Méta-analyse diagnostique
Aspects méthodologiques et statistiques

Gilles Chatellier

gilles.chatellier@egp.aphp.fr

Université Paris Descartes

HEGP

Méta-analyses d'études de tests diagnostiques. Quels objectifs ?

1. **Proposer des estimateurs résumés valides des indices de performance diagnostique.**
2. **Identifier les raisons de la variabilité inter-étude.**
3. **Améliorer la qualité des futures études originales en identifiant des erreurs de méthodologie.**

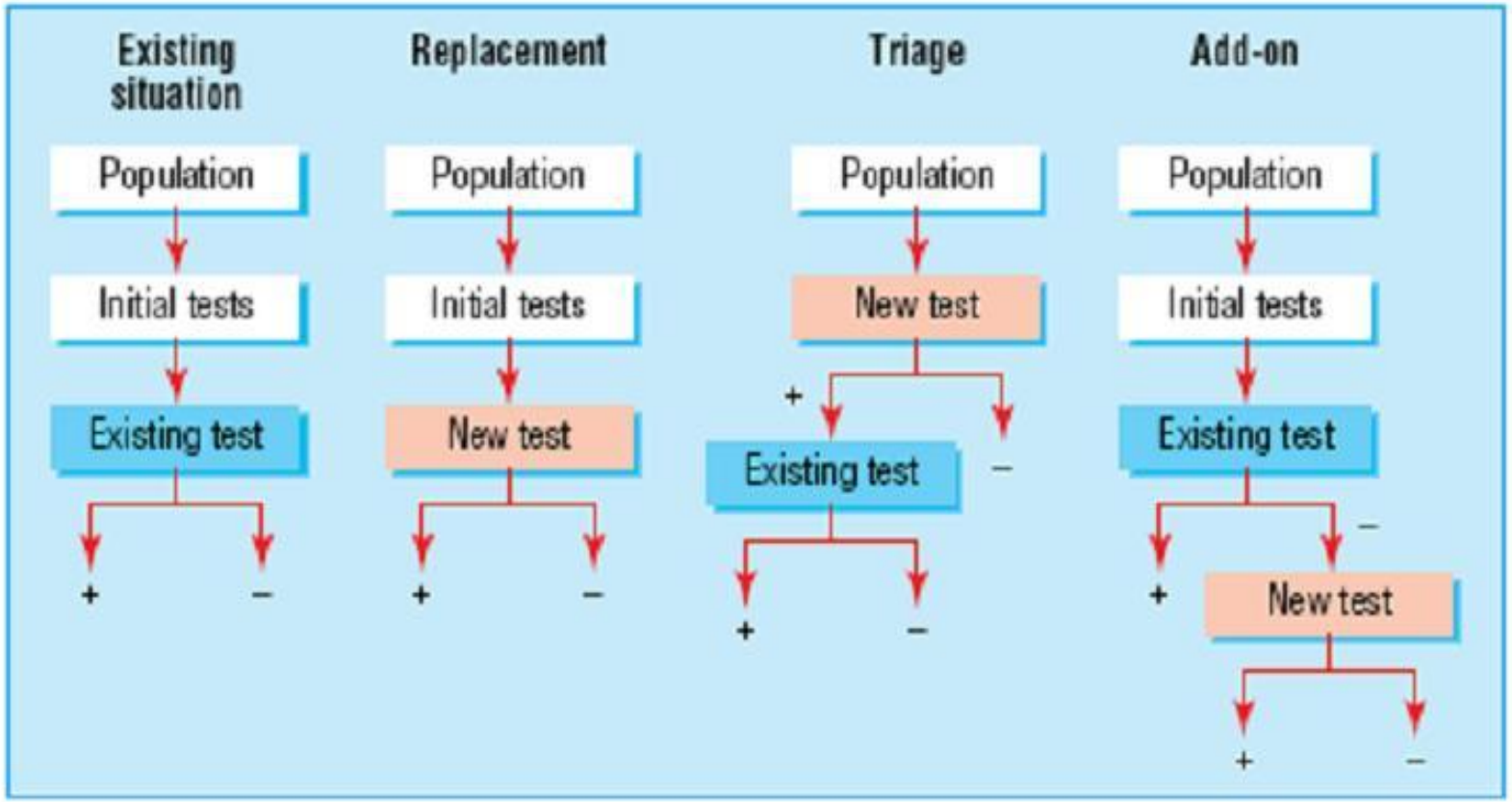
Méta-analyses diagnostiques
Architecture générale

- **Formuler la question.**
- **Identifier la stratégie de recherche.**
- **Sélection des études éligibles et évaluation de la qualité.**
- **Extraction des données et calcul des mesures individuelles de chaque étude.**
- **Choix du modèle statistique.**
- **Recherche et évaluation de l'hétérogénéité et des biais.**

Tests diagnostiques: de nombreuses utilisations!

- **Screening (douleur thoracique)**
- **Routine (biologie)**
- **Diagnostic (présence/absence d'une maladie)**
- **Staging (Cancer)**
- **Monitoring (récidive cancer…)**

Tests diagnostiques: quelle place dans le processus diagnostique?



Roles of tests and positions in existing diagnostic pathways

Bossuyt, BMJ, 2006

Tests diagnostiques: de nombreuses manières de les évaluer

Phase	Question	Design
I	Do test results in affected patients differ from those in normal individuals?	▪ Case-control study ▪ Comparison of diagnostic marker mean concentrations
II	Are patients with certain test results more likely to have the target disorder?	▪ Case-control study ▪ Cut-point selection ▪ Accuracy assessment (Se, Sp)
III	Do test results distinguish patients with and without the target disorder among those in whom it is clinically sensible to suspect the disorder?	▪ Prospective cohort study ▪ Patients in whom it is clinically sensible to suspect the disease (clinical practice) ▪ Accuracy assessment (Se, Sp)
IV	Do patients undergoing the diagnostic test fare better than similar untested patients?	▪ Randomized controlled trial ▪ Patient outcome; process of care

The architecture of diagnostic research. Sackett & Haynes. BMJ 2002;324:539–41

Evaluation des tests diagnostiques.
La complexité de l'évaluation a des conséquences sur la pratique des méta-analyses

- **Sensibilité et spécificité d'un même test vont différer selon que l'on en situation de dépistage, de soins primaires, ou de soins de référence…**
- **Cette différence tient autant aux malades (sévérité) qu'aux non-malades (maladies associées).**
- **Identifier le stade de développement du test ET le processus diagnostique dans lequel il est inclus est donc indispensable.**

Méta-analyse diagnostique
Formuler la question
PICO*→ PPTTO

- **Population**
- **Pathologie**
- **Test Index ←**
- **Test de Reference**
- **Outcome (critère de jugement diagnostique)**

PICO*: Population, Intervention, Control, Outcome, dans le contexte des essais randomisés

Métanalyse diagnostique
Recherche des articles (1)

- **Indexation et qualité des articles originaux (méthodes + rédaction des résultats) sont en général inférieures à celle observées dans les essais randomisés.**
- **Conséquence pratique: défaut de performance de la recherche → Nombre d'articles à lire important à la phase initiale de la M-A.**
- **Intérêt du travail en collaboration:**

Méthodologiste / Clinicien / Documentaliste

Métanalyse de test diagnostique
Recherche des articles (2)

- **La requête doit être formulée de manière à répondre aux particularités de la recherche sur les tests diagnostiques**
- **Combinaison de mots-clés (MeSH) et de texte libre correspondant aux 4 éléments suivants:**
  1. **Test index**
  2. **Pathologie cible**
  3. **Test de référence**
  4. **Population de patients-cibles**

Métanalyse diagnostique : exemple de recherche d'études diagnostiques dans MEDLINE

((((((((((("sensitivity and specificity"[All Fields] OR "sensitivity and specificity/standards"[All Fields]) OR "specificity"[All Fields]) OR "screening"[All Fields]) OR "false positive"[All Fields]) OR "false negative"[All Fields]) OR "accuracy"[All Fields]) OR ((((("predictive value"[All Fields] OR "predictive value of tests"[All Fields]) OR "predictive value of tests/standards"[All Fields]) OR "predictive values"[All Fields]) OR "predictive values of tests"[All Fields])) OR (("reference value"[All Fields] OR "reference values"[All Fields]) OR"reference values/standards"[All Fields])) OR ((((((((((("roc"[All Fields] OR "roc analyses"[All Fields]) OR "roc analysis"[All Fields]) OR "roc and"[All Fields]) OR "roc area"[All Fields]) OR "roc auc"[All Fields]) OR "roc characteristics"[All Fields]) OR "roc curve"[All Fields]) OR "roc curve method"[All Fields]) OR "roc curves"[All Fields]) OR "roc estimated"[All Fields]) OR "roc evaluation"[All Fields])) OR "likelihood ratio"[All Fields]) AND notpubref [sb]) AND "human"[MeSH Terms])

