30 non-ischemic

What's not to like here?

18

Spectrum Bias

- Study sample does not include the complete spectrum of patient and disease characteristics
  - Spectrum of disease presentation (severity, type, location) (affects sensitivity)
  - Spectrum of non-disease and other diseases that mimic disease of interest (affects specificity)**

**critical to this study

19

Example 2: Patient recruitment **with enrichment** *("B"-level correction)*

200 consecutive stroke patients (NIHSS $\geq$ 7)

~170 ischemic

~30 non-ischemic

Enrich with less common, difficult cases:
- Infratentorial ICH
- Perimesencephalic SAH

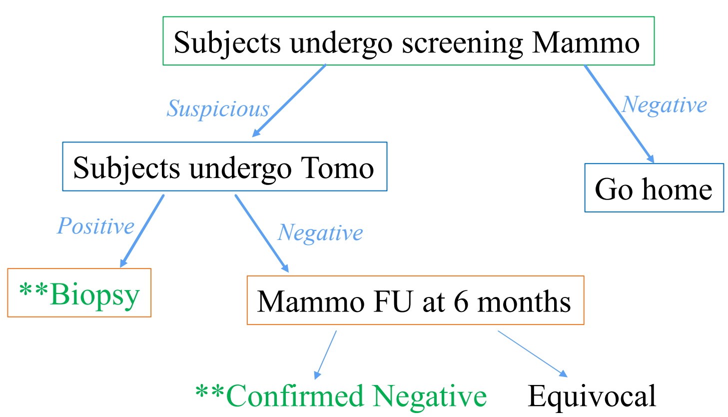
20

Example 3: Tomosynthesis study

- New tomosynthesis modality designed to replace 2D mammography
- We need to design a study to test if readers' accuracy with Tomo is superior to their accuracy with Mammo

21

Example 3: Proposed Study Plan

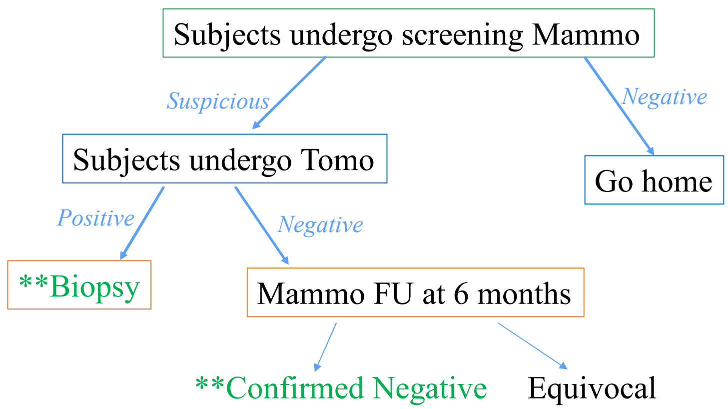


**Subjects eligible for reader study

22

Example 3: Proposed Study Plan

What's not to like here?

