Different evidence for different questions

This session

- How to frame a diagnostic question
- Design diagnostic accuracy studies
- Design impact studies

- Helen, 67 years old
- No remarkable clinical history
- Sees her GP for discrete discomfort in the chest
- Not really painful
- Worsens with exercise
- Exercise ECG?

Formulate your question

- P
- I
- C
- O

- P
- I
- R
- T

PIRT

Let's exercise

- P — How would I describe a group of patients similar to mine?
- I — Which test am I considering?
- R — What is the reference standard considered to be ideal to diagnose the target condition?
- T — Which target condition/diagnosis do I want to either rule in or rule out?

melanoma patients and to determine the diagnostic value of subsequent PET/CT and MRI of the brain in these patients.

PATIENTS AND METHODS

Patients

Between August 2006 and March 2009, 46 melanoma patients without symptoms and signs of recurrent disease were referred for total body PET/CT and MRI of the brain because of an increased S-100B. The mean age of the patients was 59 years (range 25–93 years). Serum S-100B was monitored during follow-up after the surgical treatment of regional or distant metastases or because a patient was at increased risk due to primary tumor features (Table 1).

S-100B Analysis

The S-100B concentration was determined in serum using the Elecsys S100 assay, which is an electrochemiluminescence immunoassay (ECLIA) for the in vitro quantitative determination of S100 (S100 A1B and S100 BB) in human serum (Roche Diagnostics, Mannheim, Germany). The immunoassay ECLIA is intended for use on Elecsys and cobas e immunoassay analyzers as described in detail previously.[12] In our laboratory, the upper reference value of S-100B has been established at 0.10 μg/L. In cases of an increased S-100B level, sampling and measurement of the tumor marker were repeated for confirmation within a few days. Only patients in whom the repeat value was also increased were enrolled in the study.

FDG PET/CT

A hybrid PET/CT camera (Gemini II, Philips, Eindhoven, The Netherlands) was used, and FDG was administrated in dosages of 180–240 MBq (4.9–6.5 mCi). PET/CT scans were performed after fasting for 6 hours. The interval between FDG administration and scanning was 60 minutes ± 10 minutes. Low-dose CT images (40 mAs, 5 mm slices) were acquired without oral or intravenous contrast. Generated images were displayed using an Osirix Dicom viewer in a Unix-based operating system (MAC OS X, Power G5, Apple, Cupertino, CA) and were evaluated on the basis of two-dimensional orthogonal reslicing. PET was fused to low-dose CT after correction for attenuation. The PET/CT scans were reviewed by 3 experienced nuclear medicine physicians together.

MRI

MRI was performed with a high-field strength 3.0 T scanner (Achieva, Philips, Eindhoven, The Netherlands). The protocol consisted of precontrast transversal T2-weighted imaging, axial fluid attenuated inversion recovery (FLAIR) imaging, diffusion-weighted imaging and precontrast and postcontrast coronal T1-weighted 3D-FFE imaging.

Reference Standard

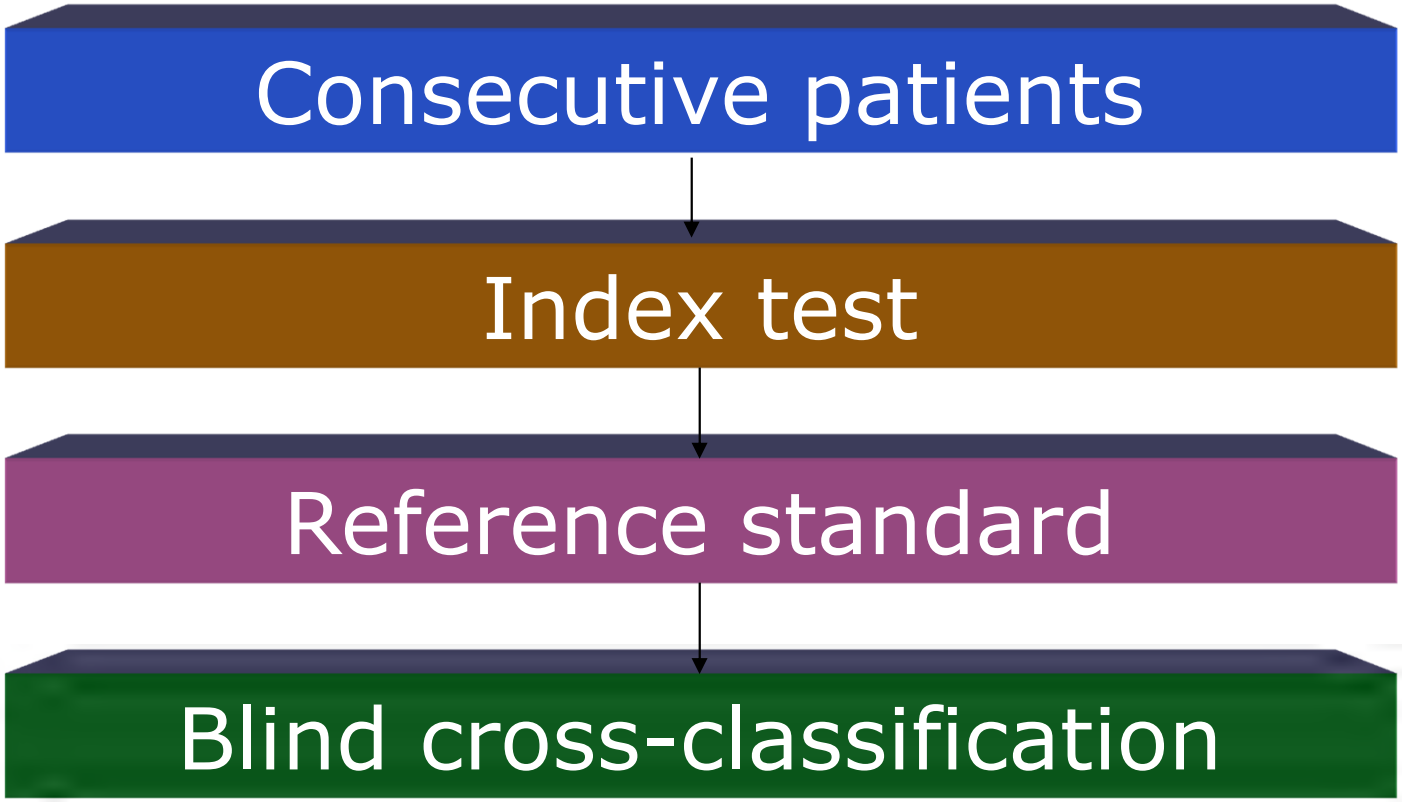
The presence or absence of melanoma recurrence was established by fine needle aspiration cytology or histological biopsy when possible. Additional imaging and the clinical course were used as the gold standard if no pathologic result could be obtained.

Statistical Analysis

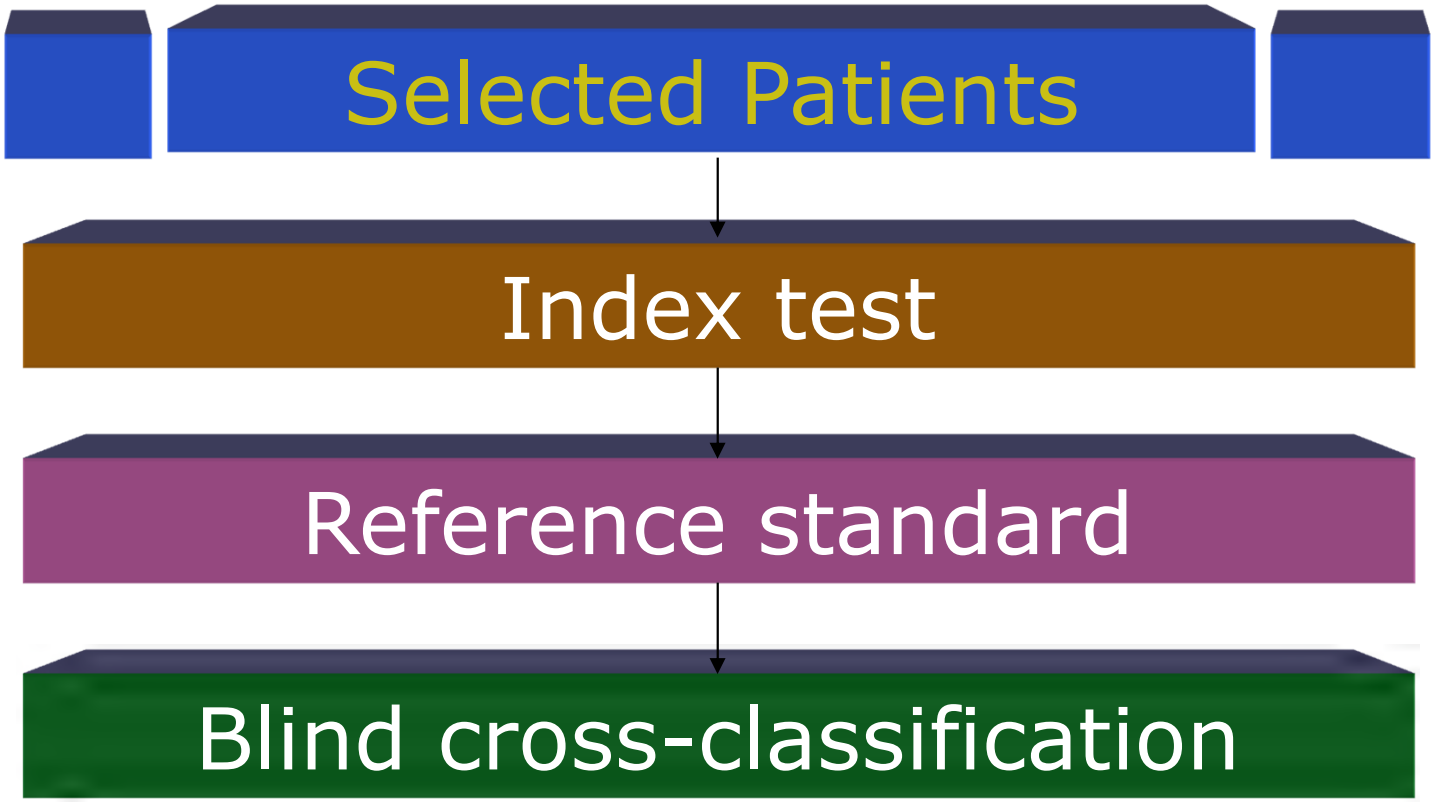
Statistical analyses were performed using SPSS 15 (Version 15, for Windows, SPSS Inc, Chicago, IL). The accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of PET/CT for the detection of local-regional recurrence or distant metastases were calculated using the standard definitions. Kaplan-Meier curves were used to analyze survival and were compared using a two-sided log-rank test. A difference was considered statistically significant if the associated *P* value was .05 or less.