Deville WL et. al. BMC Medical Research Methodology 2002, 2:9-22

M-A diagnostiques vs. M-A thérapeutiques

- **Evaluation de la qualité**
  - **Echantillon pertinent de sujets (malades ET non-malades)**
  - **Test de référence (pertinence)**
- **Méthodes statistiques**
  - **Résumés statistiques : sensibilité, spécificité, O-R diagnostique, rapports de vraisemblance**
  - **Homogénéité**
    - **Effet seuil: Courbe ROC résumée**
    - **Sinon: méthodes habituelles**

M-A diagnostique: évaluation de la qualité

QUADAS tool (*Whiting et al. Health Technology Assessment. 2004;8: 1–234)*
1. Was the spectrum of patients representative of the patients who will receive the test in practice?
2. Were selection criteria clearly described?
3. Is the reference standard likely to correctly classify the target condition?
4. Is the time period between reference standard and index test short enough to be reasonably sure that the target condition did not change between the two tests?
5. Did the whole sample or a random selection of the sample, receive verification using a reference standard of diagnosis?
6. Did patients receive the same reference standard regardless of the index test result?
7. Was the reference standard independent of the index test (i.e. the index test did not form part of the reference standard)?
8. Was the execution of the index test described in sufficient detail to permit replication of the test?
9. Was the execution of the reference standard described in sufficient detail to permit its replication?
10. Were the index test results interpreted without knowledge of the results of the reference standard?
11. Were the reference standard results interpreted without knowledge of the results of the index test?
12. Were the same clinical data available when test results were interpreted as would be available when the test is used in practice?
13. Were uninterpretable/ intermediate test results reported?
14. Were withdrawals from the study explained?

Table 2 – List of aspects to be checked in the assessment of diagnostic and prognostic studies during the systematic review and meta-analysis

Age and gender distribution of the population studied[29].
Inclusion date and follow-up period of the study[29].
Standardized reference test, adequacy of the gold standard chosen, evaluating whether this does not lead to the wrong classification of disease status[13].
Technical aspects of the performance of the test.
Evaluate the degree of missing data.
Original false and true-positive results, false and true-negative results. Occasionally, these data can be estimated from the sensitivity and specificity values as well as from the positive and negative values of the endpoint or reference test.
Reference values for the gold-standard test and for the index test, in a clear way and representative of the disease of interest[12,29].
The confidence interval and the standard error for test accuracy measurements[29].
The number of readers and their training for the index and the gold-standard test[29].
Presence of review bias: verify whether the test result in the study was evaluated blind to the endpoints and other tests (independent interpretation).
Presence of verification bias: the reference test may have been performed preferably in patients with positive tests, which is more frequent when the tests considered as a gold standard are invasive. In this case, the choice of patients for verification by the gold-standard test is not random[12].
Whether the reference test was performed in all patients. If the index and the gold-standard tests have not been performed in all patients, which is ideal, evaluate whether the choice of patients for the tests was random, thus decreasing the chance of bias[3].
Presence of clinical spectrum bias: lack of representation of the clinical spectrum of the disease of interest in the study population. Evaluate patients' demographic and clinical data such as age, gender, race, clinical characteristics, presence of symptoms, disease stage, duration, and comorbidities. The prevalence of the condition among the population studied provides a broader view of the spectrum, circumstances and potential of generalizability.
In screening tests, there may be excess diagnosis bias (when a disease that could progress asymptomatically is detected), excess representation bias (for diseases that progress slowly, making them "stand out" because of the screening), and early detection bias (which overestimates the effects of clinical benefits)[13].

Extraction des données
(M-A sur données résumées)

- **Qualité de l'extraction: .**
  - **Agrément inter-investigateurs: consensus, kappa….**
  - **Entrainement des personnes en charge de l'extraction…**
- **Patients: informations importantes (penser aux sous-groupes), modalités de sélection, exclusions secondaires**
- **Tests quantitatifs: identifier le seuil séparant malades et non-malades**
- **Résultats des tests:**
  - **Définition: positif, négatif, indéterminé, non réalisable, combinaisons de modalités (exemple: non interprétable + négatif = négatif)**
  - **Biais de pratique**

Test diagnostique: Matrice de décision

	MALADIE	
	Présente	Absente
TEST	M+	M-
Anormal (T+)	VP	FP
Normal (T-)	FN	VN

Tests diagnostiques
Métriques

- **Métriques pairées et non indépendantes**
  - **Sensibilité & Specificité**
  - **VPP& VPN**
  - **Likelihood Ratios (LR) / rapport de vraisemblance (pour un test positif & negatif)**

- **Mesures simples**
  - **Diagnostic odds ratio (DOR)**
  - **Area under the curve (AUC) of summary Receiver Operating Characteristics (sROC) curve**

Rapports de vraisemblance (Likelihood ratios)

TEST

	Présente	Absente	
Anormal (T+)	a	b	a+b
Normal (T-)	c	d	c+d
	a+c	b+d	

Rapport de vraisemblance + = p(T+si M+)/p(T+si M-)
Rapport de vraisemblance - = p(T-si M+)/p(T-si M-)

Rapport de vraisemblance + = (a/a+c) / (b/b+d)
Rapport de vraisemblance - = (c/a+c) / (d/b+d)
Utilisable si test> 2 réponses

Likelihood ratios
(Rapports de vraisemblance)

- **Les LRs positif et négatif décrivent la valeur discriminante des tests**
- **Echelle proposée par Jaeschke en 1994:**
  - **–LR+ >10 and LR- <0.1 “conclusive evidence”**
  - **–LR+ 5-10 and LR- 0.1-0.2 “strong diagnostic evidence”**
  - **–LR+ 2-5 and LR- 0.2-0.5 “weak diagnostic evidence”**
  - **–LR+ 1-2 and LR- 0.5-1 “negligible evidence”**

Méta-analyse

- **Modèle à Effet Fixe (MEF) ou Modèle à Effet Aléatoire (MEA) ?**
  - MEF: toutes les études sont supposées estimer le même effet commun sous-jacent
  - MEA: hétérogénéité inter-études

- **Hétérogénéité, quelles origines:**
  - Population source
  - Conception de l'étude
  - Technologie du test (ex: type de machine en radio)
  - Réalisation du test (ex: Expérience des radiologues)
  - Hasard!

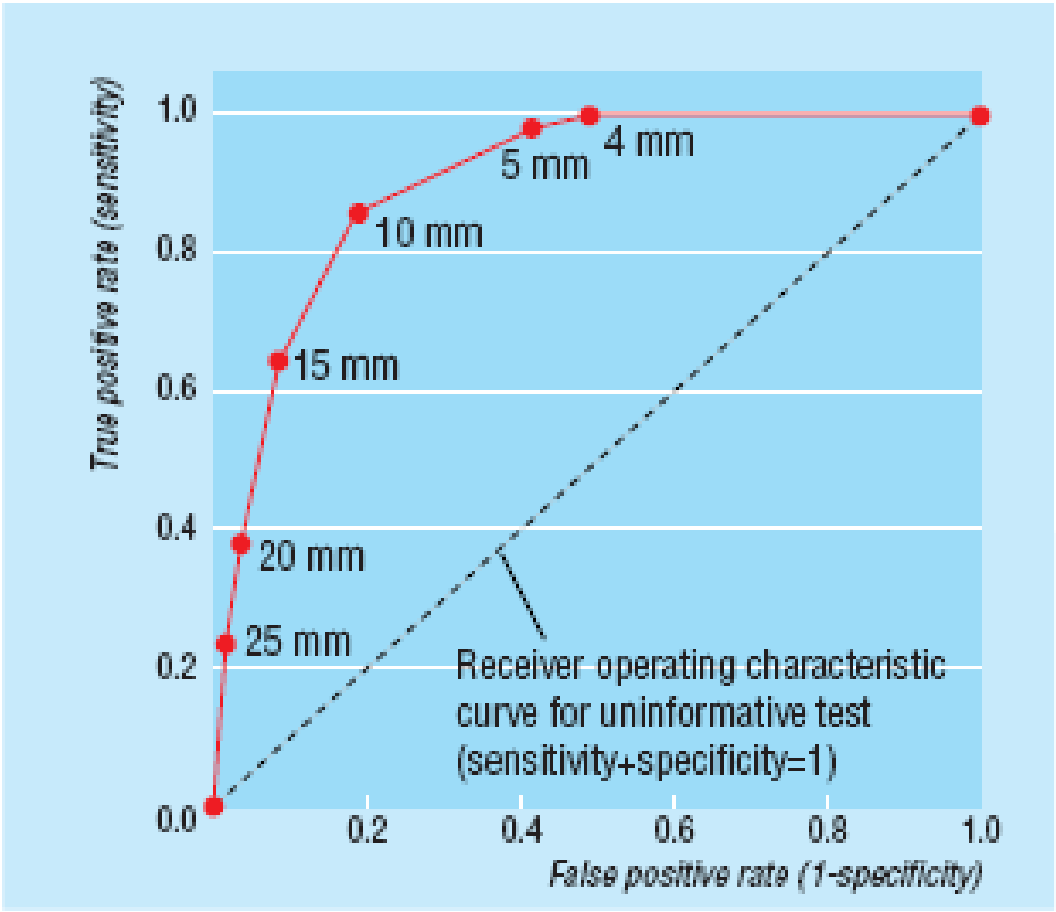
Hétérogénéité, statistiques

- **Analyse avec les méthodes habituelles: graphique, chi-2 d'hétérogénéité.**
- **Attention au I^2 de Higgins: une partie de l'hétérogénéité peut être liée à l'effet seuil.**
- **Hétérogénéité. Quelles solutions ?**
  - Rester qualitatif
  - Identifier des sous groupes
  - Modèle à effet aléatoire