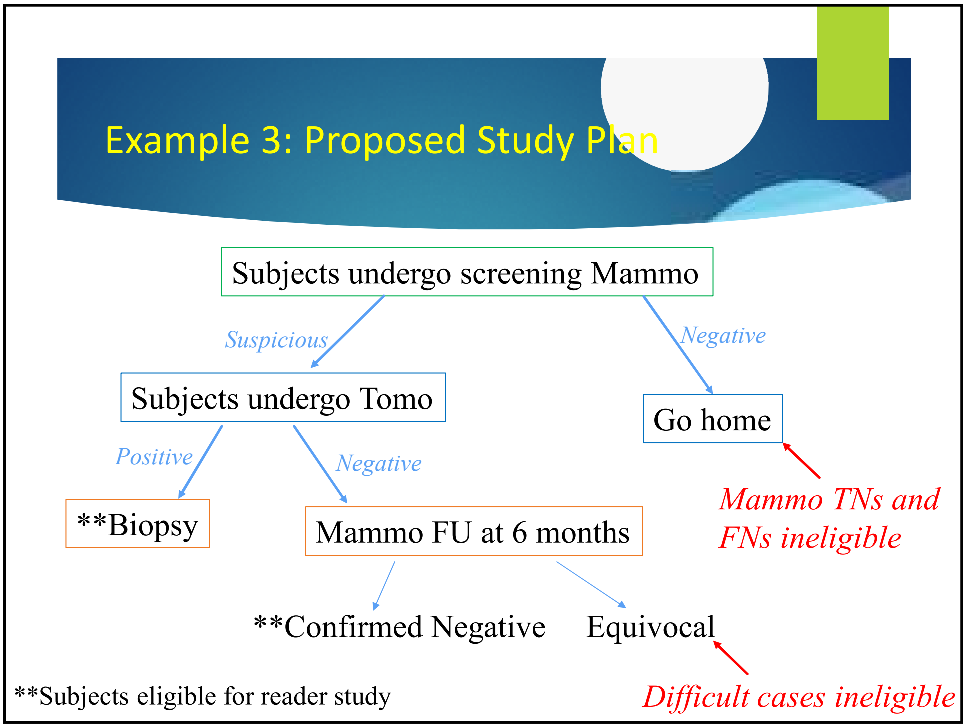


23

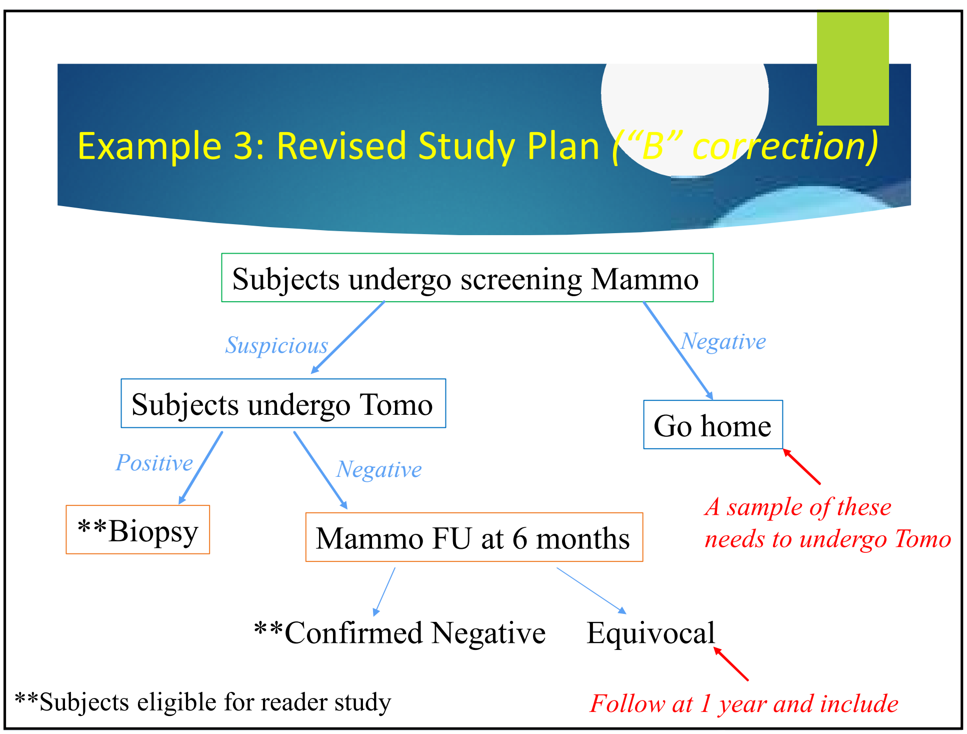
Verification Bias

- Patients with positive (or negative) test results are preferentially referred for reference standard procedure; the bias occurs when estimates of accuracy are based only on verified patients.

