Studies of Medical Tests

Thomas B. Newman, MD, MPH

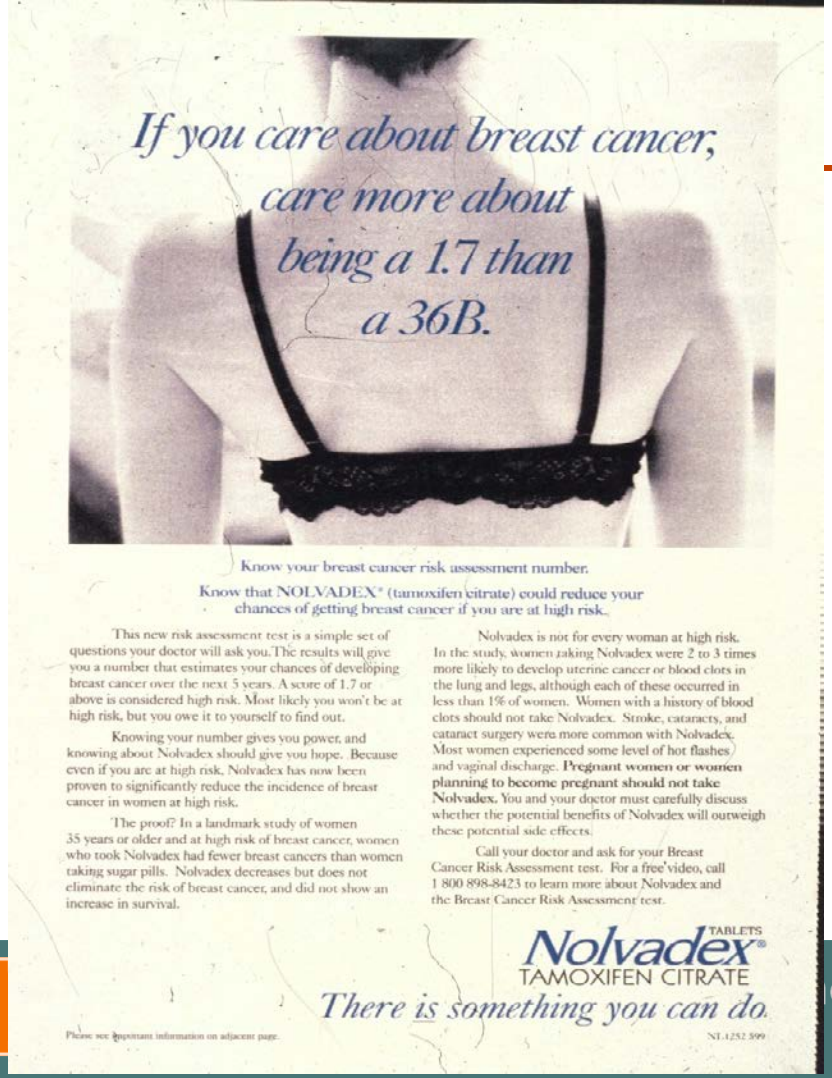
Overview

- **General Issues**
  - Contrast with other observational designs
  - Types of research questions addressed
  - “Gold standard” issues
  - Spectrum of disease and of results
  - Sampling and generalizability
- **Examples:**
  - Reproducibility and Accuracy of S_3
  - Visual assessment of jaundice

What do we mean by “tests”?

Studies, procedures, maneuvers intended to provide information about the probability of different health states, e.g.,

- Items of the history and physical examination
- Blood tests
- X-rays
- Endoscopies



"Tests" include history questions

How are studies of tests similar to other clinical research studies?

- Same basic pieces
  - Research question
  - Study design
  - Subjects
  - Predictor variables
  - Outcome variables
  - Analysis
- Same need to generalize from study subjects to populations and from measurements to phenomena of interest

How are studies of tests different?

