

Introduction to diagnostic test accuracy (DTA) reviews

Yemisi Takwoingi
October 2015

Learning objectives

- To understand the role of test accuracy
- To be familiar with the different study designs used to evaluate test accuracy
- Be able to define the components of a DTA review question

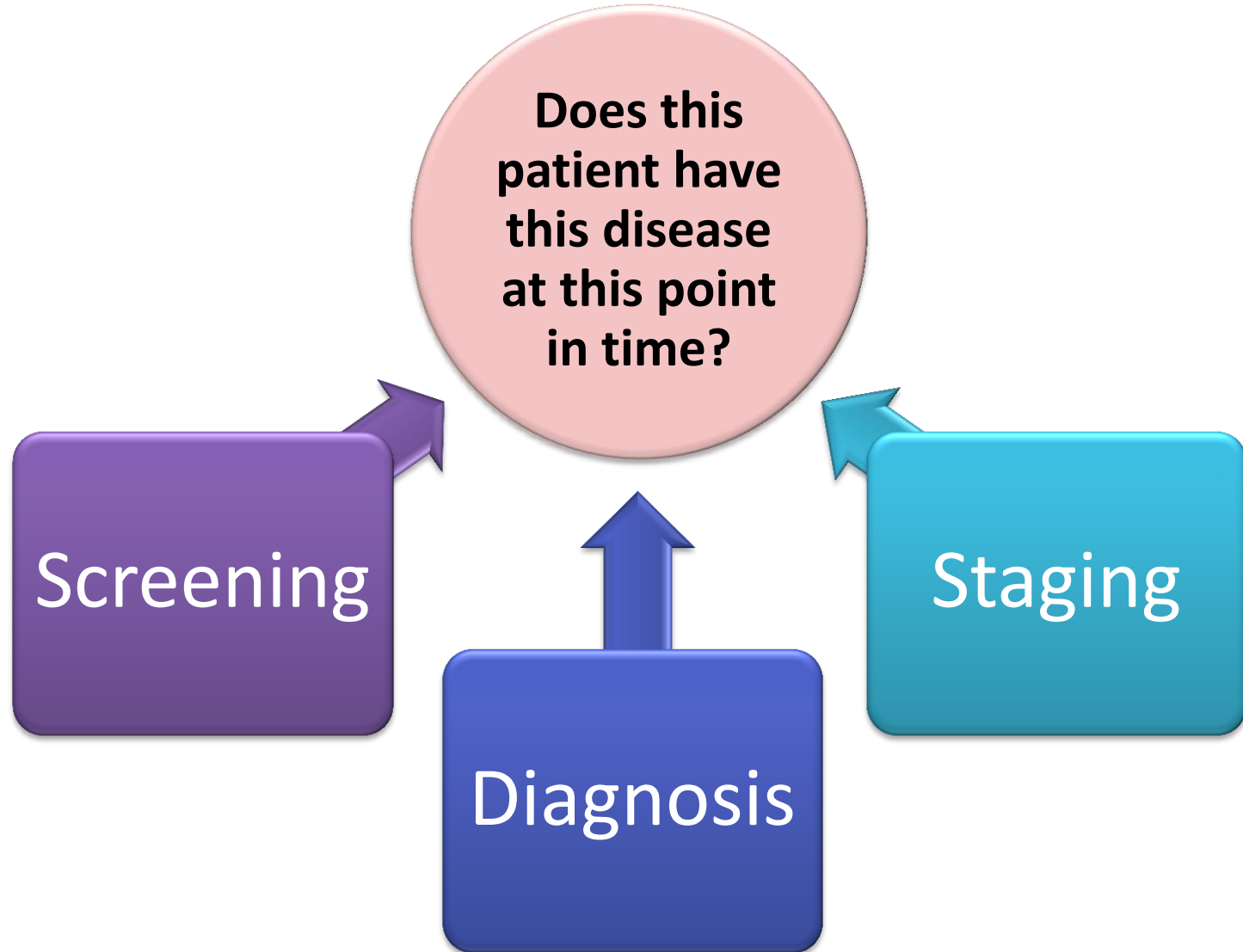
Outline

- Why test accuracy
- Study design
 - Single test accuracy study
 - Test comparison study
- Components of a DTA review question

What are tests used for?

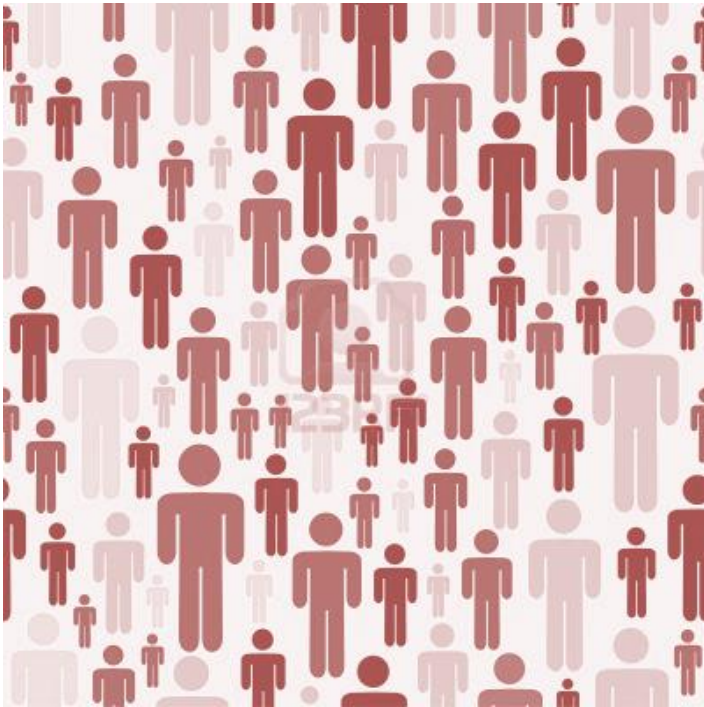
- **Predisposition** (who could develop the disease)
- **Screening** (who has asymptomatic disease)
- **Diagnosis** (who has symptomatic disease)
- **Staging** (how advanced is the disease)
- **Prognosis** (how progressive will the disease be)
- **Stratification** (who will be a responder)
- **Efficacy** (is the drug effective)
- **Monitoring** (is the disease controlled)
- **Recurrence** (relapse of disease)

What is diagnostic test accuracy?



