Cancer Care Ontario
Action Cancer Ontario

EVIDENCE-BASED GUIDELINE DEVELOPMENT FOR DIAGNOSTIC QUESTIONS

Emily Vella, Xiaomei Yao

Cancer Care Ontario's Program in Evidence-Based Care, Department of Oncology, McMaster University, Ontario, Canada

Ontario
Cancer Care Ontario
Action Cancer Ontario

Who are we?

- Cancer Care Ontario (CCO)
  - Government agency in Ontario, Canada
  - Seeks to improve cancer services
  - Provides information for patients, health professionals, and public on cancer prevention, screening, treatment, and supportive care
- Program in Evidence-Based Care
  - Part of CCO

What do we do?

- The Program in Evidence-Based Care (PEBC) produces evidence-based guidelines
- Every guideline has 3 sections
  - Section 1: Guideline Recommendations
  - Section 2: Systematic Review
  - Section 3: Development Methods, Recommendations Development, and External Review Process
- A formal standardized process to ensure the currency of each document

Workshop Objectives

1. Generating an appropriate research question
2. Developing relevant eligibility criteria
3. Critically appraising diagnostic studies using existing tools and quality criteria
4. Determining recommendations from different types of evidence and information available

Diagnostic tests

- Should reduce uncertainty
- Are cross-sectional
- Diagnostic accuracy of a test can change in different subgroups with same target condition
- Goal is to change clinician's behaviour and affect patient outcomes

1. Generating an appropriate research question

1. Generating a research question

PICOT for treatment studies can be replaced by PIRO for diagnostic studies

P = Patients/target condition

I = Index test

R = Reference standard

O = Outcomes

1. Generating a research question

Patients/target condition

What kind of patients are the target population for the guideline?

- Gender
- Age
- Presentation of target condition
- Target condition
- Other morbidities with specific treatments

1. Generating a research question

Patients/target condition

- If prevalence of target condition is different, we may need to look at subgroups of patients

Ex. patients with an undiagnosed chest nodule or mass demonstrated on imaging

1. Generating a research question

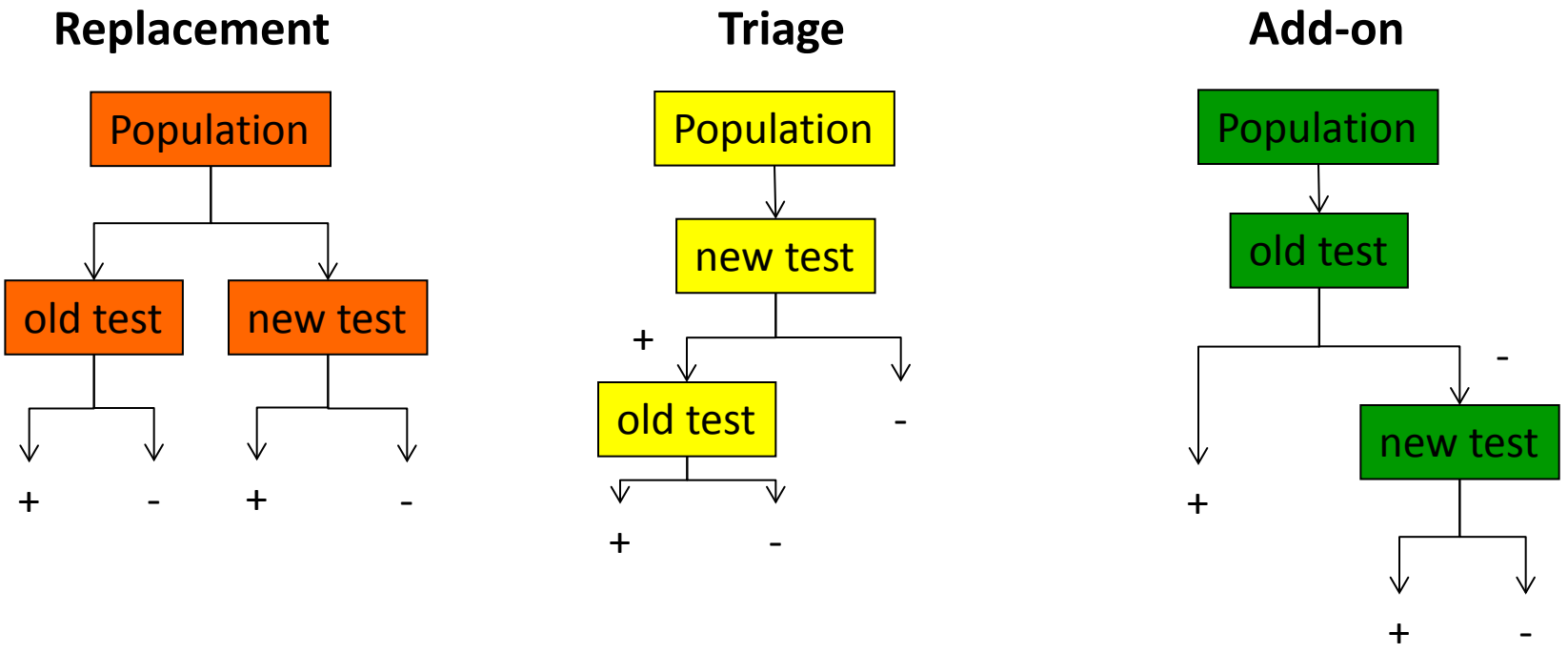
Index test

What is your index test of interest for the guideline?

- It's the test you are trying to assess
- Multiple index tests can be included

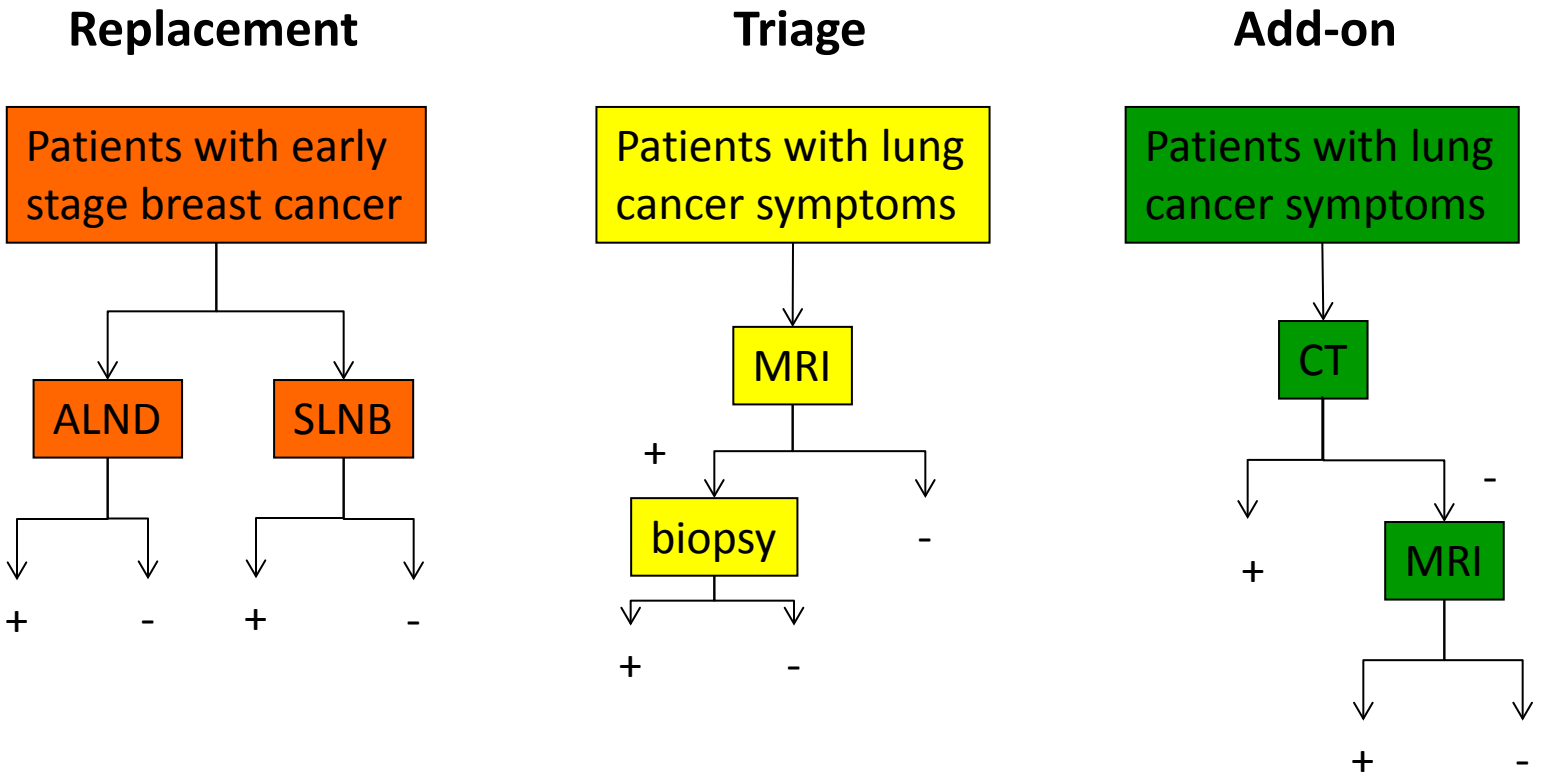
1. Generating a research question

Role of index test



Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332:1089-92.