24


Example 3: Proposed Study Plan
Subjects undergo screening Mammo
Suspicious
Negative
Subjects undergo Tomo
Go home
Positive
Negative
**Biopsy
Mammo FU at 6 months
Mammo TNs and FNs ineligible
**Confirmed Negative
Equivocal
**Subjects eligible for reader study
Difficult cases ineligible

25


Example 3: Revised Study Plan ("B" correction)
Subjects undergo screening Mammo
Suspicious
Negative
Subjects undergo Tomo
Go home
Positive
Negative
**Biopsy
Mammo FU at 6 months
A sample of these needs to undergo Tomo
**Confirmed Negative
Equivocal
**Subjects eligible for reader study
Follow at 1 year and include

26

Differential reference bias

- Occurs in studies with two (or more) reference standards, where if both were applied to a subject, they would give different results.
- In this study we have Biopsy, 6-month, and 1-year FU as our reference standards.
  - We can't possibly require biopsy for everyone, so need multiple reference standards.
  - Subjects with only 6-month FU should be followed until 1 year.

27

Example 4: CT colonography CAD to detect colon polyps

- Computer Aided Detection (CAD) to help radiologists find colon polyps
- Subjects can have multiple polyps, many possible locations
- We need to determine how to define sensitivity and specificity

28

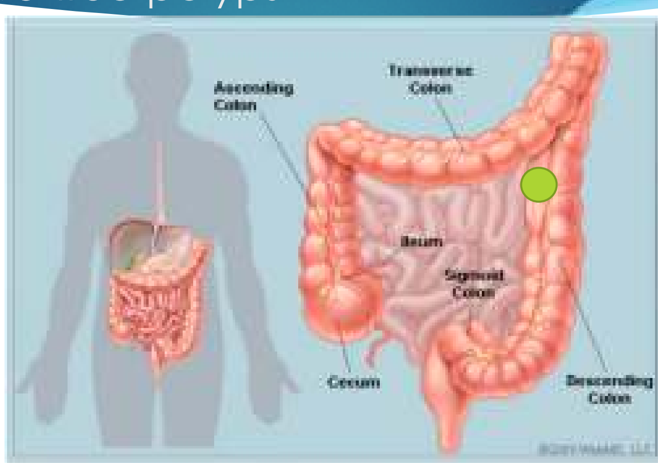
Data Analysis Biases

- Location bias
- Accuracy vs. Agreement

29

Without CAD, reader finds a FP in descending colon and misses two true polyps

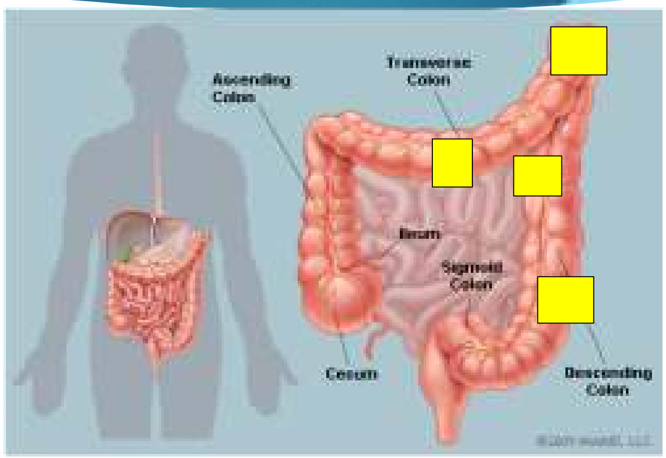
30



30

31

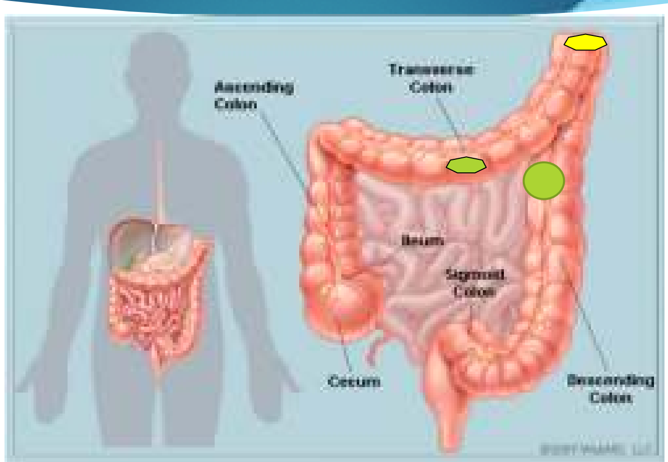
Computer-Aided Detection – marks 4 suspicious areas