Test accuracy

What proportion of those with the disease does the test correctly identify? (sensitivity)



What proportion of those without the disease does the test correctly exclude? (specificity)



Test accuracy - 2x2 table

	Reference standard positive	Reference standard negative
Index test positive	TP	FP
Index test negative	FN	TN

Measures of diagnostic accuracy

- Sensitivity TPR/TPF
- Specificity TNR/TNF
- 1-Specificity FPR/FPF
- Positive Predictive Value PPV
- Negative Predictive Value NPV
- Positive Likelihood Ratio LR+
- Negative Likelihood Ratio LR-
- Diagnostic Odds Ratio DOR

Sensitivity and specificity

	Reference standard positive	Reference standard negative	
Index test positive	TP	FP	TP + FP
Index test negative	FN	TN	FN + TN
	TP + FN	FP + TN	TP + FN + FP + TN

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{Specificity} = \text{TN} / (\text{TN} + \text{FP})$$

Compute sensitivity and specificity

	Reference standard positive	Reference standard negative	
Index test positive	5	10	15
Index test negative	5	990	995
	10	1000	1010

And write sentences explaining what they mean

Predictive values

	Reference standard positive	Reference standard negative	
Index test positive	TP	FP	TP + FP
Index test negative	FN	TN	FN + TN
	TP + FN	FP + TN	TP + FN + FP + TN

$$\text{PPV} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{NPV} = \text{TN} / (\text{TN} + \text{FN})$$

Likelihood ratios

	Reference standard positive	Reference standard negative
Index test positive	sensitivity	1-specificity
Index test negative	1-sensitivity	specificity
	1	1

$$\text{LR+} = \frac{\text{sensitivity}}{1 - \text{specificity}}$$

$$\text{LR-} = \frac{1 - \text{sensitivity}}{\text{specificity}}$$

Limitations of test accuracy?

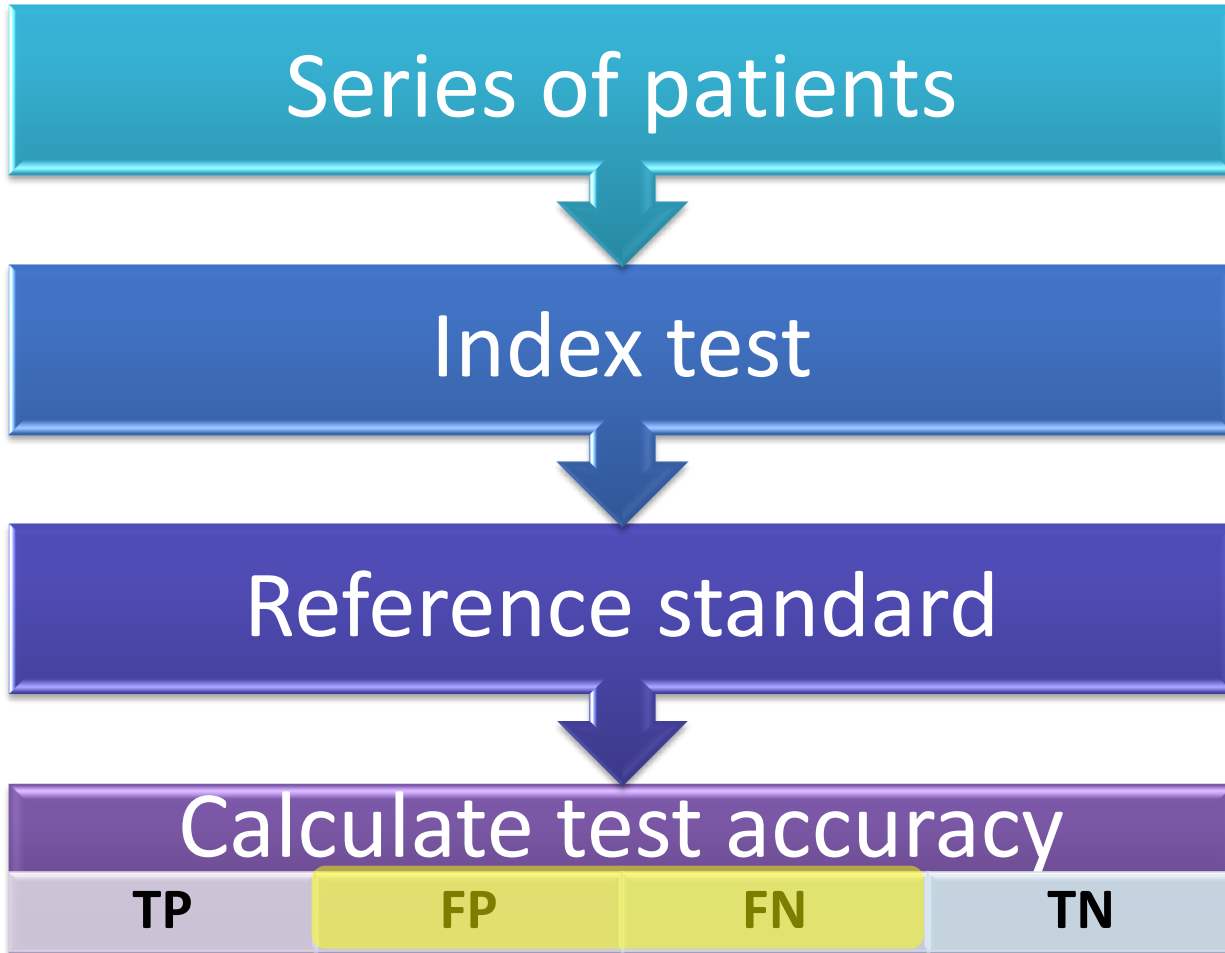
- “How well does the test identify the target condition?”
 - Does not directly assess effect of test on outcomes
 - Does not directly answer the question of whether using a test does more good than harm
 - Only possible when there is an adequate reference standard

But...