Méta-analyse : quelles métriques

- **Sensibilité et spécificité, mais:**
  - **ne sont pas indépendantes**
  - **Se et Sp dépendent de la population, mais pas de la prévalence (en théorie)**
- **VPP et VPN: éviter, car rôle de la prévalence dans la variabilité inter-étude**
- **dOR: éviter (compréhension)**
- **LRs: possible, même problèmes que Se et Sp**
- **Si test quantitatif ET absence d'effet seuil:**
  - **ROC curve**

Courbe ROC
Tests diagnostiques avec seuil

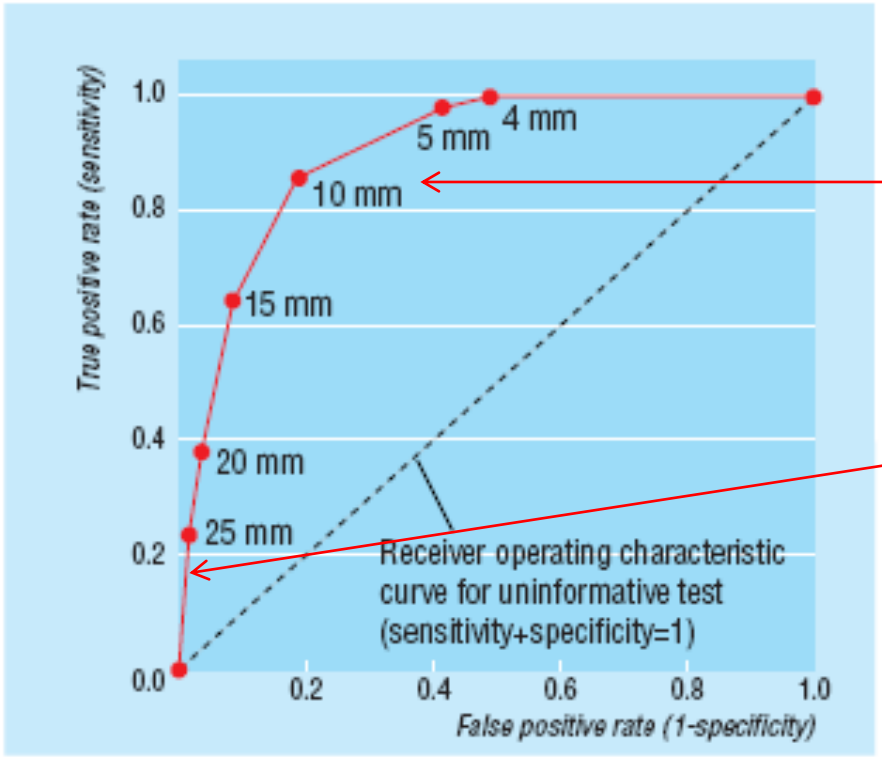


AUC:

- Valeurs: 0.5 - 1
- Plus la valeur est proche de 1, mieux c'est !
- Test parfait: AUC = 1
- Test inutile: AUC = 0.5
- Test pire que le test de référence: AUC < 0.5

BMJ, 2001

summary ROC plot



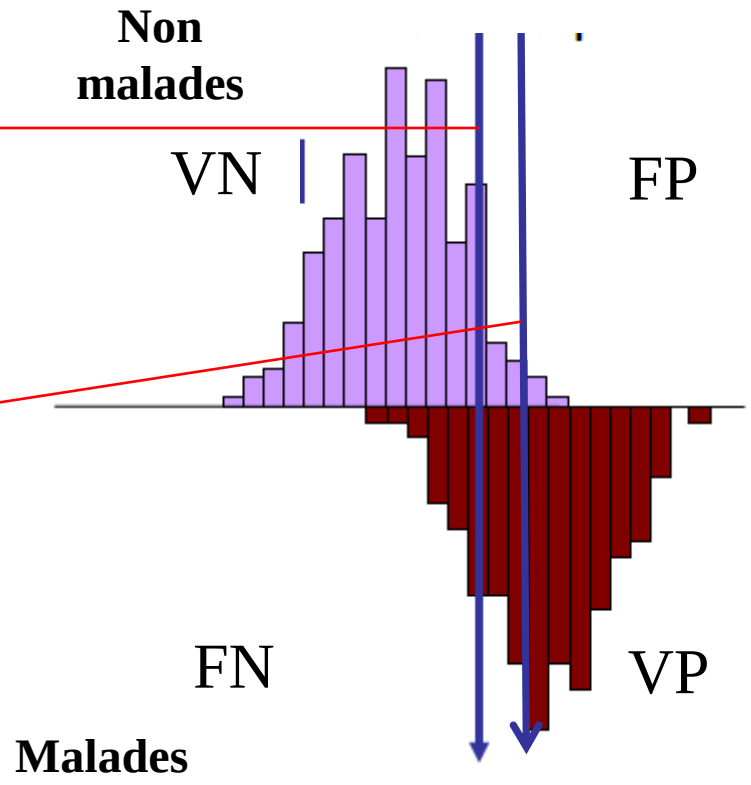
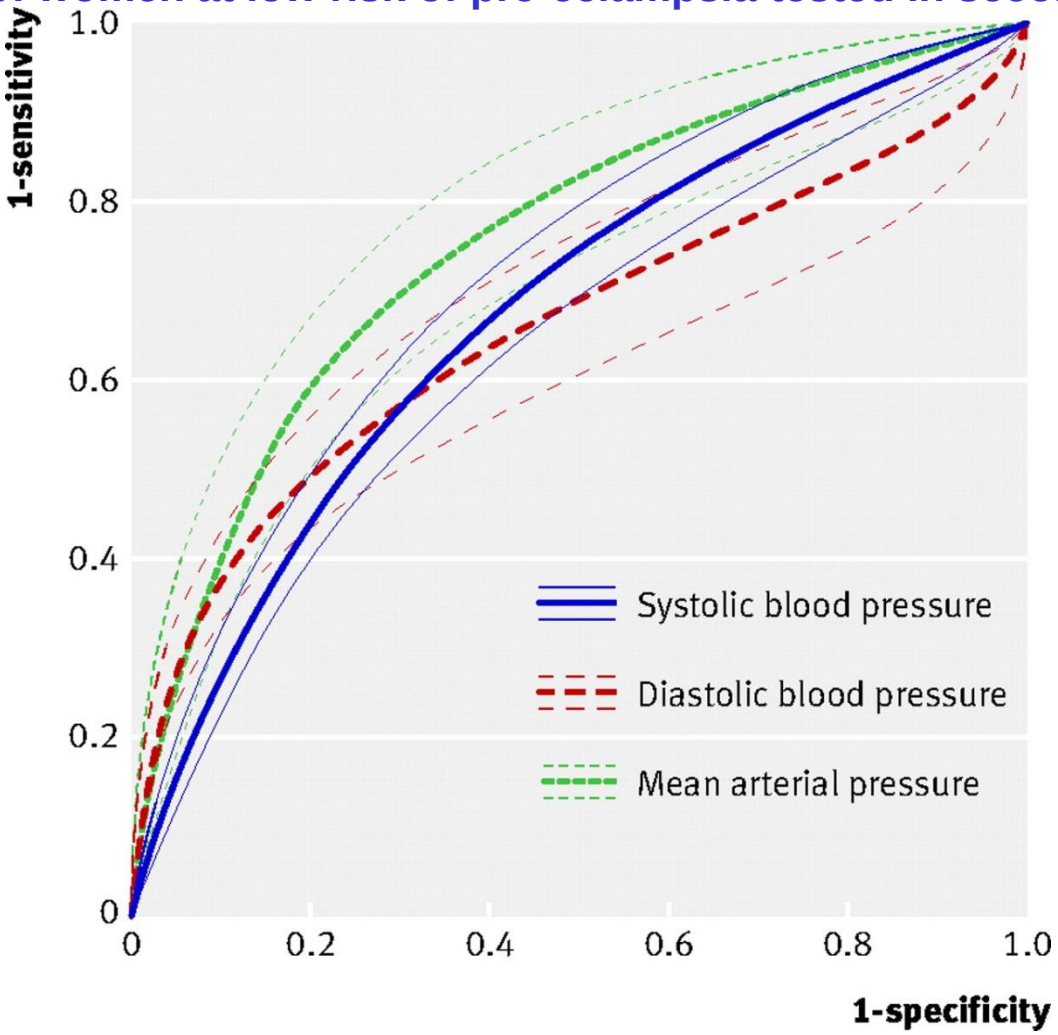