- Address different types of questions
  - Primarily descriptive
    - Confidence intervals rather than P-values
  - Causal inference may or may not be relevant
- Different biases
  - Spectrum, verification, etc.
- Different statistics used to summarize results
  - Kappa, sensitivity, specificity, ROC curves, likelihood ratios

Sample size for studies of diagnostic tests

- Because they are primarily descriptive, traditional hypothesis testing (Type 1 and Type 2 errors, null hypothesis, etc.) does not apply
- Goal is not to determine whether test result is associated with disease, but HOW STRONGLY
- Therefore, sample size of diseased subjects is based on the WIDTH of the 95% CI desired and the estimated **sensitivity**
- Sample size of nondiseased based on the WIDTH of the 95% CI desired and the estimated **specificity**

Sample size example

- You want to estimate the sensitivity of abnormal urine smell for urinary tract infection in children.
- You estimate it might be 70%.
- You want a 95% CI with width 10% (e.g. 65%-75%)
- DCR-4 Page 81 (use 0.3)
  - N = 323

Sample Size -- Confidence Interval For a Proportion

Instructions: Enter parameters in the Red cells. Answer will appear in the blue cell.

- P = Expected Proportion
- W = Total Width of Confidence Interval

Confidence Level (e.g., 0.95)	P	W
0.95	0.7	0.1

Z-Alpha	
1.96	
N	323
Recalc	

www.sample-size.net for CTSI sample size calculator

Diagnostic Test Research Questions

- How reproducible is it?
- How accurate is it?
- How much new information does it provide?
- How often do results affect clinical decisions?
- What are the costs, risks, and acceptability of the test?
- What is the effect of testing on outcomes?
- How do the answers to these questions vary by patient characteristics?

Gold Standard -1

- Needed for studies that measure accuracy
- Can’t include test being measured (Incorporation bias)
  - Example: White Blood Cell (WBC) count as a predictor of sepsis in newborns
  - “Gold standard” (+blood culture) imperfect
  - Why not include *probable* sepsis, based on judgment of treating clinicians?
  - Judgment affected by WBC count!

Gold Standard -2

- Best if applied *blindly*
  - *Prevent incorporation bias*
- Best if applied *uniformly*
  - *Prevent verification bias, double-gold standard bias*
- If imperfect, test accuracy can be under-estimated or over-estimated
  - Example: culture vs PCR for pertussis
- If nonexistent, think about WHY you want to make the diagnosis
  - Examples: Attention-deficit hyperactivity disorder (ADHD), autism

Spectrum of Disease, Non-disease and Test Results

- Disease is often easier to diagnose if severe
- “Non-disease” is easier to diagnose if patient is well than if the patient has other diseases
- Test results will be more reproducible if ambiguous results excluded

Sources of variation, generalizability and sampling

- Test characteristics like sensitivity and specificity may depend on:
  - How the specimen is obtained and processed
  - How and by whom the test is done and interpreted
- Consider whether you need to sample or stratify results at these levels (depends on the RQ)

Studies of Reproducibility

- For tests with no gold standard
- Often done as part of quality control
  - For a larger study
  - For patient care

Example 1: The Third Heart Sound

- RQs:
  - What is interobserver variability for hearing S_3?
  - How does this vary with level of experience?
- Design: cross-sectional study

ORIGINAL INVESTIGATION

Relationship Between Accurate Auscultation of a Clinically Useful Third Heart Sound and Level of Experience

Gregory Marcus, MD; Joshua Vessey, MD; Mark V. Jordan, MD; Michele Huddleston, RN; Barry McKeown, MD; Ivor L. Gerber, MD; Elyse Foster, MD; Kanu Chatterjee, MB; Charles E. McCulloch, PhD; Andrew D. Michaels, MD

Marcus et al., Arch Intern Med. 2006;166:617-622

Study Subjects

- Adults scheduled for non-emergency left-sided heart catheterization at UCSF 8/03 to 6/04
- N=100

