Outils

Un test diagnostique est un examen d'imagerie, une échelle binarisée ou un examen biologique dont le résultat doit permettre de détecter la présence ou de prédire la survenue d'un état pathologique. Un « nouveau » test doit être évalué selon une méthodologie adaptée avant d'être utilisé en routine. Cette évaluation comporte plusieurs phases et permettra *in fine* d'estimer les performances diagnostiques en quantifiant les indices de précision, notamment la sensibilité, la spécificité, et le rapport des cotes. La valorisation de ces indices ne suffit pas à valider le test. Les modifications que son intégration dans la démarche médicale entraîne et l'impact de son utilisation en routine sur l'état de santé des patients doivent aussi être quantifiés.

Behrouz Kassaï
Julie Manière
Kim-An Nguyen
Inserm, CIC 201, Service de pharmacologie clinique, Hospices civils de Lyon, Université Lyon 1

Mots clés : test diagnostique, indices, seuils
DOI : 10.1684/med.2006.0010

Qu'attendre d'un « test diagnostique » ?

Les indices de précision des tests diagnostiques (1re partie)

La méconnaissance des méthodologies adaptées à l'évaluation des tests est un obstacle important à la généralisation de la médecine factuelle au domaine diagnostique [1]. D'importants progrès ont été effectués dans le domaine de la méthodologie d'évaluation des tests ces dernières années avec de réels espoirs d'une utilisation plus rationnelle de ces examens [2]. Mais l'absence de cadre réglementaire bien défini pour l'évaluation des tests avant leur généralisation reste un obstacle majeur à la généralisation de plans de développement clinique rigoureux pour les tests diagnostiques. L'objectif de cet article est d'exposer les bases méthodologiques de l'évaluation des performances diagnostiques d'un test.

Dans cette première partie, nous nous intéressons aux résultats du test : celui-ci peut être positif (« anormal », c'est-à-dire une valeur observée au-delà du seuil) ou négatif (« normal » c'est-à-dire une valeur observée en deçà du seuil), et la maladie présente ou absente. Les résultats de l'application du test diagnostique à une population dans l'idéal ou dans le cadre de son évaluation peuvent être représentés sous la forme d'un tableau 2 × 2 *(tableau 1)*. Une série d'indices a été construite à partir des nombres de ce tableau – et parfois d'autres informations – pour caractériser un test par ses performances.

Les indices de précision et combinaisons les plus simples

– La **sensibilité** (Se) est la probabilité d'avoir un test positif quand on est malade : a / (a + c).
– La **spécificité** (Sp) est la probabilité d'avoir un test négatif quand on n'est pas malade : d / (b + d).
– La **valeur prédictive positive du test** (VPP), la probabilité d'avoir la maladie quand le test est positif : a / (a + b).
– La **valeur prédictive négative du test** (VPN), celle de ne pas avoir la maladie quand le test est négatif : d / (c + d).
– Les **rapports de vraisemblance** (en anglais **Likelihood Ratio**, LR) estiment le rapport entre la probabilité d'avoir un test positif (ou négatif) chez les sujets malades à celle d'avoir un test positif (ou négatif) chez les sujets sains. Le rapport de vraisemblance positif est donc LR(+) = (sensibilité) / (1 – spécificité) et négatif LR(–) = (1 – sensibilité) / spécificité.
– Le « **Diagnostic Odds Ratio** » (DOR) ou le rapport des cotes diagnostiques est connu comme un indice statistique dans l'épidémiologie des études

Tableau 1. Tableau 2 × 2 issu de l'évaluation d'un test diagnostique

	Maladie présente	Maladie absente	
Test +	Vrai + (VP) = a	Faux + (FP) = b	a = b
Test –	Faux – (FN) = c	Vrai – (VN) = d	c + d
Total	a + c	b + d	a + b + c + d

a est le nombre de sujets atteints et pour qui le test est « positif » ; b est le nombre de sujets non atteints par la maladie pour lesquels le test est « positif » ; c est le nombre de malades pour lesquels le test est « négatif » ; d est le nombre de sujets non atteints pour lesquels le test est « négatif ».

