Some challenges in the evaluation of medical tests

Jon Deeks

Professor of Biostatistics, University of Birmingham

Director of the Birmingham Clinical Trials Unit

Topics to cover

1. *Plan out portfolios of studies*
2. *First assess whether a test could ever work*
3. *Patient pathways and roles matter*
4. *Compared to what?*
5. *Its not all about accuracy*

1) Plan portfolios of studies

Different studies answer different questions

"If I repeat this test 10 times, how similar will all the results be?"

"How well does this test differentiate between people with and without this condition?"

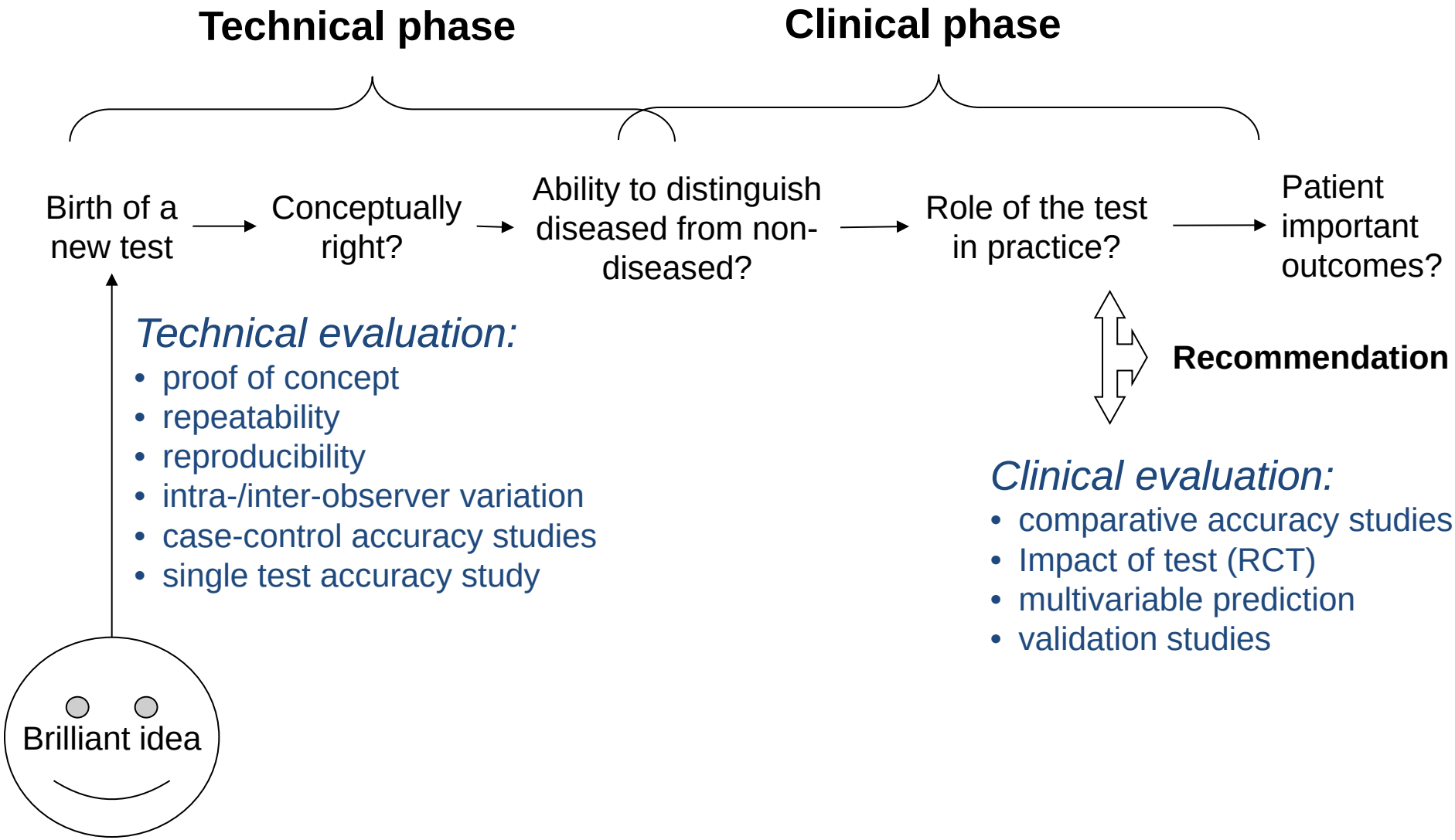
"Is the new test from company X better at ruling out patients without this condition than the one we use from company Y?"

"On top of all the clinical information that I already have, will this test add anything?"

"If I combine all the information I have, what is the chance that this patient has the condition?"

"If I introduce this new test in my patient pathways, will my patients really benefit?"