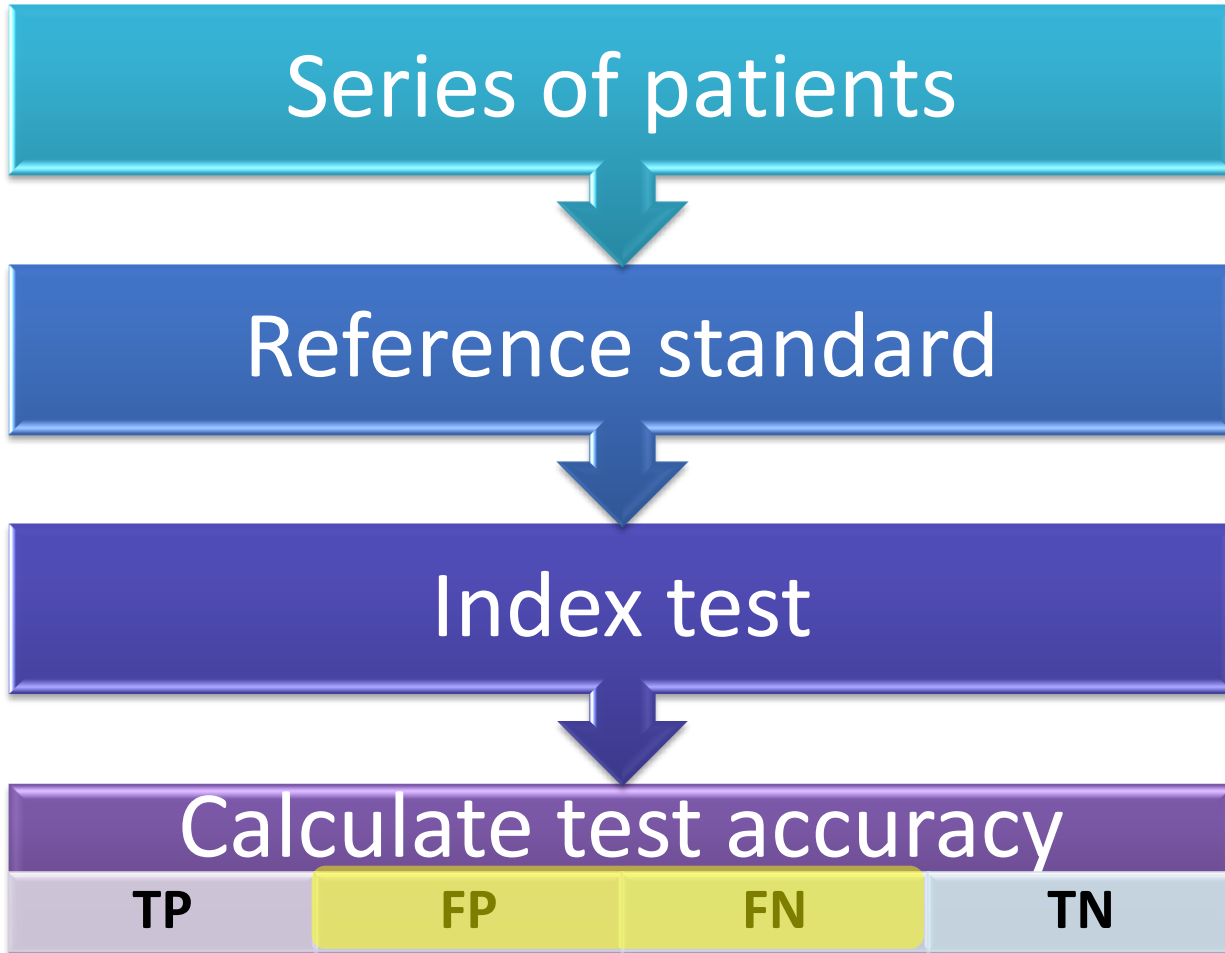
Randomized controlled trials (RCTs) of test-plus treatment strategies for evaluating the benefits of a new test relative to current best practice are not always feasible, available or necessary—sometimes evidence from accuracy studies may suffice.

- *Lord SJ, Irwig L, Simes RJ. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med 2006; 144(11):850-855.*
- *Lord SJ, Irwig L, Bossuyt PMM. Using the principles of randomized controlled trial design to guide test evaluation. Med Decis Making 2009; 29(5):E1-E12.*