cas-témoins. Il représente la force de l'association entre le facteur de risque et la maladie. Ici, il pourrait être utilisé pour montrer la force de l'association entre le résultat d'un test et la maladie. Cet indice cherche à quantifier la performance d'un test par une seule valeur. Celle-ci n'est pas influencée par la prévalence, contrairement aux indices précédents. DOR est le ratio entre la cote d'être malade (probabilité d'être malade divisée par la probabilité de ne pas être malade) lorsque le test est positif et la cote de ne pas être malade lorsque le test est négatif. DOR = (VP/FN)/(FP/VN) = ad/bc = [Se/(1 – Se)]/[(1 – Sp)/Sp] = LR+/LR– = [VPP/(1 – VPP)]/[(1 – VPN)/VPN]. La valeur du DOR varie de 0 à l'infini. Les valeurs hautes signifient une meilleure performance du test. La valeur = 1 signifie que le test n'a aucune valeur discriminante et une valeur › 1 que le test est plus souvent positif chez les malades que chez les sujets sains.

Le théorème de Bayes stipule que la cote d'être malade *après* un test (+ ou –) est égale à la cote d'être malade *avant* le test (+ ou –) que multiplie le rapport de vraisemblance (+ ou –). La prévalence de la maladie dans la population étudiée permet de calculer la cote d'être malade *avant* le test si elle est connue. La probabilité d'être malade peut être calculée à partir de la cote :

$$\text{Probabilité d'être malade (test + ou –)} = \frac{\text{cote d'être malade (test est + ou –)}}{1 + \text{cote d'être malade (test + ou –)}}$$

Fagan a proposé un nomogramme dès 1977 *(figure 1)* pour permettre de calculer la probabilité de maladie à partir des résultats d'un test [3].

Qu'en penser ?

La sensibilité et la spécificité ne sont pas des valeurs invariables et intrinsèques à prévalence de maladie donnée

En théorie, ces indices, au contraire de ce qui est enseigné habituellement dans les ouvrages pédagogiques, sont uniquement constants lorsque la positivité du test de référence (qui définit la présence de la maladie) entraîne de façon causale la positivité du nouveau test. On distingue en effet :

– **Le vrai test diagnostique :** il est relié de manière causale au test de référence (ou à la maladie). Par exemple, si l'apparition d'un cancer du côlon détecté par la colonoscopie entraînait systématiquement la présence du sang dans les selles, la sensibilité et la spécificité de l'Hémoccult® seraient constantes car le test détecterait la même proportion d'hémorragie fécale [4]. Ce cas de figure est assez rare.

– **Le test prédictif :** il prédit l'apparition de la maladie. La majorité des tests que nous utilisons sont de ce type. Un exemple de test prédictif est celui de la prédiction de complication cardiaque avant la chirurgie vasculaire par scanner au dypiramidole-thalium [5]. On constate une grande variabilité des valeurs des indices de précisions évaluées par différentes études.

– **Le test corrélationnel :** la maladie cause la positivité du test de référence et du nouveau test sans lien forcément causal entre ces deux tests [4]. Par exemple, Hlatky et al. ont montré que pour le diagnostic d'une coronaropathie, en fonction de l'âge, du sexe, du stade de l'exercice atteint, de la sévérité et de la durée des symptômes, la sensibilité du décalage du segment ST (› 1 mm) à l'électrocardiographie d'effort comparée à la coronarographique variait entre 41 et 89 % et la spécificité variait entre 70 et 100 % [6].