1. Generating a research question



SLNB = sentinel lymph node biopsy
ALND = axillary lymph node dissection
MRI = magnetic resonance imaging
CT = computed tomography

1. Generating a research question

Index test

Ex. fine need aspiration biopsy (FNAB) versus core needle biopsy (CNB)

1. Generating a research question

Reference Standard

- Best available, clinically accepted, error-free
- A series of procedures
- Sometimes the reference standard may be different for patients testing positive versus negative

Chapter 6 in the Cochrane Handbook for DTA Review
http://srdta.cochrane.org/sites/srdta.cochrane.org/files/uploads/Chapter06-Including-Studies%20(September-2008).pdf

1. Generating a research question

Reference Standard

Ex. histological confirmation for patients with positive results from biopsy or clinical follow-up for patients with negative results from biopsy

1. Generating a research question

Outcomes

(1) Diagnostic accuracy outcomes

- Sensitivity
- Specificity
- PPV
- NPV

} affected by prevalence

- Diagnostic accuracy
- Likelihood ratio
- Diagnostic odds ratio
- ROC curve
- Agreement

		reference standard +	reference standard -
index test	+	a=TP	b=FP
index test	-	c=FN	d=TN

1. Generating a research question

(2) Patient-related outcomes

- Referring clinicians' diagnostic thinking
- Change in patient management
- Patient outcomes
- Costs and benefits

Ex. diagnostic accuracy to diagnose lung cancer, patient-important outcomes, and complications of test

Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making. 1991;11:88-94

1. Generating a research question

An example of a Research Question

When compared with the reference standard (the histological confirmation from wedge biopsy, surgical resection, metastases, or autopsy; or clinical follow-up), which test has a higher diagnostic accuracy to diagnose lung cancer and lead to better patient outcomes with less complications: fine need aspiration biopsy (FNAB) versus core needle biopsy (CNB) in patients with an undiagnosed chest nodule or mass demonstrated on imaging?

Red = PATIENTS

Blue = INDEX TESTS

Purple = REFERENCE STANDARD

Green = OUTCOMES

Exercise 1 (4 mins)

Please convert the following research questions into one research question with PIRO with your group members (4 minutes).

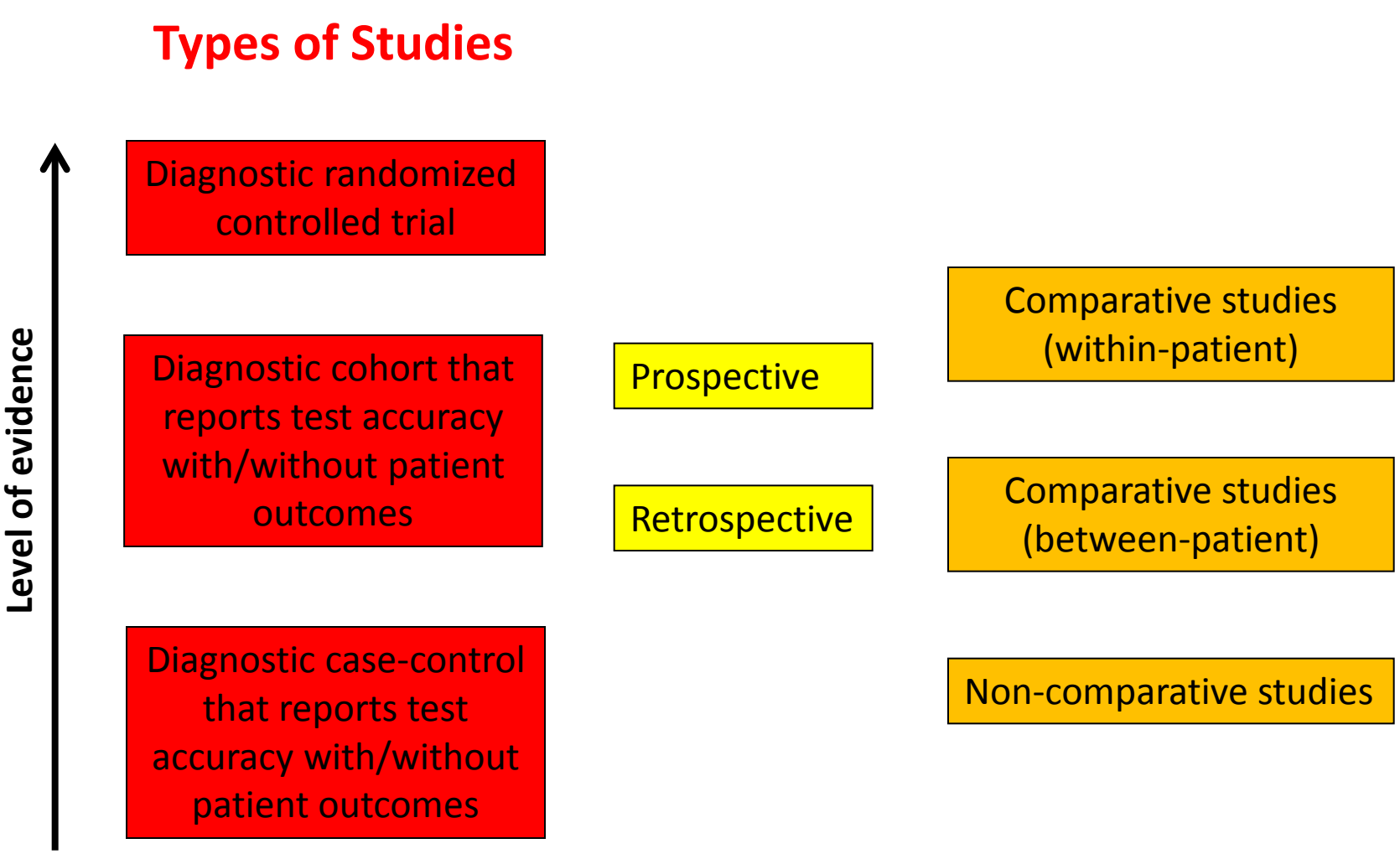
- Can axillary lymph node dissection (ALND) (the reference standard) be avoided in patients with negative findings on sentinel lymph node biopsy (SLNB) (the index test)?
- Should SLNB be the recommended standard of care for patients with clinically proven early-stage breast cancer?
- Is ALND necessary for all patients with positive findings on lymph node biopsy?
- What factors affect the success of SLNB (including low rates of complications and false-negative results)?
- What are the potential benefits and harms associated with SLNB?

When sentinel lymph node biopsy is compared with axillary lymph node dissection (the current reference standard), is lymph node biopsy accurate enough to positively change patient management (with acceptable false-negative rates) and patient outcomes (e.g., survival rate, QOL, etc.) in patients with clinically proven early-stage breast cancer without causing severe harms?

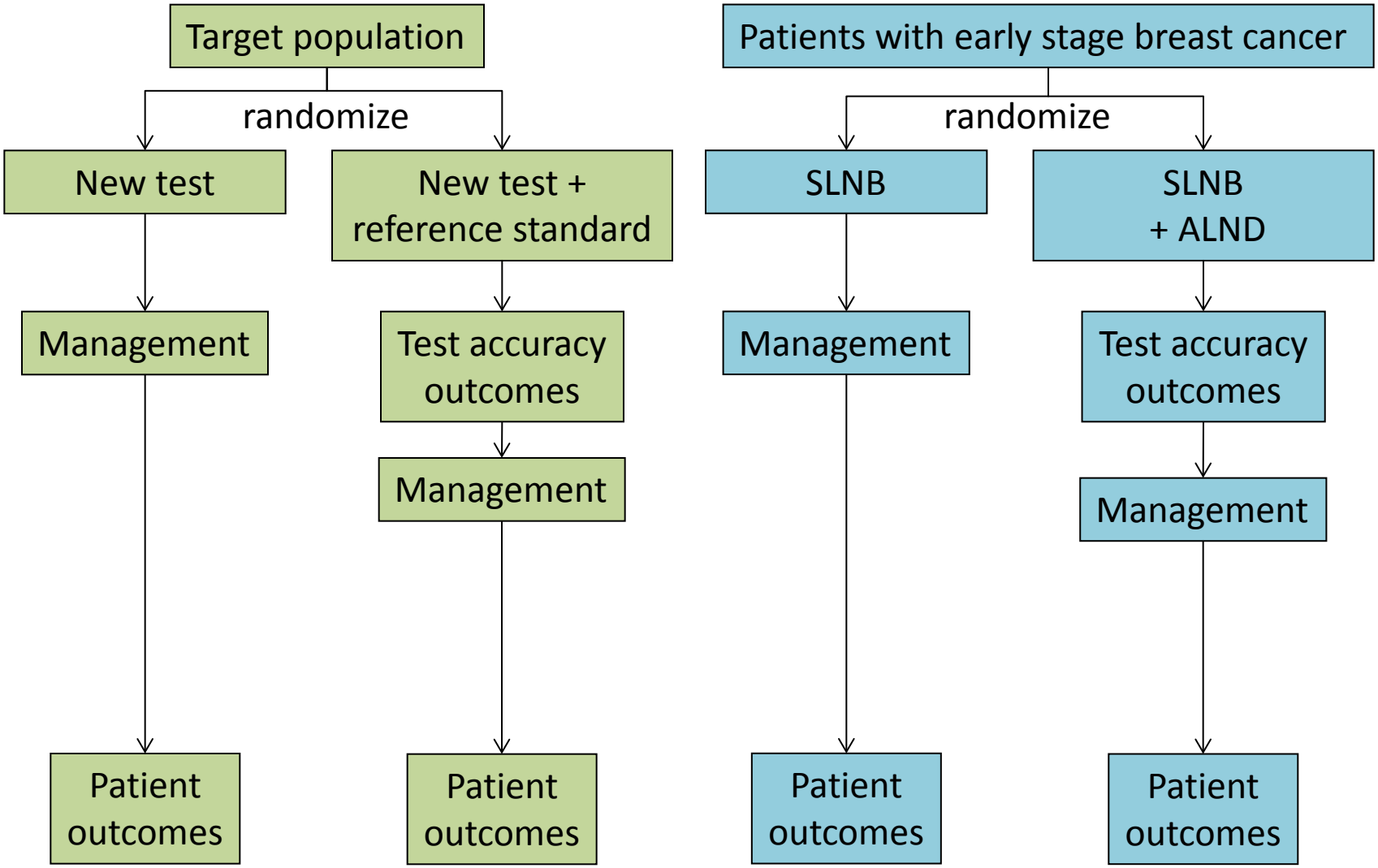
- Red = PATIENTS
- Blue = INDEX TEST
- Purple = REFERENCE STANDARD
- Green = OUTCOMES

