Ann Biol Clin 2005 ; 63 (2) : 145-54

La courbe ROC (*receiver operating characteristic*) : principes et principales applications en biologie clinique

H. Delacour[1,2]
A. Servonnet[2]
A. Perrot[1]
J.F. Vigezzi[1]
J.M. Ramirez[1]

[1] Laboratoire de biochimie, toxicologie cliniques, Hôpital d'Instruction des Armées du Val-de-Grâce, Paris

[2] Laboratoire de biochimie, toxicologie cliniques, Hôpital d'Instruction des Armées Robert Picqué, Villenave d'Ornon
<h_delacour@yahoo.fr>

Résumé. Les performances diagnostiques des tests de laboratoire sont généralement évaluées à l'aide de leur sensibilité, spécificité et valeurs prédictives positives et négatives. Malheureusement, ces indices ne reflètent qu'imparfaitement la capacité d'un test à distinguer les malades des non malades. Le recours à la courbe ROC (*receiver operating characteristic*) apparaît comme un outil de choix pour cette évaluation. Utilisée dans le domaine médical depuis les années 1960, la courbe ROC est une représentation graphique de la relation existante entre la sensibilité et la spécificité d'un test, calculée pour toutes les valeurs seuils possibles. Elle permet la détermination et la comparaison des performances diagnostiques de plusieurs tests à l'aide de l'évaluation des aires sous la courbe. Elle est aussi utilisée pour estimer la valeur seuil optimale d'un test en tenant compte des données épidémiologiques et médicoéconomiques de la maladie. Utilisée dans de nombreux domaines médicaux, cet outil statistique est facilement accessible grâce au développement de logiciels informatiques. Cet article expose les principes de construction et d'exploitation d'une courbe ROC.

Mots clés : *courbe receiver operating characteristic, sensibilité, spécificité, performance diagnostique, valeur seuil*

Abstract. Laboratory test's diagnostic performances are generally estimated by means of their sensibility, specificity and positive and negative predictive values. Unfortunately, these indices reflect only imperfectly the capacity of a test to correctly classify subjects into clinically relevant subgroups. The appeal to ROC (receiver operating characteristic) curve appears as a tool of choice for this evaluation. Used in the medical domain since the 60s, ROC curve is a graphic representation of the relation existing between the sensibility and the specificity of a test, calculated for all possible cut-off. It allows the determination and the comparison of the diagnostic performances of several tests. It is also used to consider the optimal cut-off of a test, by taking into account epidemiological and medical - economic data of the disease. Used in numerous medical domains, this statistical tool is easily accessible thanks to the development of computer softwares. This article exposes the principles of construction and exploitation of a ROC curve.

Key words: *ROC curve, sensitivity, specificity, diagnostic accuracy, cut-off*

Article reçu le 9 août 2004,
accepté le 30 novembre 2004

L'exercice de la biologie clinique est marqué par l'apparition régulière de nouveaux marqueurs ou de nouvelles techniques de dosages. Leurs performances diagnostiques sont le plus souvent évaluées à l'aide de leur sensibilité, spécificité et de leurs valeurs prédictives positives et négatives. Malheureusement ces indices ne reflètent qu'imparfaitement la capacité d'un test à distinguer les malades des non malades et ne permettent pas de le classer vis-à-vis

Tirés à part : H. Delacour

Tableau 1. Tableau de contingence d'un échantillon de N sujets classés en fonction de leur état de santé selon une méthode de référence (M^+/M^-) et le test étudié (S^+/S^-).

		Test étudié		Total
		Classés malades (S^+)	Classés non malades (S^-)	
Méthode de référence	Malades (M+)	Vrai positif (VP)	Faux positif (FP)	VP + FP
	Non malades (M-)	Faux positif (FP)	Vrai négatif (VN)	FP + VN
Total		VP + FP	FP + VN	N

des tests préexistants. Le recours à la courbe ROC (*receiver operating characteristic*) permet de pallier ces limites. Initialement développée dans les années 1950 à des fins militaires (exploitations des données Radar), son intérêt dans le domaine médical a été souligné dès 1960 par Lee Lusted [1, 2]. Depuis, cet outil statistique a été utilisé notamment dans le domaine pharmaceutique [3], en radiologie [4] et en biologie [5]. Étant parfois mal connu, il nous est paru utile de faire une mise au point sur son utilisation. Après un rappel sur les caractéristiques d'un test biologique (sensibilité, spécificité et valeurs prédictives), nous développerons les principes méthodologiques de la courbe ROC et ses applications en biologie clinique. L'objectif étant d'effectuer une présentation simple de cet outil statistique, les nombreux principes mathématiques le régissant ne seront pas abordés.

Caractéristiques d'un test biologique

Les caractéristiques d'un test sont de deux ordres : celles relevant exclusivement du test lui-même : ce sont la sensibilité (Se) et la spécificité (Sp), et celles fonction des caractéristiques intrinsèques du test (Se et Sp) et des caractéristiques de la population à qui il est appliqué (prévalence de la maladie dans la population considérée) : ce sont les valeurs prédictives positive et négative.

Sensibilité et spécificité

Considérons un échantillon de sujets extrait au hasard de la population chez qui est réalisé le test étudié. Les sujets sont classés en malade (M^+) ou non malade (M^-) à l'aide d'une méthode dite de référence ayant fait la preuve de sa valeur diagnostique (résultat d'une biopsie prostatique pour différencier un adénocarcinome d'une hypertrophie bénigne de la prostate par exemple) et en résultat positif (S^+) ou négatif (S^-) en fonction du résultat du test réalisé. Les résultats du double croisement (S^+/S^-) et (M^+/M^-) figurent dans le *tableau 1*.

Les vrais positifs (VP) sont les résultats positifs chez les sujets porteurs de la maladie, les faux positifs (FP) sont les résultats positifs chez les sujets indemnes de la maladie. De même, les vrais négatifs (VN) sont les résultats négatifs chez les sujets non malades et les faux négatifs (FN) les résultats négatifs chez les sujets malades.