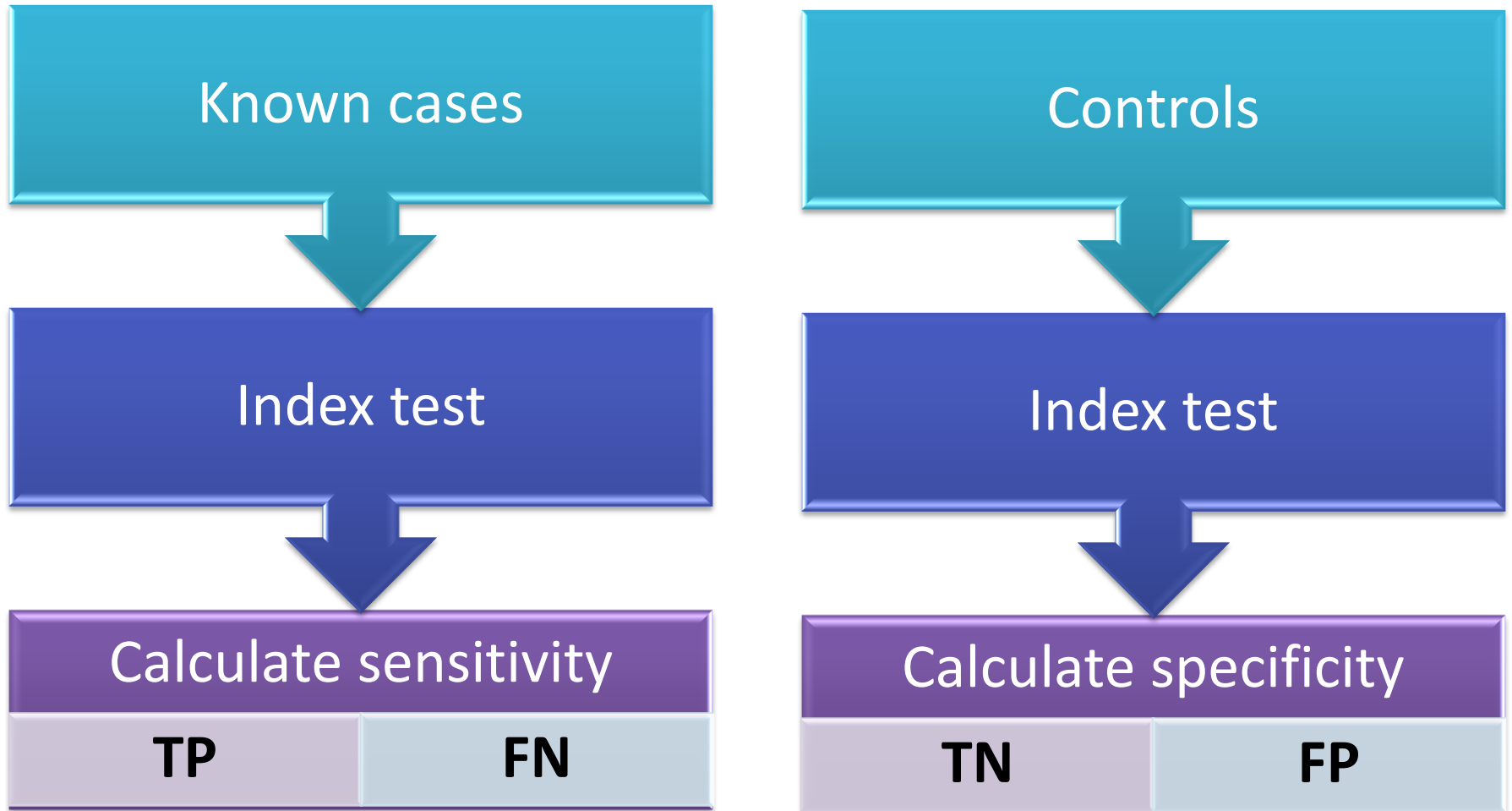
Basic design



Basic design



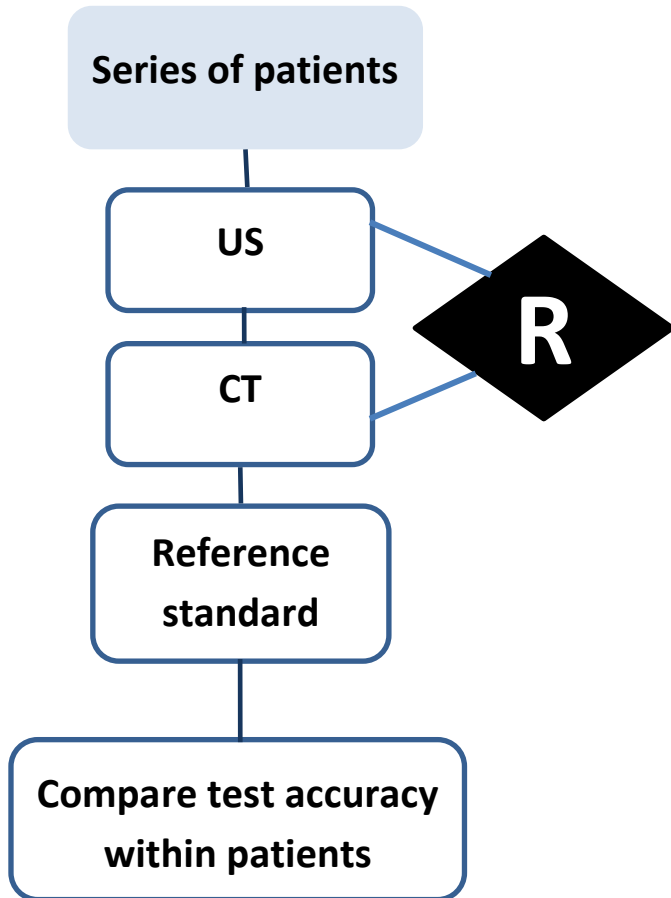
Case control design



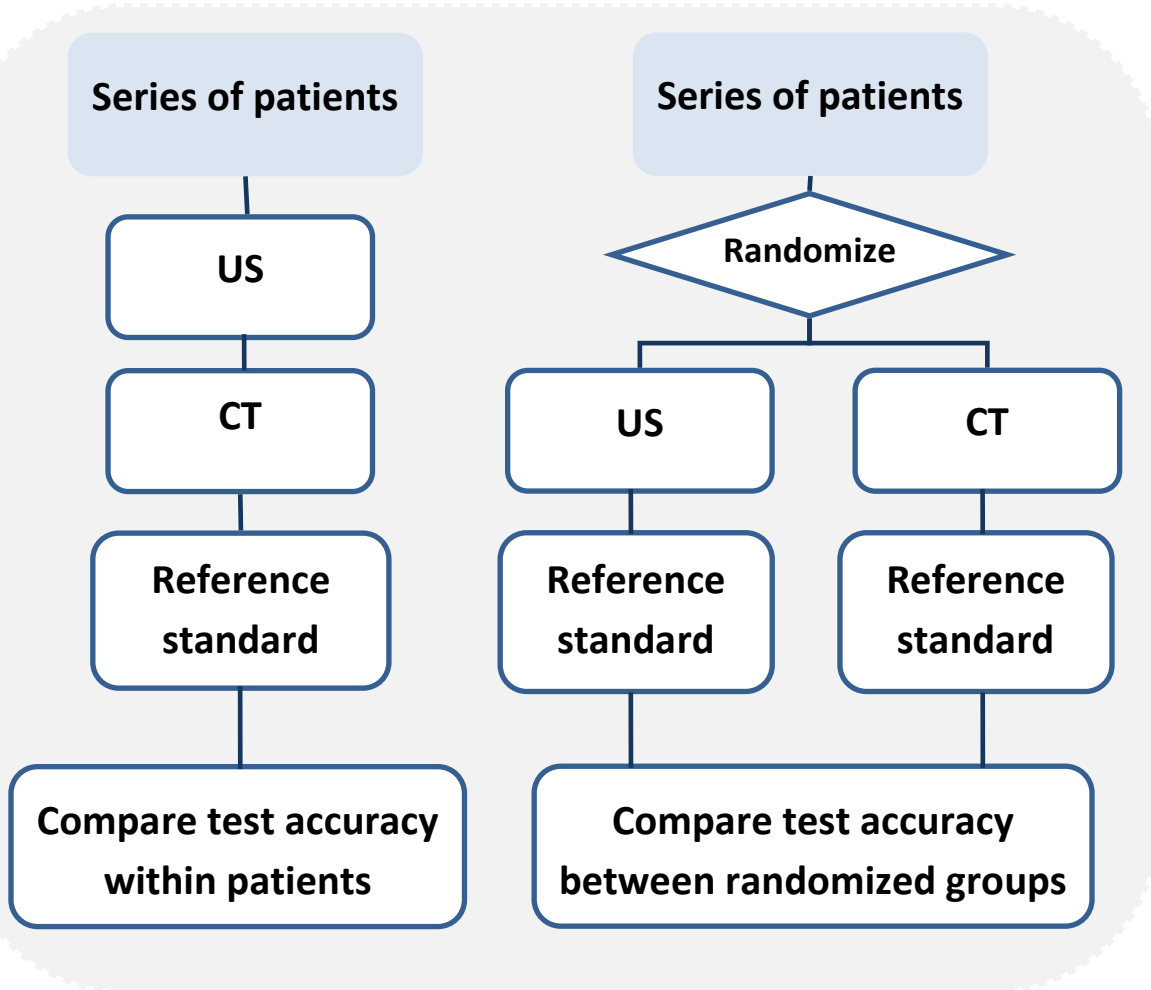
**Which test
is best?**



Test comparison designs



Test comparison designs



Robust comparative studies

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

Background: Systematic reviews that “compare” the accuracy of 2 or more tests often include different sets of studies for each test.

Purpose: To investigate the availability of direct comparative studies of test accuracy and to assess whether summary estimates of accuracy differ between meta-analyses of noncomparative and comparative studies.

Data Sources: Systematic reviews in any language from the Database of Abstracts of Reviews of Effects and the Cochrane Database of Systematic Reviews from 1994 to October 2012.

Study Selection: 1 of 2 assessors selected reviews that evaluated at least 2 tests and identified meta-analyses that included both non-comparative studies and comparative studies.

Data Extraction: 1 of 3 assessors extracted data about review and study characteristics and test performance.

Data Synthesis: 248 reviews compared test accuracy; of the 6915 studies, 2113 (31%) were comparative. Thirty-six reviews (with 52 meta-analyses) had adequate studies to compare results of non-comparative and comparative studies by using a hierarchical sum-

mary receiver-operating characteristic meta-regression model for each test comparison. In 10 meta-analyses, noncomparative studies ranked tests in the opposite order of comparative studies. A total of 25 meta-analyses showed more than a 2-fold discrepancy in the relative diagnostic odds ratio between noncomparative and comparative studies. Differences in accuracy estimates between non-comparative and comparative studies were greater than expected by chance ($P < 0.001$).