2. Developing relevant eligibility criteria

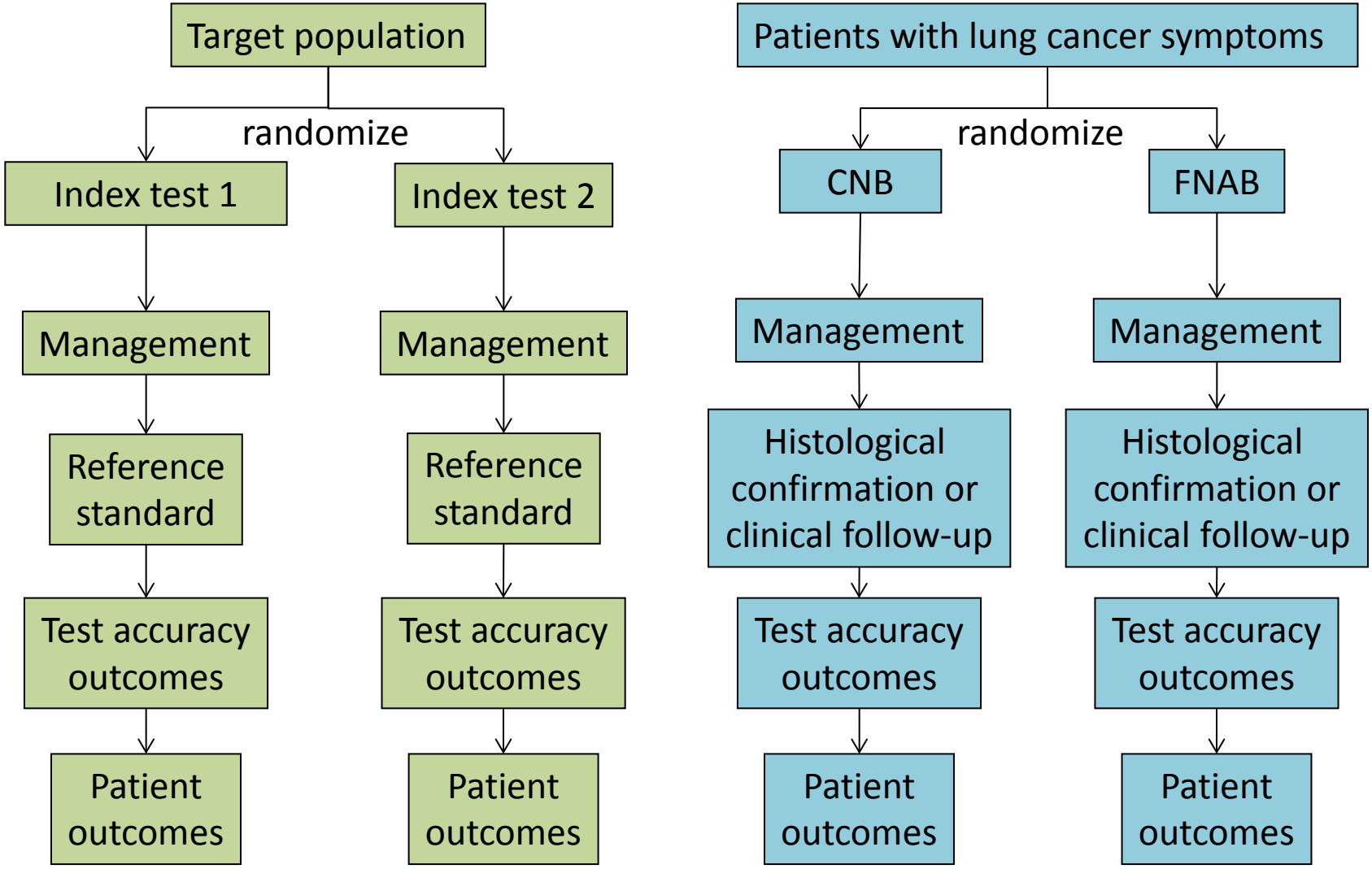
2. Developing relevant eligibility criteria