Think
…..about a diagnostic question
What's your PIRT?
SCIENCEphotoLIBRARY

Diagnostic Accuracy Study: Basic Design



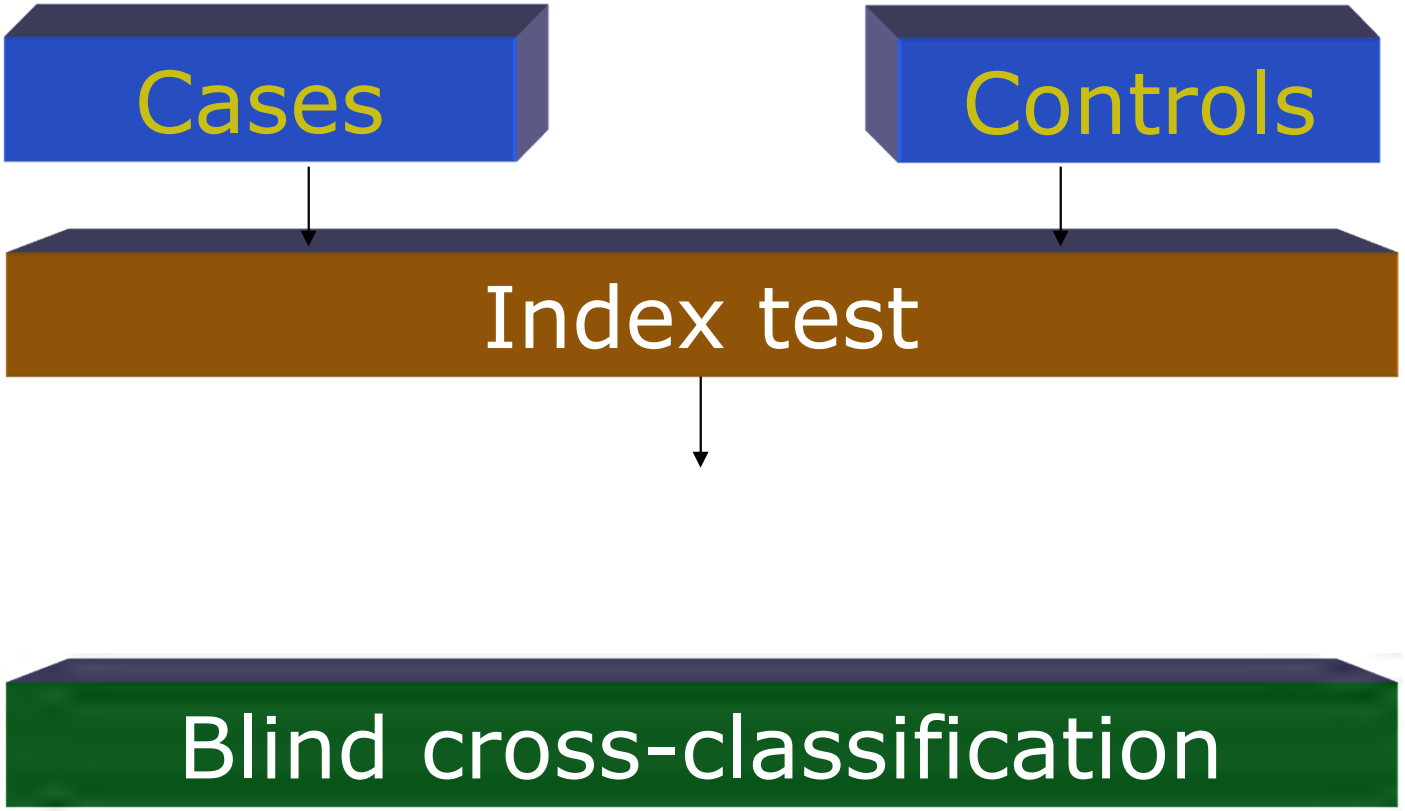
Spectrum and Selection Bias



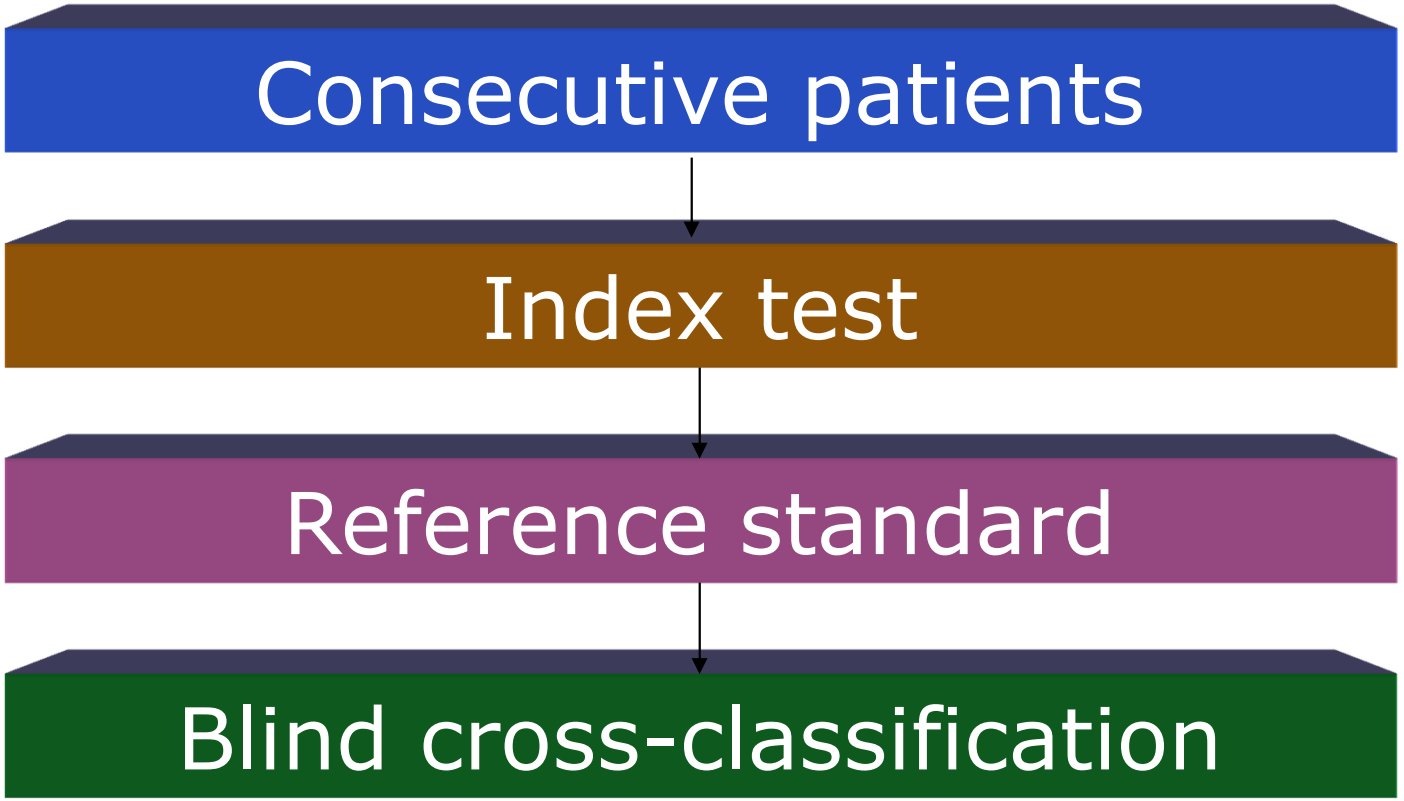

masterfile

Case-control vs consecutive

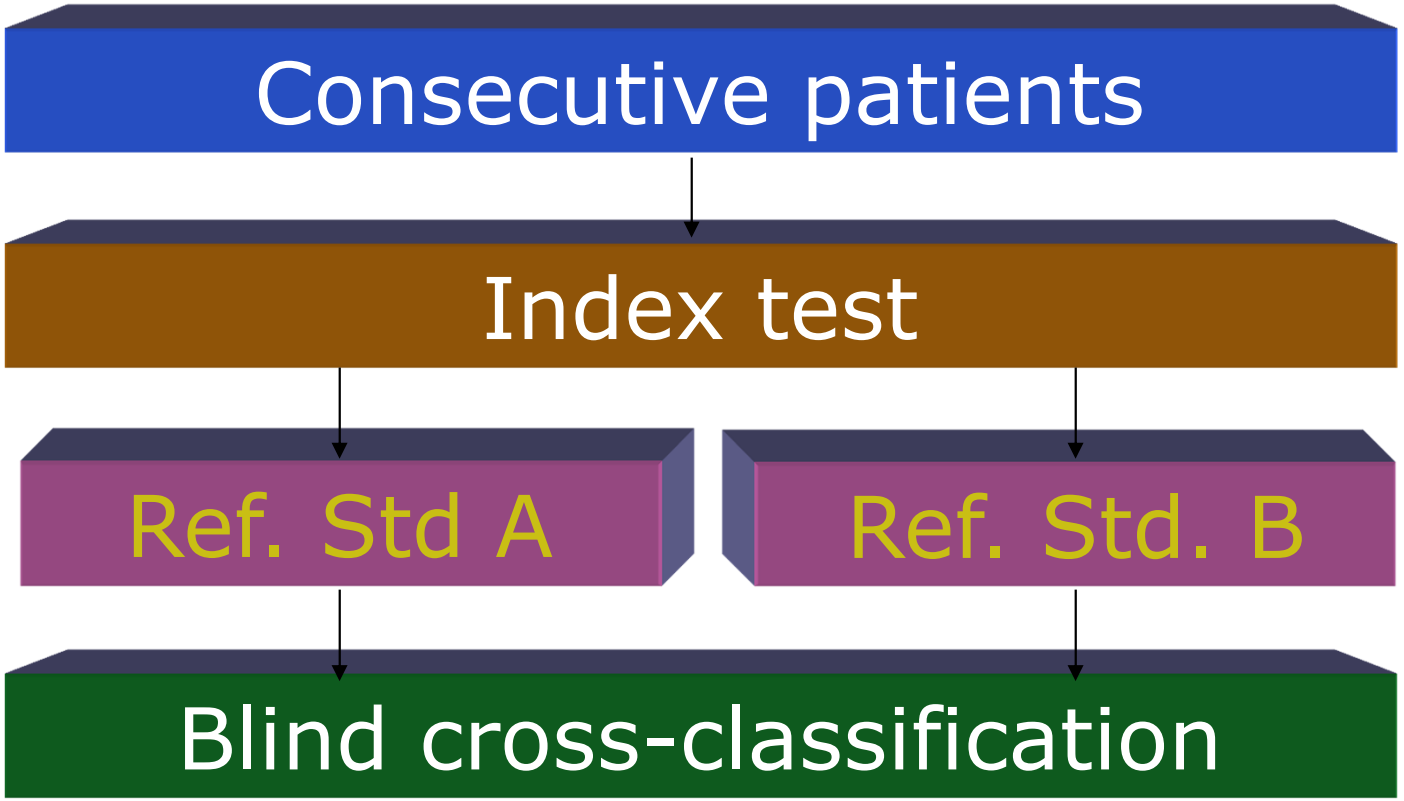
“Case-control” design



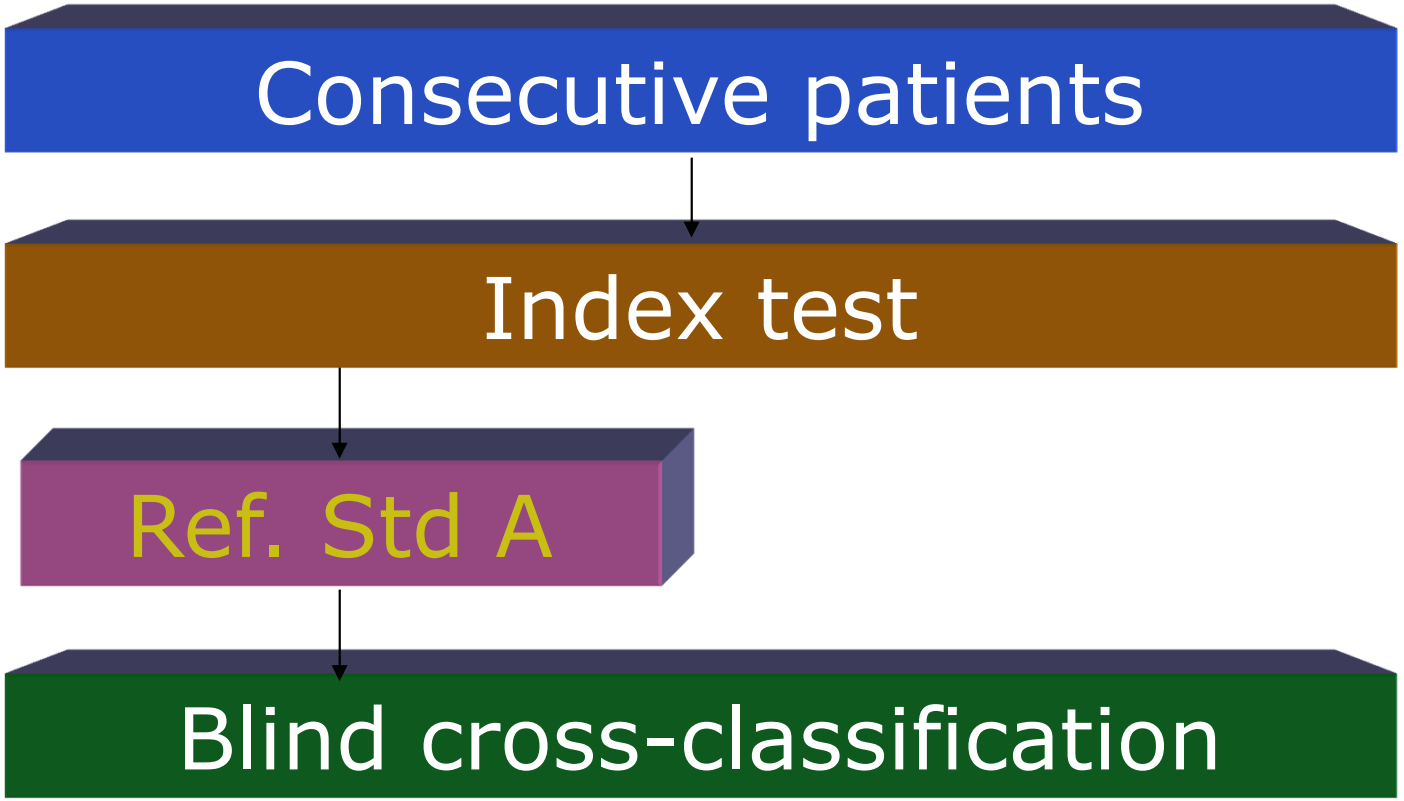
The ‘gold’ problem



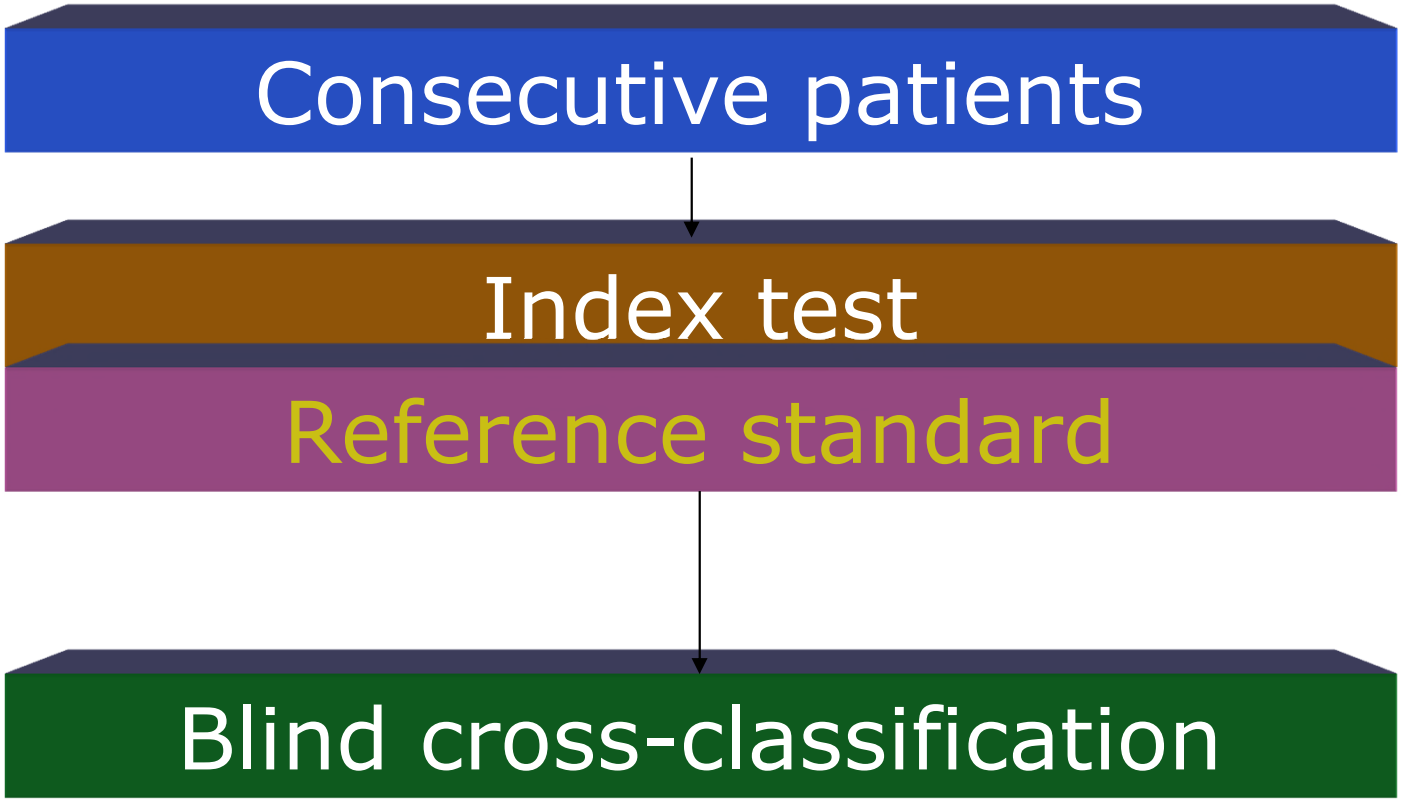
Differential Reference Bias



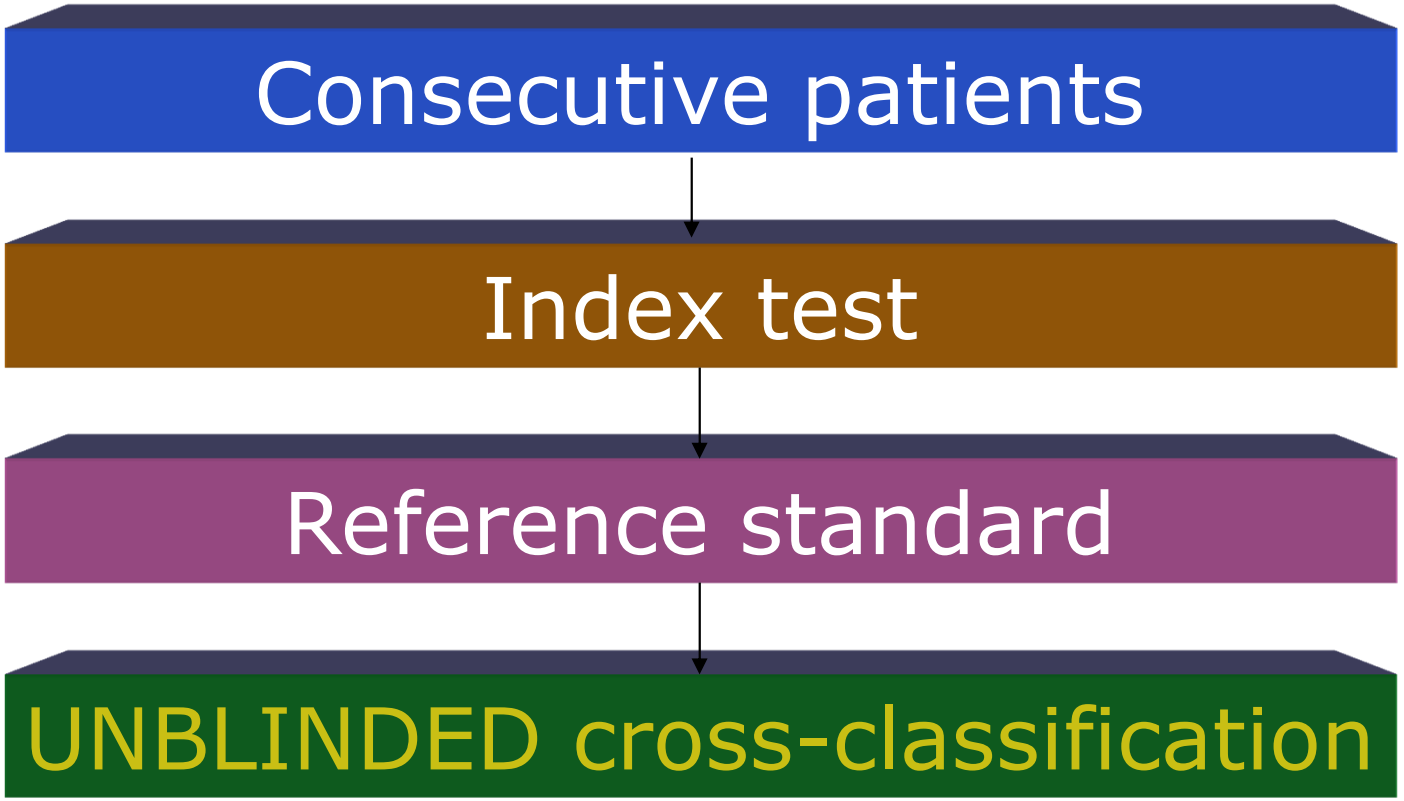
Partial Reference Bias



Incorporation bias



Observer bias



Quality published studies

Let's exercise

- Lijmer et al. JAMA 1999
- 218 studies

• Non-consecutive patient inclusion	56%
• Differential verification	22%
• Unblinded cross-classification	68%
• Unknown/retrospective data collection	55%

melanoma patients and to determine the diagnostic value of subsequent PET/CT and MRI of the brain in these patients.

PATIENTS AND METHODS

Patients

Between August 2006 and March 2009, 46 melanoma patients without symptoms and signs of recurrent disease were referred for total body PET/CT and MRI of the brain because of an increased S-100B. The mean age of the patients was 59 years (range 25–93 years). Serum S-100B was monitored during follow-up after the surgical treatment of regional or distant metastases or because a patient was at increased risk due to primary tumor features (Table 1).

S-100B Analysis

The S-100B concentration was determined in serum using the Elecsys S100 assay, which is an electrochemiluminescence immunoassay (ECLIA) for the in vitro quantitative determination of S100 (S100 A1B and S100 BB) in human serum (Roche Diagnostics, Mannheim, Germany). The immunoassay ECLIA is intended for use on Elecsys and cobas e immunoassay analyzers as described in detail previously.