The career of a medical test



2) First assess whether a test could ever work

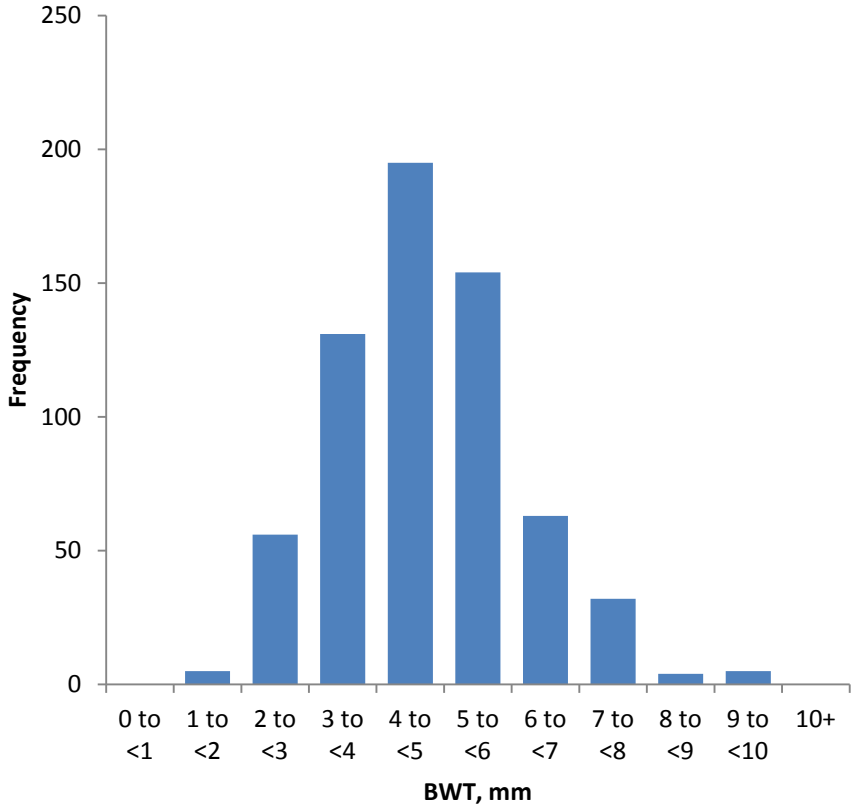
there is error in the measurements or classifications made by tests

often studies are restricted to assessing error in components of the test process

need to assess how repeatable the complete test process is and compare the magnitude of error with differences between groups

Example – US measurement of bladder wall thickness

Sub-study	Individual variability SD (mm)	Analytical variability SD (mm)	Smallest real difference (mm)
Inter-observer measures of same scans	1.23	0.35	0.97
Inter-observer measures from repeated scans	0.95	0.76	2.11



3) Patient pathways and test roles matter

Need to be clear when and how a test will be used, and ensure the studies in the review do likewise

Applicability has a higher profile in test accuracy reviews

Pulse oximetry screening for congenital heart defects in newborn infants (PulseOx): a test accuracy study

Andrew K Ewer, Lee J Middleton, Alexandra T Furmston, Abhay Bhoyar, Jane P Daniels, Shakila Thangaratinam, Jonathan J Deeks, Khalid S Khan, on behalf of the PulseOx Study Group

Lancet 2011; 378: 785–94
Published Online August 5, 2011
DOI:10.1016/S0140-6736(11)60753-8