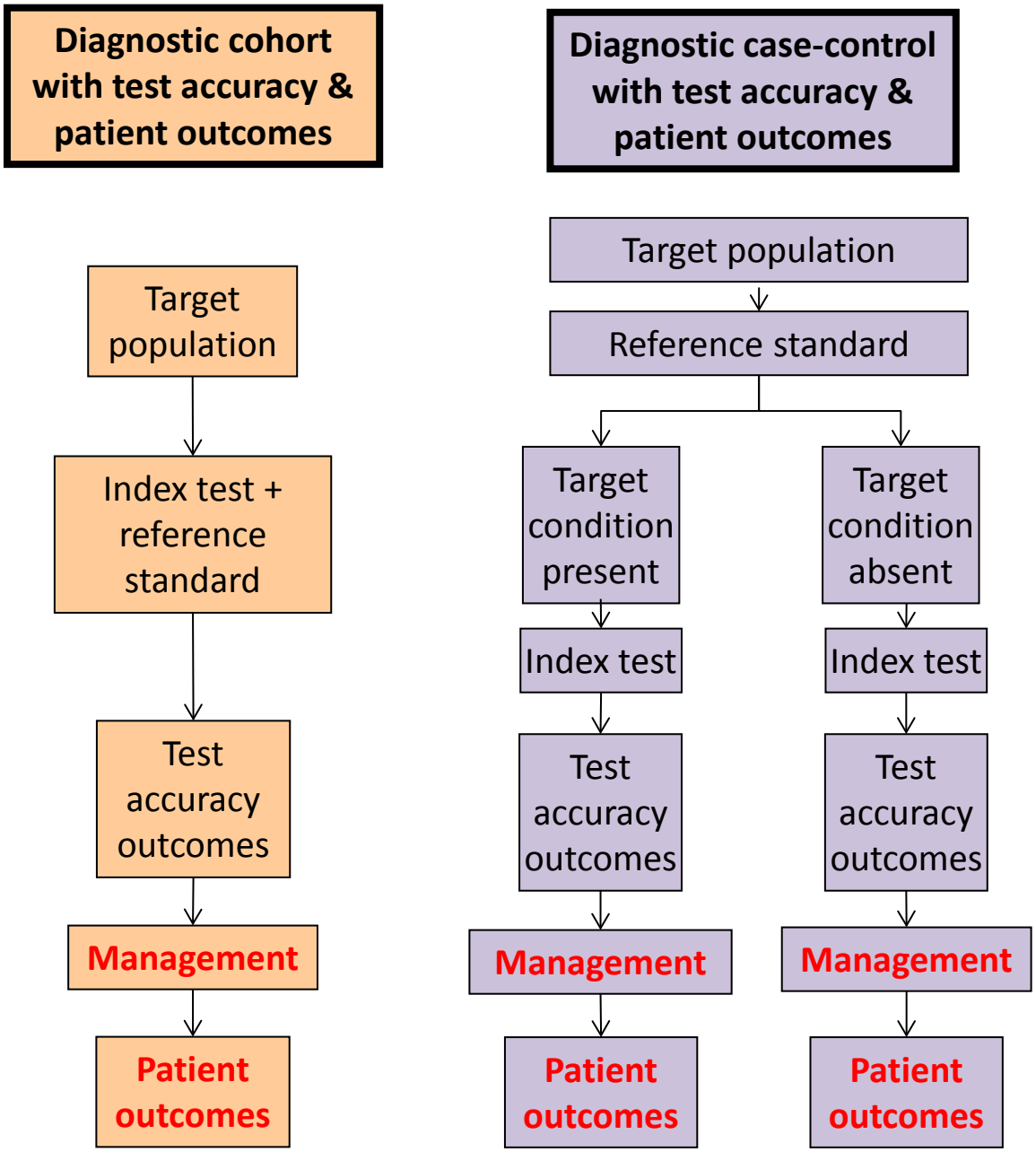
Diagnostic RCTs with one index test



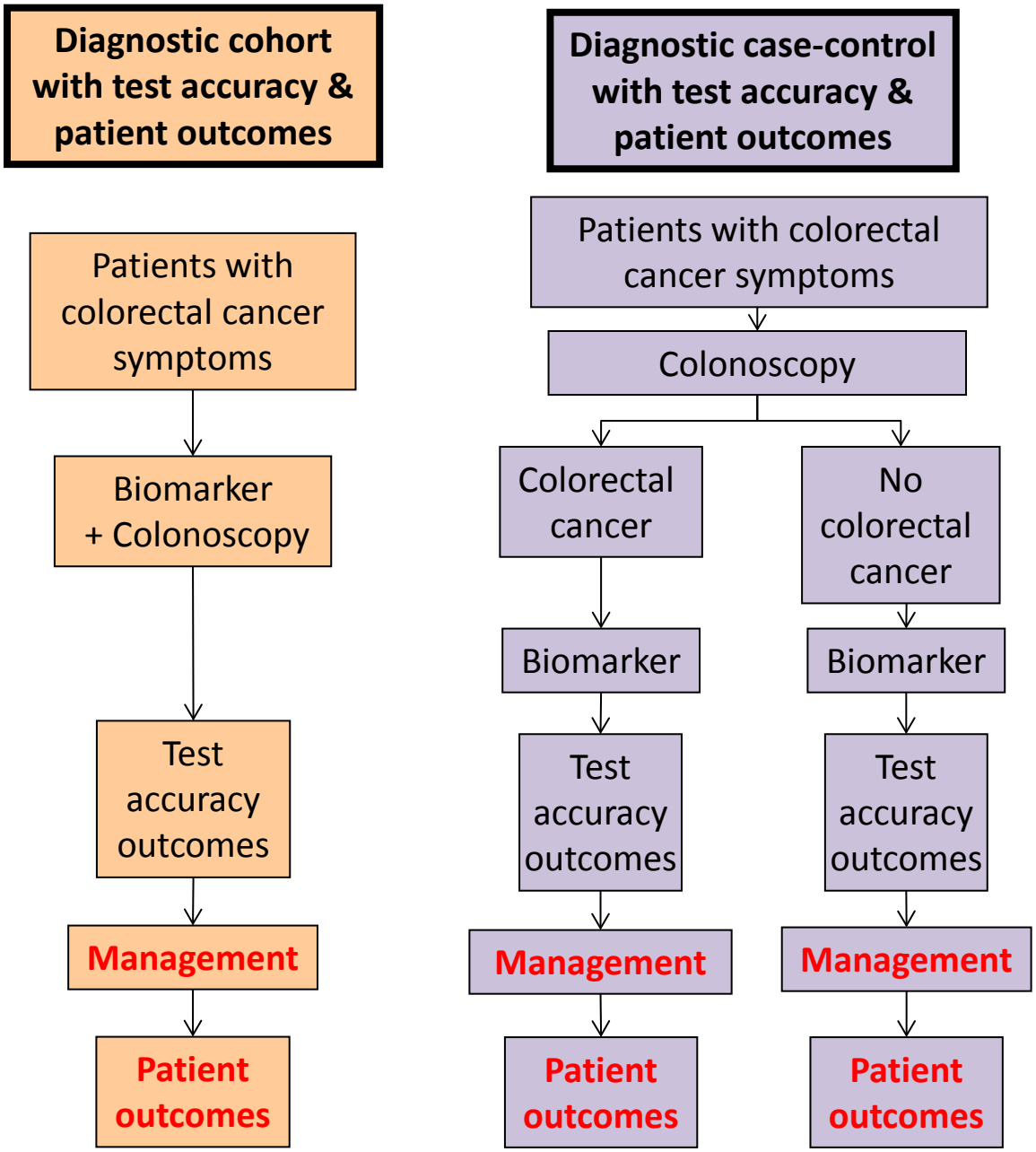
Diagnostic RCTs with multiple index tests