Marcus et al., Arch Intern Med. 2006;166:617-622

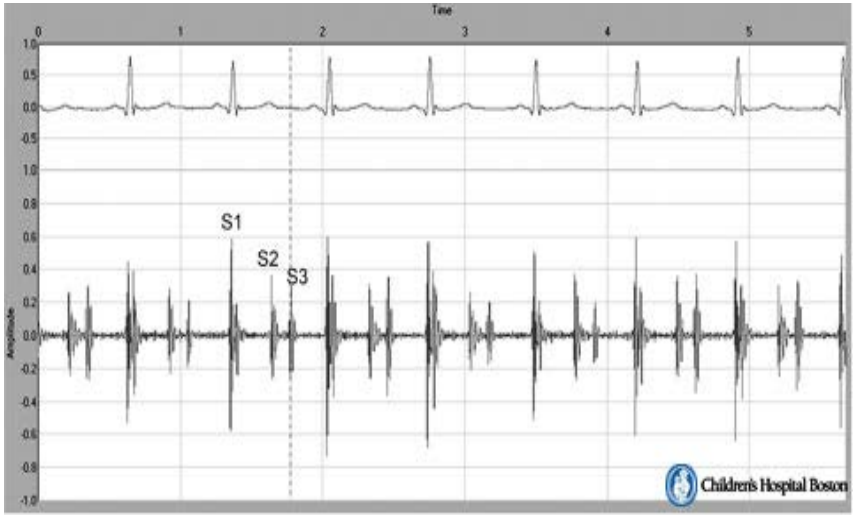
Examining Physicians

- Cardiology attendings (N=26)
- Cardiology fellows (N= 18)
- Internal medicine residents (N=54)
- Internal medicine interns (N=48)
- All from UCSF

•*Marcus et al., Arch Intern Med. 2006;166:617-622*

Measurements

- Auscultation
  - Standard procedure in quiet room
  - Examiners blinded to other information
- Phonocardiogram with computerized analysis to determine S_3



Analysis: Kappa

- Measures agreement beyond that expected by chance
- For *ordinal* variables use *weighted* kappa, which gives credit for coming close

Kappa	Agreement
0-0.2	Poor
0.2-.04	Fair
0.4-0.6	Moderate
0.6-0.8	Good
0.8-0.9	Very Good
0.9-1	Excellent

Results: Comparison of Auscultation with Phonocardiogram

	Kappa	P
Attendings	0.29	0.003
Fellows	0.37	<.001
Residents	0.13	0.11
Interns	0.04	0.36

Marcus, G. et al. Arch Intern Med 2006;166:617-622.

ARCHIVES OF INTERNAL MEDICINE

Issues: 1. Generalizability

- Were patients representative of those in whom an S_3 is relevant?
- Were study participants (MDs) representative of those who listen for S_3?
  - UCSF representative?
  - How many of the attending examinations were done by Kanu Chatterjee?

Issues: 2. Does test provide *new* information?

- Blinding observers to rest of H & P not sufficient
  - Information could still be redundant
- Options
  - Compare accuracy of prediction of (gold standard) left ventricular end-diastolic pressure (LVEDP) with and without examination for S_3
  - Record all clinical information and use multivariate techniques