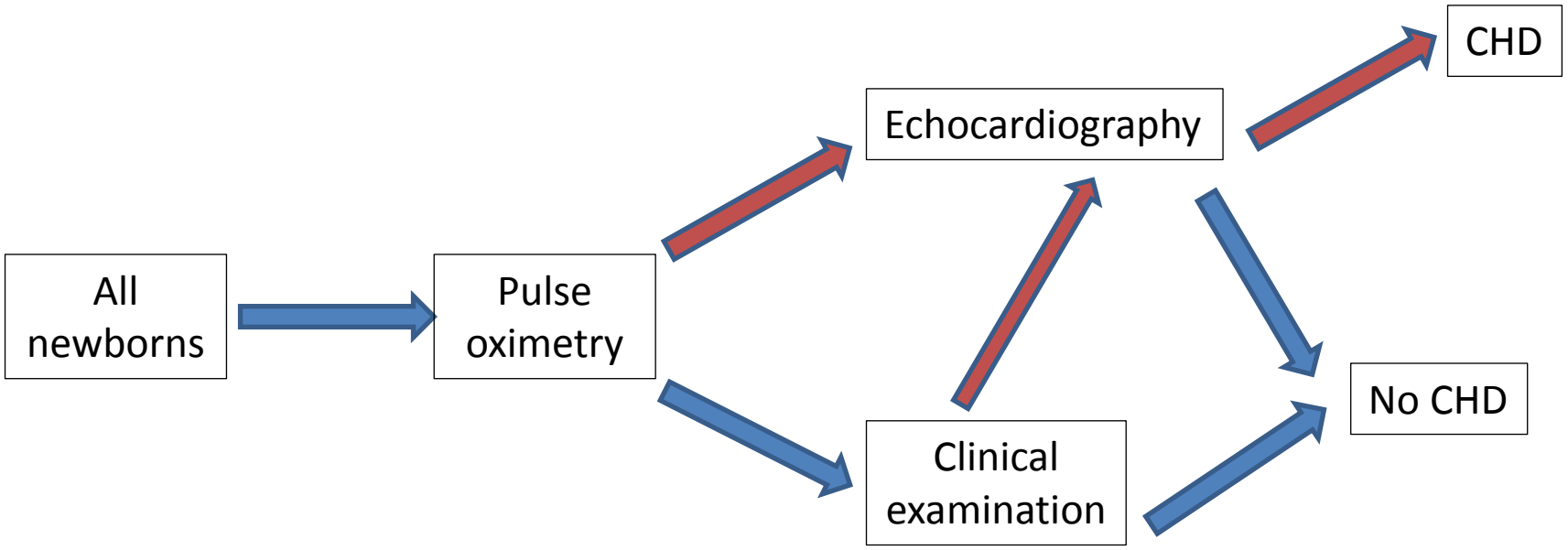
West Midlands study of 20055 newborns

	Critical CHD	No critical CHD
Low O2	18	177
Normal O2	6	19854
	24	19860

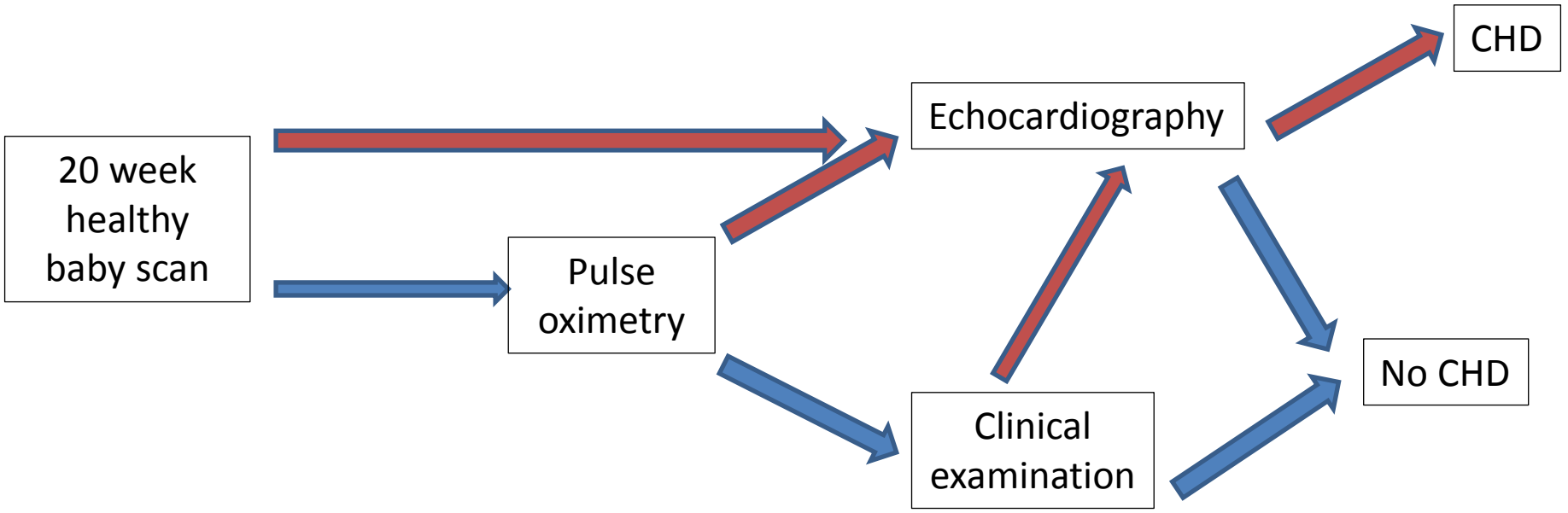
Sensitivity = 18 / 24 = 75%

Specificity = 19854/19860 = 99.97%

What is the diagnostic pathway?



What is the diagnostic pathway?



Results excluding 23 diagnosed on antenatal scan

	Critical CHD	No critical CHD
Low O2	7	170
Normal O2	5	19850
	12	19855

12 of the 24 were detected on the 20 week scan

Sensitivity = 7 / 12 = 58%

Specificity = 19850 / 19855 = 99.97%

Note: accounting for prior testing reduces sensitivity from 75% to 58%

4) Compared to what?

How accurate does a test need to be to be useful?

Is it better than the alternatives and current practice?

Much diagnostic research is 'comparison free'

Accuracy of High-Resolution Magnetic Resonance Imaging in Preoperative Staging of Rectal Cancer

Takayuki Akasu, MD[1], Gen Iinuma, MD[2], Masashi Takawa, MD[1], Seiichiro Yamamoto, MD[1], Yukio Muramatsu, MD[2], and Noriyuki Moriyama, MD[2]

ABSTRACT

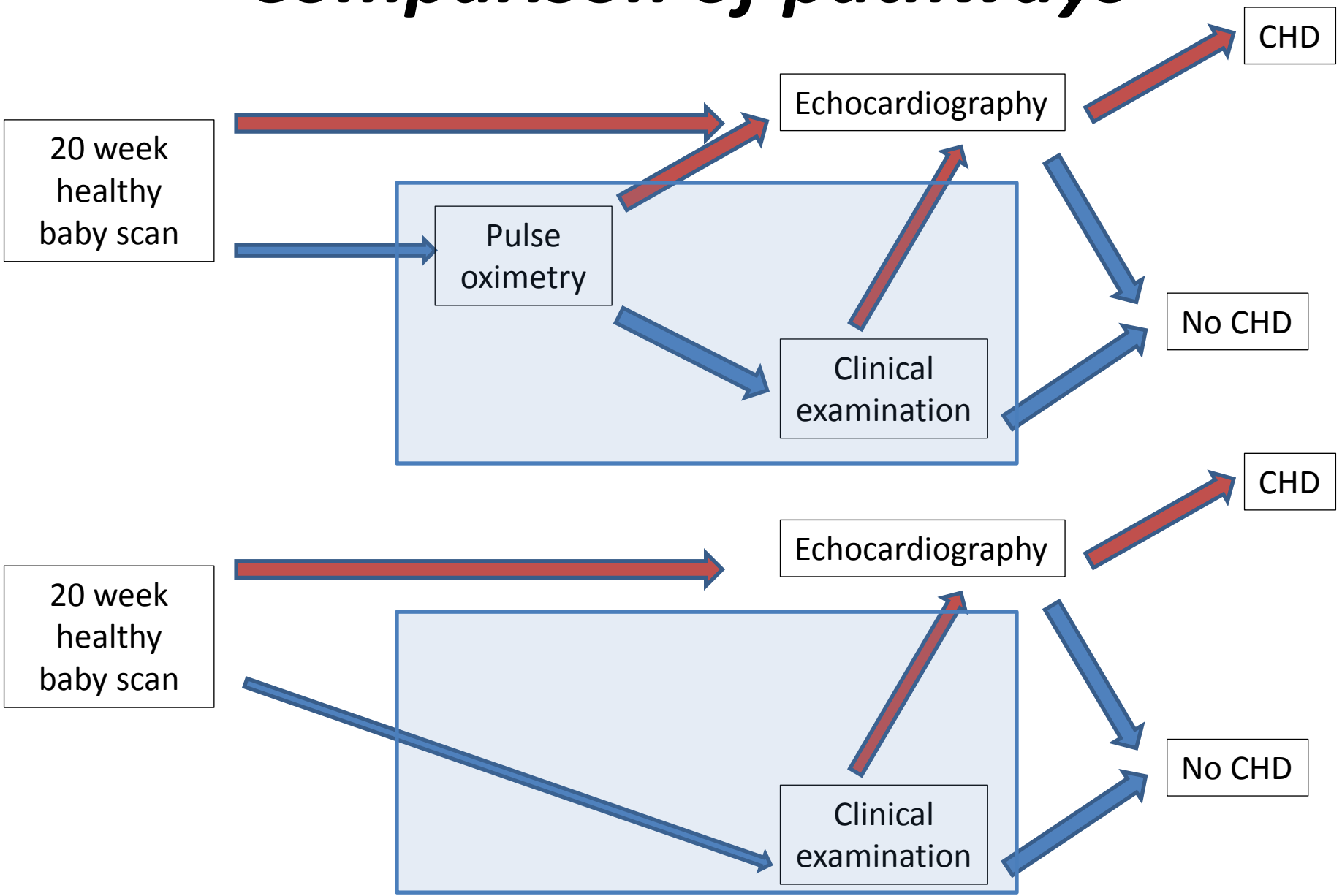
Background. To achieve better prognosis and quality of life for patients with rectal cancer, extent of surgery and neoadjuvant chemoradiotherapy should accurately reflect disease extent. The aim of this study was to evaluate accuracy of high-resolution magnetic resonance imaging (HRMRI) for preoperative staging of rectal cancer.