Limitation: A paucity of comparative studies limited exploration of direction in bias.

Conclusion: Evidence derived from noncomparative studies often differs from that derived from comparative studies. Robustly designed studies in which all patients receive all tests or are randomly assigned to receive one or other of the tests should be more routinely undertaken and are preferred for evidence to guide test selection.

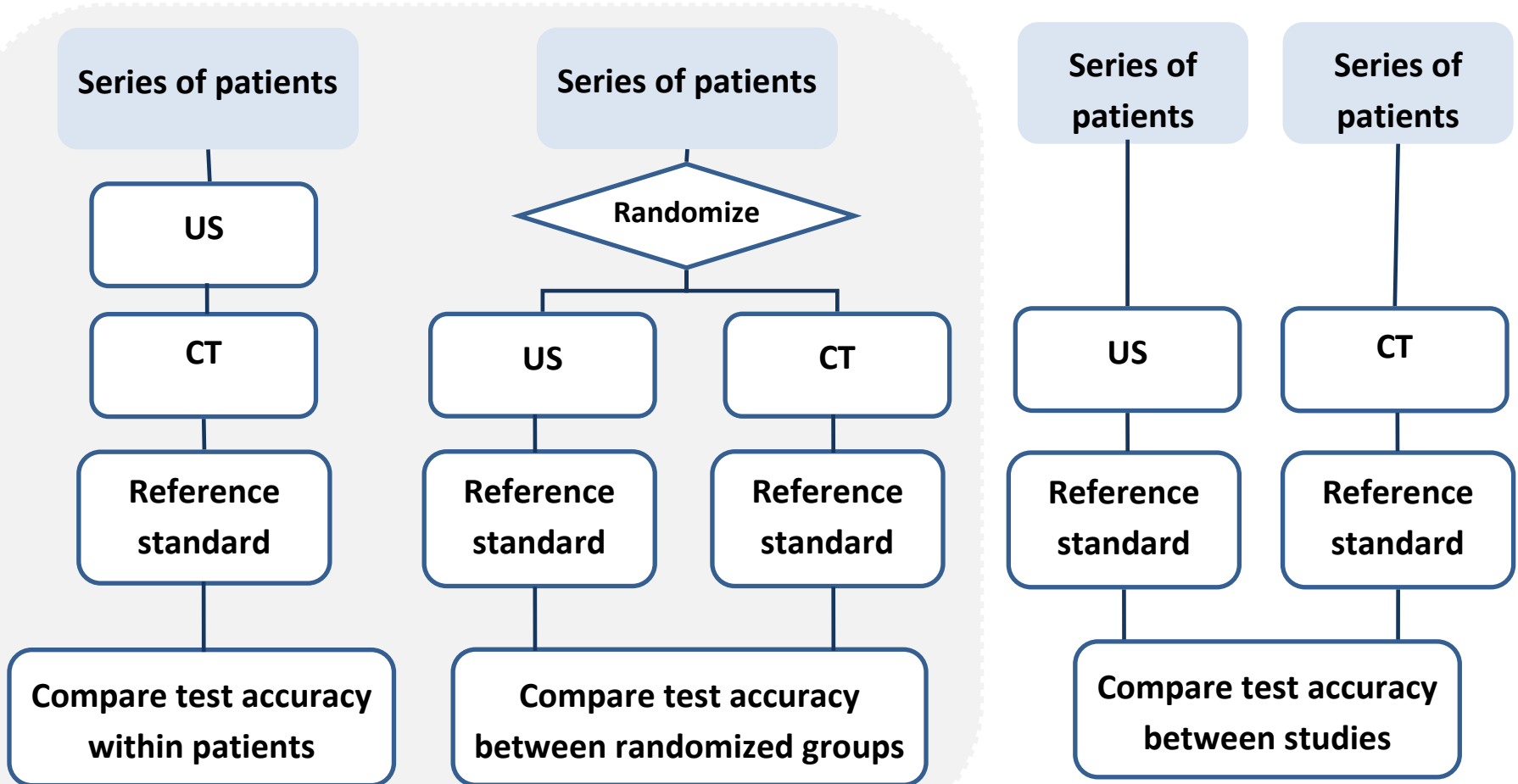
Primary Funding Source: National Institute for Health Research (United Kingdom).

Ann Intern Med. 2013;158:544-554.

For author affiliations, see end of text.

www.annals.org

Test comparison designs



Robust comparative studies

Non-comparative studies

**Are between study (indirect)
comparisons reliable?**



Well...

What's wrong with between study comparisons?

- Studies done in different time periods and places may **differ**
- Patient groups may **differ** systematically between studies
- Diagnosis verified in **different** ways
- Study methods (patient selection, blinding, etc) may **differ**

Rationale for systematic reviews

“The hundreds of hours spent conducting a scientific study ultimately contributes only a piece of an enormous puzzle.

The value of any single study is derived from how it fits with and expands previous work, as well as from the study's intrinsic properties.

Through systematic review the puzzle's intricacies may be disentangled”.

Mulrow CD. BMJ 1994;309:597-9.




Screening

Screening for alcohol problems in primary care: a systematic review

Fiellin D A, Carrington Reid M, O'Connor P G

Authors' objectives

To evaluate the accuracy of screening methods for alcohol problems in primary care.



**Who has
asymptomatic
disease?**

Diagnosis

Diagnostic Testing for Celiac Disease Among Patients With Abdominal Symptoms A Systematic Review

Daniëlle A. W. M. van der Windt, PhD

Petra Jellema, PhD

Chris J. Mulder, MD, PhD

C. M. Frank Kneepkens, MD, PhD

Henriëtte E. van der Horst, MD, PhD

JAMA. 2010;303(17):1738-1746

Context The symptoms and consequences of celiac disease usually resolve with a lifelong gluten-free diet. However, clinical presentation is variable and most patients presenting with abdominal symptoms in primary care will not have celiac disease and unnecessary diagnostic testing should be avoided.

Objective To summarize evidence on the performance of diagnostic tests for identifying celiac disease in adults presenting with abdominal symptoms in primary care or similar settings.

Who has
symptomatic
disease?

Staging

Annals of Internal Medicine

ARTICLE

Test Performance of Positron Emission Tomography and Computed Tomography for Mediastinal Staging in Patients with Non–Small-Cell Lung Cancer

A Meta-Analysis

Michael K. Gould, MD, MS; Ware G. Kuschner, MD; Chara E. Rydzak, BA; Courtney C. Maclean, BA; Anita N. Demas, MD; Hidenobu Shigemitsu, MD; Jo Kay Chan, BS; and Douglas K. Owens, MD, MS

Context

Is computed tomography (CT) or positron emission tomography with 18-fluorodeoxyglucose (FDG-PET) better for mediastinal staging of non–small-cell lung cancer?