[12] In our laboratory, the upper reference value of S-100B has been established at 0.10 μg/L. In cases of an increased S-100B level, sampling and measurement of the tumor marker were repeated for confirmation within a few days. Only patients in whom the repeat value was also increased were enrolled in the study.

FDG PET/CT

A hybrid PET/CT camera (Gemini II, Philips, Eindhoven, The Netherlands) was used, and FDG was administrated in dosages of 180–240 MBq (4.9–6.5 mCi). PET/CT scans were performed after fasting for 6 hours. The interval between FDG administration and scanning was 60 minutes ± 10 minutes. Low-dose CT images (40 mAs, 5 mm slices) were acquired without oral or intravenous contrast. Generated images were displayed using an Osirix Dicom viewer in a Unix-based operating system (MAC OS X, Power G5, Apple, Cupertino, CA) and were evaluated on the basis of two-dimensional orthogonal reslicing. PET was fused to low-dose CT after correction for attenuation. The PET/CT scans were reviewed by 3 experienced nuclear medicine physicians together.

MRI

MRI was performed with a high-field strength 3.0 T scanner (Achieva, Philips, Eindhoven, The Netherlands). The protocol consisted of precontrast transversal T2-weighted imaging, axial fluid attenuated inversion recovery (FLAIR) imaging, diffusion-weighted imaging and precontrast and postcontrast coronal T1-weighted 3D-FFE imaging.

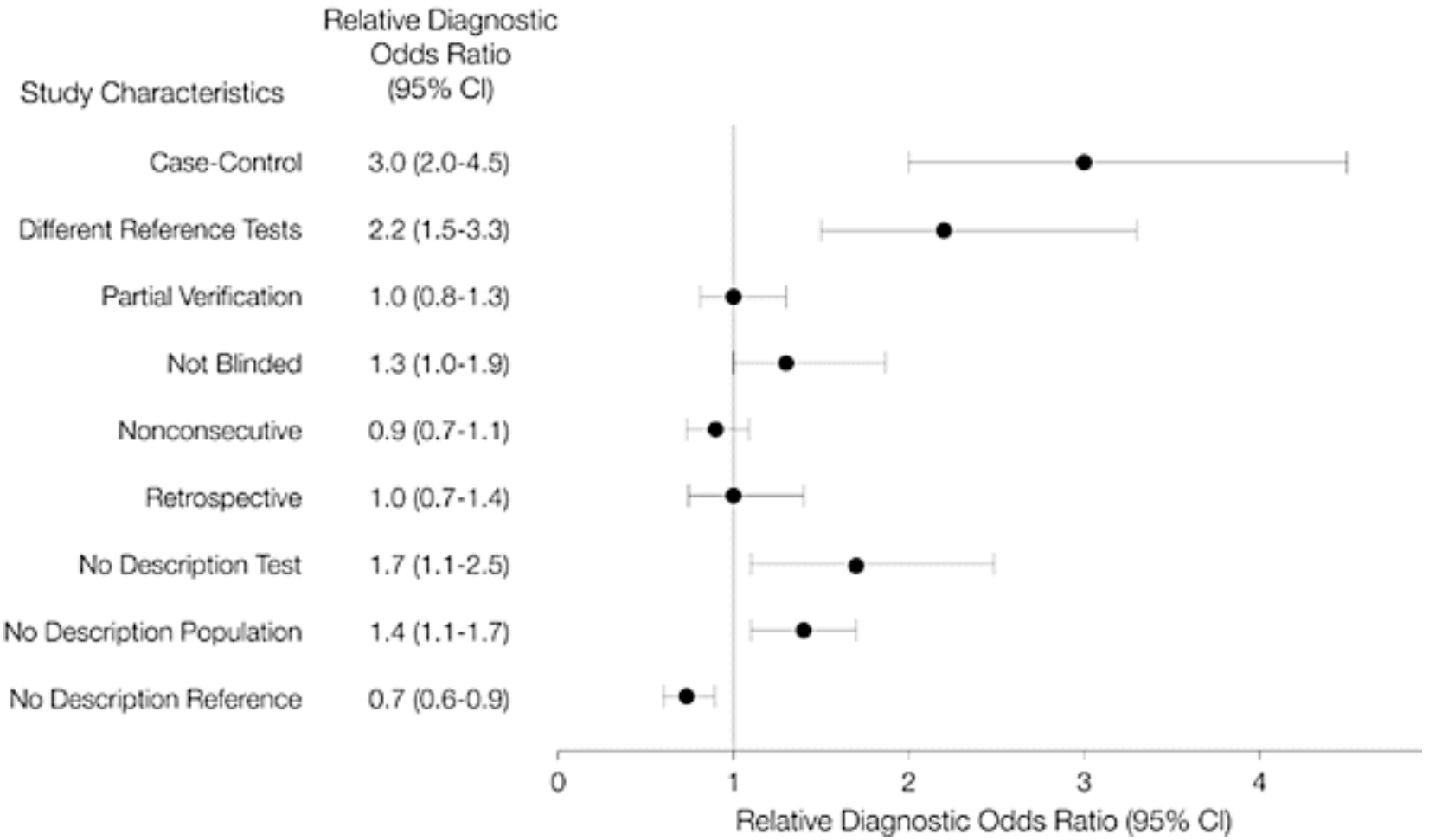
Reference Standard

The presence or absence of melanoma recurrence was established by fine needle aspiration cytology or histological biopsy when possible. Additional imaging and the clinical course were used as the gold standard if no pathologic result could be obtained.