Issues 3: Value of Information

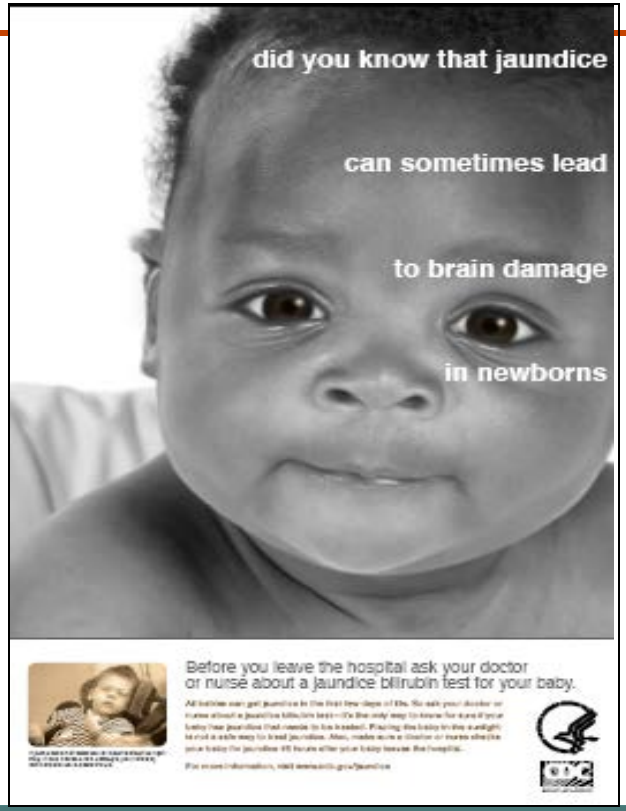
- What decision is the test supposed to help with?
- How often does the test change the decision?
- What is the effect of the change in decision on outcome?
- What is the *value* of that effect?

Example 2: Should every newborn have a bilirubin test before discharge?

- About 60% of newborns develop some jaundice
- Usually it is harmless, but can be dangerous if very high
- Proposal: check bilirubin level before discharge on all newborns
  - Recommended by some experts but insufficient evidence to recommend in guideline
  - Alternative: visual estimation

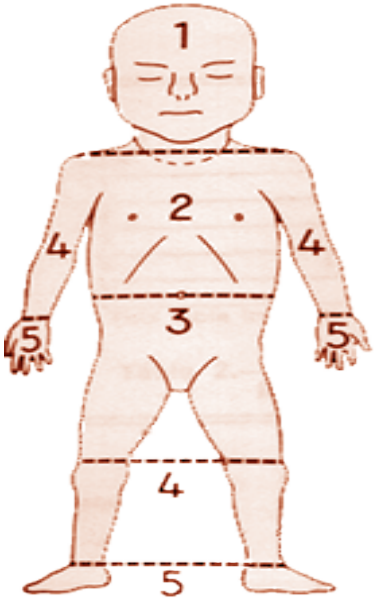
CDC Posters


did you know that jaundice
can sometimes lead
to brain damage
in newborns
Before you leave the hospital ask your doctor or nurse about a jaundice bilirubin test for your baby.


did you know that jaundice
can sometimes lead
to brain damage
in newborns
Before you leave the hospital ask your doctor or nurse about a jaundice bilirubin test for your baby.