Cependant, même pour les vrais tests diagnostiques, la stabilité de la sensibilité et de la spécificité sont théoriques. Les mesures seront variables en fonction de la pathologie et du test de référence utilisé, de la population et de la question clinique posée, de la place du nouveau test dans la démarche (en plus ou en remplacement d'un autre test) et de la nature du test (quel appareil, quelle technique, quel étalonnage) [7]. La variation de ces indices mettrait en doute le rationnel d'utilisation des tests en pratique clinique courante fondé sur l'hypothèse de Bayes et la constance de la vraisemblance d'un test. Néanmoins, cette variation est en grande partie due aux biais de sélection, d'interprétation, de vérification mais aussi à l'imperfection du test de référence [8-12].

Un peu plus sur la variabilité des indices : les seuils

En plus des éléments que nous venons de citer, une part de la variation de la précision des tests est due aux différents seuils utilisés par les observateurs pour classer les patients en positifs ou négatifs. Par exemple, lors d'une étude chez 1 168 femmes en post-ménopause présentant une hémorragie vaginale, l'ultrasonographie endovaginale a été évaluée pour le diagnostic de cancer de l'endomètre avant le curetage chirurgical [13]. Cette étude a évalué la précision de différents seuils de l'épaisseur de la tumeur pour définir le

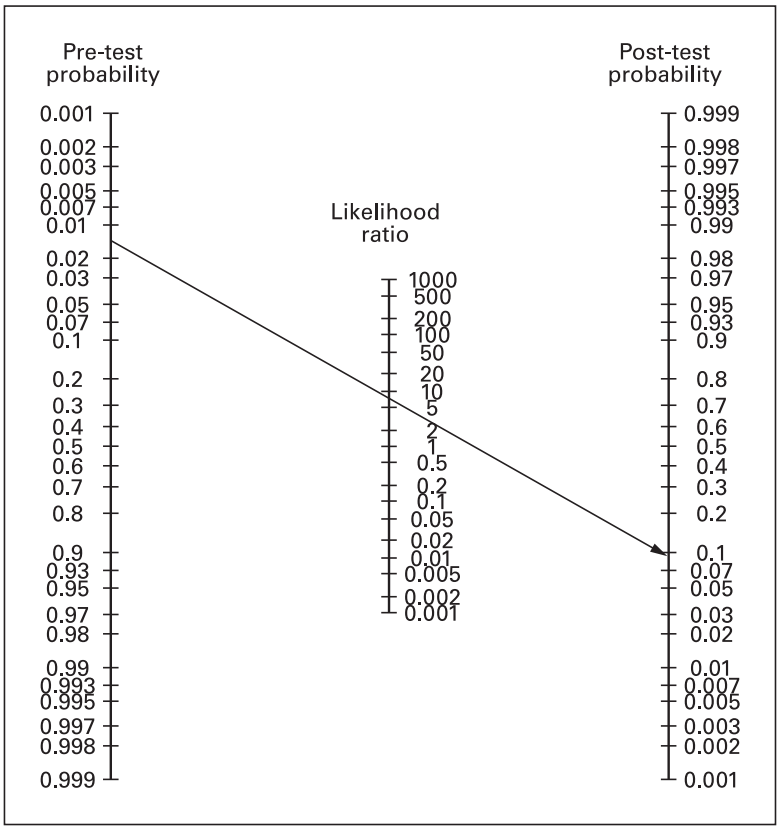


Figure 1. Le nomogramme de Fagan.