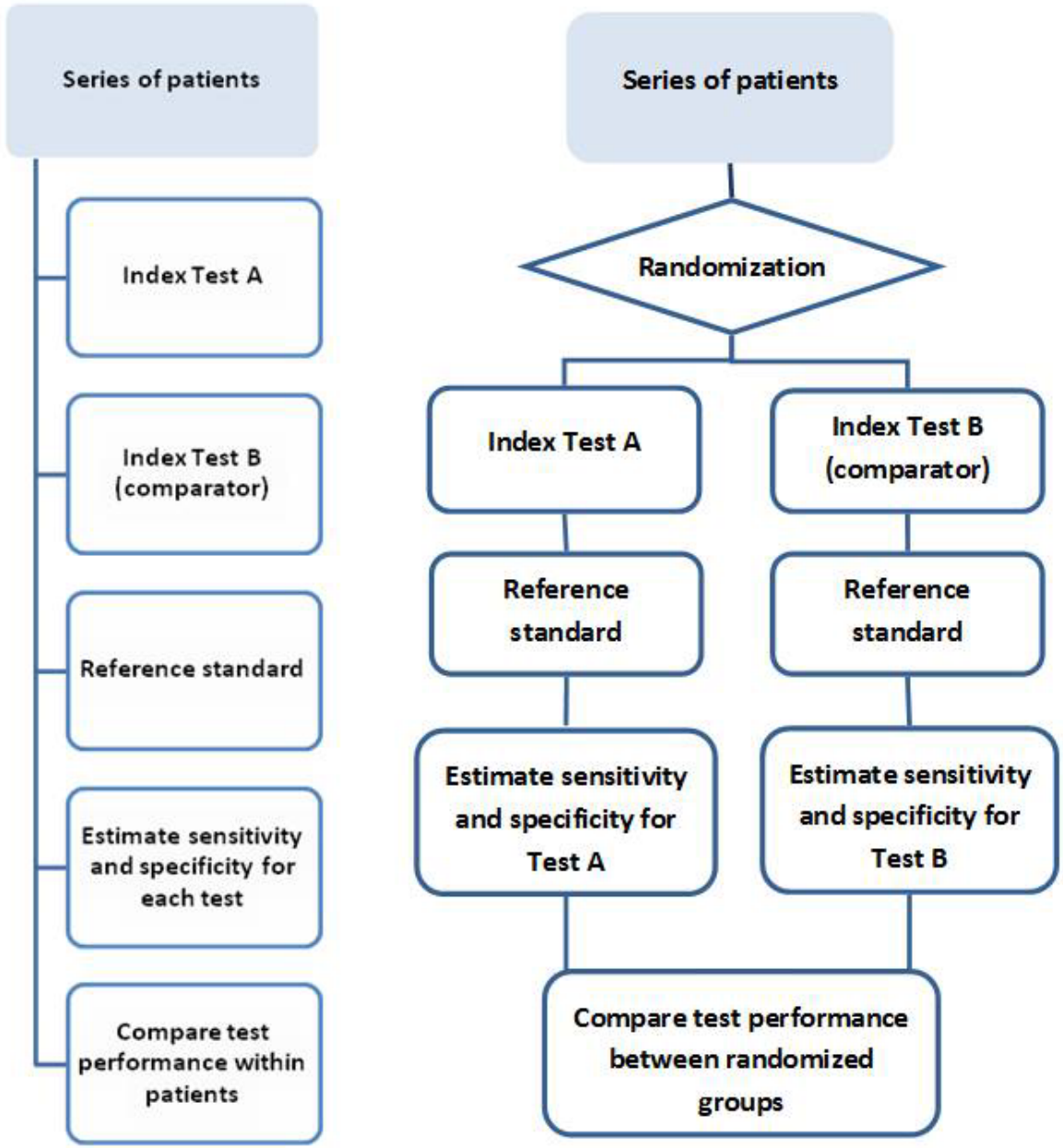
Methods. Between 2001 and 2003, 104 patients with primary rectal cancer were examined with HRMRI and underwent radical surgery. Transmural invasion depth and lymph node metastasis were assessed prospectively and classified according to the American Joint Committee on Cancer (AJCC) tumor–node–metastasis (TNM) system by both HRMRI and histopathology, and results were compared. Criteria for mesorectal and lateral pelvic lymph node involvement were short-axis diameters of ≥5 mm and ≥4 mm, respectively.

Results. There were 15 pT1, 25 pT2, 50 pT3, and 14 pT4 tumors. Overall accuracy rate for transmural invasion depth was 84%. The mesorectal fascia could be visualized in 98% of patients. Twenty-three patients had mesorectal fascia involvement and the overall accuracy rate was 96% (sensitivity, 96%; specificity, 96%). Fifty-three patients had mesorectal lymph node metastasis and the overall accuracy rate was 74% (sensitivity, 83%; specificity, 64%). Lateral pelvic lymph node metastasis was observed in 15 patients and the overall accuracy rate was 87% (sensitivity, 87%; specificity, 87%).