Fig 4 Summary receiver operating characteristic curves with 95% confidence intervals for systolic blood pressure, diastolic blood pressure, and mean arterial pressure in population of women at low risk of pre-eclampsia tested in second trimester



BMJ

Cnossen, J. S et al. BMJ 2008;336:1117-1120

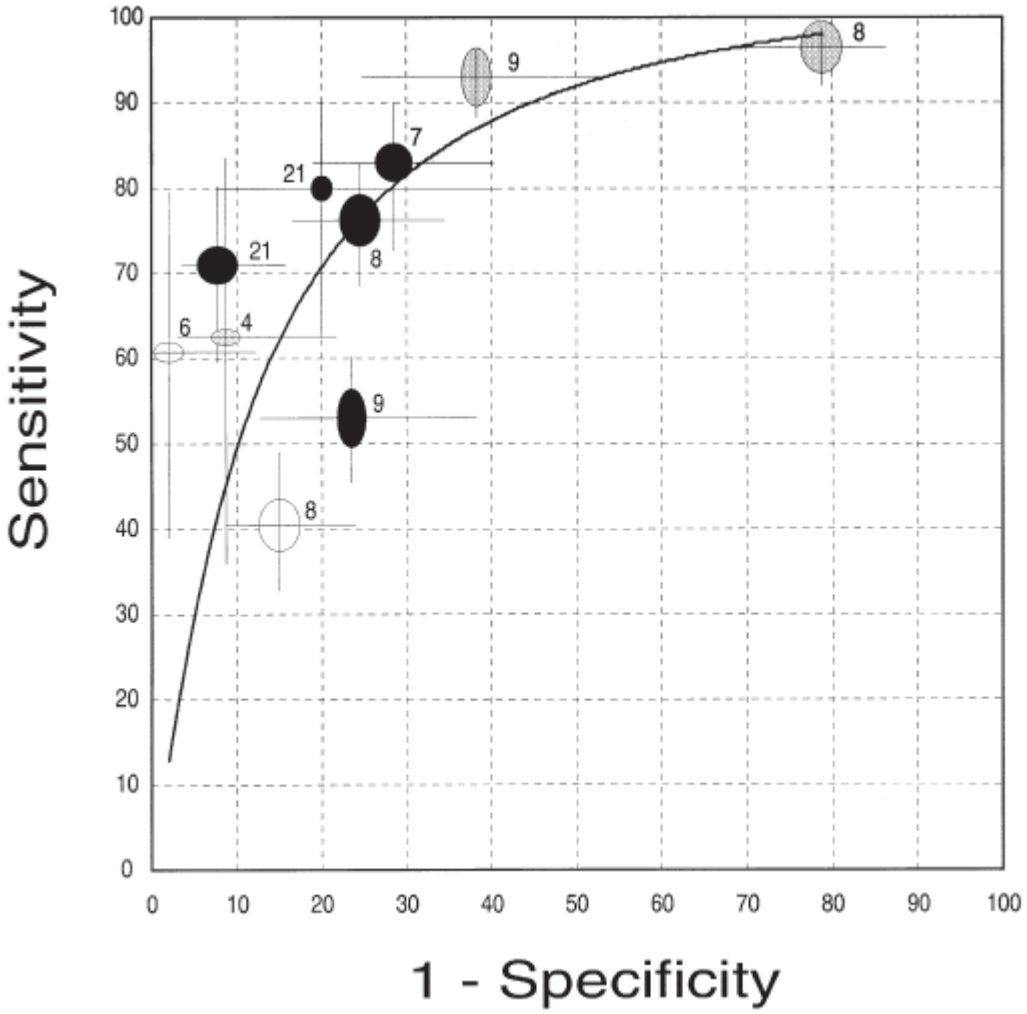
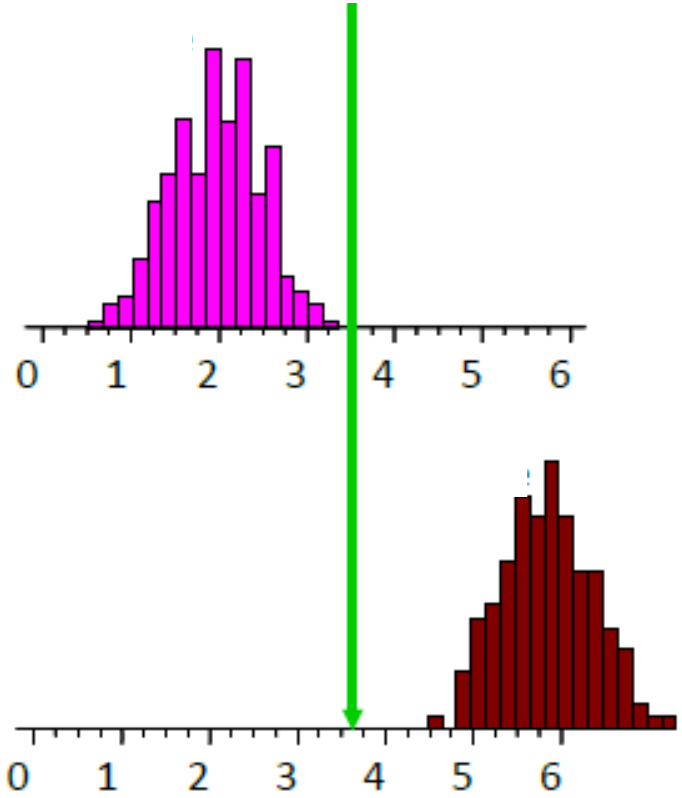


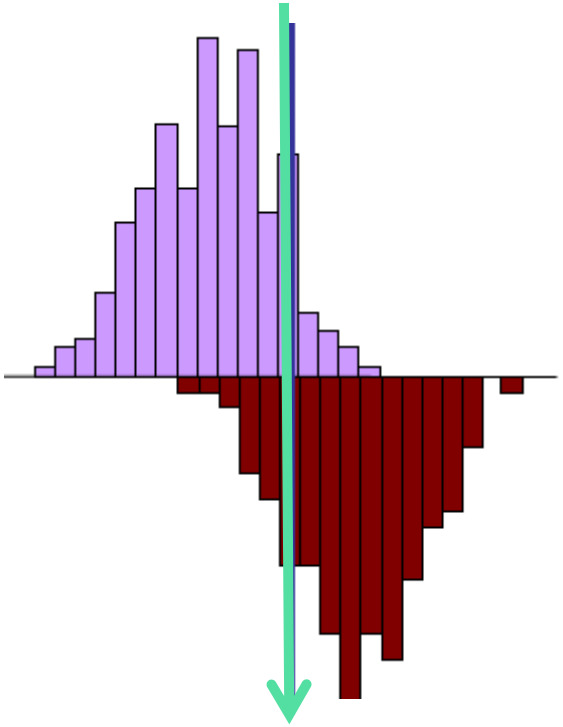
Fig. 1. Summary receiver operator characteristic curve for sinus radiography, compared with sinus puncture/aspiration for the diagnosis of acute sinusitis. Each ellipse corresponds to a study estimate of sensitivity and specificity; the area of each ellipse is proportional to the study's size, and horizontal and vertical lines associated with each ellipse correspond to 95% confidence intervals for the study's estimates. Black ellipses provide estimates using "sinus fluid or opacity" as the radiographic criterion for sinusitis. Gray ellipses provide estimates for "sinus fluid or opacity or mucous membrane thickening" as the radiographic criterion for sinusitis. The two white ellipses provide estimates for "sinus opacity" and for an unspecified diagnostic criterion. Numbers next to each ellipse refer to the study that provided data for that estimate, as identified in Table 1 and in the references.

sROC curve: effet seuil.

Pas d'effet seuil



Effet seuil



Methods Commonly Used To Calculate a Summary Point

Method	Description or Comment	Does It Have the Desired Characteristics?
Independent meta-analysis of sensitivity and specificity	• **Separate meta-analyses per metric** • **Within-study variability preferably modeled by the binomial distribution**	• **Ignores the correlation between sensitivity and specificity** • **Underestimates the summary sensitivity and specificity and wrong confidence intervals**
Joint (multivariate) meta-analysis of sensitivity and specificity based on hierarchical modeling	• **Based on multivariate (joint) modeling of sensitivity and specificity** • **Two families of models that are equivalent when there are no covariates** • **Modeling preferably using binomial likelihood rather than normal approximations**	• **The generally preferred method**

Trikalinos TA, Coleman CI, Griffith L, et al. Meta-analysis of test performance when there is a "gold standard." In: Chang SM and Matchar DB, eds. Methods guide for medical test reviews. Rockville, MD: Agency for Healthcare Research and Quality; June 2012. p. 8.1-21. AHRQ Publication No. 12-EHC017. Available at www.effectivehealthcare.ahrq.gov/medtestsguide.cfm.