La sensibilité du test est estimée par la proportion de vrais positifs chez les malades, soit :

$$Se = \frac{VP}{VP + FN}$$

La spécificité du test est estimée par la proportion de vrais négatifs chez les non malades, soit :

$$Sp = \frac{VN}{VN + FP}$$

Différents indices associant sensibilité et spécificité ont été proposés. Le plus classique est celui de Youden (Se + Sp – 1) qui vaut 1 quand l'examen est parfait. Plus un test réel approche de cette valeur, meilleur il est [6]. Un autre indice est le rapport de vraisemblance positif (*likelihood ratio "L"*), défini comme étant égal à (L = Se / (1 – Sp)). Idéalement infini, il est égal à 1 quand le test n'apporte aucune information.

Valeurs prédictives

La probabilité que le sujet soit réellement malade sachant que son test est positif s'appelle la valeur prédictive positive. De façon analogue, la valeur prédictive négative correspond à la probabilité que le sujet soit réellement indemne si son test est négatif. Ces deux probabilités peuvent se déduire de la connaissance de la sensibilité, de la spécificité et de la prévalence p de la maladie dans l'échantillon d'étude par le théorème de Bayes.

La valeur prédictive positive (VPP) est estimée par la proportion de vrais positifs parmi les sujets S^+, soit :

$$VPP = \frac{VP}{VP + FP} = \text{sensibilité} \times \frac{\text{prévalence de la maladie}}{\text{prévalence } S^+}$$

La valeur prédictive négative (VPN) est estimée par la proportion de vrais négatifs parmi les sujets S^-, soit :

$$VPN = \frac{VN}{VN + FN} = \text{spécificité} \times \frac{1 - \text{prévalence de la maladie}}{1 - \text{prévalence } S^+}$$

La valeur prédictive positive dépend donc de la sensibilité de la méthode mais aussi des prévalences de la maladie et de S^+. Le pouvoir prédictif positif est donc meilleur quand la maladie est fréquente et S^+ rare. De façon analogue, le pouvoir prédictif négatif est meilleur si la maladie est rare et S^- fréquent. (1 – VPN) est appelé taux de fausse alarme et (1 – VPP) le taux de fausse assurance. Comme pour la sensibilité et la spécificité, différents indices ont été développés comme la valeur discriminante (VD = VPP + VPN –1) ou encore l'efficience (E = P x Se + (1 – P) x Sp) qui représente le pourcentage de bons classements [7].

Effet de la valeur seuil sur les caractéristiques d'un test

Quand un test conduit à des résultats quantitatifs continus (cas de la majorité des tests biologiques), il est nécessaire de définir un seuil (ou valeur seuil) permettant de classer le résultat en normal (S^-) ou anormal (S^+). Le choix de cette valeur seuil influencera la sensibilité et la spécificité du test et donc ses valeurs prédictives.

Dans le cas hypothétique d'un test parfait, les distributions des résultats du test chez les sujets malades (M^+) et non malades (M^-) ne se superposent pas et la valeur seuil du test est située entre ces deux distributions (*figure 1*). Tous les sujets seront classés correctement à l'aide du test : la sensibilité et la spécificité sont de 100 %.

Malheureusement, pour la majorité des tests, les distributions des résultats des sujets (M^+) et (M^-) présentent une zone de chevauchement (*figure 2*). Tout choix de valeur seuil conduira dès lors à des erreurs de classifications : certains sujets malades (M^+) seront classés non malades (S^-), d'autres seront considérés comme malades (S^+) alors qu'ils ne le sont pas (M^-). Une diminution du seuil entraîne une diminution du nombre de faux négatifs (d'où une augmentation de la sensibilité) mais aussi une augmentation du nombre de faux positifs (donc une diminution de la spécificité). Inversement, une augmentation du seuil est accompagnée d'une diminution des faux positifs (augmentation de la spécificité) et d'une augmentation des faux négatifs (diminution de la sensibilité). Ainsi, sensibilité et spécificité varient inversement.

Chaque seuil possède des valeurs de sensibilité et de spécificité qui lui sont propres et qui ne décrivent en aucun cas les performances du test à d'autres valeurs seuils. Ce phénomène doit être pris en compte lors des comparaisons des tests diagnostic et est une des indications de la courbe ROC.

La courbe ROC : principes et construction

La courbe ROC est une représentation graphique de la relation existante entre la sensibilité et la spécificité d'un test pour toutes les valeurs seuils possibles. L'ordonnée représente la sensibilité et l'abscisse correspond à la quantité (1 - spécificité) (*figure 3*). Sa construction nécessite l'emploi d'un logiciel de calcul spécialisé (*tableau 2*).

Dans tous les cas, la méthode de travail est identique. Les résultats du test sont classés par ordre croissant et pour chaque valeur, un tableau de contingence identique au *tableau 1* est réalisé. Le calcul des effectifs VP, FP, VN, FN permet de déduire la sensibilité et la spécificité du test pour chaque valeur obtenue. Les couples {1 – spécificité, sensibilité} sont alors placés sur la courbe. Leur jonction par des lignes droites conduit à un tracé en marches d'escaliers reliant le coin inférieur gauche du graphique (Se = 0 et Sp = 1) au coin supérieur droit (Se = 1 et Sp = 0). La technique décrite a l'avantage d'être applicable pour toute distribution statistique des sujets indemnes et malades : on parle de courbe ROC non paramétrique.

Pour chaque valeur seuil, l'inclusion d'un vrai positif accroît la sensibilité du test. Graphiquement, la jonction du nouveau point avec le point précédemment obtenu est une ligne verticale. À l'inverse, l'inclusion d'un faux positif, à l'origine d'une diminution de la spécificité, produit une ligne horizontale. En cas d'inclusion simultanée d'un vrai positif et d'un faux positif (on parle dans ce cas de résultats ou de sujets ex aequo), la sensibilité et la spécificité du test varient conjointement. La résultante au niveau

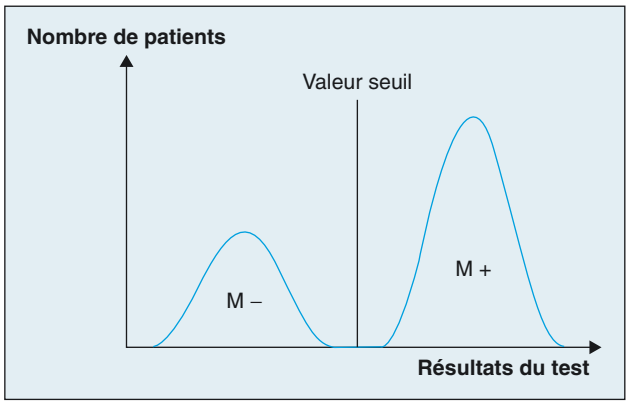

Figure 1. Distributions des résultats dans le cas d'un test parfait en fonction du caractère malade ou non des sujets (M^+/M^-). La valeur seuil se situe entre les deux distributions. Tout résultat se situant au dessus de cette valeur est considéré comme "positif", tout résultat situé en dessous de cette valeur comme "négatif".

