DFGSM2

Etudes d'évaluation des tests diagnostiques

Dr Muriel Rabilloud

UE1 – Santé Publique et LCA

Faculté de Médecine Lyon Est

OBJECTIFS

- Connaître les méthodes de quantification des performances des tests diagnostiques
- Connaître les notions de probabilités pré et post test
- Savoir quantifier l'information apportée par un test diagnostique
- Connaître les différentes phases d'évaluation d'un test diagnostique
- Connaître la méthodologie des études d'évaluation des tests diagnostiques

PLAN

- Méthodes d'évaluation d'un test diagnostique
  - Fiabilité (Reliability)
  - Exactitude (Accuracy)
- Eléments pour la lecture critique d'une étude évaluant les performances d'un test diagnostique

CONNAISSANCES ANTERIEURES

- Sensibilité, spécificité
- Courbe ROC
- Ratios de vraisemblance positif et négatif
- Valeurs prédictives positive et négative
- Probabilités pré et post-test

Test diagnostique

- « Un test diagnostique est une méthode d'exploration dont le but est de fournir une information pour faire progresser une démarche diagnostique à la recherche d'une maladie M dans une situation clinique déterminée » [Grenier 1994]

- Clinique ou para clinique

- Qualitatif ou quantitatif

Tests diagnostiques et Médecine factuelle

- Dispose-t-on d'études ayant quantifié les performances du test diagnostique ?
- Quel est le niveau de preuve de ces études ?
- Les résultats du test permettront-ils de poser le diagnostic de la maladie ou de l'éliminer ?
- Comment utiliser ce test dans votre pratique ?

Le résultat du test est-il susceptible de modifier la prise en charge du patient ?

Evaluation des performances d'un test diagnostique

Fiabilité d'un test (Reliability)

- Un test est fiable si son résultat varie peu lorsqu'on répète la mesure
- Pour les tests quantitatifs, fiabilité classiquement quantifiée par le coefficient de variation
- Pour les tests qualitatifs, fiabilité quantifiée par un coefficient mesurant la concordance

Evaluation de la fiabilité d'un test quantitatif

- Coefficient de variation
  - Mesures répétées sur une échantillon de patients
  - Calcul de l'écart type quantifiant la variabilité entre les mesures répétées
  - La variabilité est exprimée en pourcentage de la moyenne des mesures

$$CV(\%)=\frac{Ecart\ type}{Moyenne}\times 100$$

Quantification standardisée de la variabilité des mesures répétées

Evaluation de la fiabilité d'un test quantitatif

- Pour un dosage biologique
  - Répétabilité (répétitions du dosage dans les mêmes conditions) jugée acceptable si CV<10%
  - Reproductibilité (répétitions du dosage dans des conditions différentes) jugée acceptable si CV<15%

Evaluation de la fiabilité d'un test quantitatif

- Dosage sanguin d'une protéine par une méthode ELISA (méthode immuno-enzymatique)
  - Répétabilité (même plaque) : intraassay CV
  - Reproductibilité (plaques différentes) : interassay CV

	Concentration Protein X (pg/mL)				
Plate Number	Reading 1	Reading 2	Reading 3	Mean	Intra CV (%)
1	183.5	179.6	177.8	180.30	1.62
2	201.4	196.6	194.9	197.63	1.71
3	179.1	177.7	173.6	176.80	1.62

Inter CV (%)	6.03

Evaluation de la fiabilité d'un test qualitatif

- Concordance entre plusieurs évaluations du résultat positif ou négatif d'un même test
- Même mammographie évaluée plusieurs fois par le même radiologue (reproductibilité intra observateur)
- Même mammographie évaluée par plusieurs radiologues (reproductibilité inter observateurs)

Evaluation de la fiabilité d'un test qualitatif

- Concordance des résultats de mammographies évaluées par 2 radiologues

Radiologue 1

Radiologue 2	Positive	Négative	
Positive	32	3	35 0,29
Négative	8	77	85 0,71
	40 0,34	80 0,66	120

Evaluation de la fiabilité d'un test qualitatif

- Concordance observée : $C_o = \frac{32+77}{120} = 0{,}91$

- Concordance attendue par le simple hasard :

$$C_a = (0{,}34 \times 0{,}29) + (0{,}66 \times 0{,}71) = 0{,}57$$

Evaluation de la fiabilité d'un test qualitatif

- Coefficient de concordance Kappa

$$Kappa = \frac{C_o - C_a}{1 - C_a}$$

- Mesure la concordance en pourcentage de la concordance maximum non liée au hasard
- Interprétation proposée par Landis-Koch (1977)
  - Kappa > 0,8 : concordance excellente
  - Kappa entre 0,6 et 0,8 : bonne concordance
  - Kappa entre 0,4 et 0, 6 : concordance moyenne
  - Kappa < 0,4 : faible concordance