Methods Commonly Used To Calculate a Summary Line

Method	Description or Comment	Does It Have the Desired Characteristics?
Moses-Littenberg model	• Summary line based on a simple regression of the difference of logit-transformed true-positive and false-positive rates versus their average	• Ignores unexplained variation between-studies (fixed effects) • Does not account for correlation between sensitivity and specificity • Does not account for variability in the independent variable • Inability to weight studies optimally – yields wrong inferences when covariates are used
Random intercept augmentation of the Moses-Littenberg model	• Regression of the difference of logit-transformed true-positive and false-positive rates versus their average for random effects that allows for variability across studies	• Does not account for correlation between sensitivity and specificity • Does not account for variability in the independent variable

Summary receiver operator characteristic (ROC) based on hierarchical modeling	• Same as for multivariate meta-analysis to obtain a summary point — hierarchical modeling
	• Many ways to obtain a (hierarchical) summary ROC:
	  ○ Rutter-Gatsonis (most common)
	  ○ Several alternative curves
		• Most theoretically motivated method
		• Rutter-Gatsonis hierarchical summary ROC is recommended in the *Cochrane Handbook*, as it is the method that has been used most often

Trikalinos TA, Coleman CI, Griffith L, et al. Meta-analysis of test performance when there is a "gold standard." In: Methods guide for medical test reviews. Available at www.effectivehealthcare.ahrq.gov/medtestsguide.cfm.
Littenberg B, Moses LE. Med Decis Making 1993 Oct-Dec;13(4):313-21. PMID: 8246704.
Rutter CM, Gatsonis CA. Acad Radiol 1995 Mar;2 Suppl 1:S48-56; discussion S65-7, S70-1 pas. PMID: 9419705.

Représentation graphique
Forrest plot

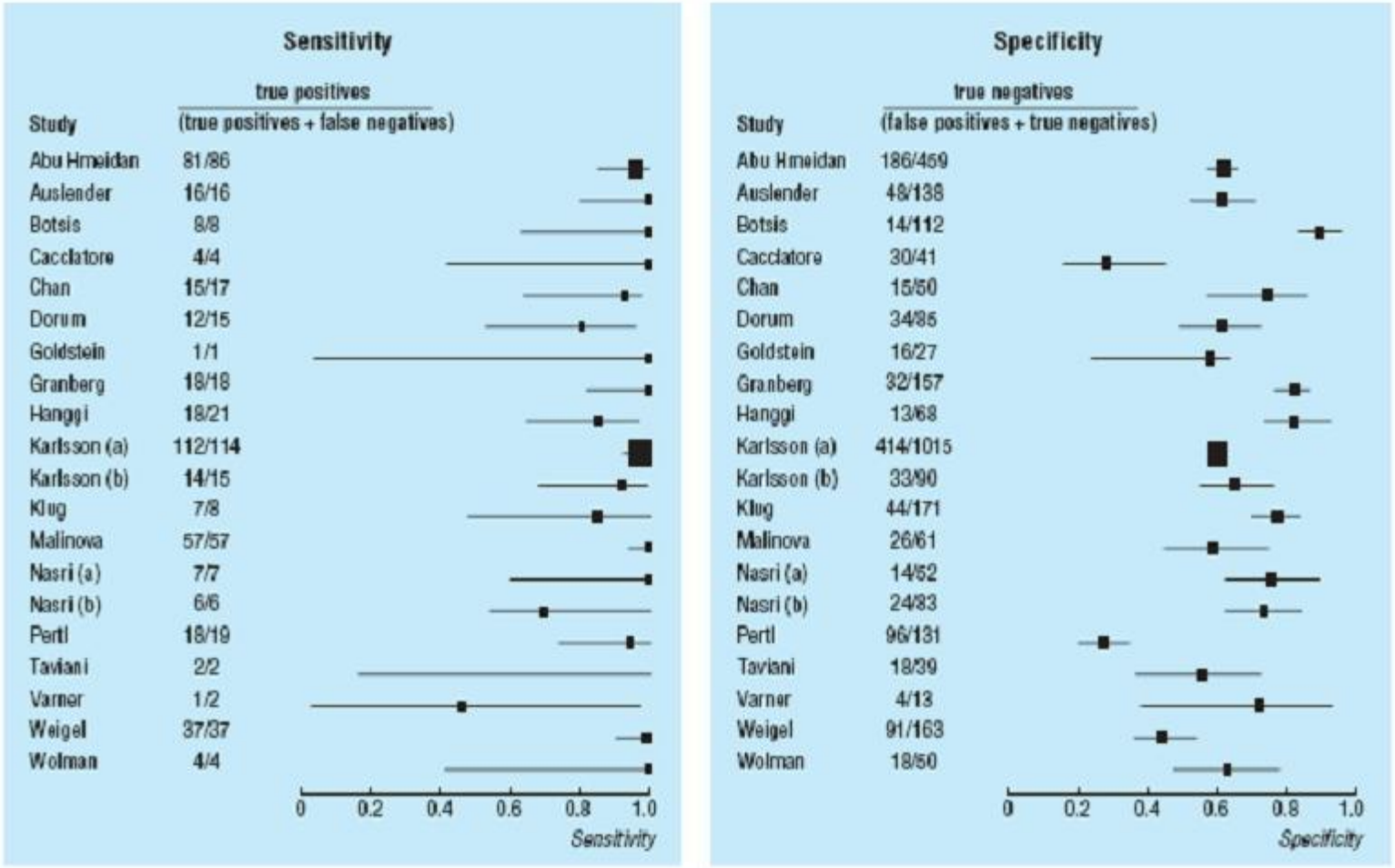
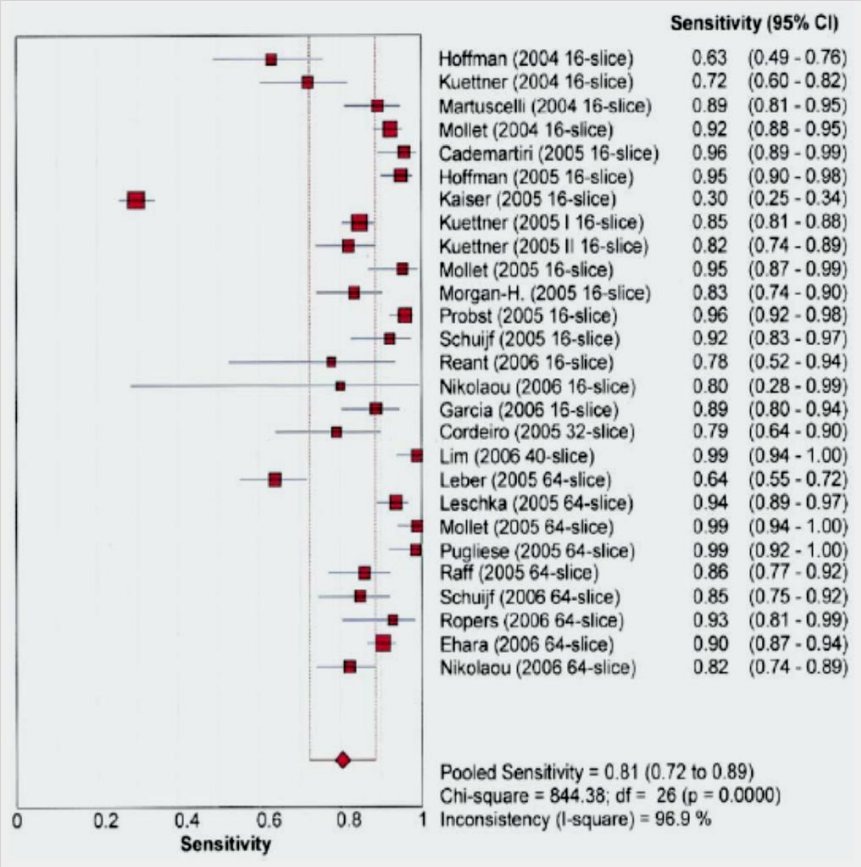