Ann Intern Med. 2003;139:879-892.



**How
advanced is
the disease?**

DTA reviews are useful

Allergy

EUROPEAN JOURNAL OF ALLERGY
AND CLINICAL IMMUNOLOGY



Allergy

REVIEW ARTICLE

The diagnosis of food allergy: a systematic review and meta-analysis

K. Soares-Weiser¹, Y. Takwoingi², S. S. Panesar³, A. Muraro⁴, T. Werfel⁵, K. Hoffmann-Sommergruber⁶, G. Roberts^{7,8,9}, S. Halcken¹⁰, L. Poulsen¹¹, R. van Ree^{12,13}, B. J. Vlieg-Boerstra¹⁴ & A. Sheikh^{3,15} on behalf of the EAACI Food Allergy and Anaphylaxis Guidelines Group*

tions. This systematic review assessed the diagnostic accuracy of tests aimed at supporting the clinical diagnosis of food allergy.

Food Allergy and Anaphylaxis Guidelines

Translating knowledge into clinical practice



Steps of a DTA systematic review

1. Define the question
2. Define objectives and eligibility criteria
3. Develop protocol
4. Search for studies and selection
5. Collect data
6. Assess bias and applicability
7. Analyse and present results
8. Interpret results and draw conclusions

Importance of DTA review question formulation

- Identify potentially relevant studies
- Select studies for inclusion (based on eligibility criteria)
- Assess applicability of included studies
- Plan analyses
- Interpret results and draw conclusions (implications for practice and for research)


Components of a question

- For **intervention** reviews
 - **P**articipants
 - **I**ntervention
 - **C**omparative intervention
 - **O**utcome

Components of a question

- For **diagnostic test accuracy** reviews
 - **P**articipants
 - **I**ndex test
 - **C**omparator test
 - **T**arget condition

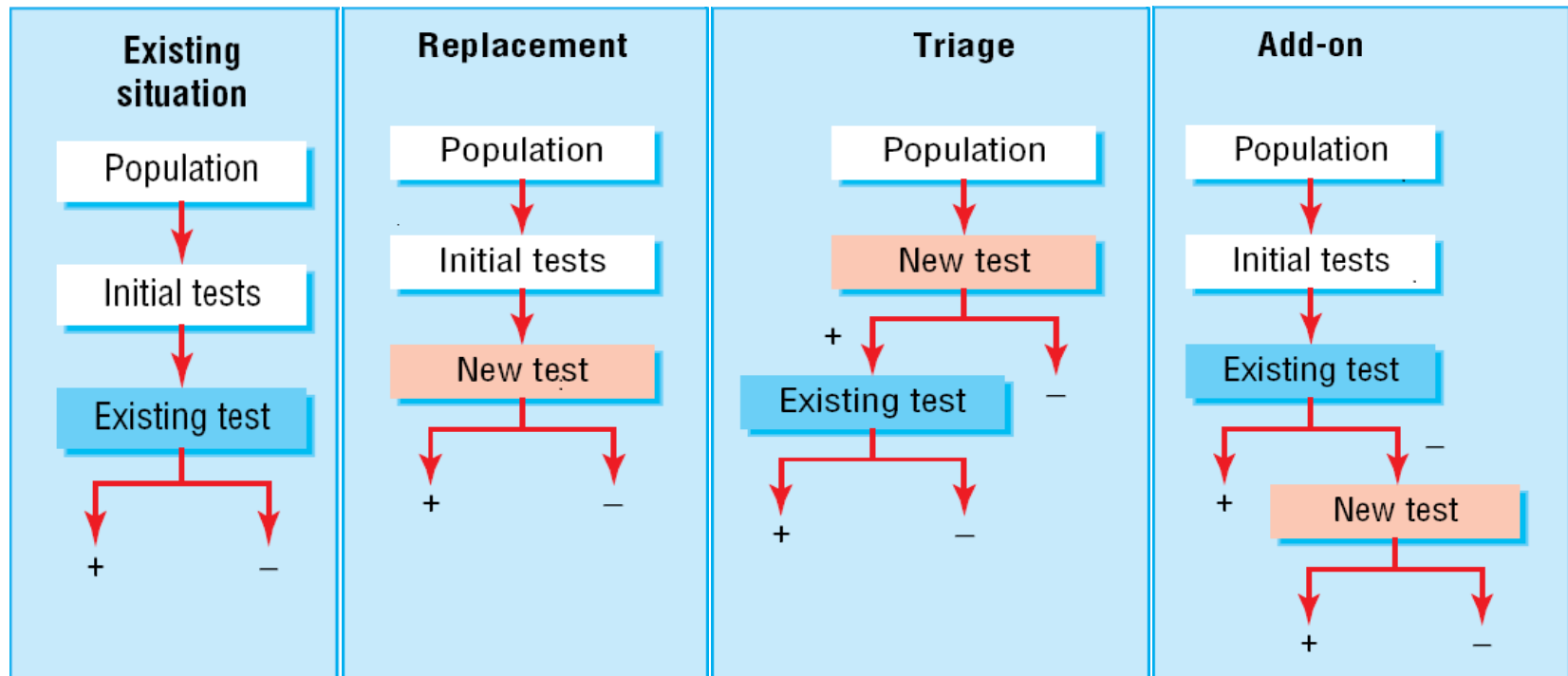
Components of a question

- For **diagnostic test accuracy** reviews
 - **P**articipants
 - **P**resentation
 - **P**rior tests
 - **I**ndex test
 - **C**omparator test
 - **P**urpose (role of test) 
 - **T**arget condition
 - **R**eference standard