Evaluation de la fiabilité d'un test qualitatif

- Coefficient Kappa mesurant la concordance entre les 2 radiologues

$$Kappa = \frac{0{,}91 - 0{,}57}{1 - 0{,}57} \approx 0{,}8$$

- Mesure de la reproductibilité inter observateurs
- Dans ce cas, ne permet pas de conclure à une reproductibilité excellente

Exactitude (accuracy) d'un test diagnostique

- Capacité d'un test diagnostique à discriminer les malades des non malades
- Nécessite d'avoir une méthode de référence (gold standard) pour déterminer le statut réel
- Les méthodes de quantification de l'exactitude du test dépendent :
  - De la nature du test qualitative ou quantitative
  - De la phase d'évaluation du test précoce ou tardive

Qualités intrinsèques d'un test diagnostique

- Sensibilité :
  - capacité du test à identifier les malades
  - Probabilité que le test soit positif chez les malades
- Spécificité :
  - capacité du test à identifier les non malades
  - Probabilité que le test soit négatif chez les non malades
- Ne dépendent pas de la prévalence de la maladie
- Peuvent dépendre des caractéristiques des sujets, des conditions de réalisation du test…

Qualités intrinsèques d'un test diagnostique

		Etat réel des sujets	
		Malade	Non-malade
TesT	Positif	VP	FP
	Négatif	FN	VN

Sensibilité : $P(Test\ positif | Malade) = \frac{VP}{VP+FN}$

Spécificité : $P(Test\ négatif | Non\ Malade) = \frac{VN}{VN+FP}$

Qualités intrinsèques d'un test diagnostique

- Test : Dosage radio-immunologique des phosphatases acides
- Maladie : cancer de la prostate
- Type d'étude : transversale sur un échantillon de malades et un échantillon de non malades
  - Echantillon de 113 malades
  - Echantillon de 240 non malades