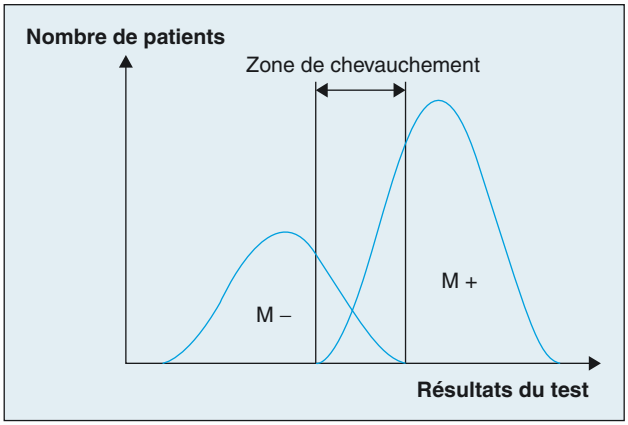

Figure 2. Distributions des résultats dans le cas d'un test "réel" en fonction du caractère malade ou non des sujets (M^+/M^-) : les résultats des tests pour les sujets sains et malades présentent un chevauchement. Le choix de la valeur seuil influencera la sensibilité et la spécificité du test.

de la courbe ROC est une diagonale, c'est-à-dire la combinaison d'une ligne verticale et d'une ligne horizontale. Il est en effet impossible de déterminer le trajet exact de la courbe : augmentation de la sensibilité puis diminution de la spécificité se traduisant graphiquement par une ligne verticale puis une ligne horizontale ou l'inverse ? La diagonale permet de faire la moyenne entre ces deux possibilités.

Si l'observation de quelques résultats ex aequo ne pose pas problème pour l'exploitation ultérieure des caractéristiques de la courbe ROC, il n'en est pas de même lorsque leur nombre devient trop important. L'aire sous la courbe, reflet des performances diagnostiques du test, est alors sous-estimée.

Une alternative est alors le recours à une courbe ROC paramétrique comme présentée sur la *figure 3*. Cependant, la construction d'une telle courbe n'est valable que si on suppose que les distributions de deux populations (sujets sains et sujets malades) possèdent une forme analytique. La modélisation la plus rencontrée en biologie est binomiale et utilise deux distributions gaussiennes [8]. Pour la construction de la courbe, les données expérimentales sont regroupées en intervalles et les couples {1 – spécificité, sensibilité} sont calculés pour chacun. Cette approche de la courbe ROC est préférée en cas d'échantillons de taille importante et lorsque le nombre d'ex aequo est important.

Le *tableau 3* liste les avantages et les inconvénients des deux approches.

L'exemple suivant illustre la construction d'une courbe ROC. Les données utilisées sont issues d'une étude des caractéristiques analytiques et cliniques des dosages de PSA total et de PSA libre par électrochimiluminescence [9]. Au total 104 prélèvements ont été analysés, 52 appartenant à des sujets présentant un adénocarcinome et 52 une hypertrophie de la prostate. Le *tableau 4* regroupe les résultats des rapports PSA libre / PSA total (Rp) obtenus.

Le calcul des coordonnées de la courbe ROC peut se faire à l'aide du *tableau 5*. La première colonne contient les valeurs seuils pour lesquelles la sensibilité et la spécificité du test sont calculées en fonction des observations. Les colonnes suivantes regroupent les effectifs des VP, FP, VN et FN. Dans notre cas, le nombre de vrais positifs est le nombre de sujets atteints d'un adénocarcinome et présentant un Rp inférieur ou égal à la valeur seuil. Les deux dernières colonnes présentent les coordonnées de la courbe ROC. Seules les coordonnées des 20 premières valeurs de Rp ont été indiquées. En poursuivant le même raisonnement pour les autres valeurs du rapport, la courbe ROC présentée *figure 4* est obtenue.

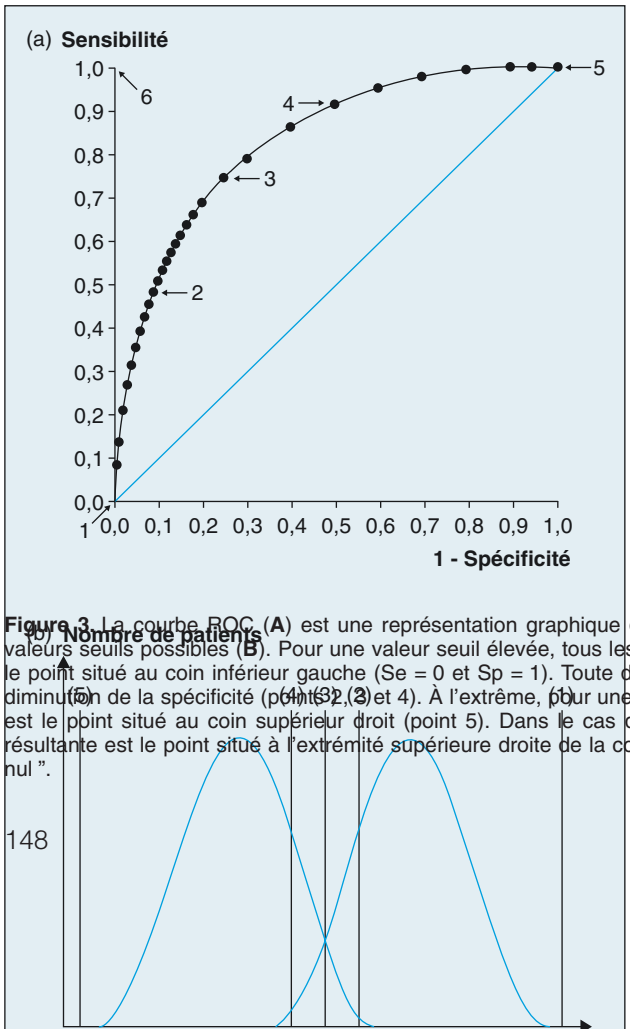

Figure 3. La courbe ROC (**A**) est une représentation graphique de la relation existante entre sensibilité et spécificité pour toutes les valeurs seuils possibles (**B**). Pour une valeur seuil élevée, tous les sujets sont classés S^- (point 1) : la résultante sur la courbe ROC est le point situé au coin inférieur gauche (Se = 0 et Sp = 1). Toute diminution du seuil entraîne une augmentation de la sensibilité et une diminution de la spécificité (points 2, 3 et 4). À l'extrême, pour une valeur seuil très basse, tous les sujets sont classés S^+ : la résultante est le point situé au coin supérieur droit (point 5). Dans le cas d'un test " parfait ", la sensibilité et la spécificité sont égales à 1. La résultante est le point situé à l'extrémité supérieure droite de la courbe ROC (point 6). La diagonale bleue représente un test " d'apport nul ".