Statistical Analysis

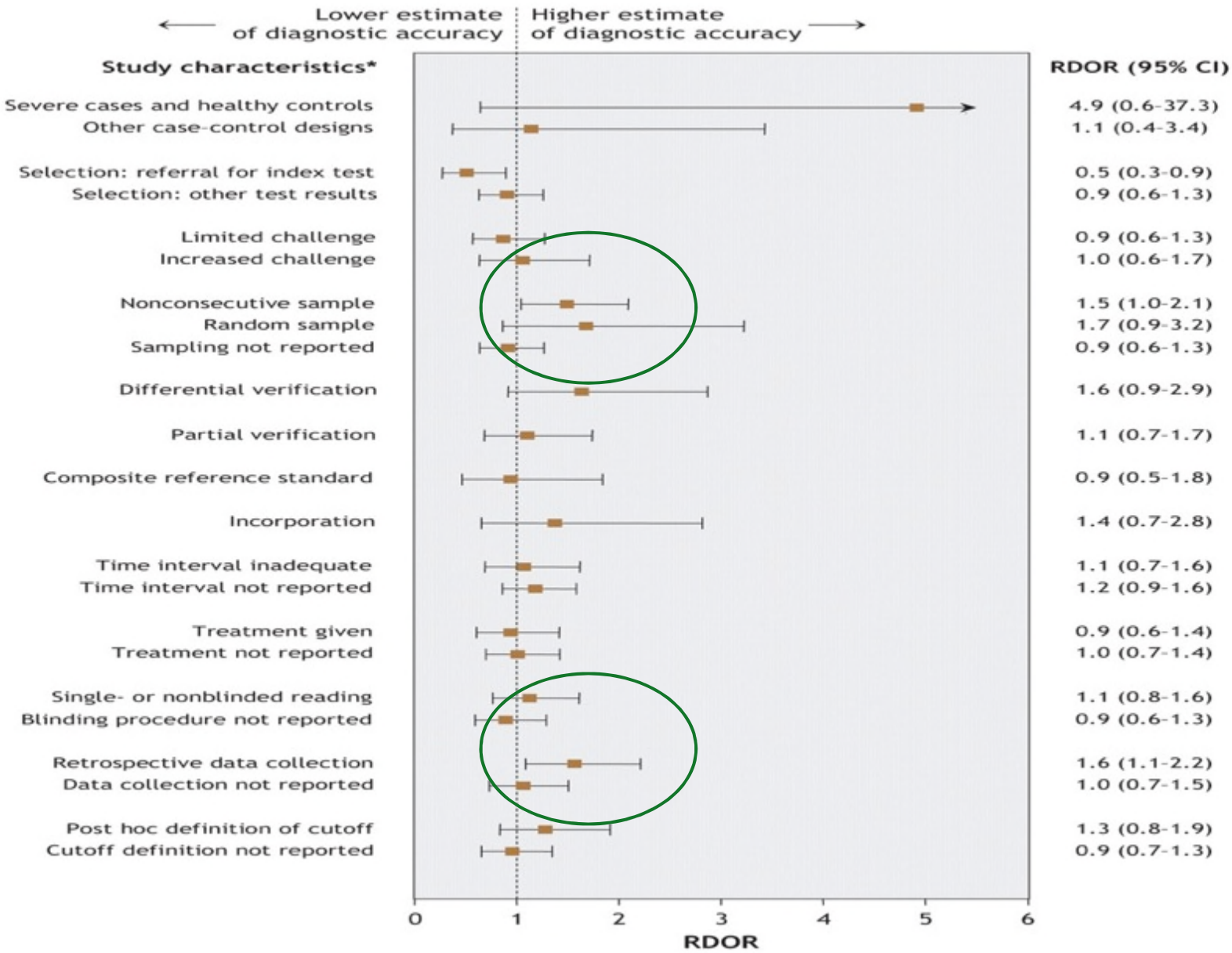
Statistical analyses were performed using SPSS 15 (Version 15, for Windows, SPSS Inc, Chicago, IL). The accuracy, sensitivity, specificity, positive predictive value, and negative predictive value of PET/CT for the detection of local-regional recurrence or distant metastases were calculated using the standard definitions. Kaplan-Meier curves were used to analyze survival and were compared using a two-sided log-rank test. A difference was considered statistically significant if the associated *P* value was .05 or less.

Relative Diagnostic Odds Ratios and 95% CIs of the 9 Study Characteristics.



Lijmer, J. G. et al. JAMA 1999;282:1061-1066

Effects of study design on diagnostic accuracy estimates



*See Appendix 2 for descriptions of the study characteristics.

Rutjes, A. W.S. et al. CMAJ 2006;174:469-476

Create your own diagnostic accuracy study

Checklist for diagnostic studies: QUADAS-2

http://www.bris.ac.uk/quadas/resources/quadas2.pdf

Phase 3: Risk of bias and applicability judgments

QUADAS-2 is structured so that 4 key domains are each rated in terms of the risk of bias and the concern regarding applicability to the research question (as defined above). Each key domain has a set of signalling questions to help reach the judgments regarding bias and applicability.

DOMAIN 1: PATIENT SELECTION

A. Risk of Bias

Describe methods of patient selection:

- ❖ Was a consecutive or random sample of patients enrolled? Yes/No/Unclear
- ❖ Was a case-control design avoided? Yes/No/Unclear
- ❖ Did the study avoid inappropriate exclusions? Yes/No/Unclear

Could the selection of patients have introduced bias? RISK: LOW/HIGH/UNCLEAR

B. Concerns regarding applicability

Describe included patients (prior testing, presentation, intended use of index test and setting):

Is there concern that the included patients do not match the review question? CONCERN: LOW/HIGH/UNCLEAR

DOMAIN 2: INDEX TEST(S)

If more than one index test was used, please complete for each test.

A. Risk of Bias

Describe the index test and how it was conducted and interpreted:

STARD Statement

STAndards for the Reporting of Diagnostic accuracy studies

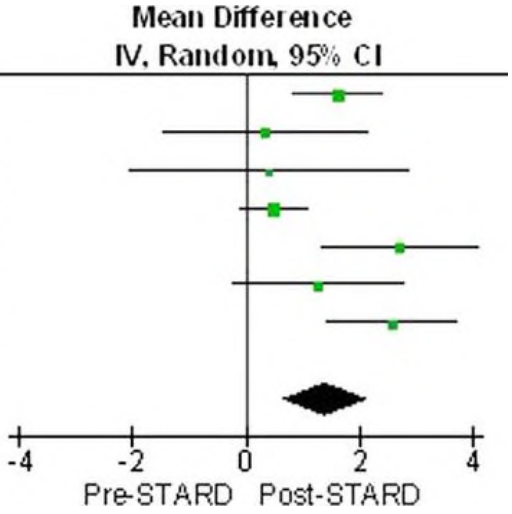
Section and Topic	Item #		On page #
TITLE/ABSTRACT/ KEYWORDS	1	Identify the article as a study of diagnostic accuracy (recommend MeSH heading 'sensitivity and specificity').	
INTRODUCTION	2	State the research questions or study aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups.	
METHODS			
Participants	3	Describe the study population: The inclusion and exclusion criteria, setting and locations where the data were collected.	
	4	Describe participant recruitment: Was recruitment based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?	
	5	Describe participant sampling: Was the study population a consecutive series of participants defined by the selection criteria in items 3 and 4? If not, specify how participants were further selected.	
	6	Describe data collection: Was data collection planned before the index test and reference standard were performed (prospective study) or after (retrospective study)?	
Test methods	7	Describe the reference standard and its rationale.	
	8	Describe technical specifications of material and methods involved including how and when measurements were taken, and/or cite references for index tests and reference standard.	

Forest plot for studies included in meta-analysis comparing adherence post-Standards for Reporting of Diagnostic Accuracy Studies (STARD) and pre-STARD.

	Post-STARD			Pre-STARD				Mean Difference	
Study or Subgroup	Mean	SD	Total	Mean	SD	Total	Weight	IV, Random, 95% CI	Year
Smidt 2006	13.55	3.2	141	11.94	3.3	124	19.4%	1.61 [0.82, 2.40]	2006
Coppus 2006	12.41	3.2	27	12.08	3.3	24	10.5%	0.33 [-1.46, 2.12]	2006
Rama 2006	15.5	3.3	19	15.1	3.3	11	7.0%	0.40 [-2.05, 2.85]	2006
Wilczynski 2008*	8.43	2.22	120	7.94	2.55	120	21.1%	0.49 [-0.11, 1.09]	2008
Areia 2010	14.3	2.9	24	11.6	3.6	86	13.5%	2.70 [1.31, 4.09]	2010
Selman 2011**	12.32	2.92	17	11.05	2.57	88	12.7%	1.27 [-0.22, 2.76]	2011
Selman 2011***	15.33	3.04	35	12.75	3.54	160	15.8%	2.58 [1.43, 3.73]	2011
Total (95% CI)			**383**			**613**	**100.0%**	**1.41 [0.65, 2.18]**	

Heterogeneity: Tau² = 0.63; Chi² = 17.70, df = 6 (P = 0.007); I² = 66%

Test for overall effect: Z = 3.61 (P = 0.0003)



Korevaar D A et al. Evid Based Med 2014;19:47-54

***Wilczynski evaluated only 13 STARD items; the other studies evaluated 25 STARD items. **Results of the studies on obstetrics. ***Results of the studies on gynaecology.**

- Helen, 67 years old
- No remarkable clinical history
- Sees her GP for discrete discomfort in the chest
- Not really painful
- Worsens with exercise
- Would an exercise ECG make it less likely Helen dies of a heart attack?

Formulate your question

- P
- I
- C
- O

- P
- I
- R
- T

PICO

Let's exercise

- P — How would I describe a group of patients similar to mine?
- I — Which test am I considering?
- C — What is the diagnostic strategy I would like to compare with?
- O — What are the outcomes that the new test could affect?

Abstract

Background: The Canadian CT Head Rule was developed to allow physicians to be more selective when ordering computed tomography (CT) imaging for patients with minor head injury. We sought to evaluate the effectiveness of implementing this validated decision rule at multiple emergency departments.

Methods: We conducted a matched-pair cluster-randomized trial that compared the outcomes of 4531 patients with minor head injury during two 12-month periods (before and after) at hospital emergency departments in Canada, six of which were randomly allocated as intervention sites and six as control sites. At the intervention sites, active strategies, including education, changes to policy and real-time reminders on radiologic requisitions were used to implement the Canadian CT Head Rule. The main outcome measure was referral for CT scan of the head.

Results: Baseline characteristics of patients were similar when comparing control to intervention sites. At the intervention sites, the proportion of patients referred for CT imaging increased from the "before" period (62.8%) to the "after" period (76.2%) (difference +13.3%, 95% CI 9.7%–17.0%). At the control sites, the proportion of CT imaging usage also increased, from 67.5% to 74.1% (difference +6.7%, 95% CI 2.6%–10.8%). The change in mean imaging rates from the "before" period to the "after" period for intervention versus control hospitals was not significant ($p = 0.16$). There were no missed brain injuries or adverse outcomes.