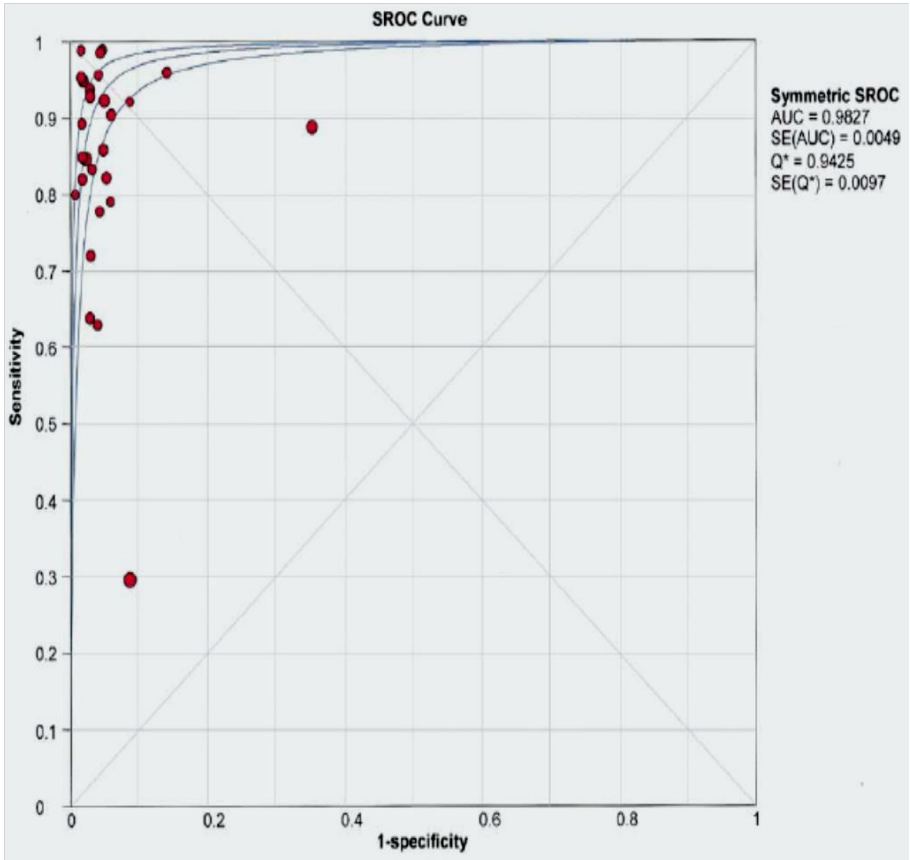
Fig 2 Estimates from 20 studies of sensitivity and specificity of measurement of endometrial thicknesses of more than 5 mm using endovaginal ultrasonography for detecting endometrial cancer.[15] Points indicate estimates of sensitivity and specificity. Horizontal lines are 95% confidence intervals for estimates. Size of points reflects total sample size

BMJ, 2001

Diagnostic Performance of Multislice Spiral Computed Tomography of Coronary Arteries as Compared With Conventional Invasive Coronary Angiography

A Meta-Analysis





Louvard et al, JACC 2006

Courbe ROC résumée (sROC curve)

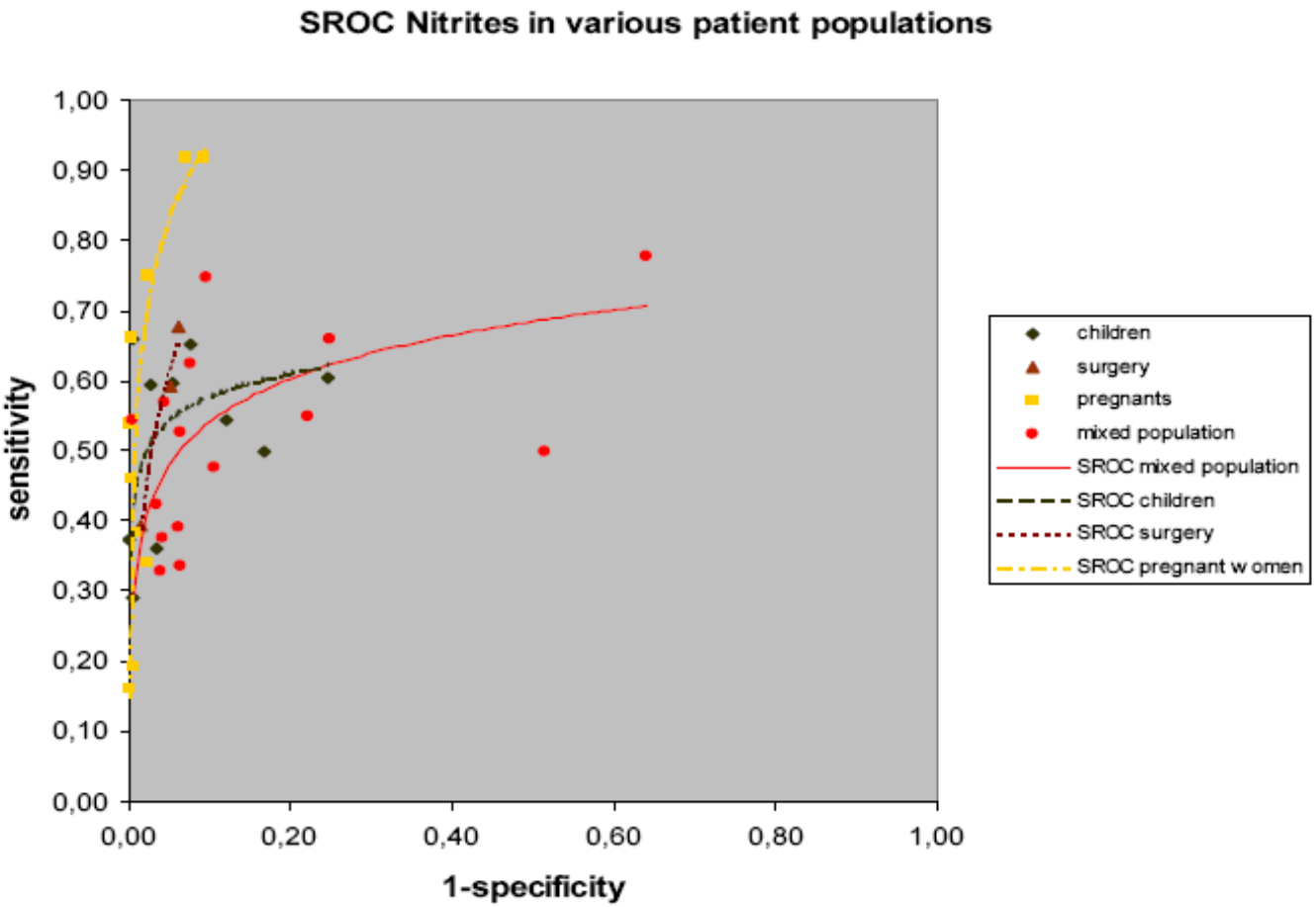
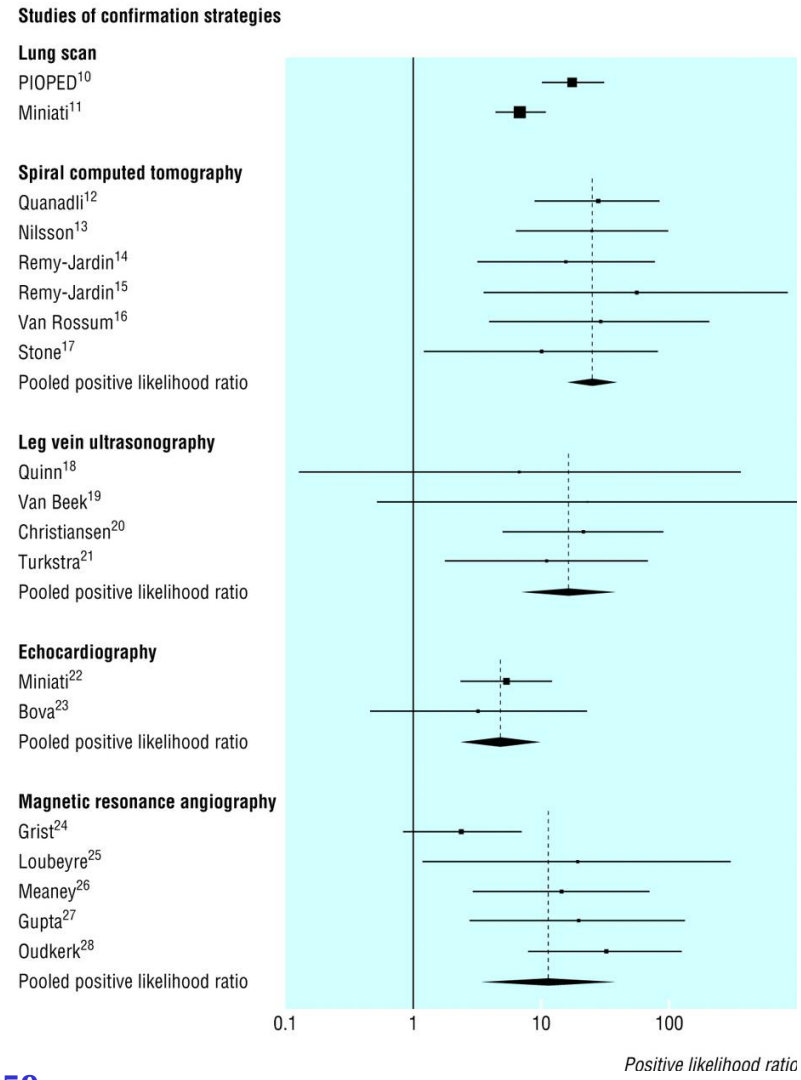


Figure 3
Summary ROC curves of nitrites in urine dipsticks for the diagnosis of bacteriuria and urinary tract infections in various homogeneous subgroups of patient populations.