Tableau 2. Références des logiciels informatiques étudiés par Stephan *et al.* [23].

	AccuROC 2.5	Analyse-It	CMDT	GraphROC 2.1
Site internet	www.accumetric.com	www.analyse-it.com	NC	www.netti.fi/~maxiw
Configuration minimale requise				
Système d'exploitation	Windows 95	Windows 95	Windows 95	Windows 3.1
Mémoire vive (méga octets)	NC	16	4	4
Espace disque dur (méga octets)	NC	6	1,2	NC
Processeur	NC	Pentium 100 Mhz	i 486	i 486
Prix	150 $ canadiens	76 – 100 £	Gratuit	61 – 297 €
	MedCalc 6.16	**MROC 1.0**	**ROCKIT 0.9 B**	**SPSS 10.0**
Site internet	www.medcalc.be	NC	NC	www.spss.com
Configuration minimale requise				
Système d'exploitation	Windows 95	Windows 95	Windows 3.1	Windows 3.1
Mémoire vive (méga octets)	8	8	4	16
Espace disque dur (méga octets)	4	5	NC	160
Processeur	i 486	i 486	NC	Pentium 90 Mhz
Prix	199 $ US	350 €	Gratuit	1 280 €

NC : non communiquée dans le travail ou adresse internet non fonctionnelle le 02 novembre 2004.

Tableau 3. Avantages et inconvénients des courbes ROC non paramétriques et paramétriques d'après Zweig *et al.* [20].

	Méthode non paramétrique	Méthode paramétrique
Avantages	Utilisation de tous les points expérimentaux. Pas d'extrapolation paramétrique des données expérimentales. Courbe passe par tous les points observés. Détermination de l'aire sous la courbe simple. Absence de biais pour la détermination de la sensibilité, de la spécificité et de l'aire sous la courbe.	Tracé "lissé". Comparaison de courbes possibles à toutes les sensibilités et spécifités.
Inconvénients	Tracé en marches d'escaliers. Construction longue avec les échantillons de taille importante. Sujets *ex aequo* source de sous estimation des performances diagnostiques du test. Comparaison des courbes uniquement aux spécificités et sensibilités observées.	Extrapolation paramétrique des données expérimentales. Détermination de l'aire sous la courbe plus complexe. Biais dans l'estimation de l'aire sous la courbe possible. Pertes de données par regroupement des points expérimentaux. Courbe ne passe pas nécessairement par les points expérimentaux.

Tableau 4. Valeurs du rapport PSA libre/PSA total (%) obtenus chez les 104 sujets (52 adénocarcinomes et 52 hypertrophies de la prostate) de l'étude des caractéristiques analytiques des dosages du PSA total et libre par électrochimiluminescence [9].

Sujets présentant une hypertrophie	6,3 7,0 7,4 7,9 8,7 9,0 9,5 9,6 9,6 9,6 9,7 9,7 10,5 10,6 11,1 11,2 11,7 12,1 12,3 12,3 13,1 13,4 13,5 13,8 14 14,1 14,3 14,3 14,4 15,4 15,5 15,8 16,3 16,5 16,9 17,3 17,8 18,4 18,9 19,1 19,1 19,8 20,2 20,5 20,7 21,4 22,5 23,9 25,7 26,5 27,6 28,9 29,4
Sujets présentant un adénocarcinome	4,1 4,9 5,0 5,1 5,7 6,0 6,2 6,3 6,4 6,7 6,7 7,0 7,2 7,3 7,5 7,6 7,7 8,1 8,6 8,8 8,9 9,0 9,0 9,2 9,5 9,7 9,9 10,2 10,5 10,5 11,1 12,2 12,6 12,9 12,9 13,3 13,5 13,7 13,8 14,0 14,6 14,6 14,9 15,5 15,5 15,9 17,3 17,7 17,7 19,3 20,1 22,8

Exploitation d'une courbe ROC et applications en biologie clinique

Détermination des performances diagnostiques d'un test

L'exploitation d'une courbe ROC débute par une étude de son allure générale et par la détermination de l'aire sous la courbe (ASC) associée. Dans le cas d'une courbe ROC non paramétrique, ce calcul consiste à sommer les surfaces des trapèzes verticaux reliant deux points successifs. La détermination des ASC des courbes ROC paramétriques est plus complexe et nécessite l'emploi de programmes informatiques. Les logiciels statistiques effectuent cette opération automatiquement et y associent le calcul de son intervalle de confiance à 95 % ($i_{95\%}$).

Tableau 5. Calcul des coordonnées de la courbe ROC de l'étude réalisée par Ramirez *et al.* [9] pour les 20 premières valeurs du rapport PSAl/PSAt. Le nombre de vrais positifs correspond au nombre de sujets atteints d'un adénocarcinome (M^+) et présentant un rapport PSAl/PSAt inférieur ou égal à la valeur seuil. Le nombre de vrais négatifs est le nombre de sujets atteints d'une hypertrophie de la prostate (M^-) et possédant un rapport PSAl/PSAt supérieur à la valeur seuil.