Qualités intrinsèques du test pour un seuil fixé

	Malades	**Non malades**	
Test +	79	31	110
Test -	34	209	243
	113	240	353

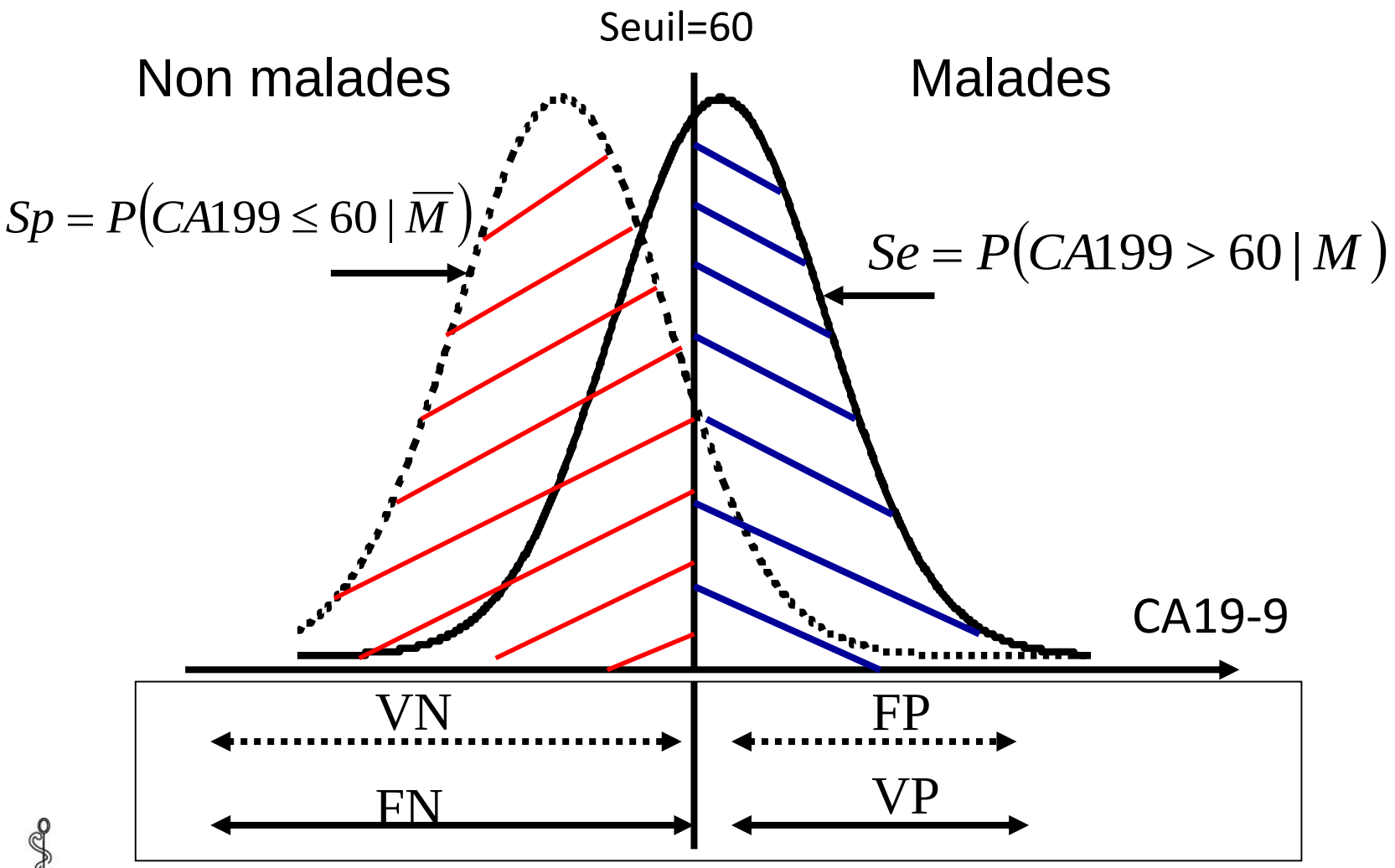
Sensibilité = P[T+|M] = 79/113 = 70%

Spécificité = P[T-|NM] = 209/240 = 87%

Capacité discriminante d'un test quantitatif

- A la phase précoce de l'évaluation d'un test quantitatif
  - Le seuil de positivité n'est pas encore fixé
  - Nécessité d'avoir un quantification globale de la capacité discriminante du test
- Exemple de l'évaluation du dosage du CA19-9 pour le diagnostic de cancer du pancréas
  - Echantillon de 90 sujets ayant un cancer du pancréas
  - Echantillon de 51 sujets n'ayant pas de cancer du pancréas
  - Dosage sanguin de CA19-9 chez les 141 sujets

Capacité discriminante d'un test quantitatif



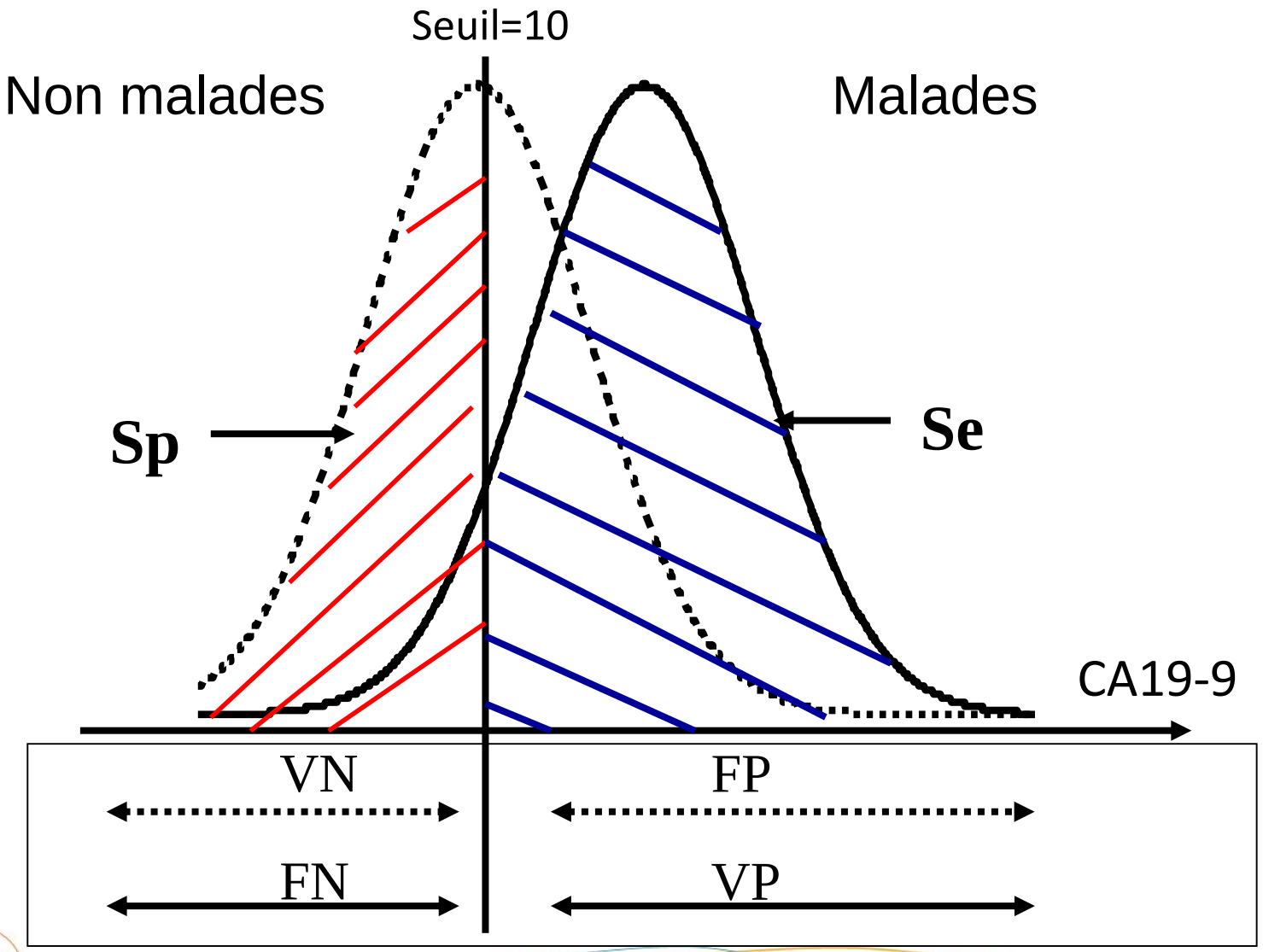
Qualités intrinsèques pour un seuil = 60