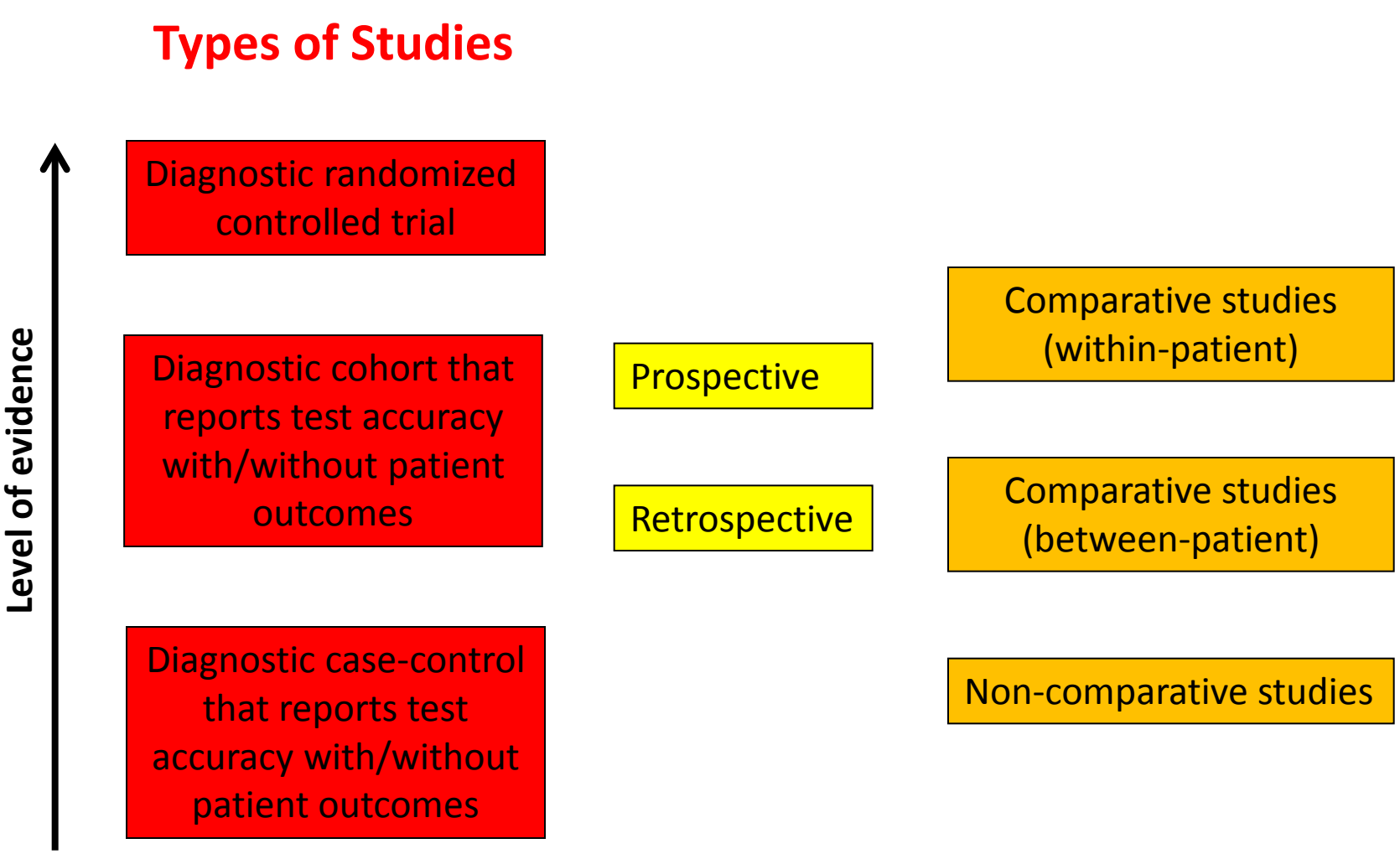
Non-RCTs with one index test



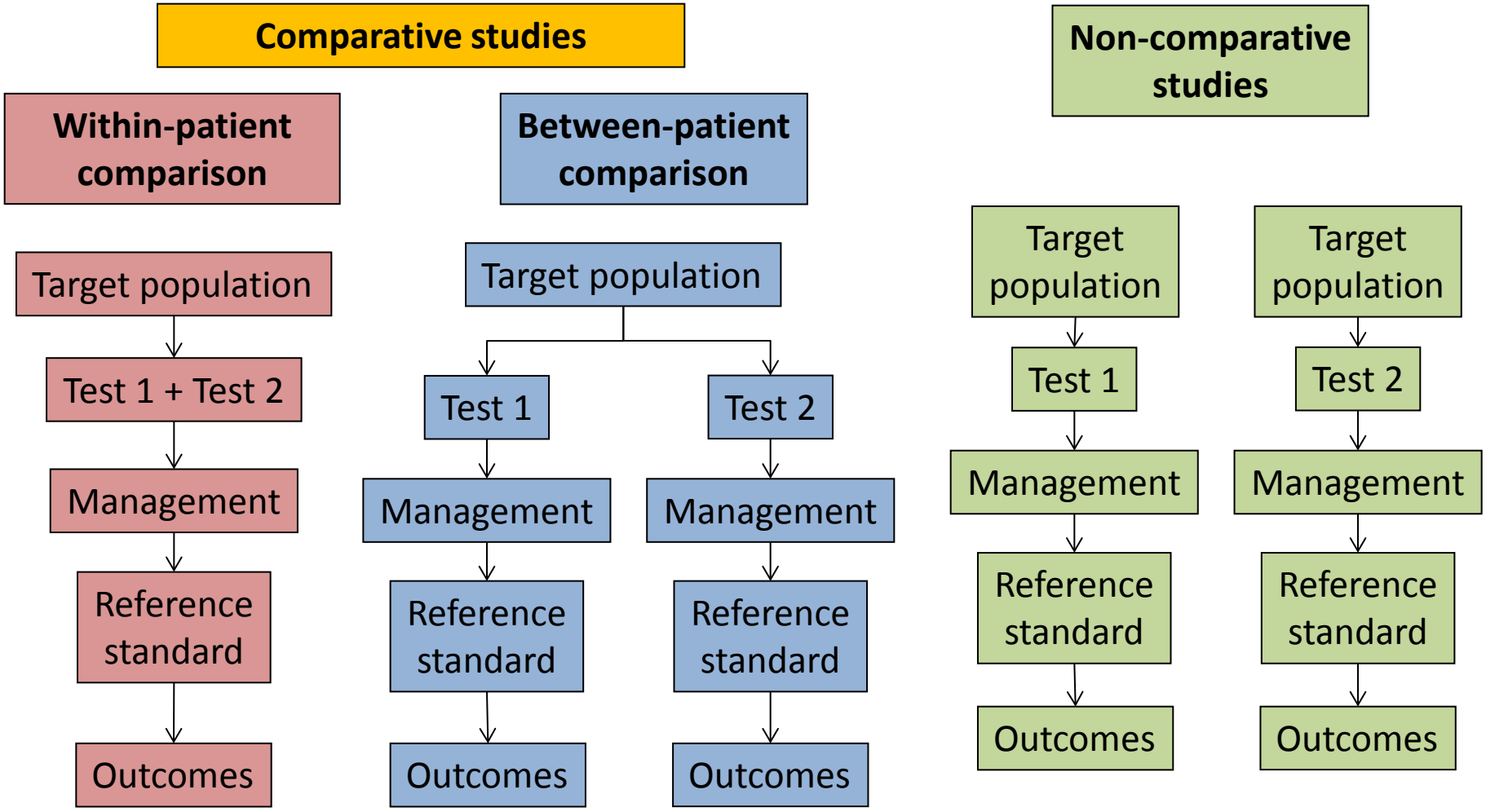
Non-RCTs with one index test



2. Developing relevant eligibility criteria

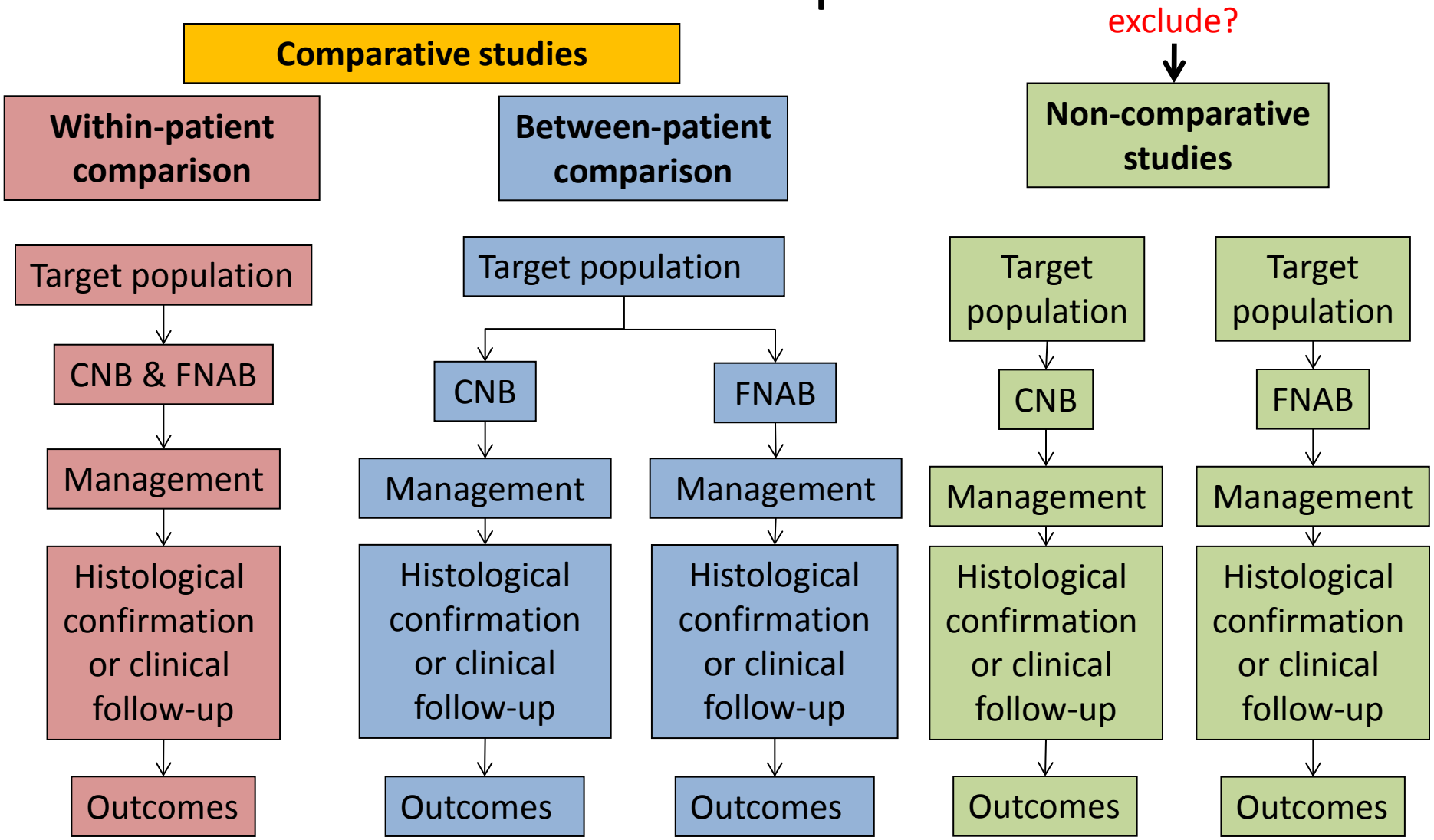


Non-RCTs with multiple index tests



- Outcomes may include diagnostic and patient outcomes

Non-RCTs with multiple index tests



- Target population = Patients with lung cancer symptoms

Hooper R Diaz-Ordaz K Takeda A Khan K. (2013) Comparing diagnostic tests: trials in people with discordant test results. Statist. Med. 32, 2443-2456.

Non-RCTs with multiple index tests

- In PEBC guidelines, we usually do not include non-comparative studies
- A current published paper supports our opinion: “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 noncomparative and comparative studies were greater than expected by chance (P = 0.001).”

Takwoingi Y et al. Empirical evidence of the importance of comparative studies of diagnostic test accuracy. Ann Intern Med 2013; 158:544-554.

2. Developing relevant eligibility criteria

Patients

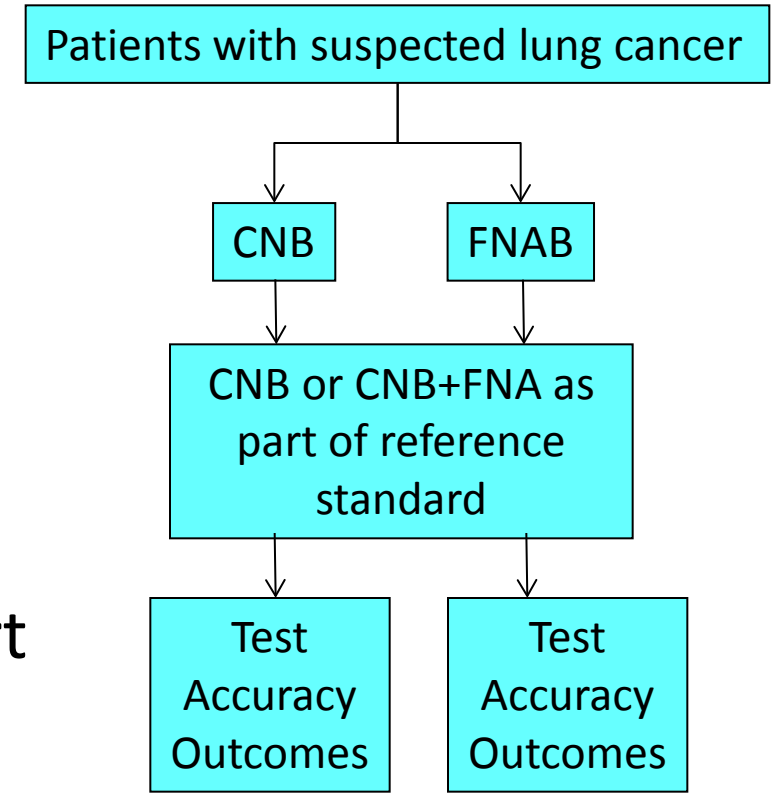
Inclusion criteria should include:

- All relevant presenting clinical characteristics
- The setting
- Any prior tests or clinical history

2. Developing relevant eligibility criteria

Index and reference tests

- Do not narrow search criteria to the specifics of the index test
- Subgroup analyses can be used if necessary
- The index test should not be the reference standard or part of the reference standard!



2. Developing relevant eligibility criteria

Outcomes

- Patient outcomes can be added in study selection criteria to narrow down the literature search results.

2. Developing relevant eligibility criteria

An Example: Inclusion Criteria

Articles were included if they met all the following criteria:

--Were published in certain time period.

--Were systematic reviews, meta-analyses, clinical practice guidelines based on a systematic review, randomized trials, or comparative studies.

--Reported or provided sufficient data to calculate at least one diagnostic characteristic for both FNAB and CNB.

--Included patients with an undiagnosed chest nodule or mass demonstrated on imaging.

--Reference standard was histological confirmation or clinical follow-up.

2. Developing relevant eligibility criteria

Exclusion Criteria

Articles were excluded if they met any of the following criteria:

--Recruited only patients who had previous or current diagnosis of lung cancer or other chest malignancies at baseline.

--Regarded the biopsy results from FNAB and/or CNB as a part of the reference standard.

--Performed FNAB and CNB on patients with completely different-sized lesions (for example, FNAB was performed on patients with <20-millimeter lesions, and CNB was performed on patients with ≥20-millimeter lesions).

3. Critically appraising diagnostic studies using existing tools and quality criteria

3. Critical appraisal

- QUADAS (Quality Assessment of Diagnostic Accuracy Studies)
  - Tool developed by Whiting in 2003
- Cochrane Collaboration recommends:
  - Modified 11 of the 14 original quality items of the QUADAS tool
  - Against using scales that yield a summary score.
- PEBC consistent with Cochrane Collaboration

Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol. 2003; 3:25.

Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ,. Chapter 9: Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* Version 1.0.0. The Cochrane Collaboration, 2009. Available from: http://srdta.cochrane.org/.

3. Critical appraisal

Key domains of modified 11 items from the Cochrane:

- Representativeness of the study sample
- Soundness of the verification procedure
- Blinding of test interpretation
- Missing data
  - uninterpretable/inconclusive/intermediate results
  - withdrawals

Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ,. Chapter 9: Assessing methodological quality. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* Version 1.0.0. The Cochrane Collaboration, 2009. Available from: http://srdta.cochrane.org/.

3. Critical appraisal

- QUADAS 2 available since 2011
  - Focused on 4 key domains:
    - Patient selection
    - Index test
    - Reference standard
    - Flow and timing
  - Changed ”Yes”, ”No”, and ”Unclear” into ”Low”, ”High”, and ”Unclear”.
  - Did not include assessment for multiple index tests
- The PEBC still use the modified 11 items from the first QUADAS tool

Whiting PF, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155:529-36

Question

- Has anyone used other tools to assess the quality of diagnostic studies?