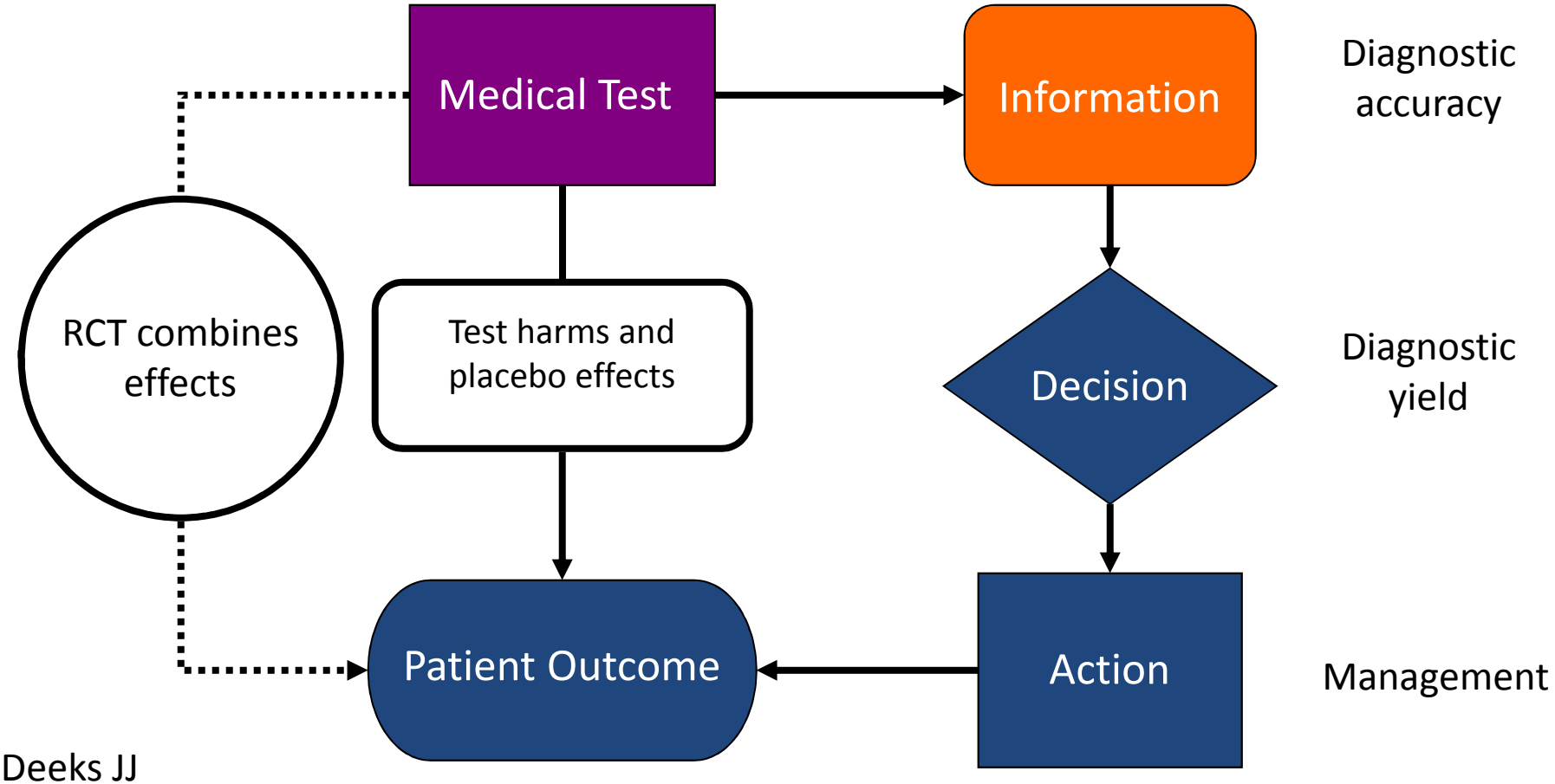
Interpretation: Our knowledge–translation-based trial of the Canadian CT Head Rule did not reduce rates of CT imaging in Canadian emergency departments. Future studies should identify strategies to deal with barriers to implementation of this decision rule and explore more effective approaches to knowledge translation. (ClinicalTrials.gov trial register no. NCT00993252)

Stiell CMAJ 2010

Studying impact of tests

- On
  - Patient outcome
  - Costs
  - Organisation of care
- Designs:
  - RCT
  - Before-after trial
  - Modelling

What is being evaluated?



Deeks JJ

Indications for diagnostic trials

- Tests detect disease earlier (screening and case-finding)
- Test itself has a harmful effect
- Interventions have harmful effects
  - Treating some non-diseased may outweigh benefits of treating diseased
- No reference standard

- Rare goods:
  - Only 37 (95% CI 35-40) diagnostic test strategies RCTs on patient outcomes per year.
  - 21,949 per year for all RCTs indexed in CENTRAL.

What is being evaluated?

Trial finds no difference:

????

- **Conditions for a test to be of diagnostic benef**
  - Test is more accurate
  - Interpretation of test results is rational and consistent
  - Management is rational and consistent
  - Treatment is effective

- **Conditions for a trial to be informative**
  - Rules for interpretation of test results are described
  - Management protocol is described

- **No descriptions given in example trials**
  - Applying the results requires faith that the behaviour of your patients and clinicians is the same as the trial

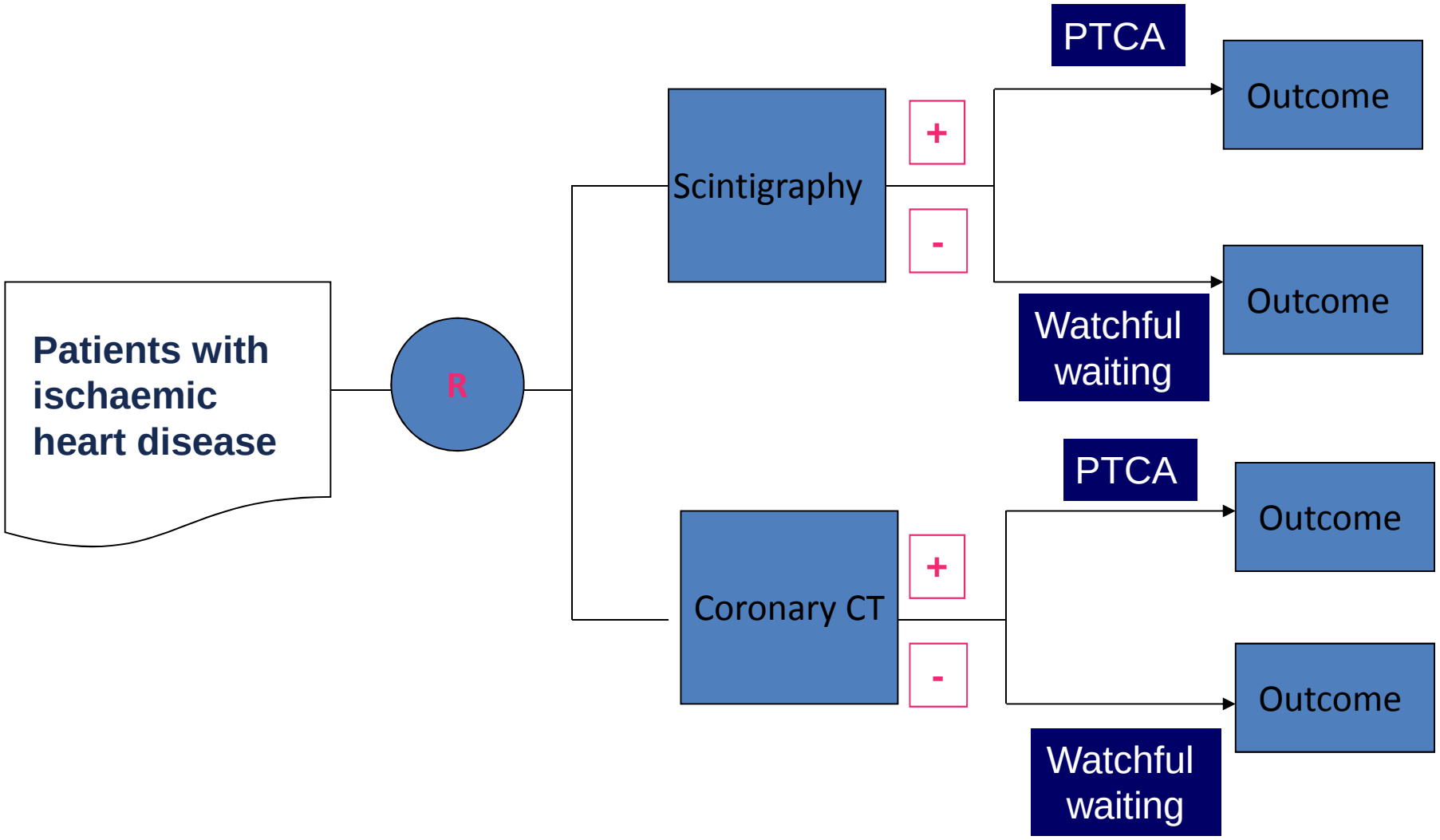
Clinically important differences

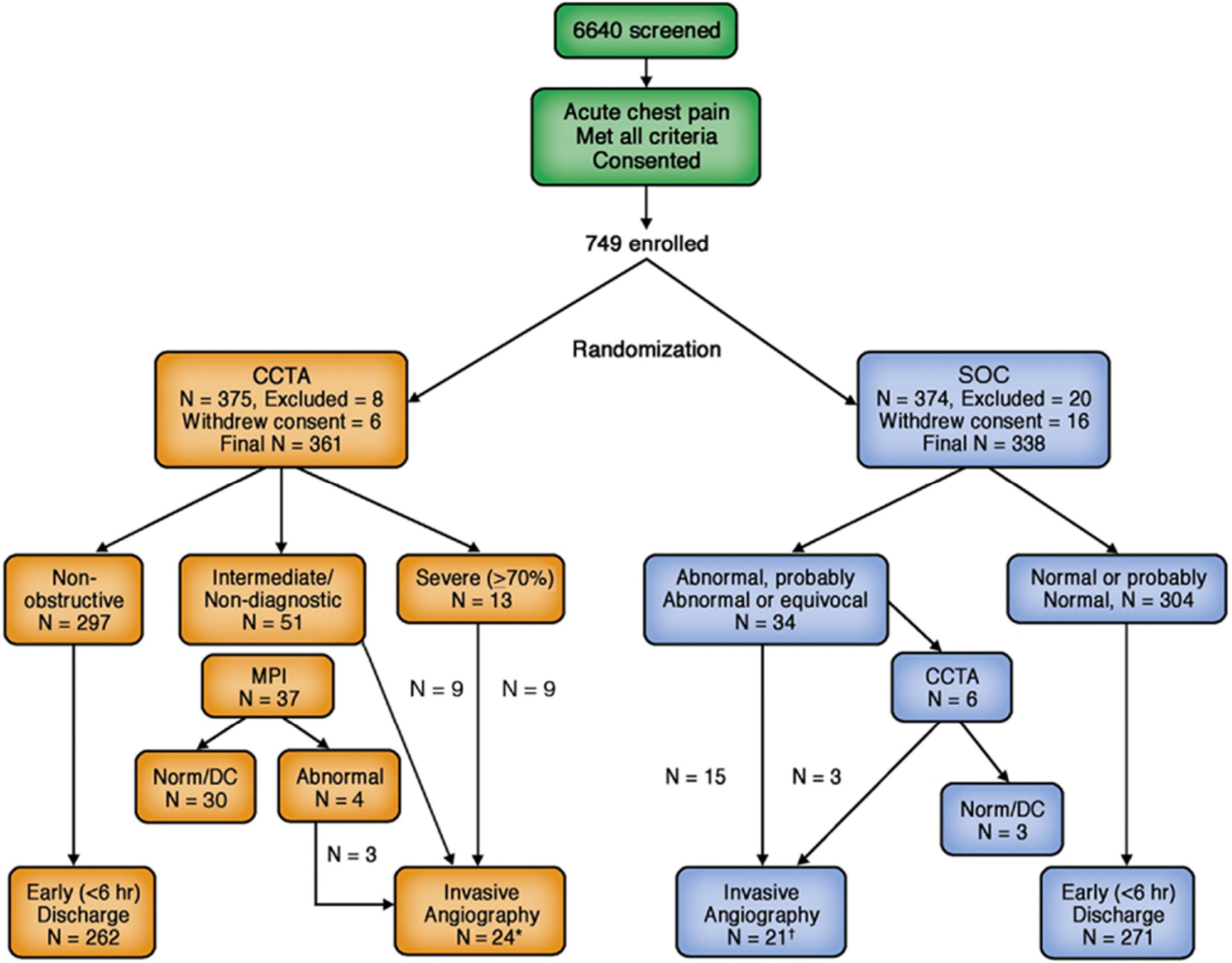
Test delivery	Timing of test
	Feasibility
	Test process
Test result	Interpretability
	Accuracy
	Timing of results
Diagnostic decision	Timing of diagnosis
	Diagnostic confidence
Treatment decision	Therapeutic yield
	Therapeutic confidence
Treatment implementation	Time to treatment
	Efficacy of treatment
	Adherence to treatment

Imagine… direct impact?