	Malades	**Non malades**	
Test +	61	2	63
Test -	29	49	78
	90	51	141

Sensibilité = P[T+|M] = 61/90 = 68%

Spécificité = P[T-|NM] = 49/51 = 96%

Capacité discriminante d'un test quantitatif



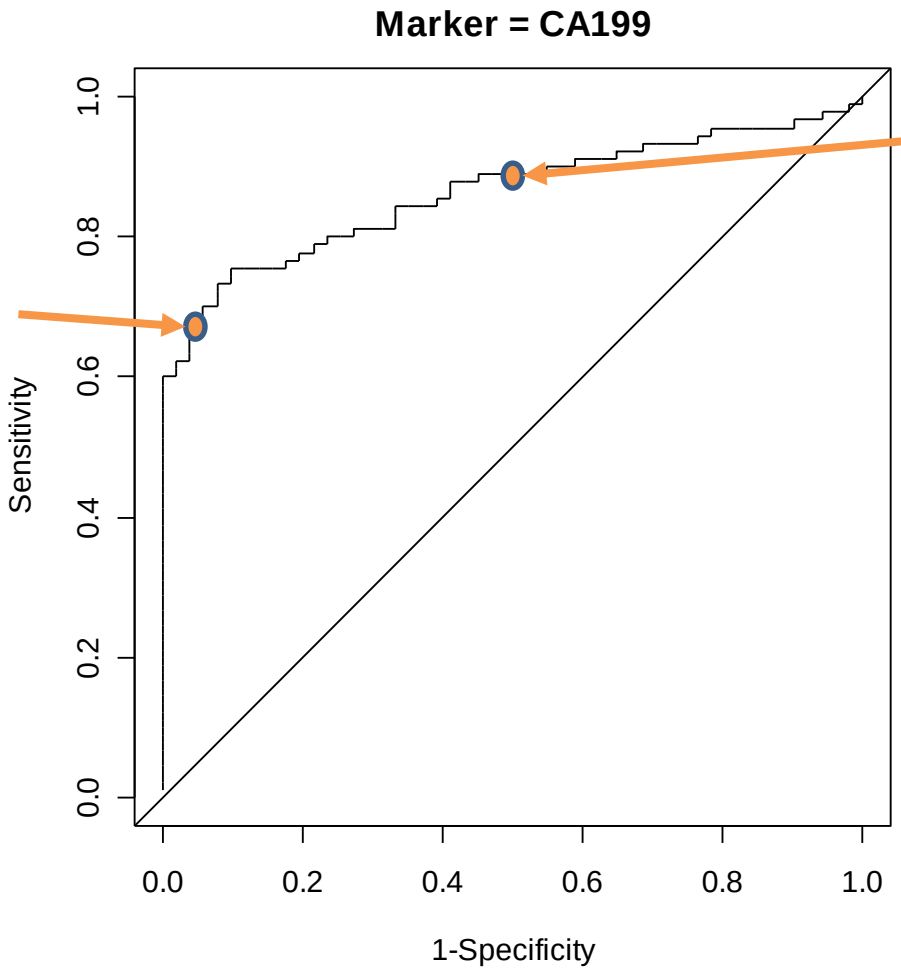
Qualités intrinsèques pour un seuil = 10