Critical Appraisal Skills Programme (CASP)

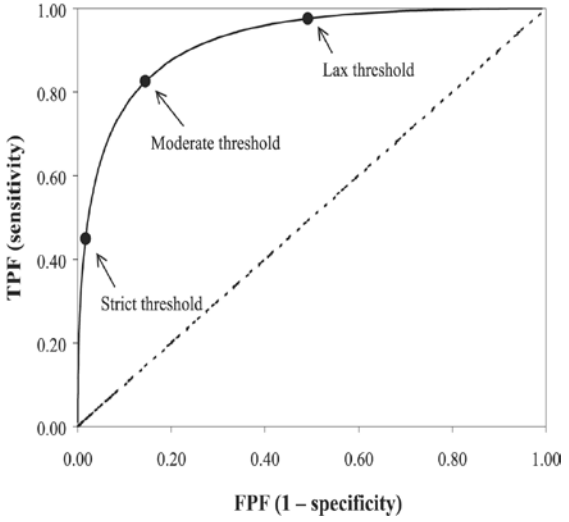
- 12 questions including 3 sections

http://www.casp-uk.net/wp-content/uploads/2011/11/CASP-Diagnostic-Test-Checklist-31.05.13.pdf

Jaesche R, Guyatt GH, Sackett DL, Users' Guides to the Medical Literature, V1. How to use an article about a diagnostic test. JAMA 1994; 271 (5): 389-391

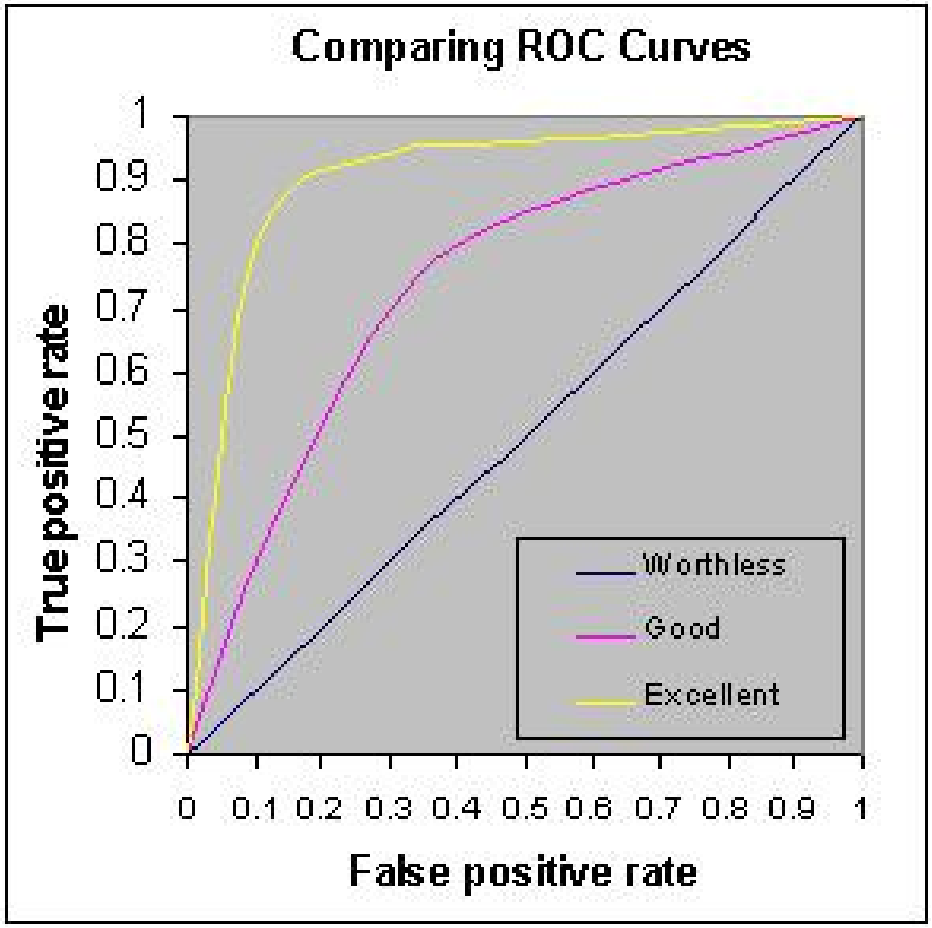
Data Analysis

- If we have sensitivity & specificity outcomes
  - Bivariate analysis
    - Use if have little variation in cutpoints among studies
    - Gives estimates of summary sensitivity and specificity of the test for a common cutpoint, or for several different common cutpoints
  - ROC (Receiver Operating Characteristic) curve
    - Use if have little consistency in cutpoints between studies
    - Describes how sensitivity and specificity trade-off with each other as cutpoints vary



Macaskill P, Gatsonis C, Deeks JJ, Harbord RM, Takwoingi Y. Chapter 10: Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors), *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* Version 1.0. The Cochrane Collaboration, 2010. Available from: http://srdta.cochrane.org/.

Data Analysis



Data Analysis

- If we have patient outcomes
  - Analyze them in the same way as intervention studies

4. Determining recommendations from different types of evidence and information available

4. Generating Recommendations

For each recommendation can include:

- Key Evidence for Benefits and Harms
- Aggregate Evidence Quality and Potential for Bias
- Values of the Working Group
- Considered Judgement

Rosenfeld RM, Shiffman RN. Clinical practice guideline development manual: a quality-driven approach for translating evidence into action. Otolaryngol Head Neck Surg. 2009;140(suppl):S1-S43.
Schunemann HJ, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336:1106-10.

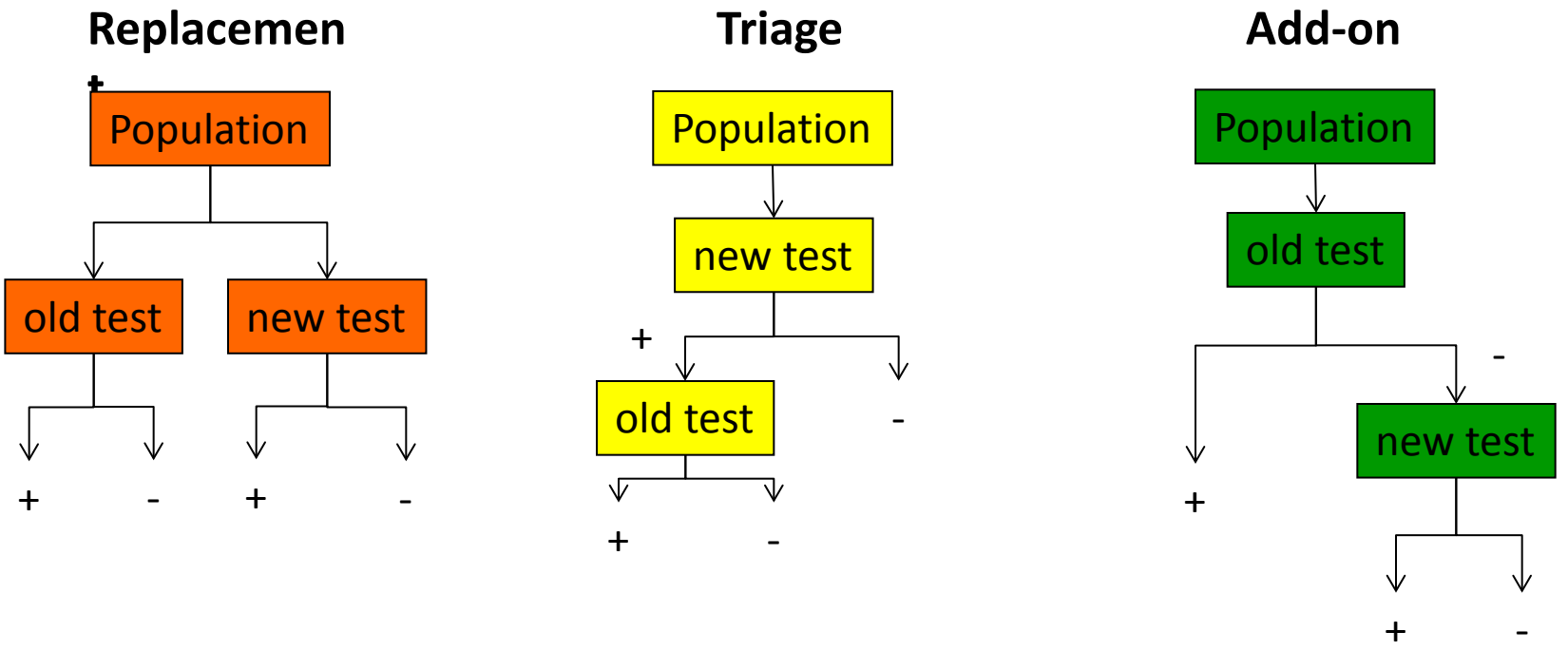
4. Generating Recommendations

Consider the benefits and harms for patients in each of the four cells of the 2 x 2 table

		reference standard +	reference standard -
index test	+	TP	FP
	-	FN	TN

4. Generating Recommendations

Role of index test



Bossuyt PM, Irwig L, Craig J, Glasziou P. Comparative accuracy: assessing new tests against existing diagnostic pathways. BMJ. 2006;332:1089-92.