Rapport PSAl/PSAt	Nombre de vrais positifs	Nombre de faux négatifs	Nombre de vrais négatifs	Nombre de faux positifs	Sensibilité (%)	1 – Spécificité (%)
4,1	1	50	52	0	1,96	0
4,9	2	49	52	0	3,92	0
5	3	48	52	0	5,88	0
5,1	4	47	52	0	7,84	0
5,7	5	46	52	0	9,80	0
6	6	45	52	0	11,76	0
6,2	7	44	52	0	13,72	0
6,3	8	43	51	1	15,68	1,92
6,4	9	42	51	1	17,65	1,92
6,7	11	40	51	1	21,57	1,92
7	12	39	50	2	23,53	3,84
7,2	13	38	50	2	25,49	3,84
7,3	14	37	50	2	27,45	3,84
7,4	14	37	49	3	27,45	5,76
7,5	15	36	49	3	29,41	5,76
7,6	16	35	49	3	31,37	5,76
7,7	17	34	49	3	33,33	5,76
7,9	17	34	48	4	33,33	7,69
8,1	18	33	48	4	35,29	7,69
8,6	19	32	48	4	37,25	7,69

Dans le cas d'un test parfait (comme celui présenté sur la *figure 1*), la courbe ROC passe par le point de coordonnées {0, 1} (Sp = 1, Se = 1), l'aire sous la courbe associée est 1. À l'inverse, si les distributions des résultats des sujets malades et des sujets sains se superposent parfaitement, la sensibilité est égale à la quantité (1 – spécificité) pour toutes les valeurs du test. La courbe ROC résultante est une diagonale reliant l'extrémité inférieure gauche à l'extrémité supérieure droite du graphique. La surface sous la courbe est de 0,5 : le test est qualifié d'apport nul. Une surface inférieure à 0,5 est obtenue pour un test dont la réponse est inversée (un sujet malade présente un résultat inférieur à un sujet non malade), une surface nulle étant caractéristique d'un test inversé parfait.

L'aire sous la courbe permet d'évaluer l'intérêt diagnostique d'un test. On distingue les tests d'apport nul (ASC = 0,5), peu informatifs (0,5 ≤ ASC < 0,7), moyennement informatifs (0,7 ≤ ASC < 0,9), très informatifs (0,9 ≤ ASC < 1) et parfaits (ASC = 1) [10]. Le même raisonnement peut être mené avec un test inversé, les seuils des aires sous la courbe étant 0,5 ; 0,3 ; 0,1 et 0. Une aire sous la courbe de 0,8, par exemple, signifie qu'un sujet malade aura un résultat pour le test supérieur à celui d'un sujet sain dans 80 % des cas.

Cette simple étude de l'ASC d'un test a permis à Ohlmann *et al.* [11] d'étudier la capacité de la troponine I (TNI) à prédire la taille de la zone nécrosée et l'apparition d'une insuffisance ventriculaire gauche suite à un infarctus du myocarde. Pour ce faire, des dosages de TNI ont été réalisés chez 63 sujets au moment de leur hospitalisation puis pendant les 72 heures suivant leur angioplastie. L'analyse des courbes ROC a démontré que les concentrations de TNI à l'admission ne sont corrélées ni avec la taille de la zone nécrosée (ASC = 0,55, $i_{95\%}$: 0,41 – 0,70), ni avec l'apparition d'une insuffisance ventriculaire gauche (ASC = 0,51, $i_{95\%}$: 0,35 – 0,73). À l'inverse, les concentrations de TNI entre la 6^e et la 72^e heure suivant l'angioplastie possèdent un caractère prédictif vis-à-vis de ces deux éléments. En effet, l'ASC associée est comprise entre 0,95 et 0,98 pour la taille de la zone nécrosée et entre 0,80 et 0,90 pour l'apparition d'une insuffisance ventriculaire gauche.

Comparaison de plusieurs tests diagnostiques

Les courbes ROC sont également employées pour la comparaison des performances diagnostiques de tests biologiques. Si l'on trace les courbes ROC associées aux différents tests évalués, celles-ci ne seront pas identiques. La question classique en statistique se pose alors : cette différence est-elle significative ? En d'autres termes peut-elle être expliquée par le hasard ou est-elle liée à un écart entre

les caractéristiques cliniques des tests ? Différentes méthodes ont été proposées pour répondre à cette question.

La première utilise la comparaison des aires totales sous la courbe. Il est important de savoir si les courbes sont obtenues à l'aide de données indépendantes (chaque test est étudié dans une population différente) ou appariées (tous les tests sont étudiés dans la même population). En biologie clinique, cette seconde possibilité est la plus souvent rencontrée. Il faut donc employer des méthodes statistiques adaptées afin de tenir compte de cette corrélation existante entre les courbes comparées. Différentes méthodes peuvent être utilisées comme celle de Hanley et MacNeil [12] ou de De Long [13] pour les courbes non paramétriques ou la méthode développée par Metz [8] pour les courbes paramétriques. Un test est considéré comme possédant des performances diagnostiques supérieures à un second si la comparaison de leur ASC aboutit à une différence significative.

La comparaison des courbes ROC peut ainsi être employée dans le cadre de transfert de techniques afin de s'assurer de l'équivalence des performances des méthodes de dosages. Ainsi, l'efficacité clinique des pourcentages de PSA libres déterminés par immunoradiobiologie (Cis Bio) et par électrochimiluminescence (Elecsys – Roche) a été évaluée. Les dosages de PSA total et de PSA libre ont été réalisés sur 104 prélèvements (52 adénocarcinomes et 52 hypertrophies bénignes) à l'aide des deux techniques étudiées puis les courbes ROC associées ont été tracées. Ces dernières présentant une allure comparable, une comparaison des aires totales a été effectuée. La différence non significative (ASC = 0,76 pour la technique isotopique et ASC = 0,74 pour la technique luminescente) traduit une efficacité clinique identique [9].

La même technique de comparaison d'aire sous la courbe a été utilisée dans la mise au point du fibrotest. Présenté comme une alternative à la biopsie hépatique pour estimer le degré de fibrose chez des sujets porteurs du virus de l'hépatite C (VHC), cet index combine les dosages sanguins de 5 paramètres biochimiques (alpha-2-macroglobuline, haptoglobine, apolipoprotéine A1, bilirubine totale et gammaglutamyltranspeptidase) avec un ajustement sur l'âge et le sexe. Initialement, trois index associant cinq, six ou dix paramètres biochimiques ont été évalués. Leurs performances diagnostiques ont été évaluées chez des sujets présentant une fibrose significative (stade F2F3F4) et ceux n'ayant pas de fibrose significative (stade F0F1). Les courbes ROC obtenues se sont révélées très similaires, leurs aires sous la courbe étant de 0,837 (5 paramètres), 0,847 (6 paramètres) et 0,851 (10 paramètres). Leur comparaison à l'aide de la technique par les aires totales a abouti à une différence non significative. La combinaison de cinq paramètres biochimiques a donc été retenue par les auteurs puisqu'elle possède les mêmes performances diagnostiques que les autres index étudiés pour un moindre coût [14].