	Malades	**Non malades**	
Test +	80	25	105
Test -	10	26	36
	90	51	141

Sensibilité = P[T+|M] = 80/90 = 89%

Spécificité = P[T-|NM] = 26/51 = 51%

Construction de la courbe ROC empirique

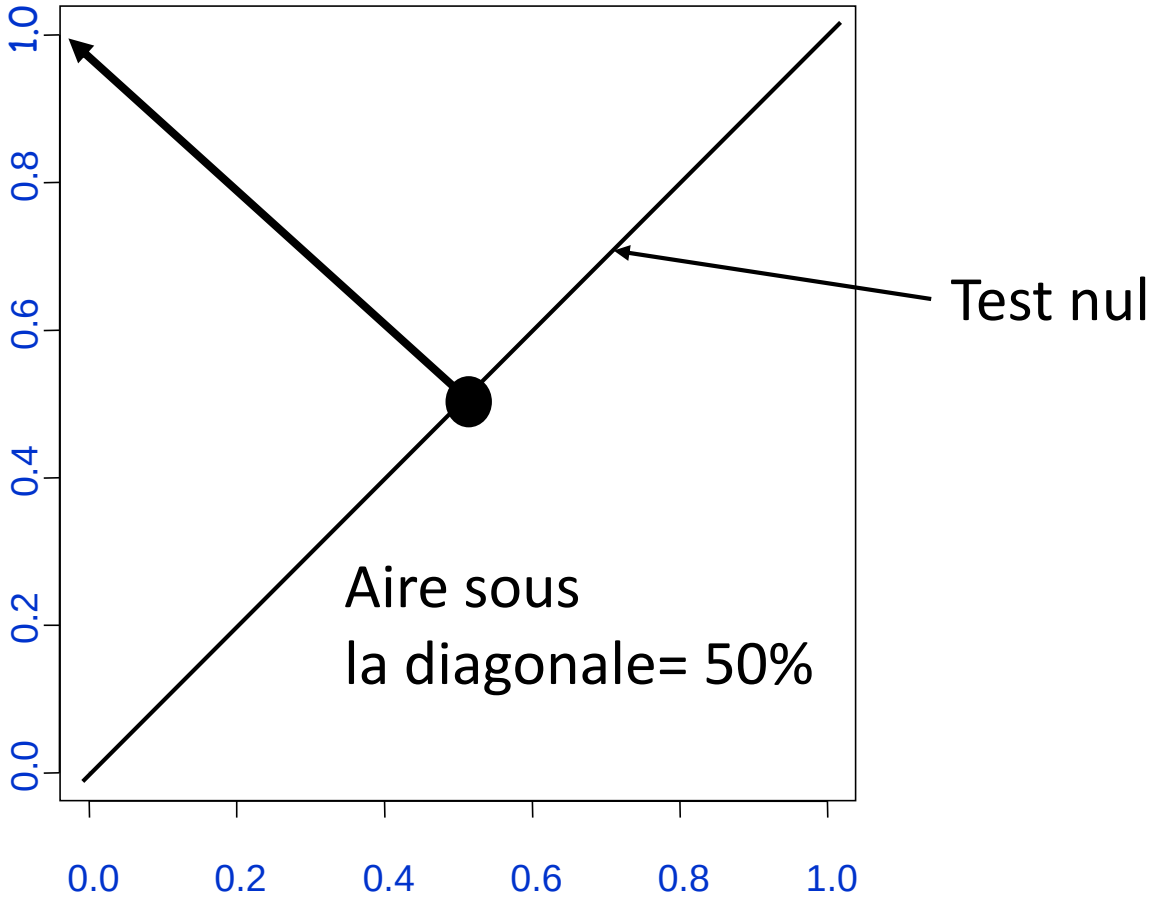


Seuil = 60

Seuil = 10

Courbe ROC

P(Test +|M)