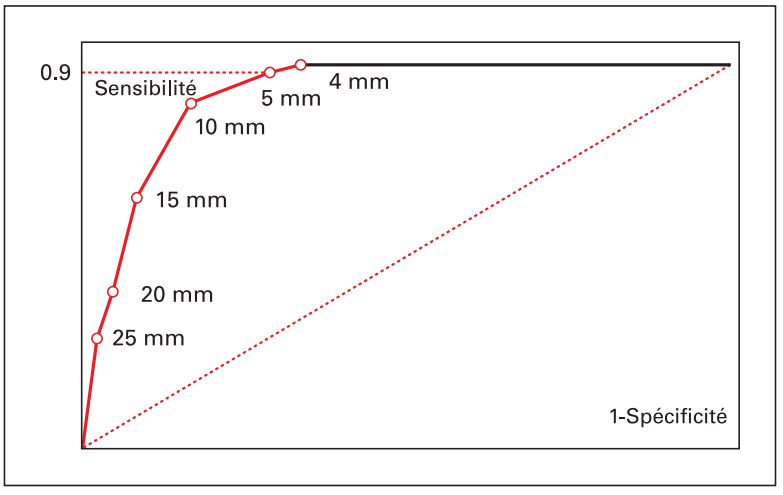
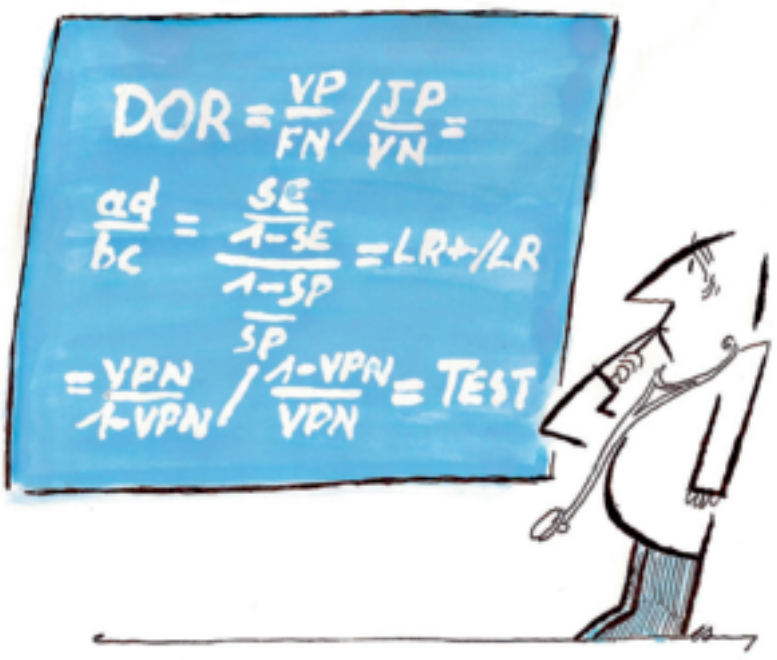


Figure 2. Courbe ROC (sensibilité en fonction de 1-spécificité ou Receiver Operating Characteristic) d'évaluation de différents seuils d'épaisseur de la tumeur mesurée par l'ultrasonographie endovaginale comparée à l'examen anatomo-pathologique après curetage.

cancer. La *figure 2* montre que lorsque le seuil de positivité est strict (< 20 mm), la spécificité est importante car la plupart des tumeurs de grande épaisseur sont des cancers (faible nombre de faux positifs). En revanche, la sensibilité est faible car un grand nombre des tumeurs avec une épaisseur < 20 mm classées en tumeurs non cancéreuses sont des faux négatifs. Un seuil plus souple (< 5 mm) sera au contraire plus sensible et moins spécifique. La courbe Receiver Operating Characteristic (ROC) permet ainsi de tenir compte de l'effet seuil et de la corrélation négative entre la sensibilité et la spécificité. La définition du seuil sera effectuée selon le contexte clinique, pour privilégier l'un ou l'autre de ces deux indices ou choisir le meilleur couple en terme de précision. Une part de subjectivité importante persiste lors de la définition d'un seuil de positivité, d'où l'importance d'utiliser une courbe ROC pour fixer ce seuil et des courbes ROC résumées pour synthétiser les résultats de plusieurs études [14].