Cependant, la comparaison des aires globales n'est applicable que si les courbes étudiées ont une forme analogue. En effet, deux courbes peuvent présenter une allure différente (et une intersection) et posséder une aire sous la courbe identique. La comparaison par les aires totales ne permet pas, dans ce cas, de mettre en évidence la supériorité d'un test par rapport à un autre. Pour pallier cette limite, une comparaison ponctuelle peut être réalisée. Elle consiste à ne prendre en compte que la partie jugée intéressante de la courbe ROC et à ne pas tenir compte du reste. La comparaison peut être menée à sensibilité ou spécificité [15-17] fixée. Jiang *et al.* [17] proposent l'emploi de l'index d'aire partielle en cas de comparaison à sensibilité fixée. Cet index est obtenu en divisant l'aire sous la courbe ROC de la portion d'intérêt par l'aire maximale possible de cette portion. L'index, allant de 0 à 1, peut être utilisé pour confronter les tests biologiques.

En 2001, Hammerer *et al.* ont comparé les performances diagnostiques du BNP, du NT-proBNP et du NT-proANP en cas d'insuffisance cardiaque [18]. Chez les sujets présentant une fraction d'éjection ventriculaire gauche inférieure à 55 %, la comparaison des aires totales sous les

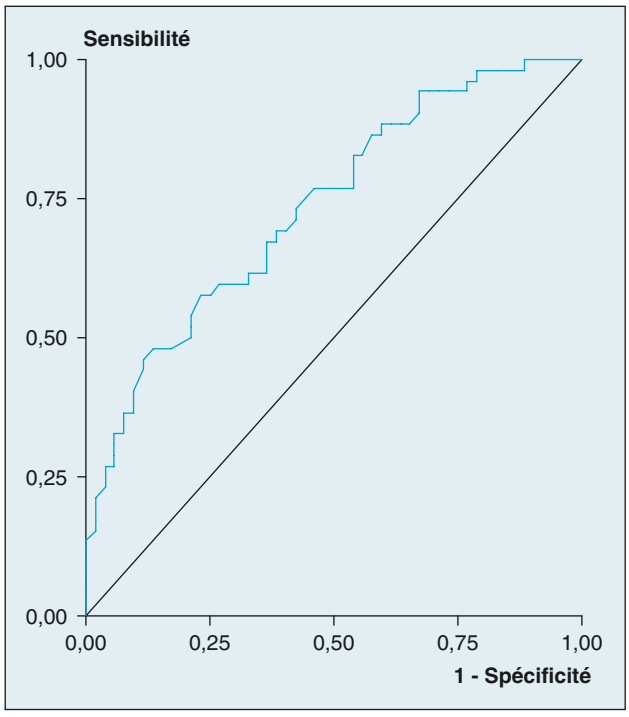

Figure 4. Courbe ROC observée lors de l'évaluation des performances cliniques des rapports PSA libre/PSA total obtenus par électrochimiluninescence. Les coordonnées de la courbe sont présentées dans le *tableau 4* [9].

courbes ROC n'a pas permis de mettre en évidence une différence significative entre le NT-proBNP (ASC = 0,67) et le NT-proANP (ASC = 0,69) ; le BNP possédant une ASC de 0,75. Cependant, les courbes ROC du NT-proBNP et du NT-proANP présentant une intersection, une comparaison ponctuelle aurait pu être menée (*figure 5*). Pour une spécificité fixée entre 0,8 et 1, une supériorité du NT-proBNP aurait été notée. À l'inverse, selon cette étude, pour une sensibilité supérieure à 0,7, le NT-proANP serait un meilleur marqueur biologique de l'insuffisance cardiaque. En résumé, la comparaison ponctuelle permet de choisir le meilleur test biologique en fonction des priorités du clinicien : limiter le nombre de faux positifs ou au contraire prévenir les faux négatifs.

Détermination de la valeur seuil optimale d'un test

La courbe ROC permet également de déterminer la valeur seuil optimale d'un test. Intuitivement, celle-ci peut être identifiée comme étant le point de la courbe le plus éloigné de la diagonale représentant le test d'apport nul. Ce point correspond également au maximum de l'indice de Youden (Se + Sp –1).

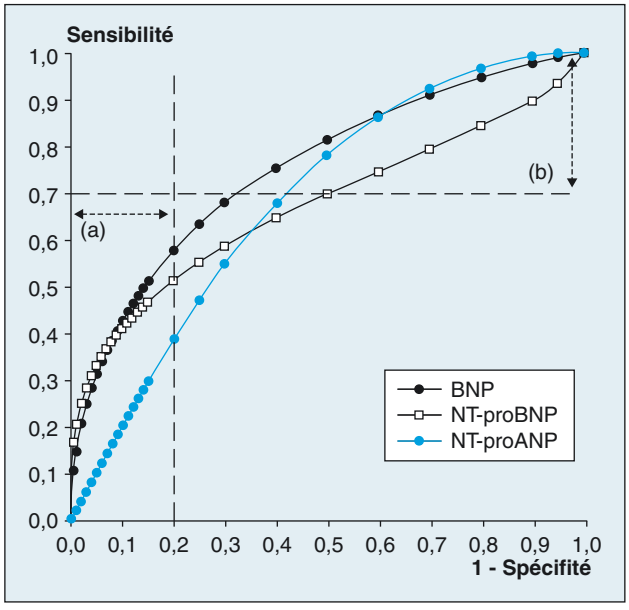

Figure 5. Courbes ROC obtenues par Hammerer *et al.* lors de la comparaison des performances diagnostiques du BNP, du NT-proBNP et du NT-proANP dans l'insuffisance ventriculaire gauche [18]. La comparaison des aires totales sous les courbes ROC n'a pas permis de mettre en évidence une différence significative entre le NT-proBNP (ASC = 0,67) et le NT-proANP (ASC = 0,69) ; le BNP possédant une ASC de 0,75. Cependant, les courbes ROC du NT-proBNP et du NT-proANP présentant une intersection, une comparaison ponctuelle aurait pu être menée. Pour une spécificité fixée entre 0,8 et 1 (zone a), une supériorité du NT-proBNP aurait été notée. Pour une sensibilité supérieure à 0,7 (zone b), le NT-proANP serait un meilleur marqueur biologique de l'insuffisance cardiaque.