Replacement tests

Compared with old test, if new test has:

(1) Similar sensitivity and specificity (with similar agreement)

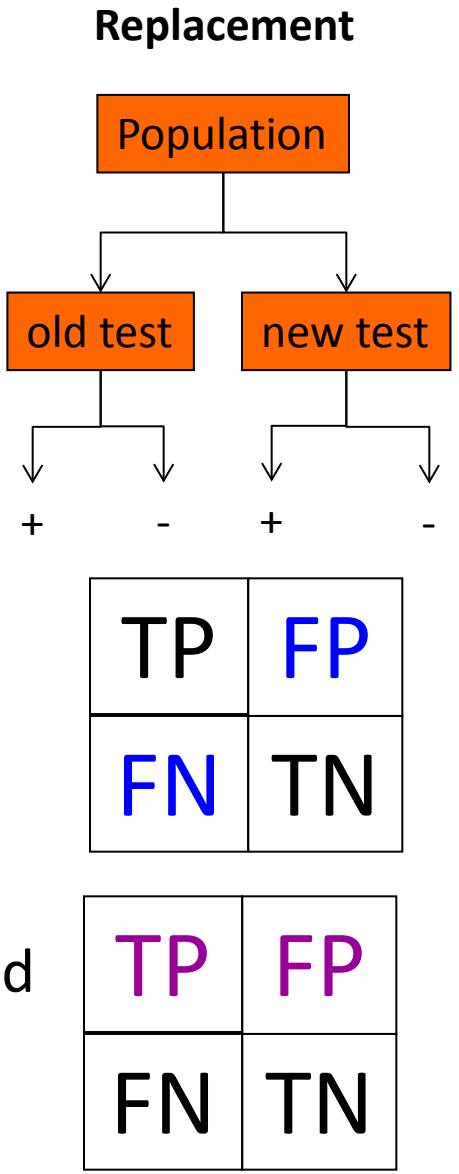
- Then may not need to look at patient outcomes (however adverse events and cost etc. of index test should be considered)

(2) Lower sensitivity and/or lower specificity but other positive attributes – ex. non-invasive and cheaper etc.

- Then may need to assess trade-off

(3) Higher sensitivity and/or higher specificity

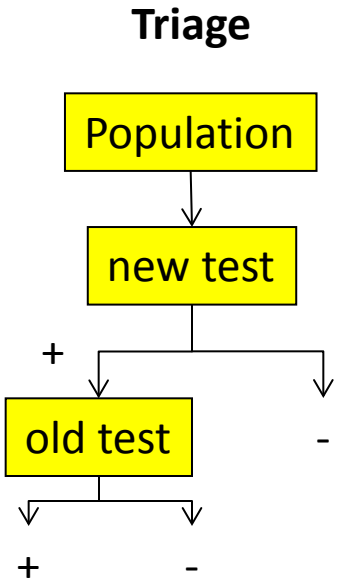
- Then patient outcomes of extra true positives and fewer false-positives may need to be examined



Lord SJ, et al. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med. 2006;144:850-5.

Triage tests

- They aim to reduce the use of existing tests that are more invasive, troublesome, or expensive rather than to improve the diagnostic accuracy.
- Patient outcomes of false-negative cases may be examined to avoid substantial harm.

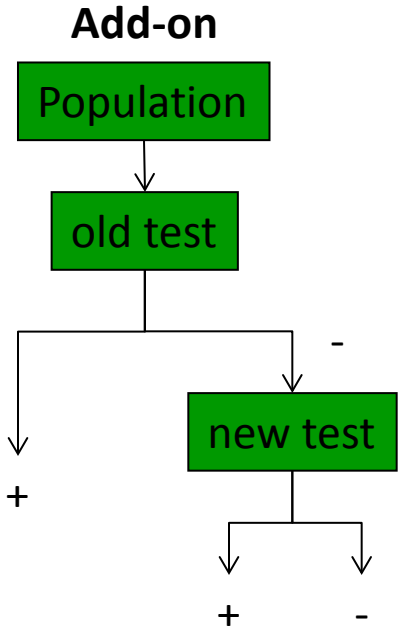


TP	FP
FN	TN

Add-on tests

- They can improve sensitivity and/or specificity.

- Patient outcomes of positive cases (including true positives and false positives) should be examined.



TP	FP
FN	TN

Exercises 2-3 (10 mins)

Exercise 2

The guideline panel wants to know if chest CT (new test) is better than chest X-ray (old test) at detecting lung metastasis in patients with rectal cancer. The reference standard for detecting lung metastasis is biopsy or clinical follow-up.

- What is the role of the index test – replacement, triage, or add-on?
  - Replacement
- Most studies compared CT with X-ray did not do biopsy or clinical follow-up for all patients. Some studies compared X-ray against CT results. Would you include this study in the systematic review?
  - No, only select studies that confirm diagnosis with biopsy, and do not use CT as reference standard
- Chest CT is shown to be more sensitive but less specific and has more inconclusive results than chest X-ray. Patients that have lung metastasis have a poor prognosis. What things should the guideline panel consider when developing their recommendations?
  - Consider the burden of verifying inconclusive results on patients and system (there was no agreement as to how these patients should be managed)
  - Consider how effective the treatments are for the extra positive cases and what harms the extra false-positive cases will encounter
  - Guideline panel recommended either test

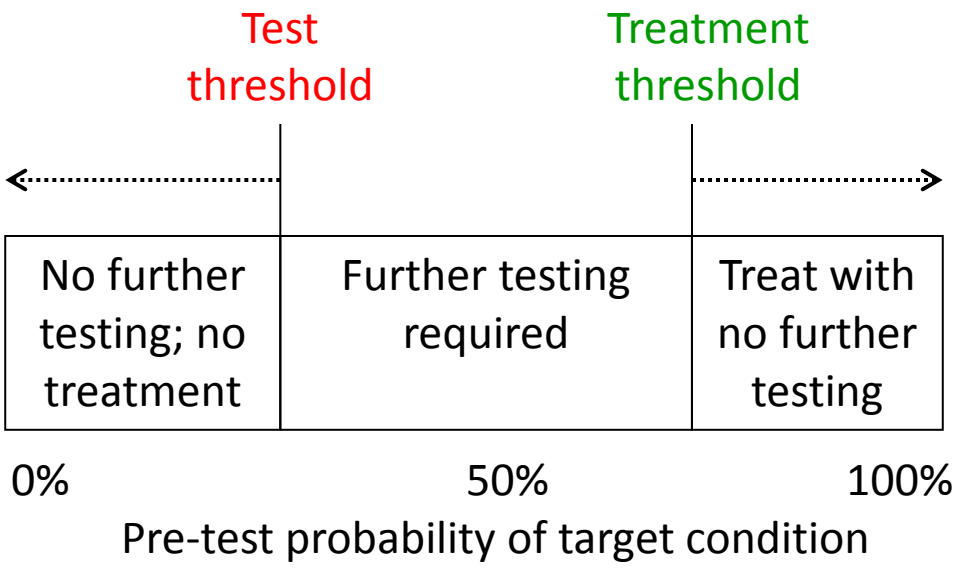
Exercise 3

The guideline panel wants to know which imaging technique (CT, MRI, sonography) to use to identify masses suspicious for ovarian cancer in patients. Patients that are positive on biopsy (the reference standard) will undergo surgical treatment for ovarian cancer.

- What is the role of the index test – replacement, triage, or add-on?
  - Replacement for triage tests
- All techniques were found to be equally sensitive and specific in detecting ovarian cancer compared with the reference standard. Do you think a diagnostic RCT is necessary in this case? Are patient outcomes necessary?
  - No diagnostic RCT are necessary because all positives identified from the 3 tests are similar
  - Not necessary because only positives from biopsy will get surgical treatment (management will not change based on the test results)
- What factors would you consider in developing recommendations to select the most appropriate imaging technique(s)?
  - Consider cost, side effects, burden to patient

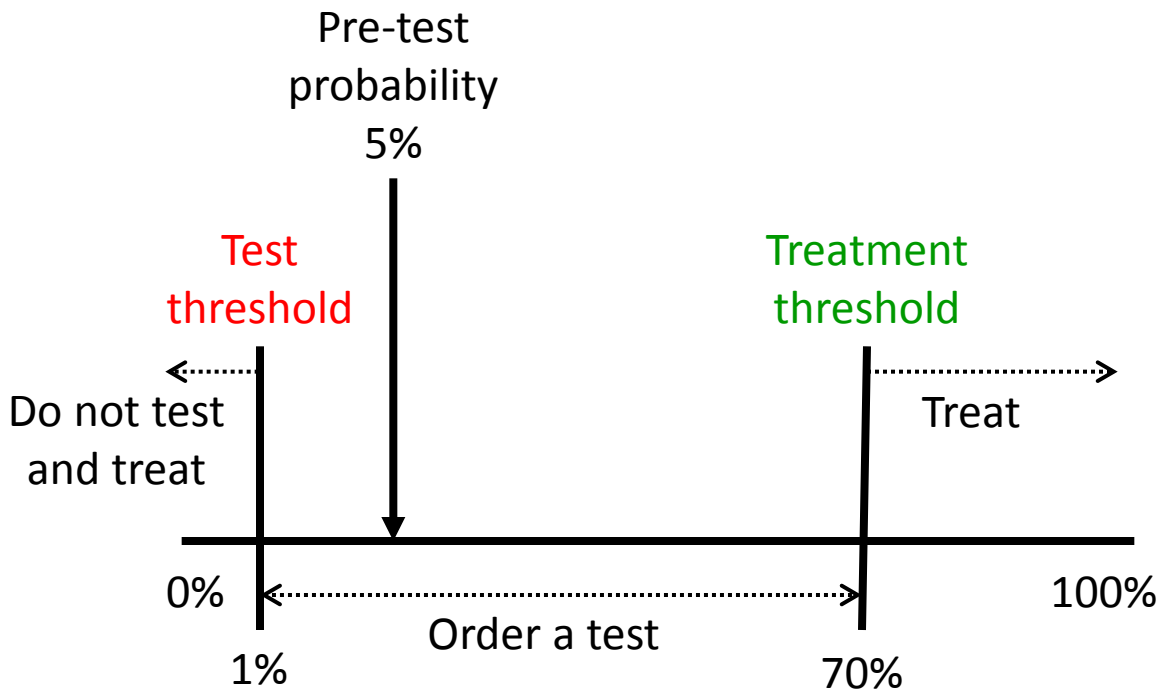
4. Generating Recommendations

- Test threshold = Probability of target condition is so low you wouldn't test and treat
- Treatment threshold = Probability at which you would start treatment
- Pre-test probability = Prevalence in the target population



Hsu J, et al. Application of GRADE: Making evidence-based recommendations about diagnostic tests in clinical practice guidelines. Implementation Science 2011, 6:62.