Le problème du test de référence

Pour évaluer un test diagnostique, sa précision doit être comparée à un test de référence, l'« étalon-or » (« gold standard »), qui déterminera la maladie de façon fiable. Cependant, les tests de référence ne sont jamais parfaits et entraînent donc des erreurs de classification de la maladie. Les biais causés par ces erreurs de classification dépendent de la prévalence de la maladie et de la corrélation entre les



erreurs du nouveau test et celles du test de référence. Les erreurs des deux tests sont indépendantes quand elles ne surviennent pas sur les mêmes patients, et à l'inverse, sont corrélées lorsqu'elles surviennent chez les mêmes patients. Quand la précision du test de référence et la relation entre les erreurs des deux tests sont connues, ces biais pourraient être corrigés [15] mais les méthodes de correction avec les erreurs corrélées ne sont pas encore développées. Malheureusement, la précision du test de référence est souvent inconnue car il a mal été évalué. La précision d'un test, ou robustesse, est sa capacité à reproduire le même résultat si on répétait le test chez un même patient. Par exemple, le test de référence pour le diagnostic de thrombose veineuse profonde est la phlébographie qui est invasive, douloureuse, et moyennement fiable avec un faible taux de concordance entre deux observateurs (Kappa variant de 0,57 à 0,90) [16]. La précision de la phlébographie n'a pas été évaluée.

Nous reviendrons dans la seconde partie de cet article, dans le prochain numéro de Médecine, sur les différents problèmes méthodologiques que pose l'évaluation des tests diagnostiques.

Références :

1. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. Jama. 1995;274(8):645-51.
2. Buntinx F, Knottnerus JA. Are we at the start of a new era in diagnostic research? J Clin Epidemiol. 2006;59(4):325-6.
3. Fagan TJ. Letter: Nomogram for Bayes theorem. N Engl J Med. 1975;293(5):257.
4. Choi BC. Causal modeling to estimate sensitivity and specificity of a test when prevalence changes. Epidemiology. 1997;8(1):80-6.
5. Wong T, Detsky AS. Preoperative cardiac risk assessment for patients having peripheral vascular surgery. Ann Intern Med. 1992;116(9):743-53.
6. Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med. 1984;77(1):64-71.
7. Irwig L, Bossuyt P, Glasziou P, Gatsonis C, Lijmer J. Designing studies to ensure that estimates of test accuracy are transferable. BMJ. 2002;324(7338):669-71.
8. Diamond GA. Reverend Bayes' silent majority. An alternative factor affecting sensitivity and specificity of exercise electrocardiography. Am J Cardiol. 1986;57(13):1175-80.
9. Diamond GA. Affirmative actions : can the discriminant accuracy of a test be determined in the face of selection bias? Med Decis Making. 1991;11(1):48-56.
10. Mulherin SA, Miller WC. Spectrum bias or spectrum effect? Subgroup variation in diagnostic test evaluation. Ann Intern Med. 2002;137(7):598-602.
11. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med. 1978;299(17):926-30.
12. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epidemiol. 1992;45(10):1143-54.
13. Karlsson B, Granberg S, Wikland M, Ylostalo P, Torvid K, Marsal K, et al. Transvaginal ultrasonography of the endometrium in women with postmenopausal bleeding – a Nordic multicenter study. Am J Obstet Gynecol. 1995;172(5):1488-94.
14. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med. 1993;12(14):1293-316.
15. Walter SD, Irwig L, Glasziou PP. Meta-analysis of diagnostic tests with imperfect reference standards. J Clin Epidemiol 1999;52(10):943-51.
16. Picolet H, Leizorovicz A, Revel D, Chirossel P, Amiel M, Boissel JP. Reliability of phlebography in the assessment of venous thrombosis in a clinical trial. Haemostasis. 1990;20(6):362-7.

En résumé : qu'attendre d'un test diagnostique ?

Les performances d'un même test diagnostique varient en fonction :

- de la prévalence de la maladie
- du lien physiopathologique avec le test de référence
- du test de référence utilisé
- du seuil, donc de l'étalonnage
- de sa place dans la démarche