Cependant, la recherche d'un seuil nécessite idéalement la prise en compte de données épidémiologiques (prévalence de la maladie) et médicoéconomiques (coût du traitement, coût des effets indésirables du traitement...). Si une maladie possède un traitement onéreux aux effets secondaires potentiellement graves, il convient de limiter au maximum le nombre de faux positifs, donc de choisir une spécificité élevée. La valeur seuil sera située dans ce cas dans la partie inférieure gauche de la courbe ROC. À l'inverse, certaines maladies possèdent des complications graves qui peuvent être évitées si un traitement simple est mis en place précocement : le test doit posséder une sensibilité élevée. La valeur seuil se situera au niveau de la partie supérieure droite de la courbe ROC.

Si C_{VP}, C_{VN}, C_{FP} et C_{FN} sont les coûts associés respectivement à un sujet classé vrai positif, vrai négatif, faux positif et faux négatif et si p est la prévalence de la maladie dans la population, alors la valeur seuil optimale peut être calculée à l'aide de la formule :

$$S = \frac{(1-p)}{p} \times \frac{(C_{FP} - C_{VN})}{C_{FN} - C_{VP}}$$

où S est la pente de la tangente à la courbe ROC à la valeur seuil optimale du test [19]. Cependant, cette dernière n'est pleinement applicable qu'avec une courbe ROC paramétrique.

Si $C_{FP} = C_{FN}$ et $C_{VP} = C_{VN}$ et que p = 0,5, alors S est égal à 1. La valeur seuil optimale obtenue correspond au point le plus éloigné de la diagonale représentant le test d'apport nul. Le seuil déterminé intuitivement est donc une approximation de cette formule de calcul.

La *figure 6* illustre le choix d'une valeur seuil à l'aide des deux méthodes présentées. La valeur seuil 1 (sensibilité = 0,79, spécificité = 0,81) est obtenue en recherchant le point le plus éloigné de la diagonale représentant le test sans intérêt. La détermination de la valeur seuil 2 est effectuée à l'aide de la formule précédente. Si la prévalence de la maladie est de 0,75 et que $C_{FP} = C_{FN}$ et $C_{VP} = C_{VN}$ alors S est égal à 3. La sensibilité et la spécificité associées à ce seuil sont respectivement de 0,43 et de 0,93. Ce choix de valeur seuil augmente le nombre de faux négatifs au profit d'une diminution du nombre de faux positifs.

Discussion

La courbe ROC s'est imposée en biologie clinique depuis plusieurs années. En 1993, Zweig *et al.* ont présenté cet outil statistique et ses principales applications pour notre profession [20]. Elle est incluse dans la liste éditée dans l'*American association for clinical chemistry* regroupant les paramètres à étudier lors de l'évaluation d'un test bio-

logique [21]. Dès lors, son utilisation est fréquente dans les travaux anglo-saxons. Obuchowski *et al.* ont ainsi relevé 58 publications originales mentionnant la courbe ROC entre janvier 2001 et décembre 2002 dans le journal *Clinical chemistry* [22].

Cependant, 38 (soit 65,5 %) présentaient des biais méthodologiques. Les erreurs les plus couramment rencontrées sont l'absence de comparaisons statistiques pour évaluer la supériorité d'un test par rapport à un autre, un mauvais choix de test statistique pour effectuer cette comparaison (test non adapté aux données appariées), un mauvais choix de tests de référence permettant de classer les sujets en malades et non malades. Comment expliquer cette proportion d'erreurs méthodologiques ?

Malgré sa facilité d'approche, l'emploi de la courbe ROC et plus précisément la comparaison de courbes ROC est un exercice délicat, le choix des tests statistiques à employer étant fondamental. Dès lors l'utilisation d'un logiciel informatique spécialisé est recommandée.

En 2003, Stephan *et al.* ont comparé les performances de huit programmes informatiques (*tableau 2*) [23]. Si tous les programmes calculent l'aire sous la courbe associée à un test (les valeurs obtenues étant équivalentes), ils diffèrent par les fonctionnalités associées :

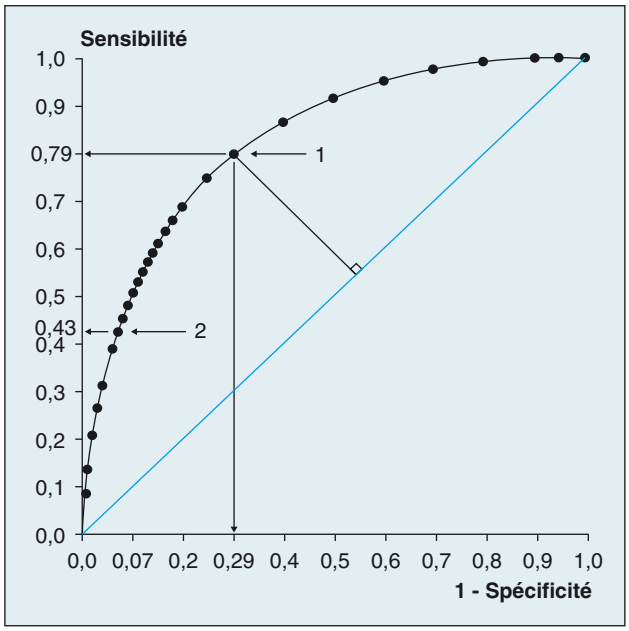

Figure 6. Exemple de choix d'une valeur seuil optimale à l'aide d'une courbe ROC. La valeur seuil 1 (sensibilité = 0,79 et spécificité = 0,71) est obtenue en recherchant le point de la courbe le plus éloigné de la diagonale représentant le test " d'apport nul ". La valeur seuil 2 (sensibilité = 0,43 et spécificité = 0,93) prend en compte des données épidémiologiques (prévalence de la maladie = 0,25) et médico-économiques ($C_{FP} = C_{FN}$ et $C_{VP} = C_{VN}$). La pente de la tangente à la courbe ROC à cette valeur seuil du test est égale à 3.