Clinical/diagnostic pathway

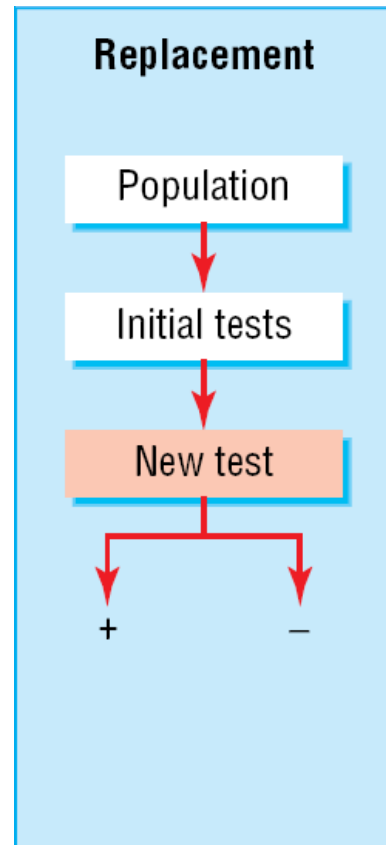
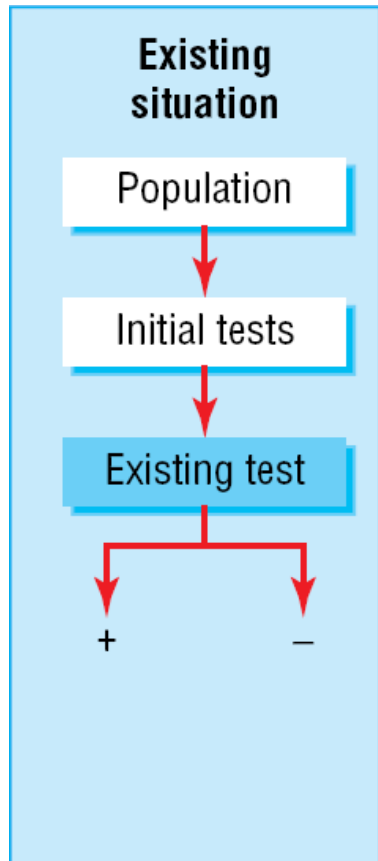
- What is currently being done to get to a diagnosis?
What is the patient's 'diagnostic journey'?
- Where does your index test fit in? What's the role of your index test?
- What happens with the patient after a diagnosis has been made? What are the consequences of (false) positive index tests and (false) negative index tests?

Roles of tests and positions in existing diagnostic pathways



Bossuyt PM et al. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332:1089-92

Replacement

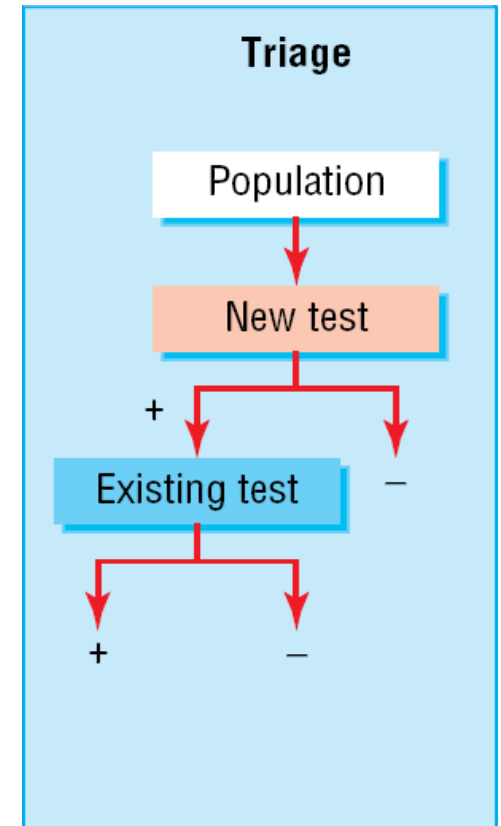


Replacement

- Replace test A with test B, because test B is
 - more accurate
 - less invasive, easier to do, less risky
 - less uncomfortable for patients
 - quicker to yield results
 - technically less challenging
 - more easily interpreted
- Compare accuracy and downstream consequences of both tests

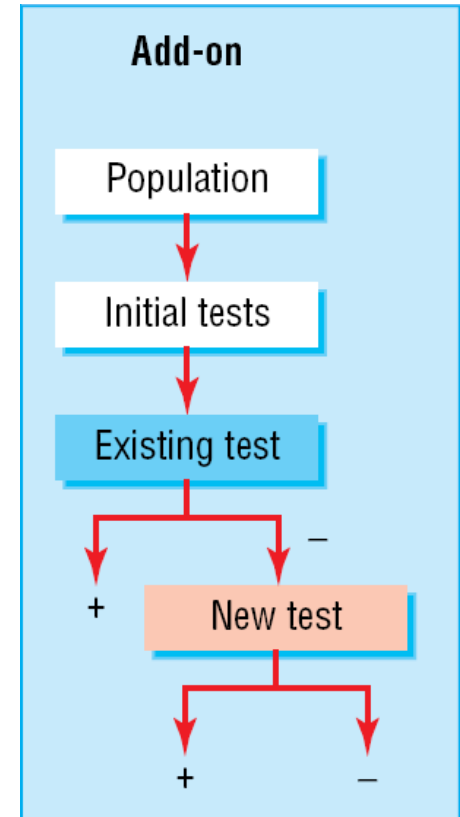
Triage

- New test positioned before the existing test pathway (= comparator)
- Purpose: to select patients for further testing (or not)
- Triage tests may be less accurate than existing tests; they may have other advantages (like simplicity or low cost)
- Compare accuracy and downstream consequences of both test strategies



Add-on

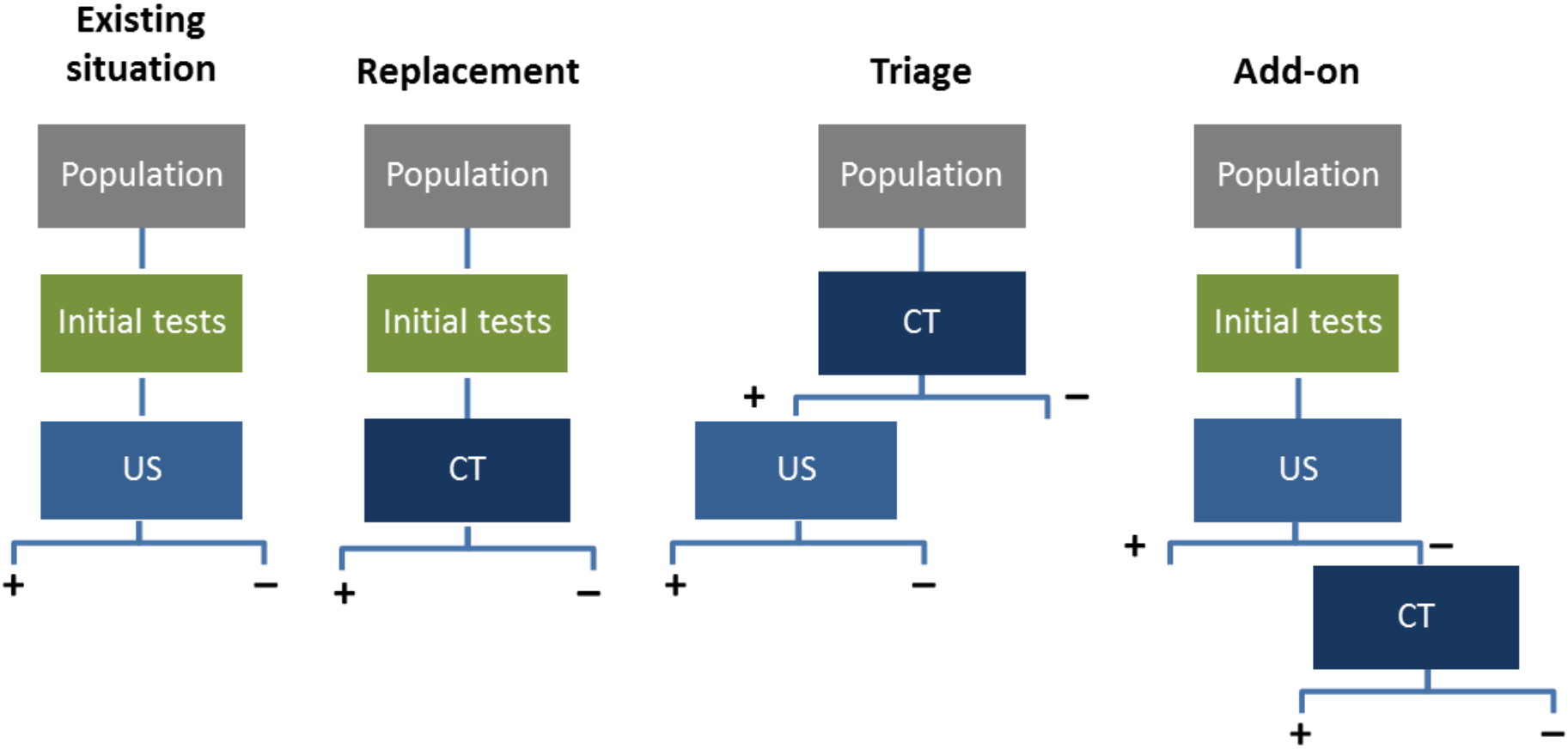
- New test positioned after the existing test pathway (= comparator)
- Purpose: to detect patients not identified by existing test(s) (= FNs)
- New test more accurate but otherwise less attractive than existing tests, e.g. costs, invasiveness, availability, etc.
- Compare accuracy and downstream consequences of both test strategies