BMC, 2002

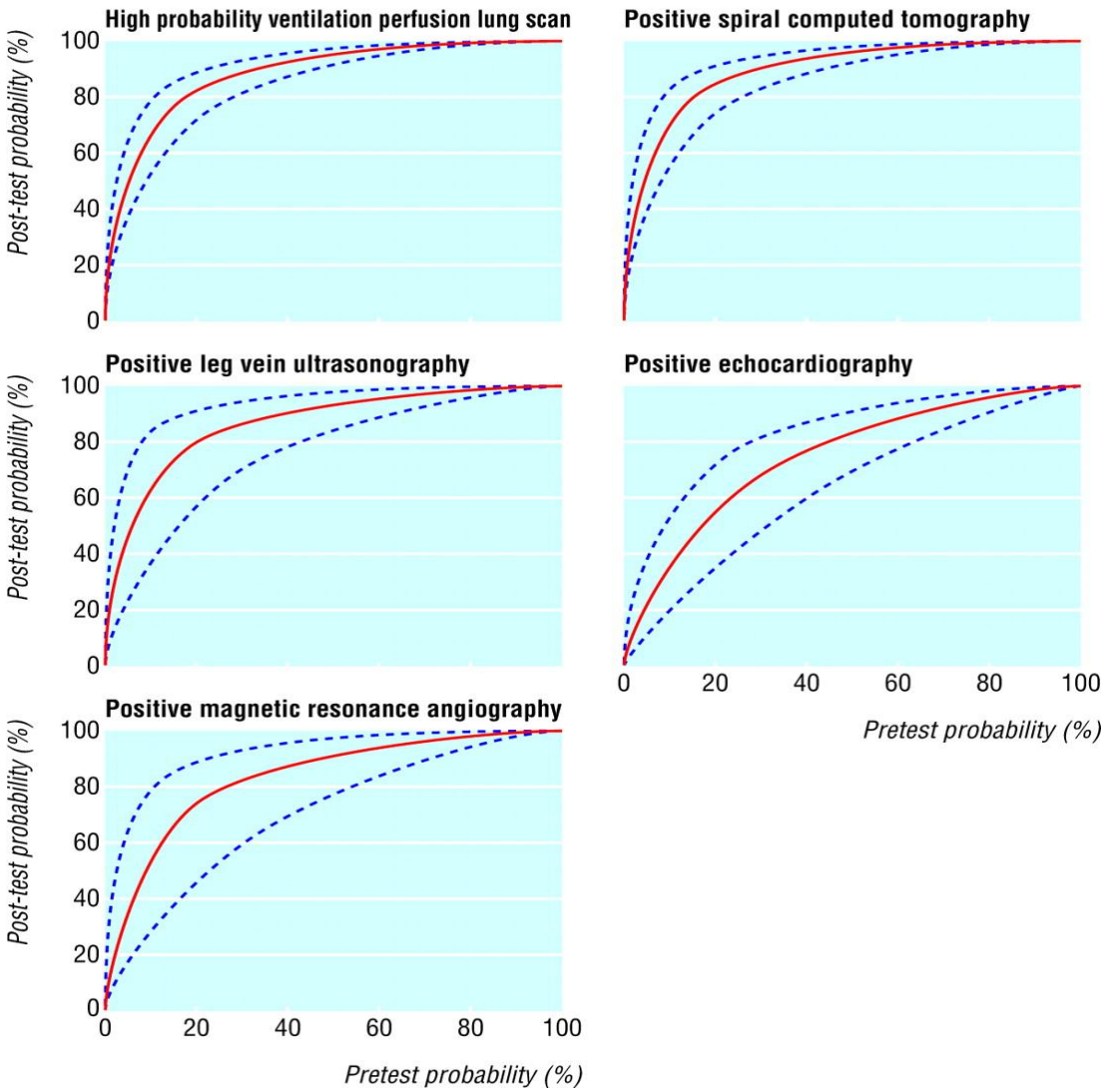
Fig 1: Positive likelihood ratios (squares) and 95% CI for strategies used to confirm a diagnosis of pulmonary embolism. Size of square is related to variance of study. Broken line represents pooled positive likelihood ratio, and limits of diamond represents 95% CI of pooled ratios



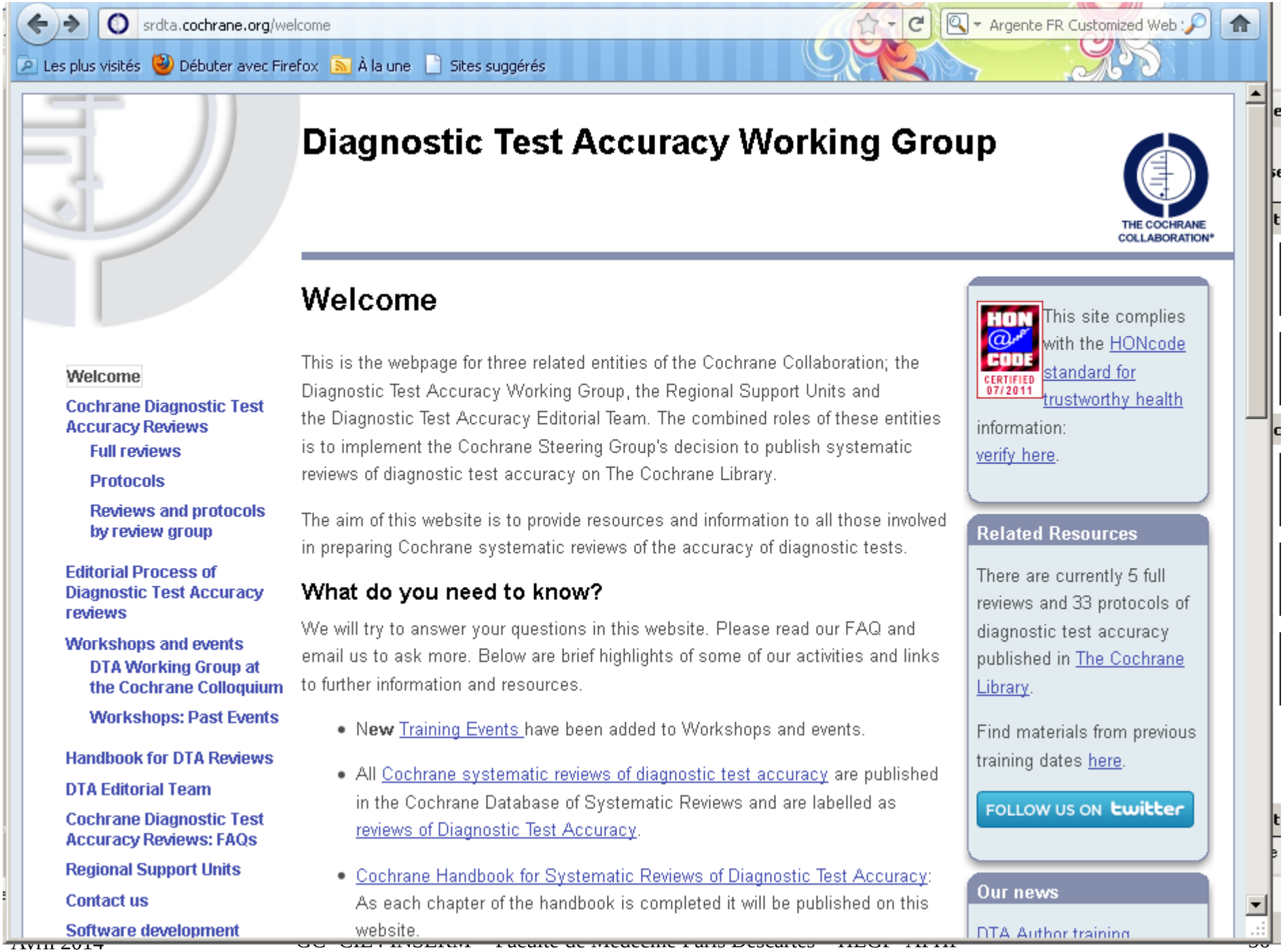
Roy PM. et al. BMJ 2005;331:259

BMJ

Fig 4 Post-test probability according to pre-test probability and pooled values (solid line) or limits of 95% confidence intervals (broken lines) of the positive likelihood ratio