31

32

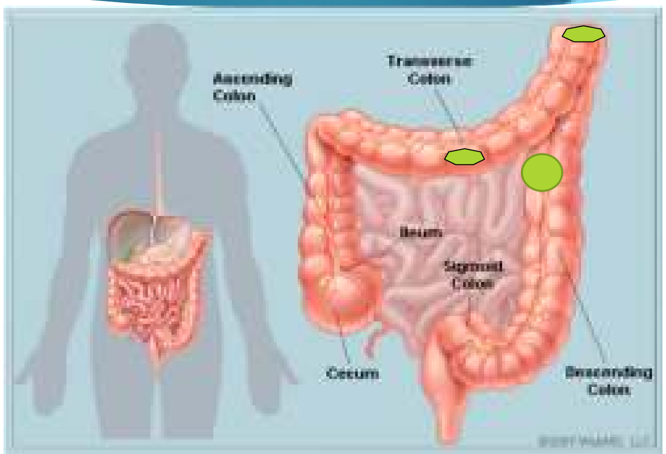
With CAD, reader finds one true polyp but misses the second



32

Or…
With CAD, reader finds both true polyps

33



33

How do we define sensitivity and specificity?

Options	Sensitivity	Specificity
Subject-level	weak	weak
Colon segment-level	better	better
Lesion-level	best	undefined

34

How do we define sensitivity and specificity?

Is detection of 1 polyp sufficient?
Is detection of a FP sufficient?

Options	Sensitivity	Specificity
Subject-level	weak	weak
Colon segment-level	better	better
Lesion-level	best	undefined

Should 10 CAD FPs in a subject be treated the same as 1 CAD FP?

35

Localization bias

- Give credit to reader for finding something in a subject with disease, even if the detected finding is not the real disease.
- It overestimates the accuracy of tests and obscures benefits of an improved test.

36

Example 4: Defining Accuracy

Segment-level Sensitivity: proportion of segments with true polyps where reader correctly detects and locates the true polyp

Segment-level Specificity: proportion of segments without true polyps where reader correctly reports no findings

37

Example 4: Defining Accuracy

Segment-level Sensitivity: proportion of segments with true polyps where reader correctly detects and locates the true polyp

Segment-level Specificity: proportion of segments without true polyps where reader correctly reports no findings

Advantages:

1. *Provides opportunity to see benefit of CAD*
2. *Strict control of potential CAD FPs*
3. *Good statistical power!*

38

Example 5: Measuring accuracy of QIB

- Challenge study of 10 image analysis software systems to measure volume of lung masses
- We need to compare the accuracy of the 10 systems