– le nombre maximal de courbes ROC par graphique : de 1 (ce qui empêche les comparaisons visuelles) à 3 ;
– la possibilité de comparer statistiquement différentes courbes (6 programmes sur 8) ;
– l'impression et l'exportation vers d'autres logiciels bureautiques des résultats statistiques (4/8) et des graphiques (4/8) ;
– le calcul de la sensibilité et de la spécificité associées à une valeur seuil (3/8).

En outre, aucun ne semble permettre une comparaison ponctuelle à sensibilité ou spécificité fixées. Pour ces auteurs, une analyse complète d'une courbe ROC nécessite l'emploi de différents programmes, aucun ne répondant à tous les critères étudiés. Cependant, l'utilisation de trois logiciels était recommandée : Analyse-it, Accu-Roc et MedCalc.

Conclusion

Généralement, les tests biologiques conduisent à des résultats quantitatifs continus. Les distributions des résultats des sujets malades et non malades présentent, le plus souvent, une zone de recouvrement. Dès lors il est nécessaire de définir une valeur seuil permettant de répartir les résultats obtenus en pathologiques et non pathologiques. Tout choix de seuil conduit à des erreurs de classification ayant des répercussions sur les valeurs de sensibilité et de spécificité qui lui sont associées. Représentation graphique de la sensibilité en fonction de (1- spécificité) pour toutes les valeurs seuil possibles, la courbe ROC est une méthode de choix pour l'étude de l'efficacité clinique d'un test biologique. En effet, la comparaison des aires sous la courbe permet d'apprécier et de classer les performances diagnostiques de plusieurs tests, mieux que la simple étude des couples sensibilité – spécificité. La courbe ROC peut également être utilisée pour déterminer la valeur seuil optimale d'un test tout en prenant en compte les données épidémiologiques et médicoéconomiques de la maladie. Le développement de nombreux logiciels informatiques, d'inégale valeur, rend plus simple l'utilisation de cet outil statistique.

Références

1. Lusted LB. Logical analysis in roentgen diagnosis. *Radiology* 1960 ; 74 : 178-93.

2. Lusted LB. Signal detectability and medical decision making. *Science* 1971 ; 171 : 1217-9.

3. Joe Au YH, Eissa S, Jones BE. Receiver operating characteristic analysis for the selection of threshold values for detecion of capping in powder compression. *Ultrasonics* 2004 ; 42 : 149-53.

4. Van Erkel AR, Pattynama P. Receiver operating characteristic (ROC) analysis : basic principles and applications in radiology. *Eur J Radiol* 1998 ; 27 : 88-94.

5. Greiner M, Pfeiffer D, Smith RD. Principles and practical application of the receiver–operating characteristic analysis for diagnostic tests. *Prev Vet Med* 2000 ; 45 : 23-41.

6. Landais P, Besson C, Jais JP. Evaluation of the diagnostic contribution of a test. Main information indices. *J Radiol* 1994 ; 75 : 141-50.

7. Bouyer J, Hémon D, Cordier S, *et al.* In : *Épidémiologie, principes et méthodes quantitatives*. Paris : Inserm, 1995 : 338-40.

8. Metz CE, Hermann BA, Shen JH. Maximum likehood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Med Decis Mak* 1998 ; 17 : 1033-53.

9. Ramirez JM, Vigezzi JF, Delacour H, Gidenne S, Clerc Y. Étude analytique des dosages de PSA total, PSA libre sur l'immunoanalyseur Elecsys 2010® Roche Diagnostics et de l'efficicacité cliniques des pourcentages de PSA libre. *Immunoanal Biol Spec* 2002 ; 17 : 257-63.

10. Swets JA. Mesuring the accuracy of diagnosis system. *Science* 1988 ; 240 : 1285-93.

11. Ohlmann P, Monassier JP, Michotey MO, *et al.* Troponin I concentrations following primary percutaneous coronary intervention predict large infarct size and left ventricular dysfunction in patients with ST-segment elevation acute myocardial infarction. *Atherosclerosis* 2003 ; 168 : 181-9.

12. Hanley JA, McNeil BJ. A method for comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983 ; 148 : 839-43.

13. De Long ER, DeLong DM, Clarke Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves : a non parametric approach. *Biometrics* 1988 ; 44 : 837-45.

14. Imbert-Bismuth F, Ratziu V, Pieroni L, Charlotte F, Benhamou Y, Poynard T. Biochemical markers of liver fibrosis in patients with hepatitis C virus infection : a prospective study. *Lancet* 2001 ; 357 : 1069-75.

15. Thomson ML, Zucchini W. On the statistical analysis of ROC curves. *Stat Med* 1989 ; 8 : 1277-90.

16. McClish D. Analyzing a portion of the ROC curves and their analysis. *Med Decis Mak* 1989 ; 9 : 190-5.

17. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial index for highly sensitive diagnostic tests. *Radiology* 1996 ; 201 : 745-50.

18. Hammerer-Lercher A, Neubauer E, Müller S, Pachinger O, Puschendorf B, Mair J. Head-to-head comparaison of N-terminal pro-brain natriuretic peptide, brain natriuretic peptide and N-terminal pro-atrial natriuretic peptide in diagnosing left ventricular dysfunction. *Clin Chim Acta* 2001 ; 310 : 193-7.

19. Smith RD. Evaluation of diagnostic tests. In : *Veterinary clinical epidemiology. A problem oriented approach*. Boca Raton : CRC Press, 1992 : 31-43.

20. Zweig MH, Campbell G. Receiver operating characteristic (ROC) plots : a fundamental evolution tool in clinical medicine. *Clin Chem* 1993 ; 39 : 561-77.

21. Bruns D, Huth EJ, Magid E, Younf DS. Toward a checklist for reporting if studies of diagnostic accurancy of medical tests. *Clin Chem* 2000 ; 46 : 893-5.

22. Obuchowski NA, Lieber ML, Wians FH. ROC curves in clinical chemistry : uses, misuses, and possible solutions. *Clin Chem* 2004 ; 50 : 1118-25.

23. Stephan C, Wesseling S, Schink T, Jung K. Comparaison of eight computer programs for receiver-operating characteristic analysis. *Clin Chem* 2003 ; 49 : 433-9.