Roy, P.-M. et al. BMJ 2005;331:259


srdta.cochrane.org/welcome
Argente FR Customized Web
Les plus visités
Débuter avec Firefox
À la une
Sites suggérés
Diagnostic Test Accuracy Working Group
THE COCHRANE COLLABORATION®
Welcome
Cochrane Diagnostic Test Accuracy Reviews
Full reviews
Protocols
Reviews and protocols by review group
Editorial Process of Diagnostic Test Accuracy reviews
Workshops and events
DTA Working Group at the Cochrane Colloquium
Workshops: Past Events
Handbook for DTA Reviews
DTA Editorial Team
Cochrane Diagnostic Test Accuracy Reviews: FAQs
Regional Support Units
Contact us
Software development
Welcome
This is the webpage for three related entities of the Cochrane Collaboration; the Diagnostic Test Accuracy Working Group, the Regional Support Units and the Diagnostic Test Accuracy Editorial Team. The combined roles of these entities is to implement the Cochrane Steering Group's decision to publish systematic reviews of diagnostic test accuracy on The Cochrane Library.
The aim of this website is to provide resources and information to all those involved in preparing Cochrane systematic reviews of the accuracy of diagnostic tests.
What do you need to know?
We will try to answer your questions in this website. Please read our FAQ and email us to ask more. Below are brief highlights of some of our activities and links to further information and resources.
New Training Events have been added to Workshops and events.
All Cochrane systematic reviews of diagnostic test accuracy are published in the Cochrane Database of Systematic Reviews and are labelled as reviews of Diagnostic Test Accuracy.
Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy: As each chapter of the handbook is completed it will be published on this website.
HON CODE CERTIFIED 07/2011
This site complies with the HONcode standard for trustworthy health information: verify here.
Related Resources
There are currently 5 full reviews and 33 protocols of diagnostic test accuracy published in The Cochrane Library.
Find materials from previous training dates here.
FOLLOW US ON twitter
Our news
DTA Author training

Summary ROC regression

- **If we can assume that the diagnostic odds ratio (dOR) is constant for every positivity threshold**
  - **This constant dOR defines a set of points in the ROC space= summary ROC curve (sROC curve)**
  - **Symmetrical around the line defined by: Se = Sp**
  - **SROC curve defined by the equation**

$$1 \Big/ 1 + \frac{1}{dOR \times \left(\frac{1 - Sp}{Se} \right)}$$

Summary ROC regression

For each study:

D_i = logit(true positive rate) – logit(false positive rate)

= logit(dOR) **accuracy measure**

S_i = logit(true positive rate) + logit(false positive rate)

proxy for positivity threshold

$$D = a + b \cdot S$$

Fixed effect model adjusted by linear regression

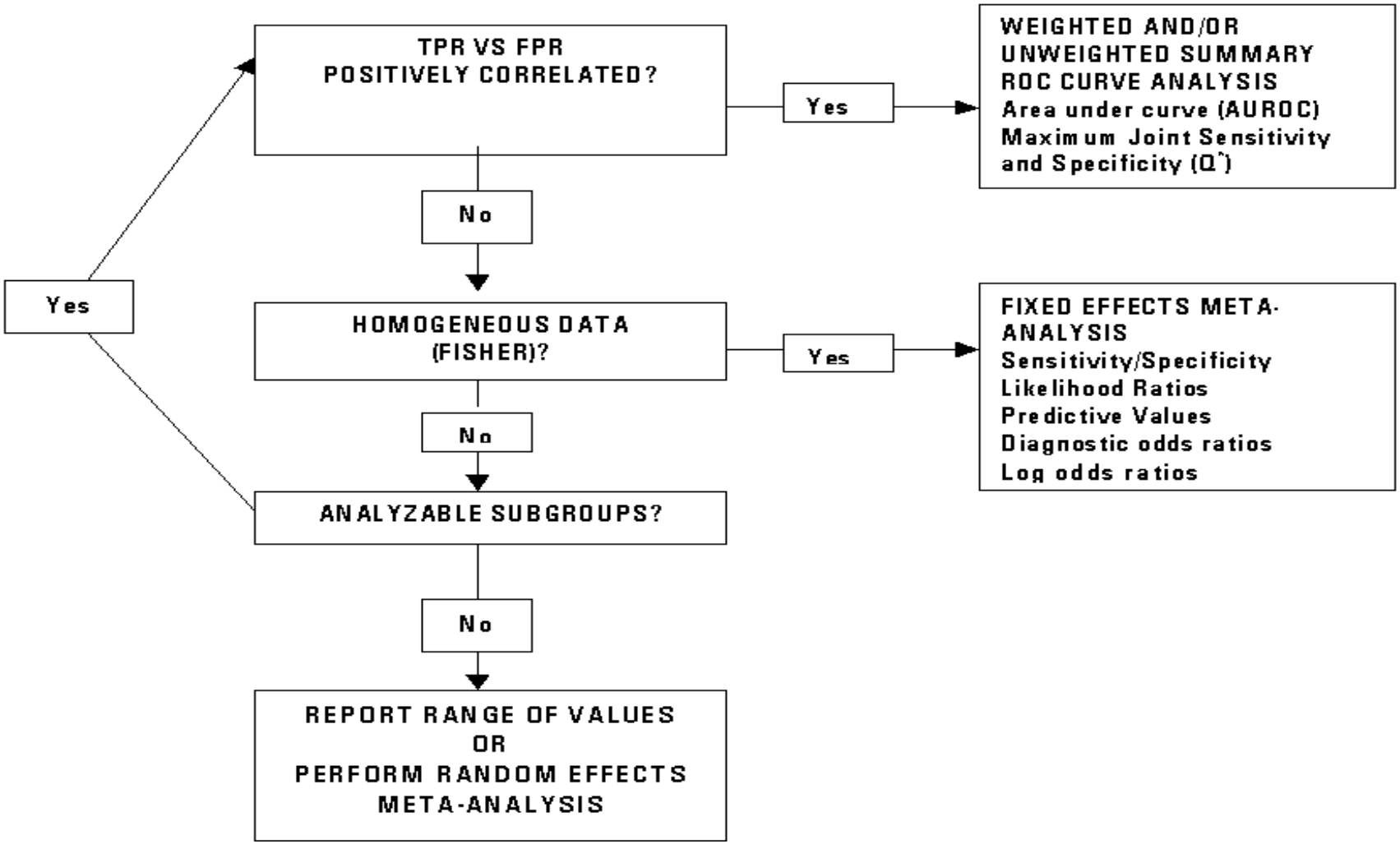
If b ~ 0, we can assume that dOR is not dependant of threshold

Moses LE et al. Stat Med 1993

Table 2 – List of aspects to be checked in the assessment of diagnostic and prognostic studies during the systematic review and meta-analysis

Age and gender distribution of the population studied[29].
Inclusion date and follow-up period of the study[29].
Standardized reference test, adequacy of the gold standard chosen, evaluating whether this does not lead to the wrong classification of disease status[13].
Technical aspects of the performance of the test.
Evaluate the degree of missing data.
Original false and true-positive results, false and true-negative results. Occasionally, these data can be estimated from the sensitivity and specificity values as well as from the positive and negative values of the endpoint or reference test.
Reference values for the gold-standard test and for the index test, in a clear way and representative of the disease of interest[12,29].
The confidence interval and the standard error for test accuracy measurements[29].
The number of readers and their training for the index and the gold-standard test[29].
Presence of review bias: verify whether the test result in the study was evaluated blind to the endpoints and other tests (independent interpretation).
Presence of verification bias: the reference test may have been performed preferably in patients with positive tests, which is more frequent when the tests considered as a gold standard are invasive. In this case, the choice of patients for verification by the gold-standard test is not random[12].
Whether the reference test was performed in all patients. If the index and the gold-standard tests have not been performed in all patients, which is ideal, evaluate whether the choice of patients for the tests was random, thus decreasing the chance of bias[3].
Presence of clinical spectrum bias: lack of representation of the clinical spectrum of the disease of interest in the study population. Evaluate patients' demographic and clinical data such as age, gender, race, clinical characteristics, presence of symptoms, disease stage, duration, and comorbidities. The prevalence of the condition among the population studied provides a broader view of the spectrum, circumstances and potential of generalizability.
In screening tests, there may be excess diagnosis bias (when a disease that could progress asymptomatically is detected), excess representation bias (for diseases that progress slowly, making them "stand out" because of the screening), and early detection bias (which overestimates the effects of clinical benefits)[13].

M-A diagnostique
Choix du modèle



M-A diagnostique: critères de qualité

Sont particulièrement importants dans les études diagnostiques originales:

- **Sélection des patients: échantillon de patients consécutifs ou non, définition de l'échantillon.**
- **Méthode: cohorte ou cas-témoin ?**
- **Test de référence:**
  - **Méthode, cut-off …**
  - **Correspond à un gold standard arfait ou imparfait ?**
- **Test index: méthode, cut-off …**
- **Interprétation: test diagnostique et test de référence interprétés indépendamment (« en aveugle »)**
- **Biais de vérification: pratique différentielle du test de référence selon le résultat du test étudié**