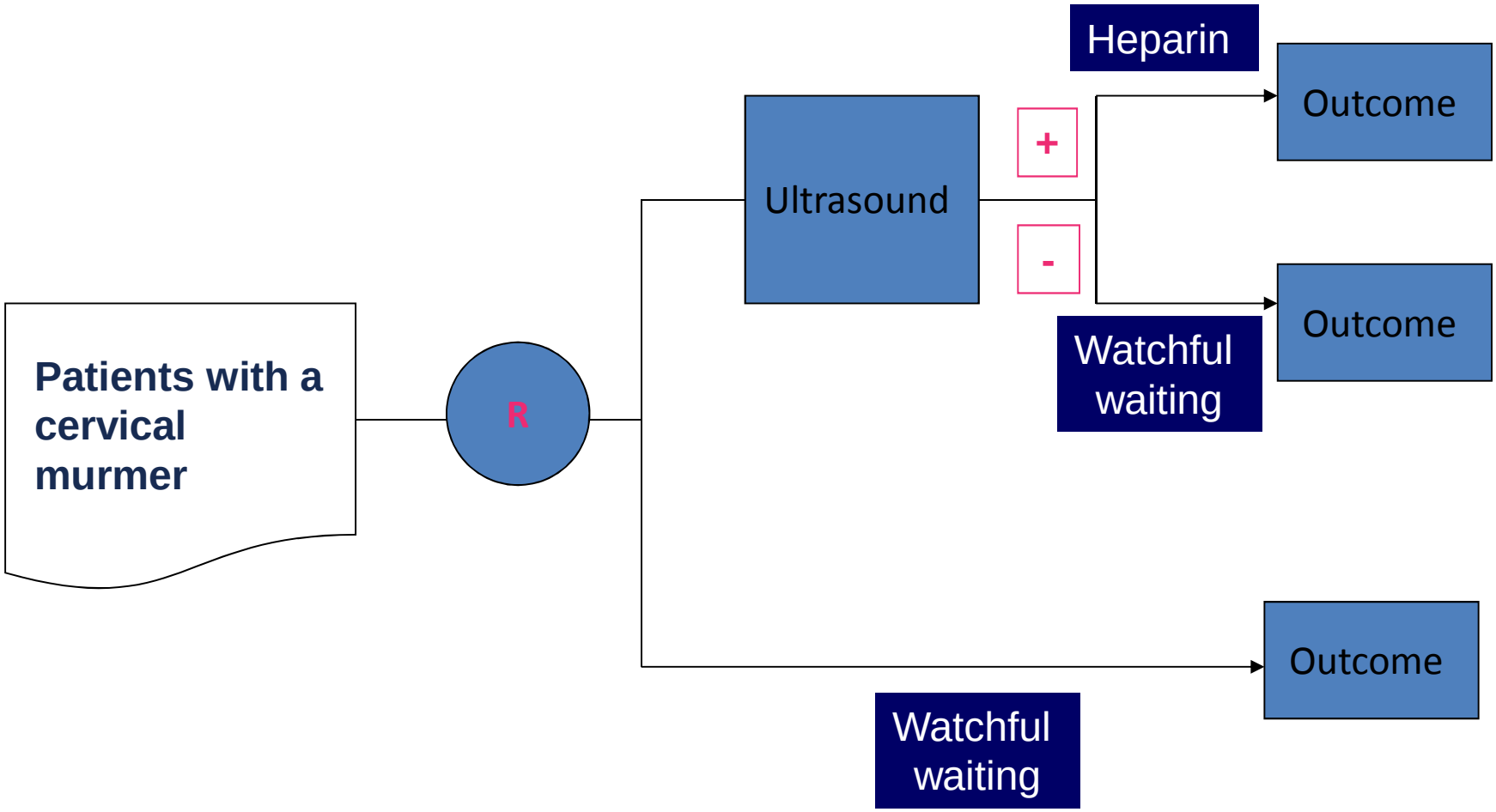
Effects of testing	What this means	Effects on health
Emotional	Test causes harmful or beneficial changes in levels of anxiety, depression, stress, psychological well being.	Increased anxiety and stress occurring following a positive test on screening that has not been confirmed with a reference standard. Reassurance and improved overall well-being after a negative test.
Social	Effects of testing on social roles, social functions, sexual relationships, social relationships.	Social isolation and stigmatisation after a positive test. Problems with employment or insurance coverage. Genetic testing results may cause guilt about passing on a genetic predisposition.
Cognitive	Patients' beliefs, perceptions and understanding about the test result and the condition.	May understand disease better – what causes it, how long it lasts etc., or affect adherence to therapy.
Behavioural	The combinations of emotional, social and cognitive effects can affect patient behaviour. Positive and negative tests can prompt change in behaviour.	Adherence to clinical intervention may be increased or decreased. Greater or less engagement with other health related behaviours, e.g. increased exercise after having cholesterol measured. Perceptions of risks from screening and repeated screening.

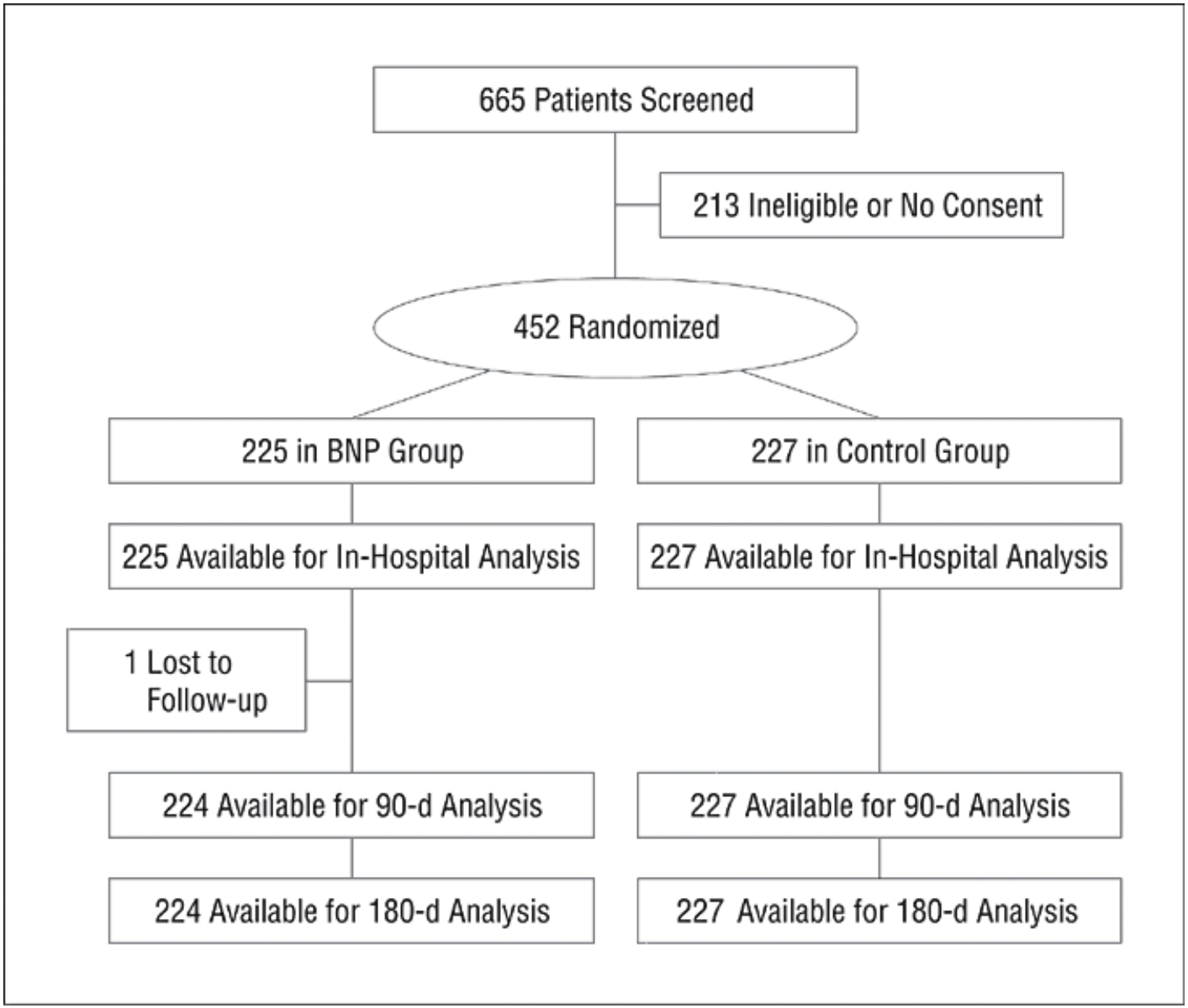
RCT architecture



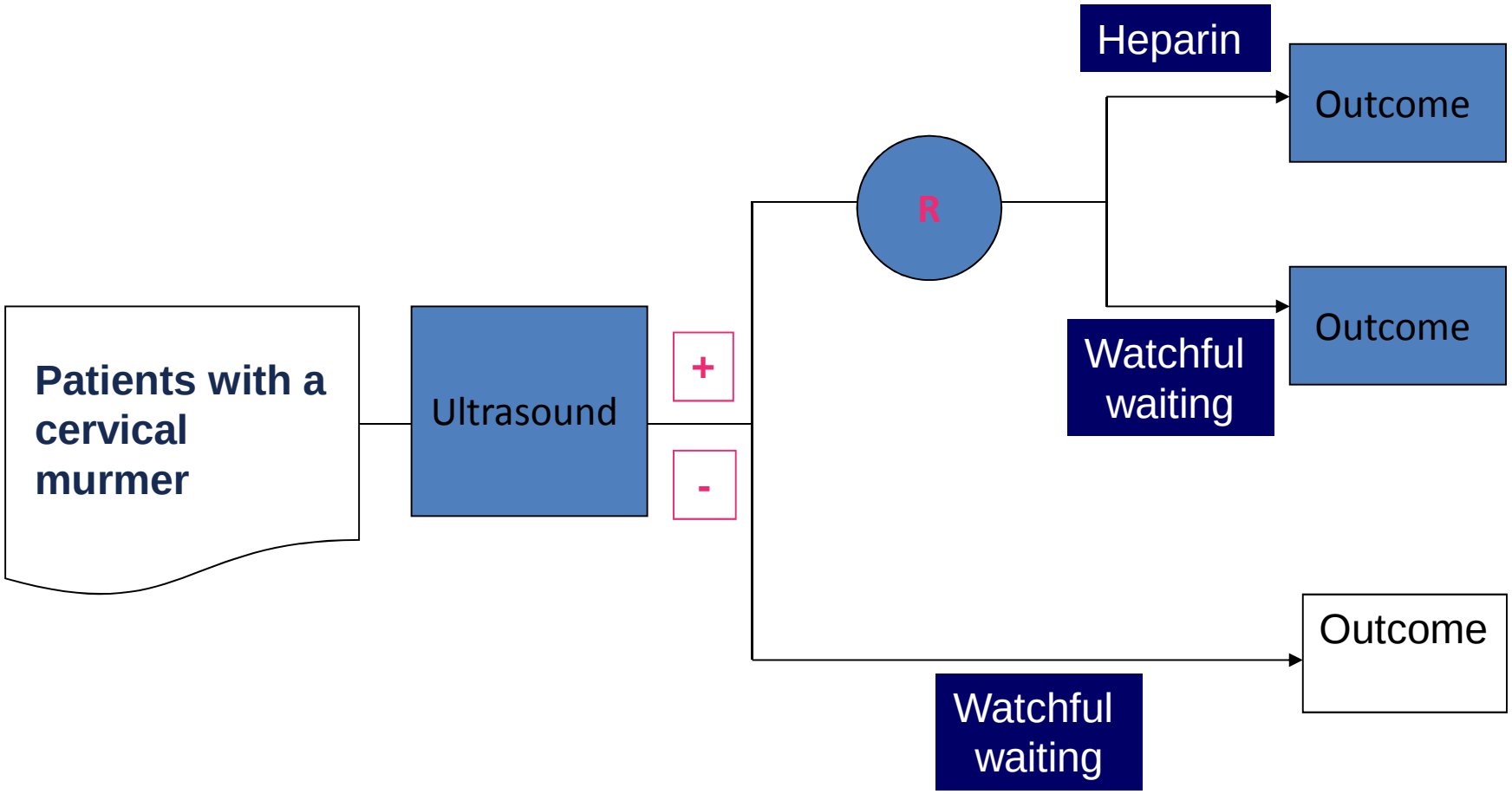

6640 screened
Acute chest pain
Met all criteria
Consented
749 enrolled
Randomization
CCTA
N = 375, Excluded = 8
Withdrew consent = 6
Final N = 361
SOC
N = 374, Excluded = 20
Withdrew consent = 16
Final N = 338
Non-obstructive
N = 297
Intermediate/
Non-diagnostic
N = 51
Severe (≥70%)
N = 13
MPI
N = 37
Norm/DC
N = 30
Abnormal
N = 4
N = 3
N = 9
N = 9
Early (<6 hr)
Discharge
N = 262
Invasive
Angiography
N = 24*
Abnormal, probably
Abnormal or equivocal
N = 34
Normal or probably
Normal, N = 304
CCTA
N = 6
N = 15
N = 3
Norm/DC
N = 3
Invasive
Angiography
N = 21†
Early (<6 hr)
Discharge
N = 271