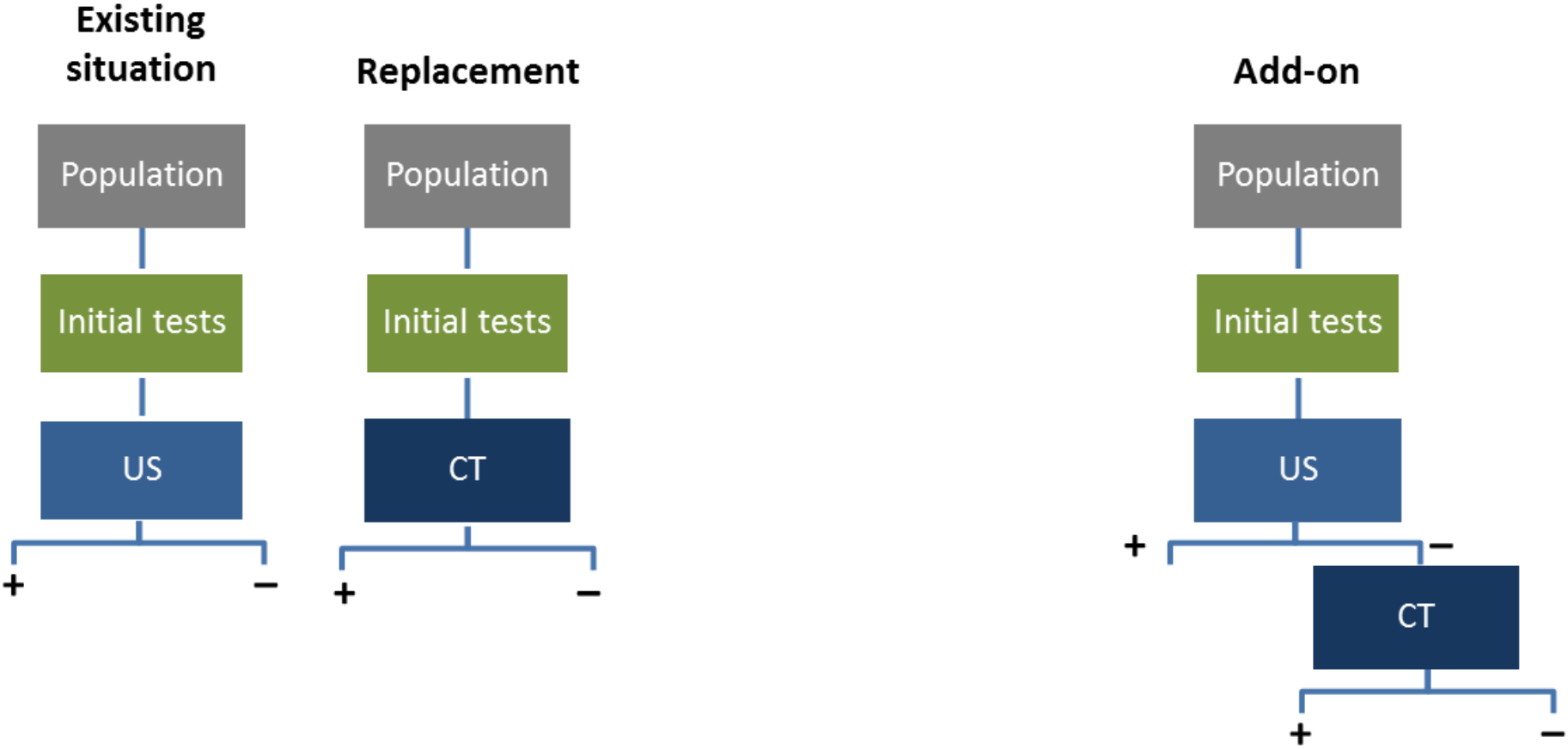
Role of CT in diagnosing acute appendicitis



Is CT or US better for diagnosing acute appendicitis?



Is CT or US better for diagnosing acute appendicitis?



Take home message (1)

- Different uses of tests
- Not every question can be answered by diagnostic accuracy
- Carefully consider – is this a test accuracy question?
- Different test accuracy study designs
 - Case control studies are prone to bias
 - Comparative studies ideal for test comparisons

Take home message (2)

- Careful formulation of DTA review questions underpins:
 - effective and efficient search for studies
 - selection of studies
 - assessment of applicability / interpretation of results
- Delineating the clinical testing pathway is a crucial component of question formulation
- Consider the downstream consequences of FPs and FNs

ACKNOWLEDGEMENTS

Materials for this presentation are based in part on material adapted from members of the Cochrane Screening and Diagnostic Test Methods Group

See <http://dta.cochrane.org/dta-author-training-online-learning>
for additional training materials

Calculation and interpretation

- Sensitivity $5/10 = 50\%$
 - Half of the patients with disease will be detected
- Specificity $990/1000=99\%$
 - Only 1 in 100 patients without disease erroneously will receive a false positive result