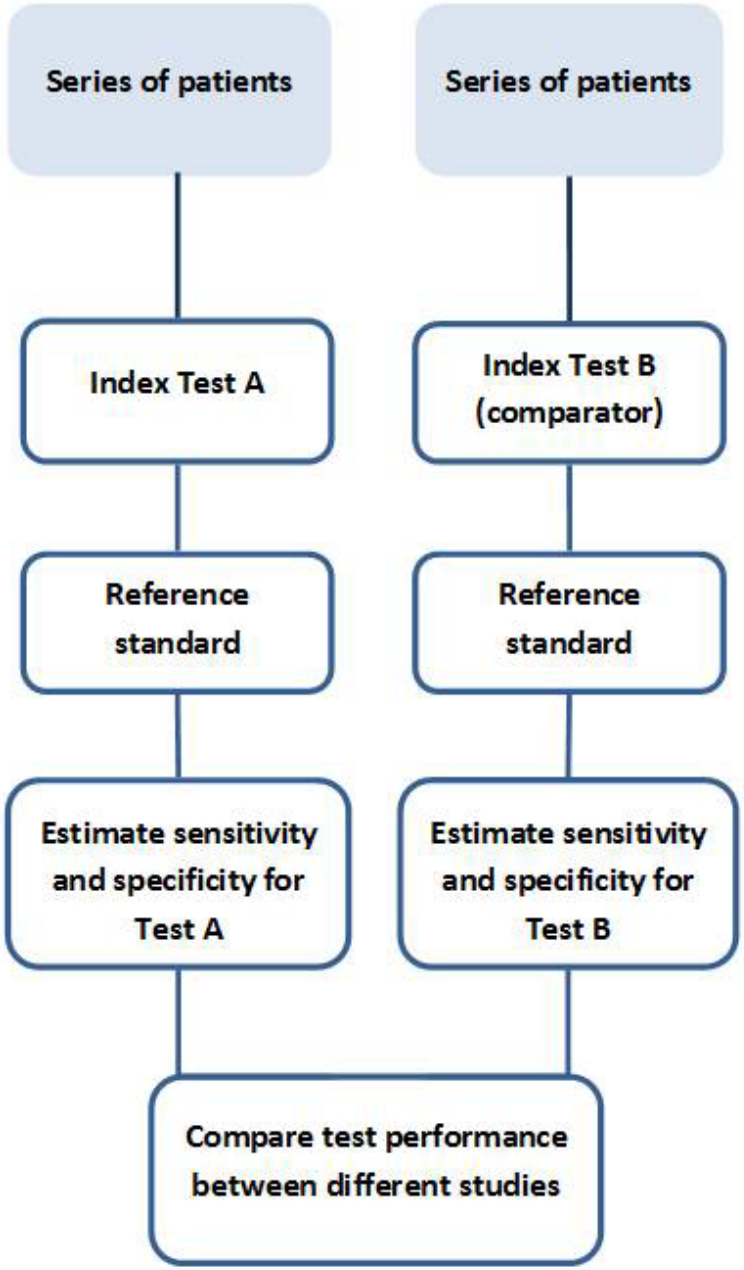
Conclusions. HRMRI was moderately accurate for prediction of mesorectal lymph node metastasis and highly accurate regarding transmural invasion depth, and mesorectal fascia and lateral pelvic node involvement. Therefore, HRMRI appears useful for preoperative decision-making in rectal cancer treatment.

Comparison of pathways





(a) Multiple test comparison (b) Randomized test comparison



(c) Between study test comparison

Annals of Internal Medicine | Research and Reporting Methods

Empirical Evidence of the Importance of Comparative Studies of Diagnostic Test Accuracy

Yemisi Takwoingi, DVM; Mariska M.G. Leeflang, PhD; and Jonathan J. Deeks, PhD

- **248 comparative reviews including 6915 studies**
  31% of the studies were comparative

- **Median [IQR] studies per review**
  6 [2, 11] comparative
  14 [4, 28] 'non-comparative'
  43 (17%) reviews found no comparative studies
  177 (71%) mixed studies
  Only 28 (11%) included only comparative studies

- **Differences between tests were found to be on average larger (P=0.001) when based upon non-comparative than comparative studies**

Are we guilty of sloppy science?

Like-with-like comparisons are known to be important for interventions ...

... why don't we insist on them for tests?

- Reasons why it is not done
  - no standard comparator can be identified
  - logistically difficult
  - the reference standard might be the comparator
  - we're not brave enough to try

- But often we should be doing better

5) Its not all about accuracy

Improved accuracy can lead to patient benefit if more patients receive effective treatment