RCT architecture



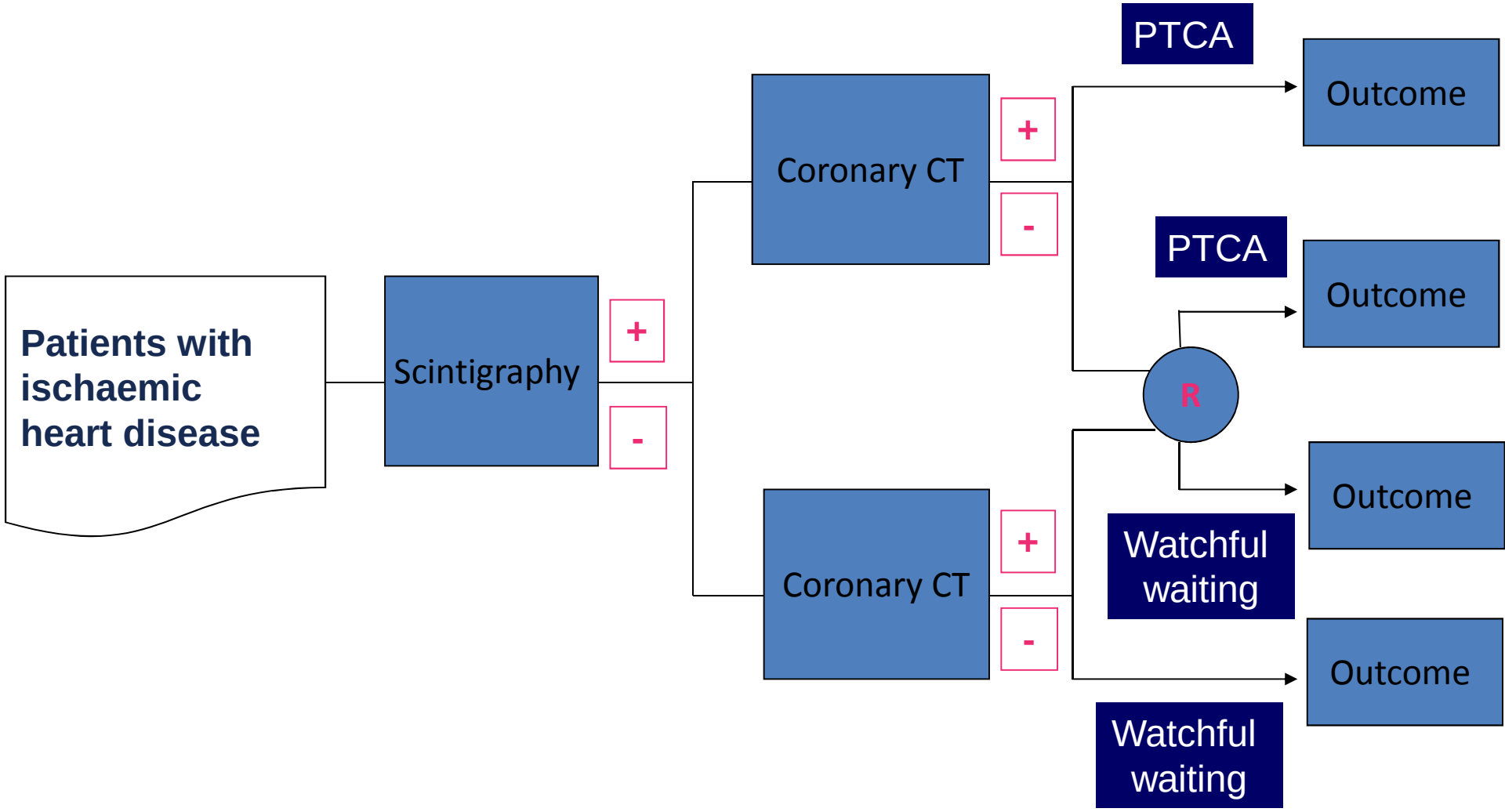

665 Patients Screened
213 Ineligible or No Consent
452 Randomized
225 in BNP Group
227 in Control Group
225 Available for In-Hospital Analysis
227 Available for In-Hospital Analysis
1 Lost to Follow-up
224 Available for 90-d Analysis
227 Available for 90-d Analysis
224 Available for 180-d Analysis
227 Available for 180-d Analysis

RCT architecture



RCT architecture

Let's exercise



Methods

This study was a pragmatic, cluster randomised, factorial, controlled trial. While recognising certain limitations,[39] we chose a cluster randomisation design to optimise the pragmatic nature of the study and to minimise contamination: once general practitioners within a practice had been trained in new communication skills they could not switch at random between using these skills and usual consulting practice. A 2×2 factorial design was used to assess the effect of each intervention and to explore the effect of the interventions combined.[40] Such trials require a prespecified factorial analysis plan with assessments for treatment interactions. We selected this design because we planned to test two treatment hypotheses. The four allocated groups were general practitioners' use of C reactive protein testing (1), training in enhanced communication skills (2), the interventions combined (3), and usual care (4). The groups were combined for analysis as follows: factor A, C reactive protein test (cells 1 and 3) compared with no test (2 and 4) (controlling for the effect of general practitioners' training in enhanced communication skills in the model); and factor B, training in enhanced communication skills (2 and 3) compared with no training (1 and 4) (controlling for the effects of C reactive protein testing in the model).

Outcomes, sample size, and randomisation

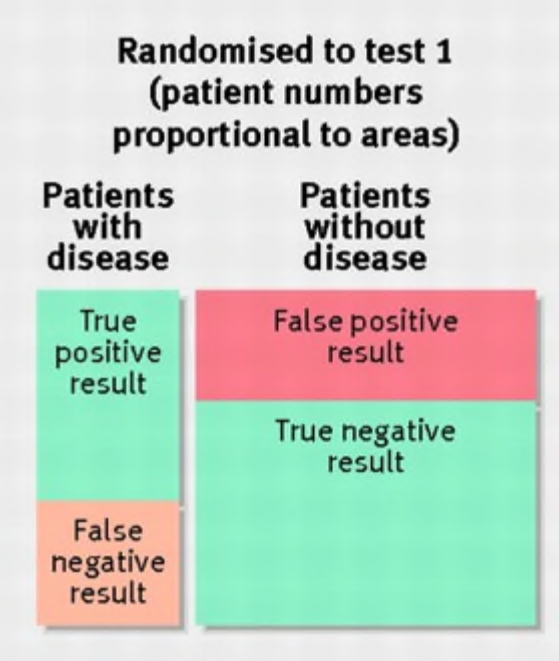
The primary outcome was antibiotic prescribing in the index consultation. Our study required 400 patients with lower respiratory tract infection to detect a reduction in antibiotic prescribing from 80% to 60% (power 80%, α 0.05, follow-up 90%) when adjusted for clustering at practice level (intracluster coefficient 0.06). The sample size was for the main effects only and assumed no interaction between the two interventions. Secondary outcomes were antibiotic prescribing during 28 days' follow-up, reconsultation, clinical recovery, and patients' satisfaction and enablement. Cost effectiveness will be reported separately. We planned to recruit 20 general practices with two participating general practitioners per practice within a large suburban region of the Netherlands. All practices and general practitioners were recruited and provided written consent before randomisation.

Practices were randomised into two groups of 10 practices per intervention, balanced for recruitment potential, resulting in four trial arms (fig 1⇓). The balancing factor used for randomisation was the amount of general practitioners' consultation time (expressed as full time equivalent) that the practice was contributing to the study, and this equated to between one and two full time equivalents for clinical contact time. The randomisation was balanced for those with 1.5 or less full time equivalents and those with more than 1.5 full time equivalents. The Dutch guideline for managing acute cough, including diagnostic and therapeutic advice for lower respiratory tract infection, is distributed to all general practitioners in the Netherlands and informs usual care.[41]

Validity Concerns

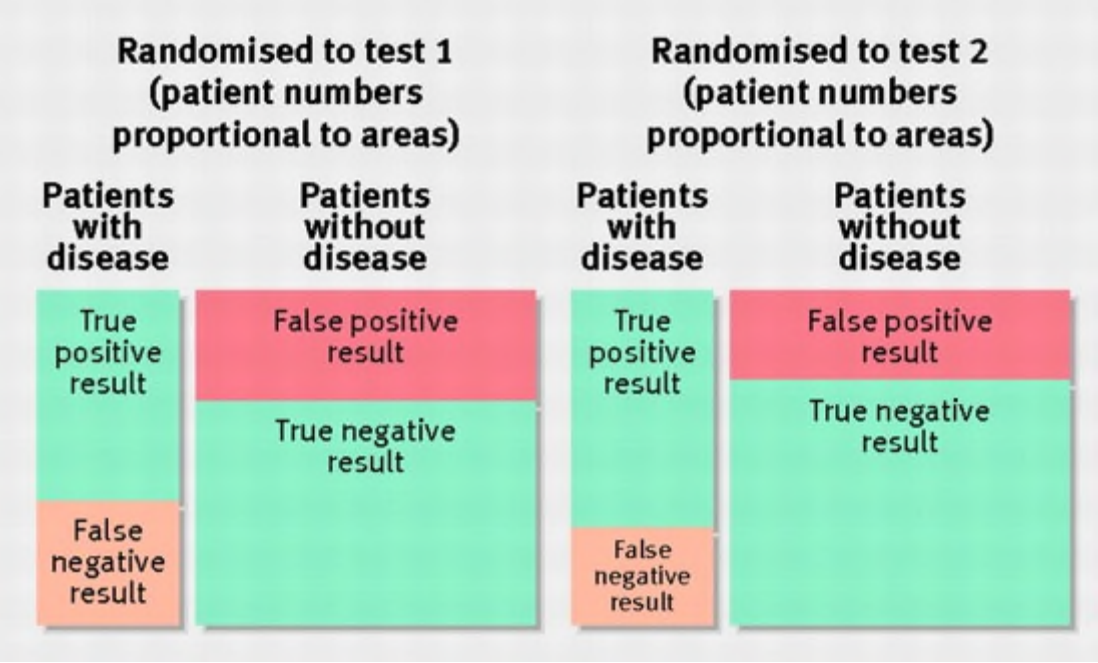
- **Blinding**
  - Rare in diagnostic trials (cluster randomisation!)
- **Drop-out**
  - Lack of blinding can induce differential drop-out
  - More stages at which drop-out occurs
- **Compliance**
  - Lack of blinding and complexity in strategies can reduce compliance
- **Power calculations**

Sample size calculations for test-treatment randomised controlled trials.



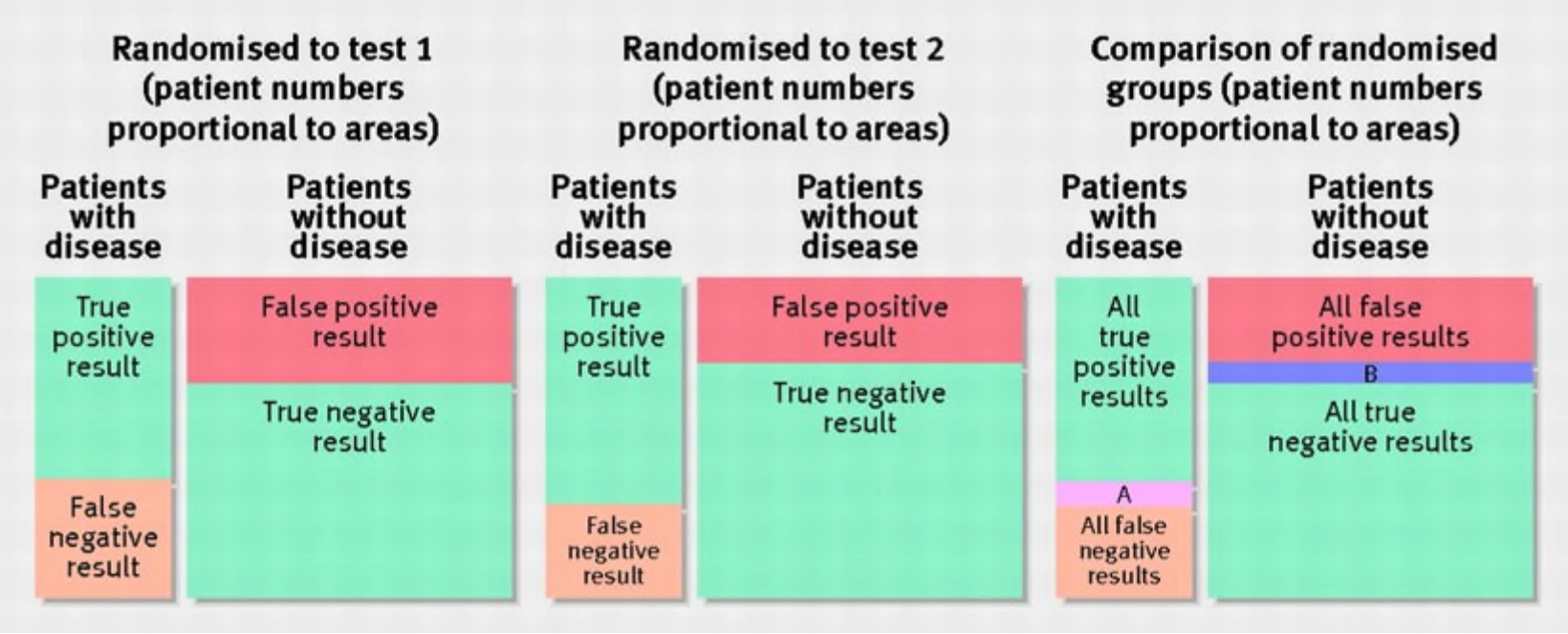
Ferrante di Ruffano L et al. BMJ 2012;344:bmj.e686

Sample size calculations for test-treatment randomised controlled trials.



Ferrante di Ruffano L et al. BMJ 2012;344:bmj.e686

Sample size calculations for test-treatment randomised controlled trials.