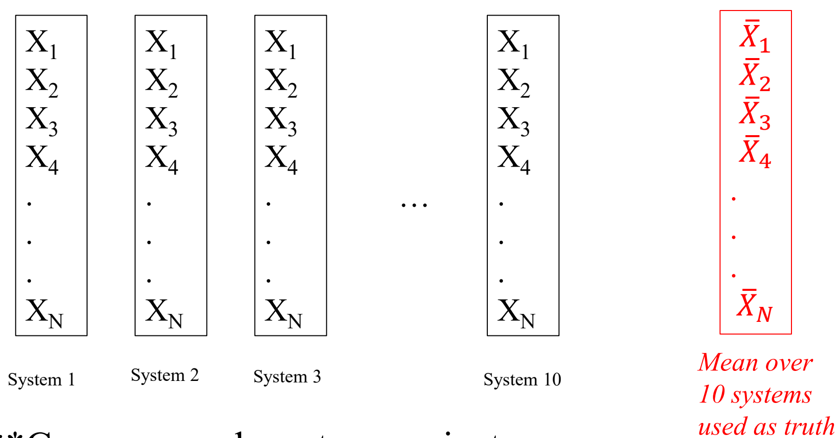
39

Example 5: Proposed accuracy estimate

System 1	System 2	System 3	…	System 10	*Mean over 10 systems used as truth*
X_1	X_1	X_1		X_1	$\bar{X}_1$
X_2	X_2	X_2		X_2	$\bar{X}_2$
X_3	X_3	X_3		X_3	$\bar{X}_3$
X_4	X_4	X_4		X_4	$\bar{X}_4$
.	.	.	…	.	.
.	.	.		.	.
.	.	.		.	.
X_N	X_N	X_N		X_N	$\bar{X}_N$

40

Example 5: Proposed accuracy estimate



**Compare each system against mean

41

Accuracy or Agreement?

- The mean of the 10 systems does not represent ground truth.
- We can't call comparisons of each system to this mean as the system's **accuracy.**
- This is an assessment of **agreement** among the systems (and maybe not even an optimal assessment of agreement)

42

Accuracy or Agreement?

- The mean of the 10 systems does not represent ground truth.
- We can't call comparisons of each system to this mean as the system's **accuracy.**
- This approach overestimates "accuracy" of poor performers and underestimates "accuracy" of good performers
  - Get a true reference standard
  - Report agreement, not accuracy

43

Reporting Biases

- Uninterpretable Test Result Bias
- Non-significant Results Bias

44

Uninterpretable test results

- These are technically unacceptable images, but they are not the same as “equivocal” results.
- Examples:
  - Abdominal gas interfering with pelvic US interpretation
  - Dense breast tissue interfering with mammo interpretation

45

Uninterpretable test results

You can’t exclude these cases!

Two Questions to Ask:

1. Could you repeat the test and potentially get a valid result?
2. Are uninterpretable results more common among diseased patients, or one modality than another?

46

Uninterpretable test results

You can't exclude these cases!

Two Questions to Ask:

1. Could you repeat the test and potentially get a valid result? If not repeatable, then treat like it's another test result: +, -, uninterpretable
2. Are uninterpretable results more common among diseased patients, or one modality than another? Use info to diagnose the patient. Report differences betw modalities.

47

Reporting accuracy results

There are many ways to report diagnostic accuracy:

ROC:
Subject-level
Segment-level

Sensitivity:
Subject-level
Segment-level
Lesion-level

Specificity:
Subject-level
Segment-level
FPs/subject

48

Reporting accuracy results

There are many ways to report diagnostic accuracy:

ROC:
Subject-level
Segment-level

Sensitivity:
Subject-level
Segment-level
Lesion-level

Specificity:
Subject-level
Segment-level
FPs/subject

You can't do them all and report whatever you like best!

49

Specify detailed analysis plan apriori

There are many ways to report diagnostic accuracy:

ROC:
Subject-level
Segment-level

Sensitivity:
Subject-level
Segment-level
Lesion-level

Specificity:
Subject-level
Segment-level
FPs/subject

Primary objective

Secondary objective

50

Biases

Study Design
1. Incorporation bias
2. Spectrum bias
3. Verification bias
4. Differential bias

Analysis
1. Location bias
2. Accuracy/agreement

Reporting
1. Indeterminates
2. Non-significant results

51

5 Key Take-Aways

1. Designing a study requires balance betw scientific rigor and practical considerations. Level "B" adjustments are critical.
2. Consider your reference standard(s) carefully. Be ready to defend it scientifically and statistically (avoid incorporation and verification biases).
3. While we might use suboptimal reference standard for training AI, the usual rigor must be applied to the reference standard for AI validation.
4. When the task is to find abnormalities (e.g. lesions), correct detection and localization are important.
5. Prepare details of your study analysis plan *apriori* so there's no confusion/opportunity for dredging / mis-reporting.

52