Advancement of Dermal Icterus in the Jaundiced Newborn

Kramer LI, AJDC 1969;118:454



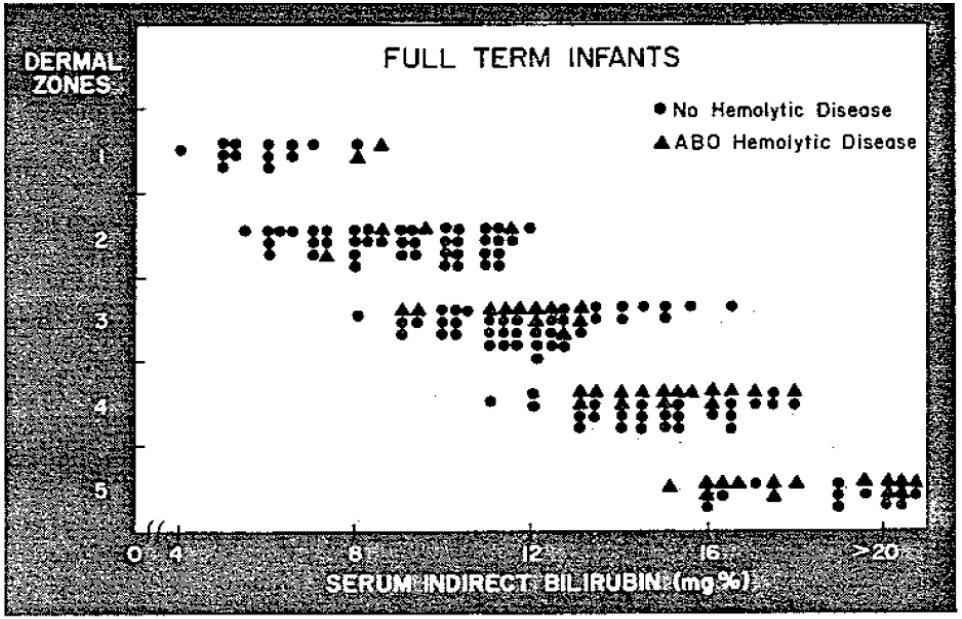


Fig 2.—Relationship of dermal icterus to serum bilirubin in full term infants.

Accuracy of Clinical Judgment in Neonatal Jaundice*

- RQ: How well can clinicians estimate bilirubin levels in jaundiced newborns?
- Study Design: cross-sectional study
- Subjects: 122 healthy term newborns (mean age 2 days) whose total serum bilirubin (TSB) was measured in the course of standard newborn care

*Moyer et al., Archives Peds Adol Med 2000; 154:391

Accuracy of Clinical Judgment in Neonatal Jaundice*

- Measurements:
  - Jaundice assessed by attendings, nurse practitioners and pediatric residents (absent/slight/obvious) at each body part and Total Serum Bilirubin (TSB) estimated
  - TSB levels measured in clinical laboratory
- Analysis
  - Agreement for jaundice at each body part by Weighted Kappa
  - Sensitivity and specificity for TSB ≥ 12 mg/dL

*Moyer et al., Archives Peds Adol Med 2000; 154:391

Results: 1.

Agreement Between Observers for Jaundice in Infants

Site	Weighted κ	95% Confidence Interval
Face and neck	0.16	−0.02 to 0.34
Neck to nipple line	0.15	0.01 to 0.29
Nipple line to umbilicus	0.23	0.09 to 0.38
Umbilicus to groin	0.19	0.05 to 0.34
Upper legs	0.20	0.06 to 0.35
Lower legs	0.16	−0.01 to 0.34
Soles	−0.06	−0.20 to 0.08
Arms	0.02	−0.14 to 0.18
Palms	0.05	−0.16 to 0.25
Conjunctiva	0.10	−0.04 to 0.23
Tip of nose	0.08	−0.07 to 0.22
Palate	0.11	−0.03 to 0.25

Moyer et al., Arch Peds Adol Med 2000; 154:391

Results: 2

- Sensitivity of jaundice below the nipple line for TSB ≥ 12 mg/dL = 97%
- Specificity = 19%

Editor's Note: The take-home message for me is that no jaundice below the nipple line equals no bilirubin test, unless there's some other indication.

--Catherine D. DeAngelis, MD

Moyer et al., APAM 2000; 154:391

Issues: 1

- No information on the numbers of different types of examiners or their years of experience
  - Generalizability uncertain
- No 95% CI around sensitivity and specificity
  - Sensitivity based upon 67/69
  - 95% CI: 90% to 99.6%

Issues: 2

- Verification bias
  - Infants NOT jaundiced below the nipples not likely to have a TSB measured
  - Sensitivity (97%) too high, specificity (19%) too low

	TSB >= 12	TSB <12
Jaundice below nipples	a	b
No jaundice below nipples	c ↓	d ↓

Extreme verification bias: Does This Child Have Appendicitis?