Ferrante di Ruffano L et al. BMJ 2012;344:bmj.e686

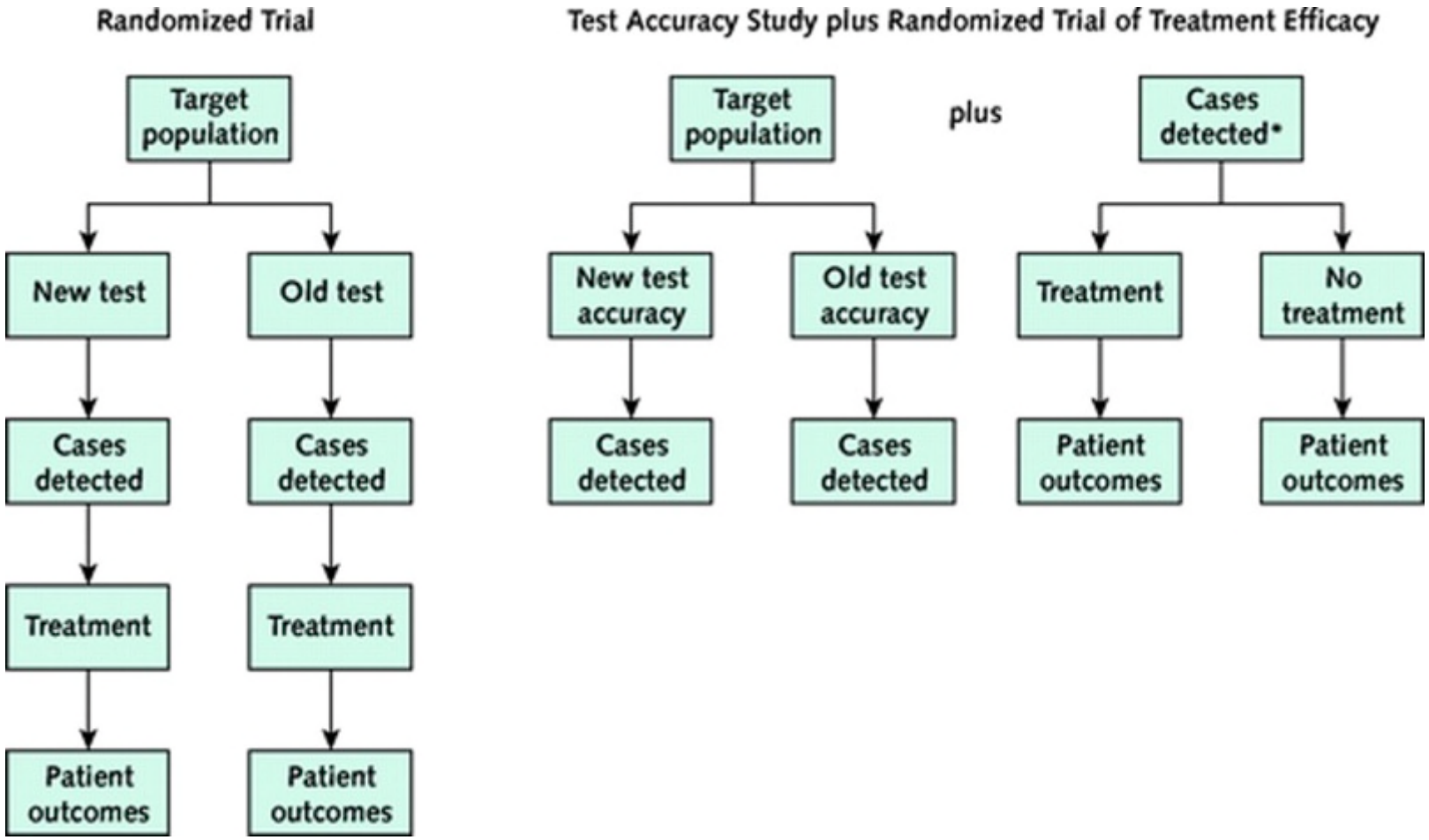
Create your own trial

Modelling

- New test affects patient outcome?
- Only diagnostic accuracy studies
- No trials

- ➔ model impact on patient outcome

Trial evidence versus linked evidence of test accuracy and treatment efficacy



Lord, S. J. et. al. Ann Intern Med 2006;144:850-855

Annals of Internal Medicine

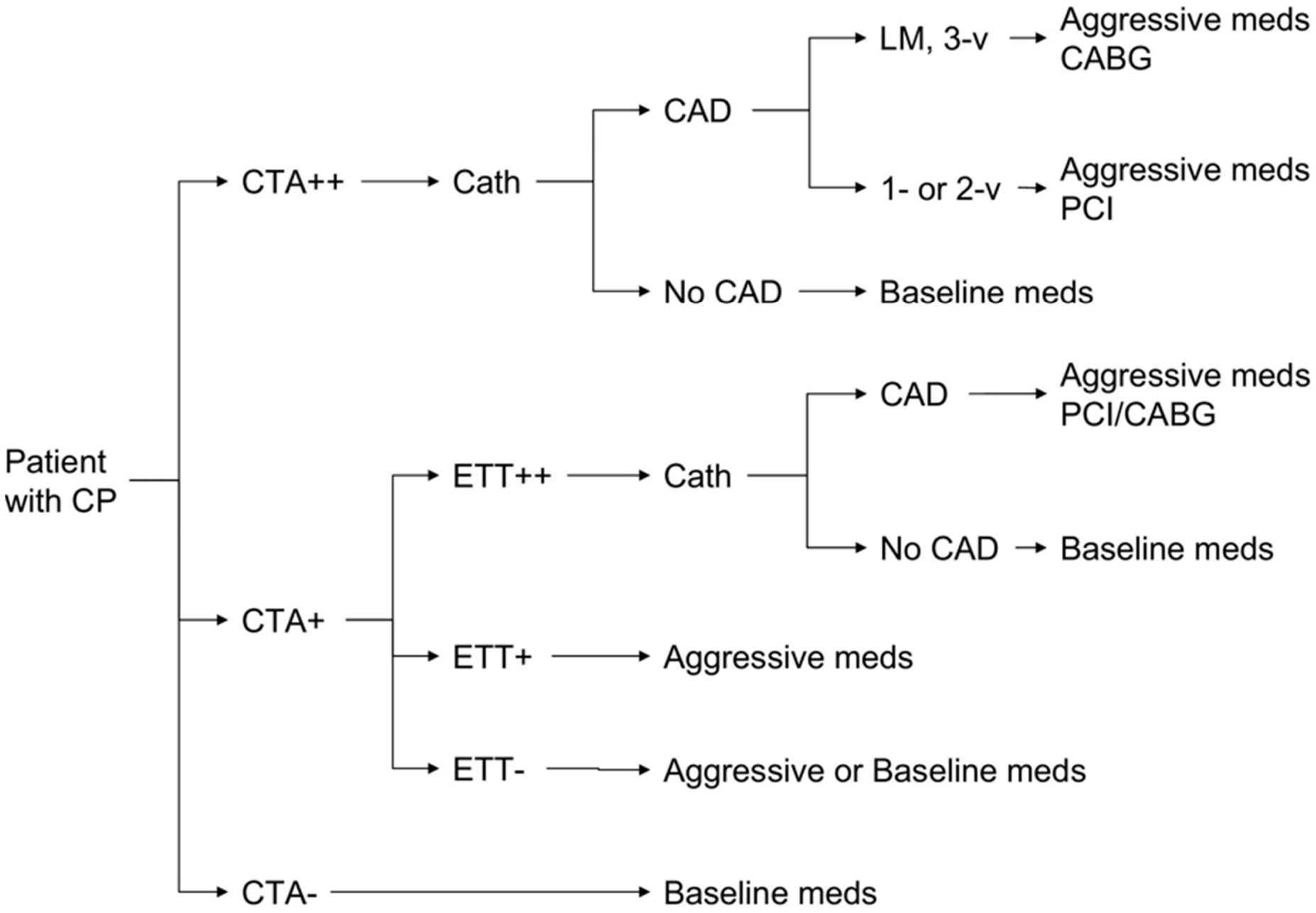

Patient with CP
CTA++
Cath
CAD
LM, 3-v
Aggressive meds CABG
1- or 2-v
Aggressive meds PCI
No CAD
Baseline meds
CTA+
ETT++
Cath
CAD
Aggressive meds PCI/CABG
No CAD
Baseline meds
ETT+
Aggressive meds
ETT-
Aggressive or Baseline meds
CTA-
Baseline meds

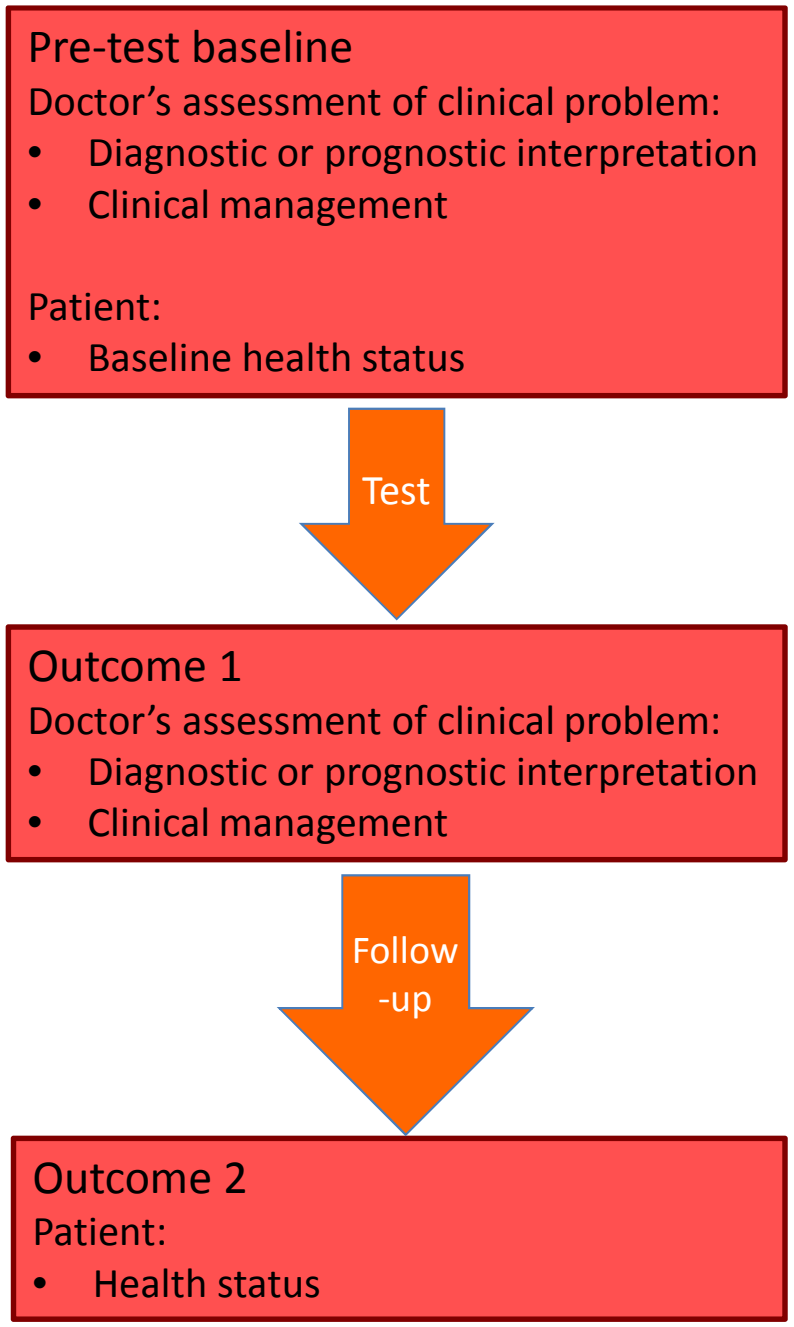
Table 4 Clinical Outcomes in 55-Year-Old Men and Women With Chest Pain

	Nonfatal MI*		Nonfatal Stroke*		Life Expectancy, yrs		QALYs	
Test Strategy	Men	Women	Men	Women	Men	Women	Men	Women
CTA–stress ECG	341	192	57	33	77.361	81.633	13.632	16.605
Stress ECG–CTA	350	198	59	34	77.165	81.548	13.552	16.571
CTA	341	192	57	33	77.36	81.633	13.631	16.604
Stress ECG	350	196	59	33	77.198	81.582	13.566	16.582
Stress echocardiography	347	195	59	33	77.247	81.584	13.586	16.585
Stress SPECT	343	193	57	33	77.331	81.628	13.62	16.6
Cardiac catheterization	339	192	57	33	77.316	81.601	13.605	16.588
No exam	380	211	66	37	76.622	81.364	13.33	16.5

*Lifetime prevalence/1,000 patients undergoing diagnostic testing; adverse events only tracked in patients with CAD.
Cath = invasive cardiac catheterization; QALY = quality-adjusted life-year; other abbreviations as in Table 1.

Diagnostic Before-and-After Studies

- To evaluate clinical impact of single or additional testing
- Change in doctor's assessment and management plan
- Impact on clinical course more difficult
  - Long follow-up
  - Interfering factors
- Alternative if RCT impossible, infeasible or unethical


Pre-test baseline
Doctor's assessment of clinical problem:
• Diagnostic or prognostic interpretation
• Clinical management
Patient:
• Baseline health status
Test
Outcome 1
Doctor's assessment of clinical problem:
• Diagnostic or prognostic interpretation
• Clinical management
Follow -up
Outcome 2
Patient:
• Health status

Assessing clinicians' behaviours

- Documentation and standardisation of decision-making
  - Particularly difficult when the comparison group is standard practice
- Assessing behaviour observed in a trial may not be representative
  - Future behaviour will depend on the trial results
  - Learning curves may affect compliance
    - Becoming acquainted with a test
    - Ascertaining how best to use it
    - Gaining confidence in its findings
    - Allowing it to replace other investigations

In conclusion

Think very carefully about your research question

Choose optimal design for that question