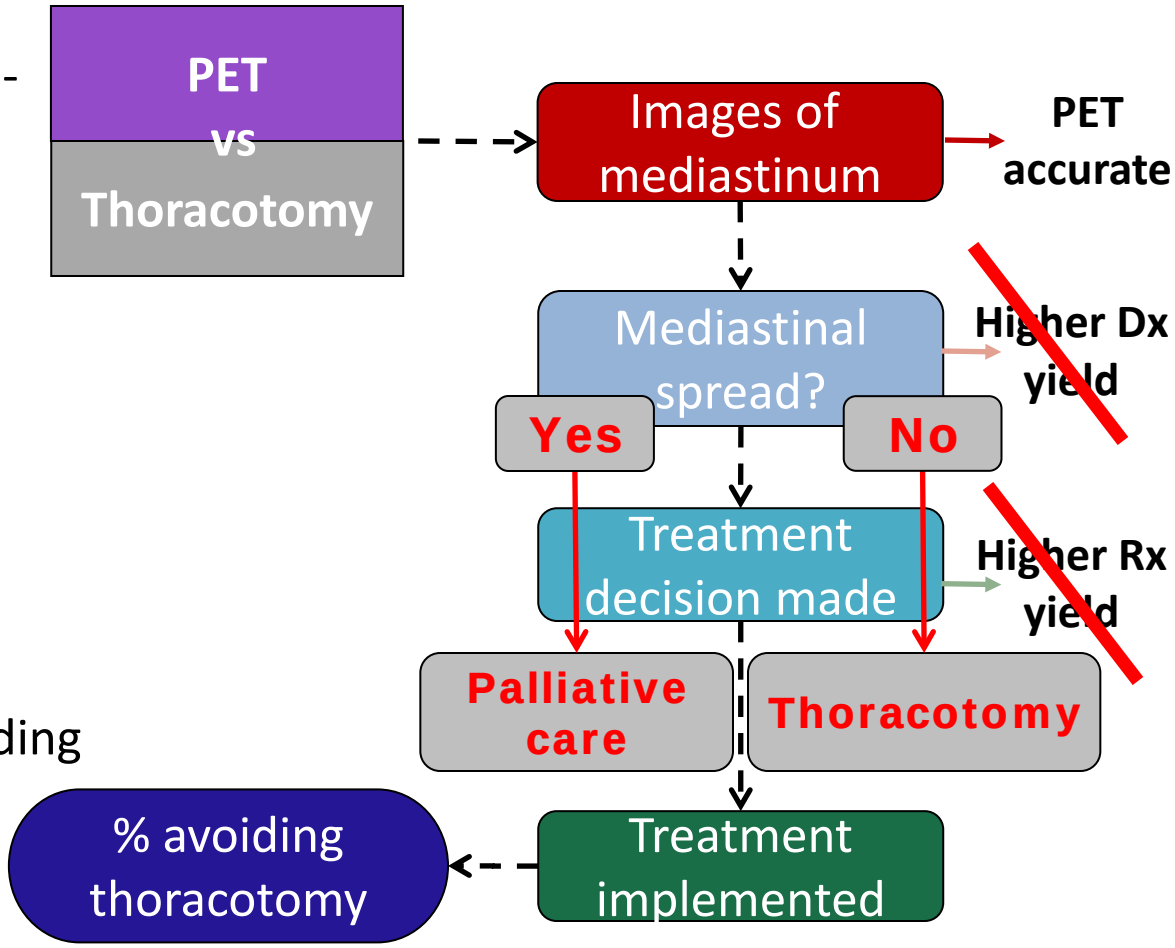
but tests can affect patients in other ways

more than accuracy needs to be considered

PET for staging NSCLC

Viney et al. J Clin Oncol 2004;22:2357-2362

- **Patients:**
  Histologically established non-small cell lung cancer
- **Target Condition:**
  Metastasised to mediastinum
- **Triage Comparison:**
  PET ± Thoracotomy
  vs.
  Thoracotomy
- **Findings:**
  No change to % patients avoiding thoracotomy:
  4% vs. 2% (p=0.2)

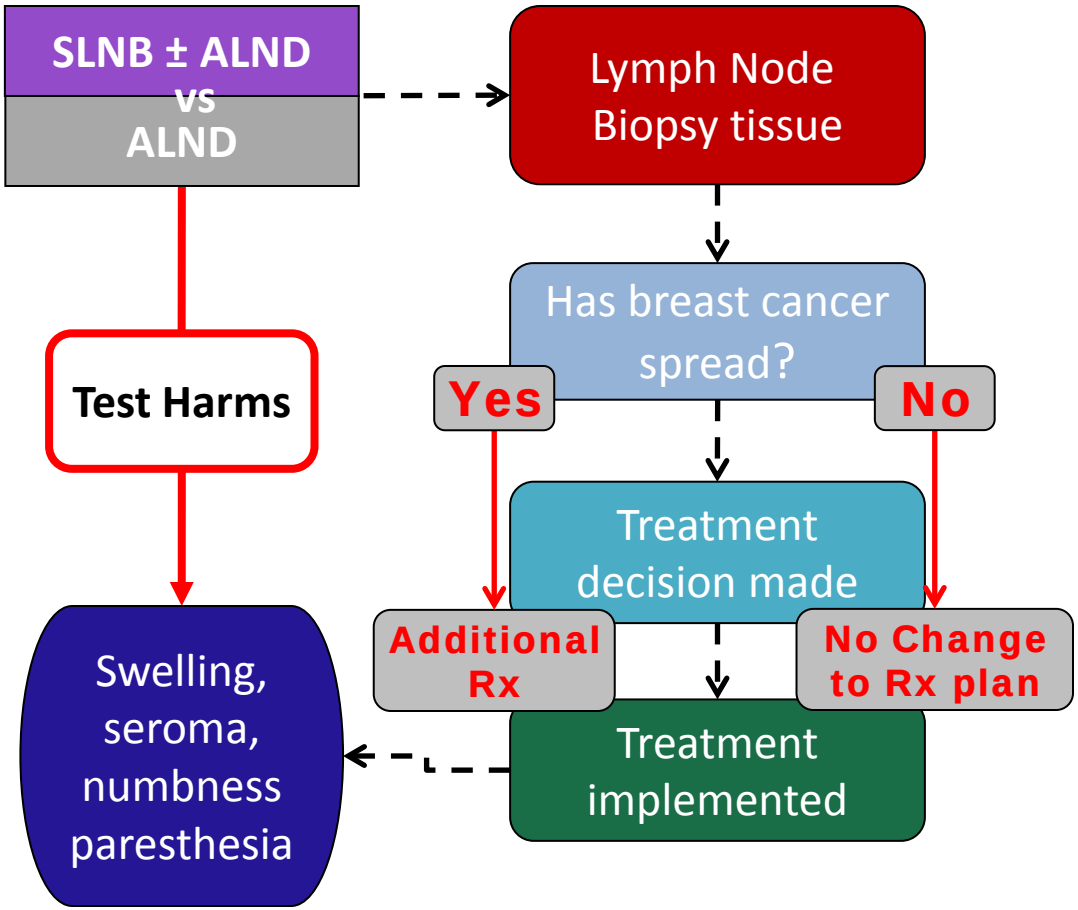


Enhanced accuracy failed to affect Dx yield as surgeons were unwilling to rely on PET results for definitive diagnosis.