JAMA. 2007;298:438-451

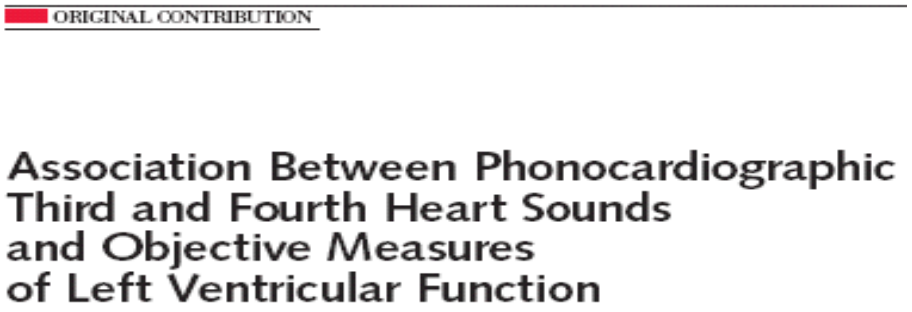
Table 2. Accuracy of Symptoms

Symptoms by Source	Sensitivity	Specificity	Positive Likelihood Ratio (95% Confidence Interval)	Negative Likelihood Ratio (95% Confidence Interval)
Pain Symptoms				
Symptom duration <24 h				
O'Shea et al,[35] 1988	0.50	0.40	0.83 (0.55-1.2)	1.3 (0.82-1.9)
Lintula et al,[48] 2005	0.44	0.46	0.82 (0.56-1.2)	1.2 (0.85-1.7)
Right lower quadrant pain				
Summary, level 3 studies			1.2 (1.0-1.5)	0.56 (0.43-0.73)
Pearl et al,[20] 1995	0.96	0.05	1.0 (0.98-1.0)	0.73 (0.35-1.5)
Kharbanda et al,[37] 2005	0.77	0.44	1.4 (1.2-1.6)	0.52 (0.38-0.72)
Mollitt et al,[58] 1988	0.62	0.63	1.7 (0.97-3.0)	0.59 (0.33-1.1)

- RLQ Pain: Sensitivity = 96% Specificity = 5%
- RLQ pain was present in 96% of those with appendicitis and 95% of those without appendicitis.
- Likelihood Ratio =1.0

Do S_3 and S_4 matter?

- RQ: How well do S_3 and S_4 predict abnormal (≥15 mm Hg) LVEDP?
- Design: cross-sectional study



JAMA. 2005;293:2238-2244

Study Subjects

- Adults scheduled for non-emergency left-sided heart catheterization at UCSF 8/03 to 6/04
  - Excluded if poor phonocardiographic quality (N=8) or paced rhythm (N=2)

Measurements

- Test: S_3 (Y/N) and S_3 “confidence score” from computer analysis of phonocardiogram
- “Gold Standard”: Left ventricular end-diastolic pressure ≥ 15 mm/Hg at cath

Results: S_3 present/absent

	LVEDP ≥15	LVEDP < 15	Total
S3 present	17	4	21
No S3	24	45	69
Total	41	49	90

Sensitivity = 17/41 = 41%;95% CI: (26%, 58%)

Specificity = 45/49 = 92%;95% CI (80%, 98%)