4. Generating Recommendations



Hsu J, et al. Application of GRADE: Making evidence-based recommendations about diagnostic tests in clinical practice guidelines. Implementation Science 2011, 6:62.

4. Generating Recommendations

Probabilities to determine before examining the evidence:

- Test threshold
  - Determine threshold to rule out target condition while considering potential consequences
- Treatment threshold
  - Determine threshold to rule in target condition while considering potential consequences
- Pre-test probability/prevalence
  - Determined through literature

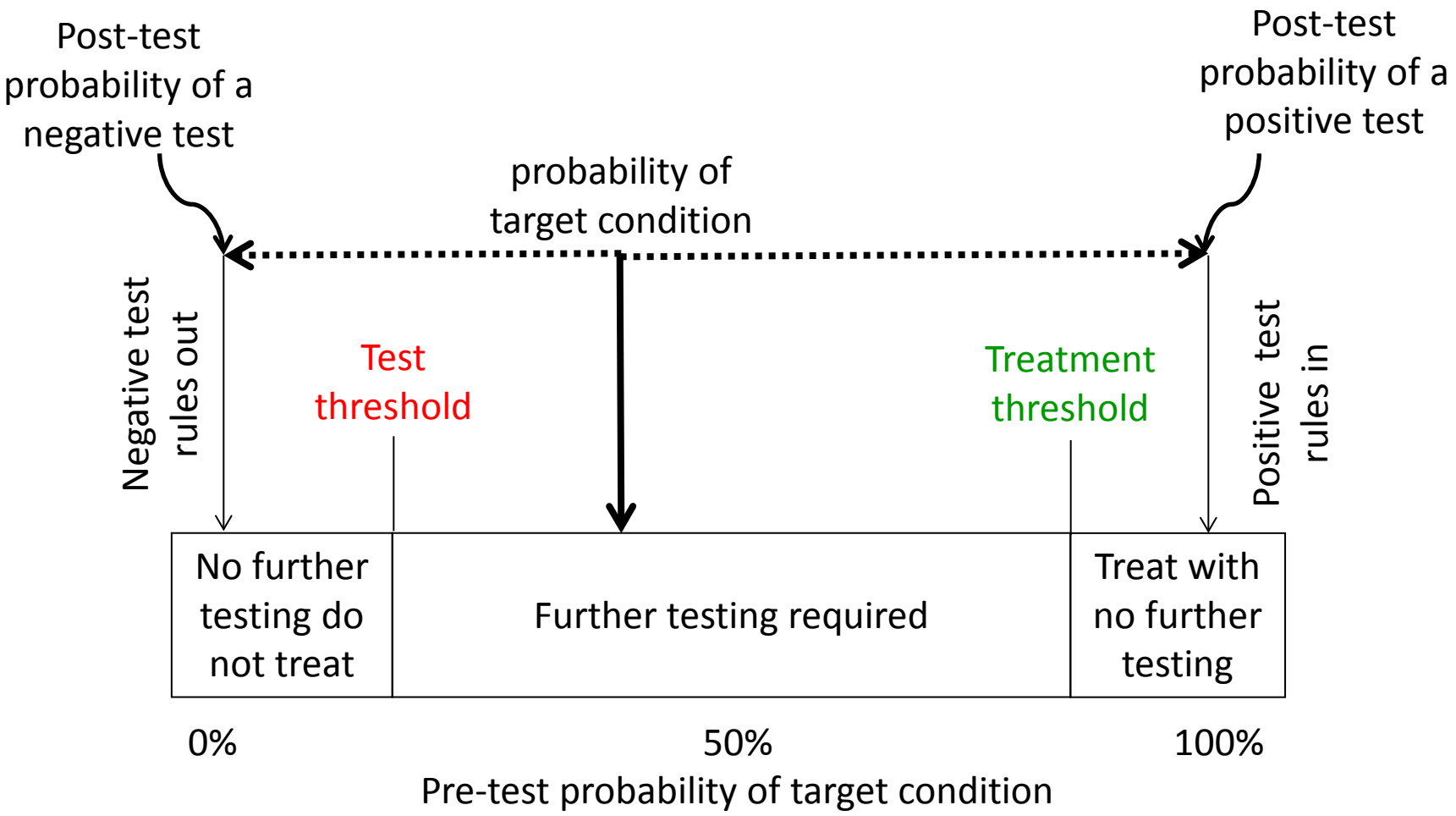
4. Generating Recommendations

Probabilities determined from evidence:

- Post test probability of a positive test
  - Derived from pre-test probability & positive likelihood ratio
  - Estimate of PPV
  - Should be higher than the treatment threshold
- Post test probability of a negative test
  - Derived from pre-test probability & negative likelihood ratio
  - Estimate of 1-NPV
  - Should be lower than the test threshold

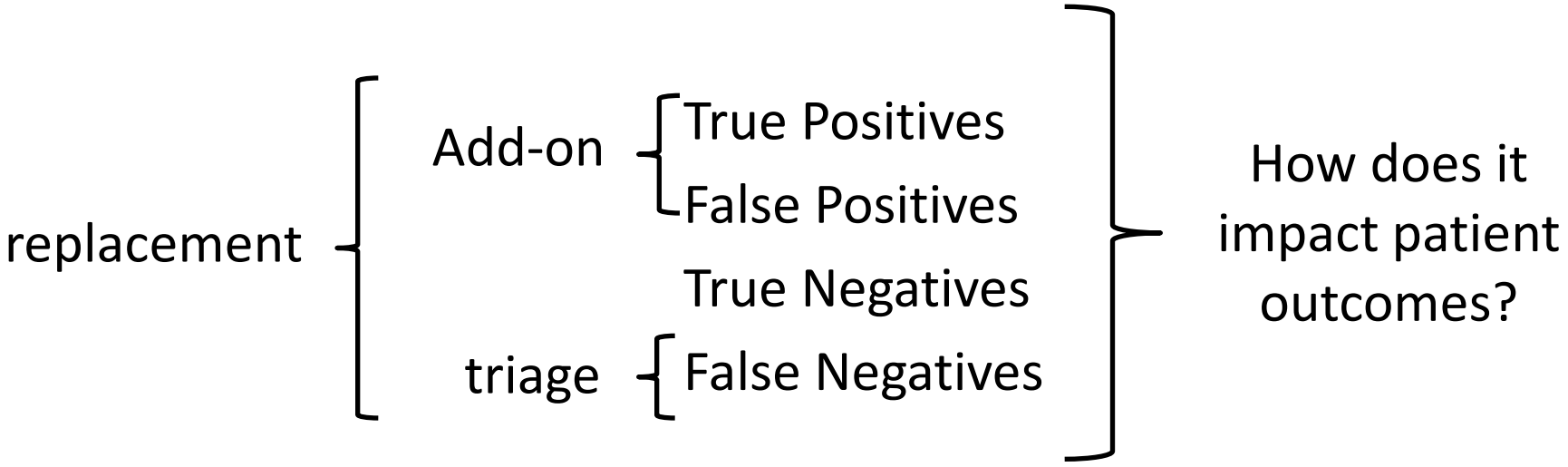
TP	FP
FN	TN

4. Generating Recommendations



4. Generating Recommendations

However, the post-test probabilities do not take into account the effect of the diagnostic result on patient outcomes.



Question

- Do you think determining the test/treatment thresholds is feasible?

4. Generating Recommendations

Why are patient outcomes important?

- Diagnostic tests' clinical value depends on whether they can provide useful information to improve patient outcomes.
- No matter the role of the index test is (replacement, triage, or add-on), test accuracy outcomes are surrogates for patient outcomes.

Lord SJ, et al. When is measuring sensitivity and specificity sufficient to evaluate a diagnostic test, and when do we need randomized trials? Ann Intern Med. 2006;144:850-5.

Schunemann HJ, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336:1106-10.

4. Generating Recommendations

When are patient outcomes not necessary?

- New test has similar summary sensitivity and specificity (with similar agreement) to old test, and other important attributes are not worse than old test (e.g., adverse events and cost)

4. Generating Recommendations

What should we do if patient outcomes are not available?

- The guideline panel has to make inferences about the likely impact on patient outcomes based on the available diagnostic test accuracy outcomes.

4. Generating Recommendations

What should we do if it's not possible to do a meta-analysis to get summary diagnostic accuracy outcomes?

- The guideline panel has to weigh the benefits versus the harms of a test
- Maybe only weak recommendations can be generated or no recommendations are possible

4. Generating Recommendations

Consider also the following outcomes:

- Consequences of inconclusive results
  - Ex. CT chest can lead to indeterminate results
- Complications of a new test and a reference standard
  - Ex. FOBT
- Resource use (cost)
  - Ex. MRI versus PET

Hsu J, et al. Application of GRADE: Making evidence-based recommendations about diagnostic tests in clinical practice guidelines. Implementation Science 2011, 6:62.

Take home messages

- Have research questions with PIRO components
- Confirm the spectrum of patients and the reference standard are appropriate
- Studies evaluating diagnostic tests can include test accuracy and more importantly patient outcomes
- Determine the role of the diagnostic test (replacement, triage, add-on)
- The role of the test will help you determine which outcomes are important
- Summary sensitivity with summary specificity, and post-test probabilities are two ways to interpret test accuracy outcomes