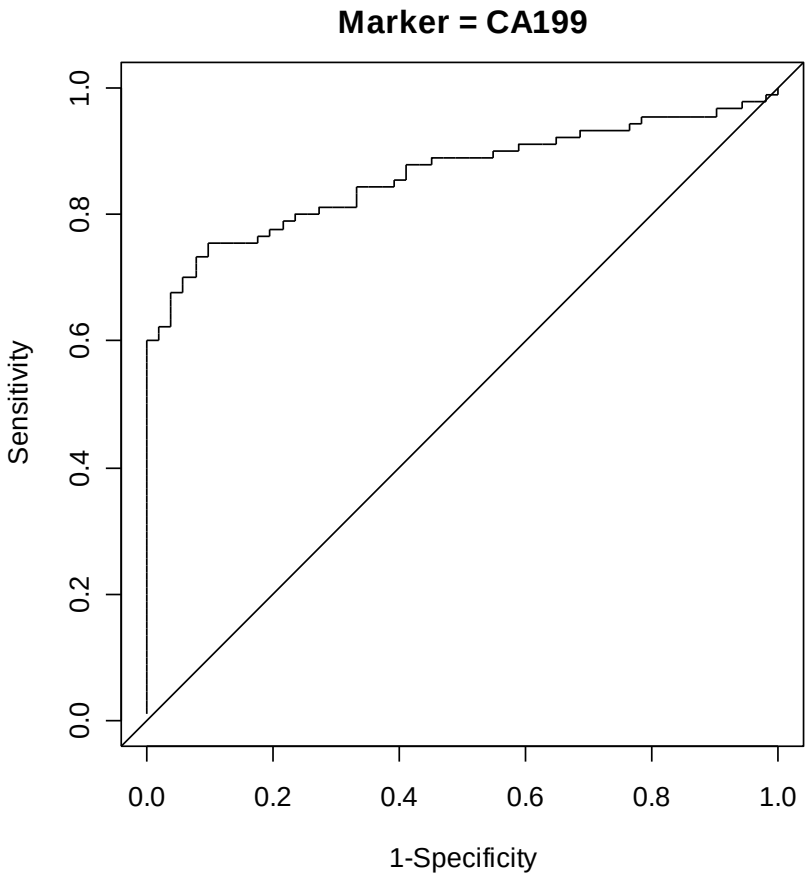
P(Test +|NM)

Capacité discriminante d'un test quantitatif

- Construction de la courbe ROC
- Estimation de l'aire sous la courbe ROC (ASC)
- Test parfait : ASC = 100%
- Test non discriminant : ASC = 50%
- Interprétation de l'ASC :

Probabilité qu'un sujet malade est une valeur du test supérieure (inférieure) à celle d'un sujet non malade

Aire sous la courbe ROC



Estimation de l'ASC = 86%
Intervalle de confiance à 95% : [80% ; 92%]

Le test a un intérêt diagnostique

Sa capacité à discriminer les sujets qui ont un cancer du pancréas de ceux qui n'en ont pas est bonne

Détermination du seuil de positivité

- Détermination du seuil à utiliser en pratique à la phase tardive de l'évaluation du test
- Seuil optimal dépend :
  - De la prévalence de la maladie dans la population cible
  - Des conséquences de ne pas poser le diagnostic chez un malade (traitement retardé)
  - Des conséquences de poser le diagnostic à tort chez un sujet non malade (complications des traitements)

Qualités extrinsèques d'un test diagnostique

- Valeur prédictive positive (VPP) :
  - Capacité du test à affirmer la présence de la maladie en cas de test positif
  - Probabilité d'avoir la maladie lorsque le test est positif
- Valeur prédictive négative (VPN) :
  - Capacité du test à éliminer la présence de la maladie en cas de test négatif
  - Probabilité de ne pas avoir la maladie lorsque le test est négatif

Qualités extrinsèques d'un test diagnostique

- VPP et VPN dépendent :
  - De la sensibilité et de la spécificité
    - Plus le test est sensible meilleure est la VPN
    - Plus le test est spécifique meilleure est la VPP
  - De la prévalence de la maladie dans la population
    - Plus la prévalence est élevée meilleure est la VPP
    - Plus la prévalence est faible meilleure est la VPN

VPP et VPN dépendent de la population dans laquelle on utilise le test

Exemple de l'étude CASS (New England of Medicine 1979)

- Performances de la douleur thoracique pour faire le diagnostic de maladie coronarienne
- Echantillon de 1465 patients adressés pour suspicion de maladie coronarienne
- Gold standard = coronarographie
- Interrogatoire des patients inclus à la recherche de douleur thoracique
- En aveugle du résultat de la coronarographie