Positive PV = 17/21= 81%

Negative PV = 45/69 = 65%

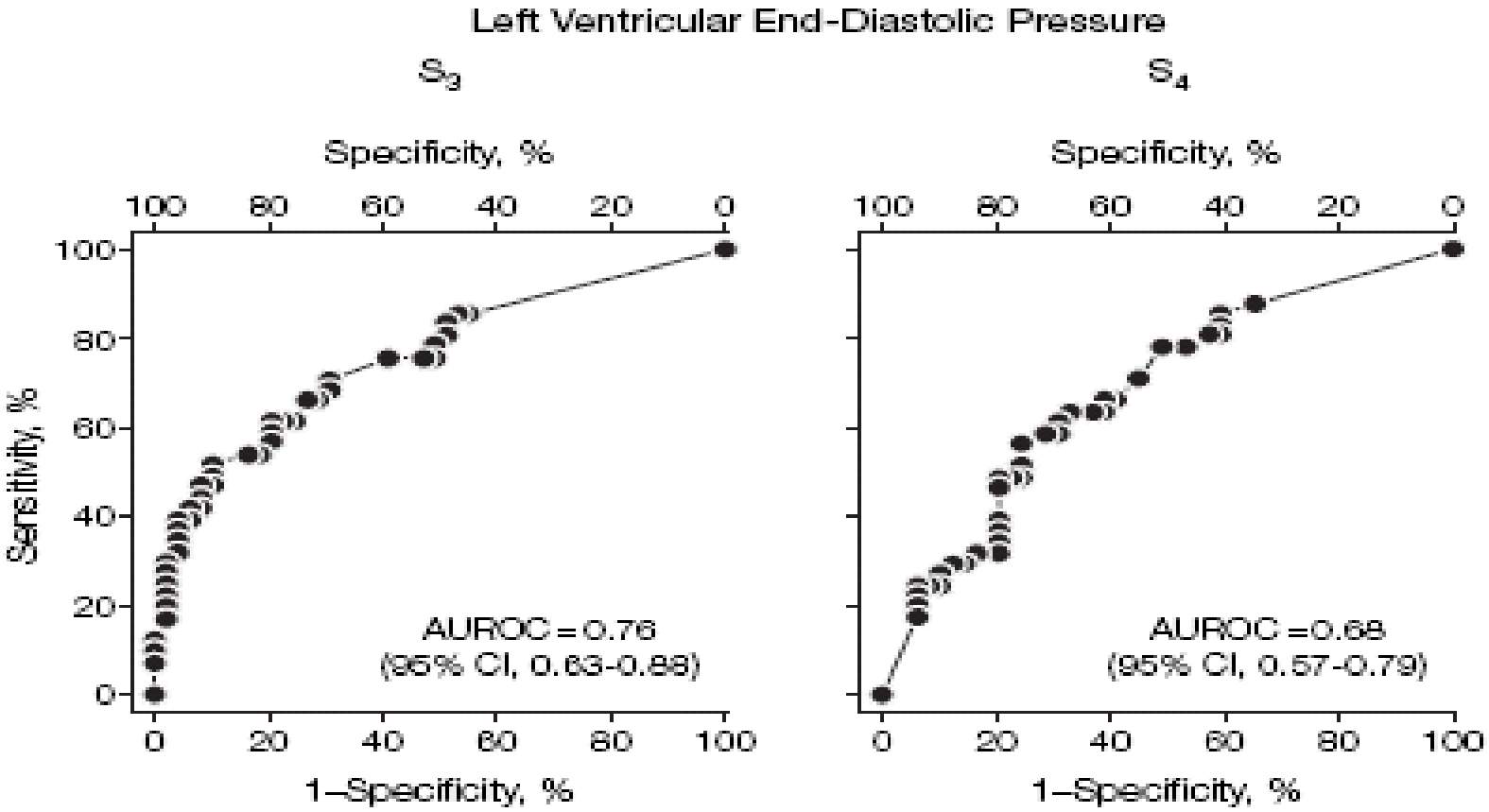
Results: “Confidence Scores”

- Many “dichotomous” tests not really dichotomous, e.g.:
  - Definite
  - Probable
  - Possible
  - Absent
- Phonocardiogram software generates “confidence scores” for S_3 and S_4

Analysis: ROC Curve

- ROC = “Receiver Operating Characteristics”
- Illustrate tradeoff between sensitivity and specificity as the cutoff is changed
- Discrimination of test measured by area under the curve (AUROC = c)
  - Perfect test 1.0
  - Worthless test 0.5

Results: S_3 & S_4 Confidence Scores



TIP

- If you are doing a study of test accuracy, Google STARD Checklist
- STARD= Standards for Reporting of Diagnostic Accuracy
- (Like CONSORT for clinical trials)

Summary: Think about

- The *question* you are trying to answer and why.
- *Sampling* of subjects, and maybe of people doing or interpreting the test
- Measurements – optimal or “real life”?
- Analysis – Kappa, Weighted Kappa, Sensitivity, Specificity, Likelihood Ratios, ROC curves, with confidence intervals
- Acknowledge limitations, think about the effect they would have on results

Good Luck!