Sentinal Lymph Node Biopsy for BC

- **Patients:**
  Early breast cancer
- **Target Condition:**
  Stage of cancer
- **Triage Comparison:**
  Sentinal lymph node biopsy ± Axillary Lymph Node Dissection
  vs.
  Axillary Lymph Node Dissection
- **Findings:**
  Significant reduction in % suffering from postoperative arm swelling, seroma formation, numbness and paresthesia

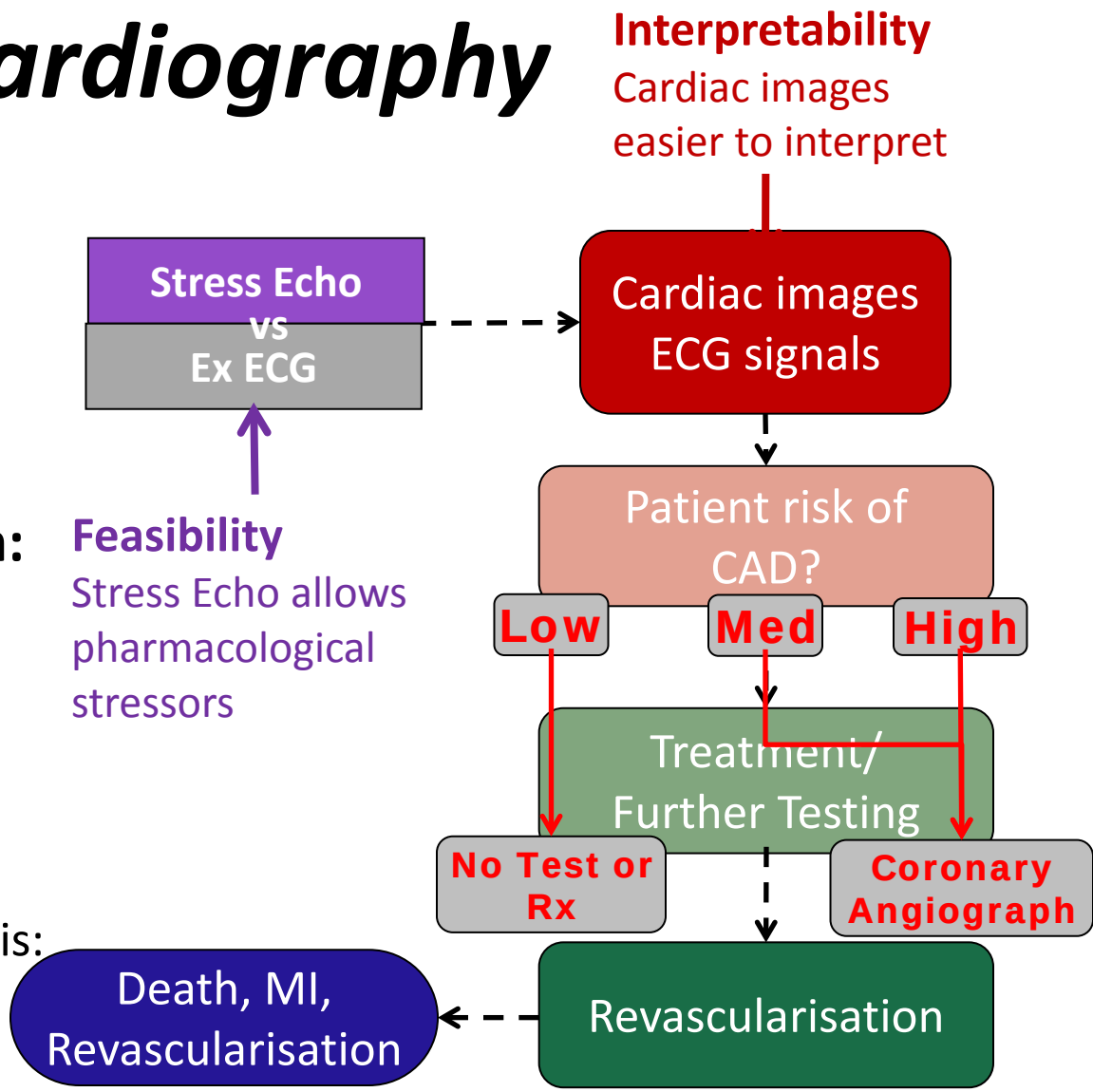
Purushotham et al. J Clin Oncol 2005;23:4312-4321



Less invasive triage test spares test-negative patients harms of more invasive test with small loss of accuracy

Stress echocardiography

- **Patients:**
  Acute chest pain
- **Target Condition:**
  Coronary Artery Disease
- **Replacement Comparison:**
  Stress Echocardiography
  vs.
  Exercise Electrocardiography
- **Findings:**
  SE doubles feasibility rate.
  Increases % conclusive diagnosis:
  3% vs. 47% (p<0.001)
  Decreases % further testing:
  52% vs. 16% (p<0.0001)



Jeetley et al. Eur J Echocard 2006;7:155-164

Lower technical failure rates and clearer test results increase definitive diagnoses and decrease referrals for further testing