Estimation des qualités intrinsèques du test

		Etat réel des sujets		
		M	NM	
TesT	T+	969	245	1214
	T-	54	197	251
		1023	442	1465

$$\text{Sensibilité} = P[T+ \mid M] = \frac{969}{1023} \approx 94{,}7\,\%$$

$$\text{Spécificité} = P[T- \mid NM] = \frac{197}{442} \approx 44{,}6\,\%$$

Estimation des qualités extrinsèques du test

		Etat réel des sujets		
		M	NM	
TesT	T+	969	245	1214
	T-	54	197	251
		1023	442	1465

$$\text{Prévalence} = \frac{1023}{1465} \approx 70\,\%$$

$$VPP = P[M \mid T+] = \frac{969}{1214} \approx 80\,\%$$

$$VPN = P[NM \mid T-] = \frac{197}{251} \approx 78{,}5\,\%$$

Qualités extrinsèques du test diagnostique

- La prévalence (70%) = **probabilité pré-test** de maladie chez un sujet issu de la population étudiée
- VPP (80%)= **probabilité post test** de la maladie chez un sujet qui a un test positif
- 1-VPN (21,5%)= **probabilité post test** de la maladie chez un sujet qui a un test négatif
- **L'information apportée par le test** permet de passer de la **probabilité pré test à post test**

Information apportée par le test diagnostique

- Ratio de vraisemblance positif de la douleur thoracique :

$$RV+ = \frac{P[T+ \mid M]}{P[T+ \mid NM]} = \frac{Se}{1-Sp} = \frac{0{,}947}{1-0{,}446} = 1{,}7$$

Information apportée par le test quand le résultat est positif

RV+ ≥ 1

Plus il est élevé plus il permet d'augmenter la probabilité pré test.

Information apportée par le test diagnostique

- Ratio de vraisemblance négatif :

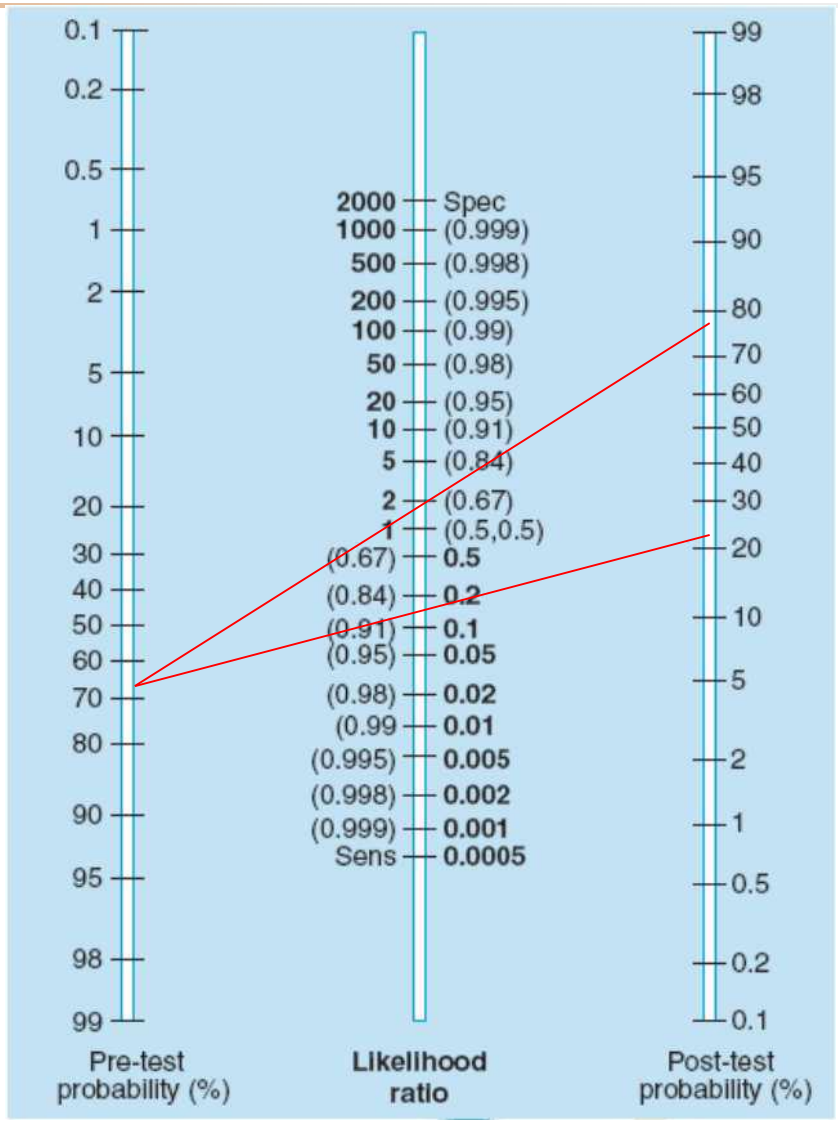
$$RV - = \frac{P[T - | M]}{P[T - |NM]} = \frac{1 - Se}{Sp} = \frac{1 - 0{,}947}{0{,}446} = 0{,}12$$

Information apportée par le test quand le résultat est négatif

$0 \leq$ RV- ≤ 1

Plus il est proche de 0 plus il permet de diminuer la probabilité pré test.

Normogramme de Fagan



Lecture critique d'une étude évaluant un test diagnostique

Détermination de l'objectif de l'étude

- **Phase 1 : phase exploratoire**

Le nouveau test a-t-il un intérêt diagnostique ?

- **Phase 2 : phase de challenge**

Evaluation des performances du test dans différents sous-groupes

Comparaison aux tests existants

- **Phase 3 : phase clinique**

Evaluation des performances dans un échantillon représentatif de la population cible

Détermination du seuil pour les tests quantitatifs

Type d'étude

- Etude transversale sur un échantillon de malades et un échantillon de non malades (phase 1, phase 2)
- Etude transversale ou de cohorte sur un échantillon représentatif d'une population (phase 3)
- Essai randomisé pour comparer des stratégies diagnostiques et thérapeutiques (phase 3)

Mesure du test diagnostique

- En aveugle du statut vis-à-vis de la maladie
- Standardisée
- Par un échantillon de médecins représentatifs des médecins qui seront amenés à interpréter le test (phase 3)

Choix du gold standard

- Gold standard parfait ou imparfait
- Définition précise de la maladie
- Détermination du statut vis-à-vis de la maladie de façon indépendante du résultat du test à évaluer
- Mesure chez tous les sujets de la même façon

Identification de la population

- Déterminer les caractéristiques de la population cible
- En phase 3, privilégier les études multicentriques
- Identifier les contre-indications à la réalisation du test

Biais potentiels

- Biais de sélection (échantillon non représentatif, spectre des patients insuffisants)
- Biais lié à l'utilisation d'un gold standard imparfait
- Biais lié à l'absence d'indépendance entre la détermination du statut vis-à-vis de la maladie et le résultat du test à évaluer (workup bias, incorporation bias)
- Biais de vérification (gold standard utilisé préférentiellement chez les sujets les plus à risque d'avoir la maladie)

Analyse statistique et résultats

- Choix du critère de mesure de la performance diagnostique
  - Etude de phase 1 et 2 : sensibilité, spécificité, comparaison de moyennes, courbe ROC, Aire sous la courbe ROC
  - Etude de phase 3 : seuil optimal, ratios de vraisemblance, valeurs prédictives, critère de résultat clinique

Présentation des résultats

- A qui généraliser les résultats ? : description de la population
- Quand estimation sensibilité, spécificité : seuil de positivité utilisé quand test quantitatif
- Quand estimation des valeurs prédictives : pour quelle prévalence de la maladie ?
- Fréquence des tests ininterprétables et causes

STAndards for reporting of Diagnostic Accuracy (STARD)

- Initiative du groupe de travail de la Cochrane
- Objectif : améliorer la qualité des articles d'évaluation de méthodes diagnostiques
- Site web : www.stard-statement.org

A RETENIR

- On prescrit un test diagnostique si son résultat est susceptible de modifier la prise en charge
- Un test est fiable si son résultat varie peu lorsqu'on répète la mesure
- L'exactitude d'un test dépend de sa capacité à discriminer les malades des non malades
- La sensibilité et la spécificité d'un test peuvent dépendre des caractéristiques des malades et des non malades

A RETENIR

- Les valeurs prédictives dépendent de la prévalence de la maladie
- Les ratios de vraisemblance permettent de passer de la probabilité pré test à la probabilité post test de la maladie
- Les différentes phases d'évaluation d'un test diagnostique et les schémas d'étude associés
- Les biais potentiels spécifiques aux études d'évaluation des tests diagnostique

MOTS EN ANGLAIS

- Fiabilité : reliability
- Exactitude : accuracy
- Concordance : agreement
- Sensibilité, spécificité : sensitivity, specificity
- Aire sous la courbe ROC : Area under the ROC curve
- Ratio de vraisemblance : likelihood ratio
- Valeur prédictive : predictive value

REFERENCES

- Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. Ed John Wiley & Sons, New York 2002.
- Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ 2002; 324: 539-41
- Haynes RB, Sackett DL, Guyatt GH, Tugwell P. Clinical epidemiology. LWW, Phyladelphia 2006

Des questions

Muriel Rabilloud
muriel.rabilloud@chu-lyon.fr