RESEARCH METHODS & REPORTING

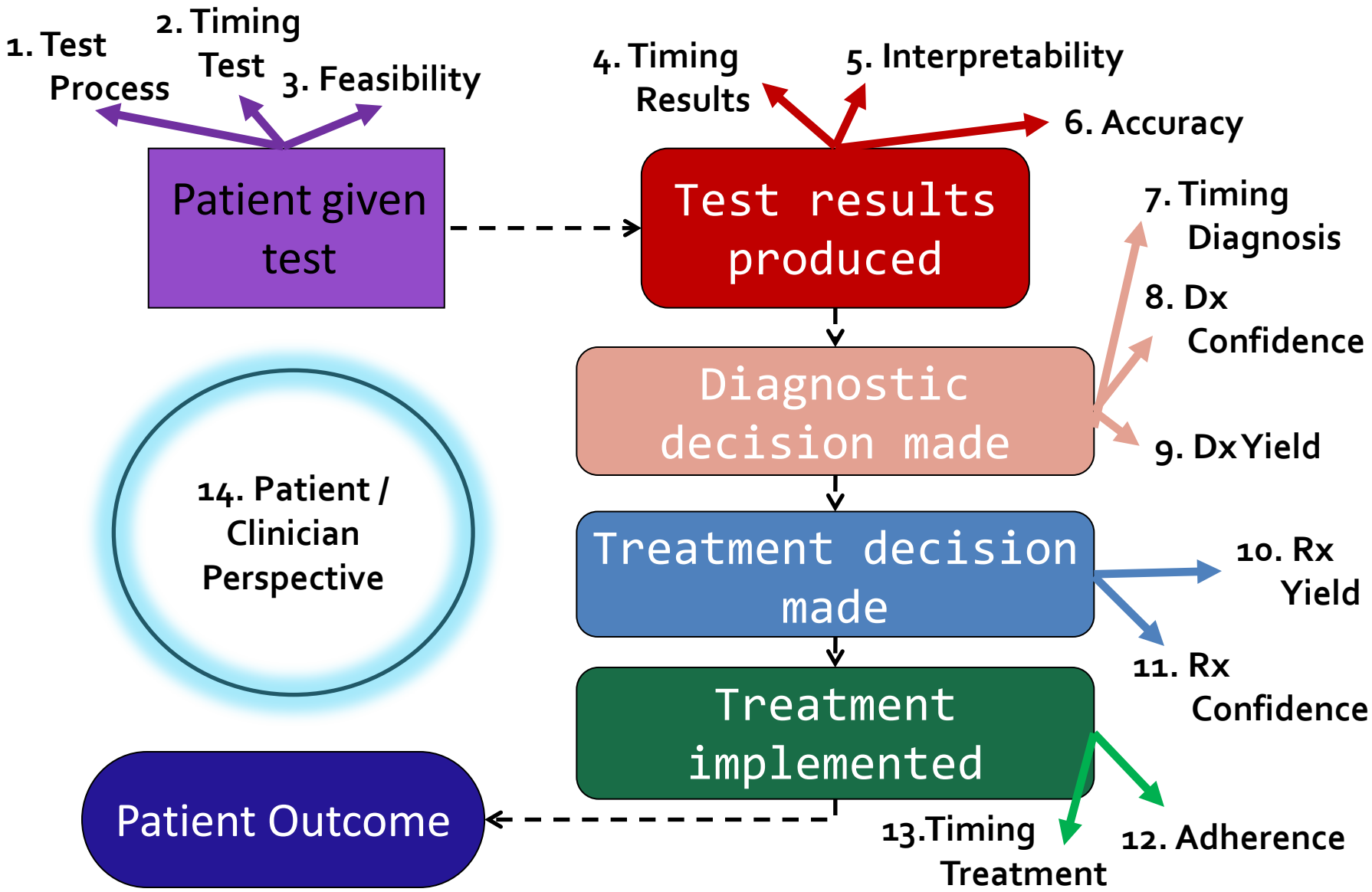
Assessing the value of diagnostic tests: a framework for designing and evaluating trials

The value of a diagnostic test is not simply measured by its accuracy, but depends on how it affects patient health. This article presents a framework for the design and interpretation of studies that evaluate the health consequences of new diagnostic tests

Lavinia Ferrante di Ruffano *research fellow*[1], Christopher J Hyde *professor of public health and clinical epidemiology*[2], Kirsten J McCaffery *associate professor and principal research fellow*[3], Patrick M M Bossuyt *professor of clinical epidemiology*[4], Jonathan J Deeks *professor of biostatistics*[1]

[1]Department of Public Health, Epidemiology, and Biostatistics, School of Health and Population Sciences, University of Birmingham, Birmingham B15 2TT, UK; [2]PenTAG, Institute for Health Services Research, Peninsula College of Medicine and Dentistry, University of Exeter, Exeter, UK; [3]Screening and Test Evaluation Program, School of Public Health, University of Sydney, Sydney, Australia; [4]Department of Clinical Epidemiology and Biostatistics, Academic Medical Centre, University of Amsterdam, Amsterdam, Netherlands

Outcomes Framework: 14 Mechanisms



Main domains to consider

- **Decision making**
  - Test accuracy, impact and confidence
- **Timing of tests**
  - Benefits of earlier treatment
- **Direct effect**
  - Feasibility, acceptability, direct harms
- **Patient and clinical factors and preferences**

Summary

- Diagnostic evidence needs to assess tests as they are **used in pathways**, **make comparisons** with alternatives and assess how test how tests **impact on patients**

- Carefully planning of **technical evaluation** studies to understand **total variability** can be very valuable

- Test-treatment **RCTs** are an obvious way of evaluating impact but often **don't deliver**, and face **many challenges – few are done**

Summary

- Alternative evaluative approaches need **portfolios** of evidence, making comparisons to address the likely **mechanisms of action**, potentially combined in a **decision analytical model**

- Studies of **diagnostic accuracy** are a ***necessary*** but not ***sufficient*** part of this

4th International Symposium

Methods for Evaluating Medical Tests and Biomarkers

University of Birmingham

19th-20th July 2016

Key themes likely include:

Methods for primary studies, systematic reviews and meta-analysis;
Evaluation of impact of tests on patients;
Industry, regulation and test development;
Translating findings into practice: guidelines and technology appraisals.