Dans le premier article, nous nous sommes intéressés aux résultats des tests diagnostiques (indices de précision et combinaisons les plus simples). Dans ce second article, nous analysons les questions qui se posent aux différents stades de développement du test. Pour chaque étape de conception d'un test diagnostique, une question et un objectif clair doivent être identifiés. Une méthodologie adaptée pour répondre à chaque question doit être utilisée pour éviter les biais potentiels. Malgré les avancées, l'évaluation des tests diagnostiques ne permet pas de les exclure tous. Il est donc nécessaire de répéter ces évaluations et de s'assurer de la précision du test diagnostique dans le contexte dans lequel il va être utilisé.

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, évaluation, méta-analyse
DOI : 10.1684/med.2006.0034

Qu'attendre d'un « test diagnostique » ? (2e partie)

L'évaluation des tests diagnostiques

Les phases de développement clinique du test

Une évaluation non biaisée des tests diagnostiques (ou biomarqueurs) doit répondre à des questions précises en fonction du stade de développement clinique du test. La méthodologie doit permettre de répondre à ces différentes questions [1] :

Phase I : Le test a-t-il des résultats différents chez les patients et chez les sujets sains ?

Phase II : La maladie est-elle plus vraisemblable chez les sujets aux résultats positifs ?

Des études cas-témoins doivent répondre à ces deux questions.

La validité technique de l'information fournie par le test doit aussi être vérifiée : *qualité* technique (de l'ECG par exemple) ; *spécificité* technique (par exemple, la glycémie est-elle bien mesurée ?) ; *fiabilité* technique (erreur de mesure systématique, précision ; variabilité de l'interprétation et concordance : intra- et inter-observateur).

Phase III : Parmi les sujets à risque, quelle est la précision du test ?

C'est une étude transversale, prospective, qui permet de répondre à cette troisième question, avec interprétation des résultats par deux évaluateurs en insu.

La *population incluse* doit être représentative des sujets malades, traités ou non, à différents stades de gravité et diagnostics différentiels. Dans chaque centre participant à l'étude, les patients doivent être inclus de façon consécutive. Ils doivent pouvoir subir l'examen à valider et l'examen de référence, « l'étalon-or », s'il est validé et permet de définir les malades. Si celui-ci n'existe pas, soit le participant doit être suivi jusqu'à l'apparition de la maladie, soit il faut remplacer l'examen de référence par une série d'examens [2]. Le test doit en outre être placé dans la stratégie diagnostique habituelle, et chaque patient faire les deux examens. Les résultats doivent être interprétés en insu sans que le médecin ait connaissance des résultats de l'autre examen.

Phase IV : L'examen diagnostique améliore-t-il l'état de santé des patients ?

L'essai clinique randomisé en groupe parallèle avec une randomisation imprévisible permet de répondre à cette question.

Exemple : le dosage des peptides natriurétiques

Nous reprenons ici l'exemple présenté par Sackett [2]. Aux États-Unis, on recense 4,8 millions de patients atteints d'un dysfonctionnement ventriculaire gauche (DVG), dont 400 000 à 700 000 nouveaux cas par an.

Le diagnostic clinique des formes asymptomatiques est difficile. Le risque de mortalité toute cause diminue grâce à un traitement par des inhibiteurs de l'enzyme de conversion. La sécrétion du peptide natriurétique de type B (BNP) par le ventricule gauche semble augmentée lors de dysfonctionnement ventriculaire et a une relation avec le pronostic de cette pathologie. Comment étayer la valeur diagnostique du BNP ?

– *Phase I : Est-ce que les patients ayant un DVG ont des valeurs de BNP supérieures aux individus sains ?*
Une étude cas-témoins [3] avec 91 patients consécutifs (dont 15 témoins sains vérifiés par l'échocardiographie et 17 patients avec DVG) a montré une différence statistiquement significative de la moyenne de BNP : 493,5 fentomol/mL (intervalle de confiance à 95 % : 248,9-909) pour les patients avec DVG et 129,4 fentomol/mL (IC95 : 53,6-159,7) pour les sujets sains.
– *Phase II : Est-ce que les patients ayant des valeurs hautes de BNP ont plus de risque d'avoir un DVG que les sujets sains ?*
Une étude cas-témoins [4] comprenant 27 témoins et 101 patients avec DVG retrouvait les résultats suivants : rapport de vraisemblance positif = 13 (IC95 : 3,5-50). Interprétation : devant un test positif, il est 13 fois plus probable d'être malade que non malade. Le rapport de vraisemblance négatif était 0,03 (0,0003-0,19), soit : devant un test négatif, il est 33 fois plus probable (1/0,003) d'être sain que malade. Ainsi les études de phase I et II permettent d'éliminer le test diagnostique non prometteur rapidement et d'éviter la réalisation d'autres études de phase III et IV plus coûteuses (financièrement et éthiquement). Ces études permettent aussi de mieux comprendre le mécanisme de la maladie et la physiopathologie nécessaire à la démarche de validation des critères intermédiaires.

– *Phase III : Parmi les sujets à risque, est-ce que le test BNP peut distinguer qui est malade de qui ne l'est pas ?*
Une étude transversale réalisée sur 126 sujets consécutifs a comparé le taux de BNP aux constatations de l'échographie cardiaque, test de référence [5]. Avec un seuil de positivité pour le BNP de 18 pg/mL la sensibilité a été de 88 % et la spécificité de 34 %.
Les études de Phase III doivent être répétées pour permettre une vraie validation du test [6]. Pour les médicaments, il est habituellement demandé qu'au moins deux essais cliniques randomisés soient réalisés avant l'obtention de l'autorisation de mise sur le marché. Devant les difficultés méthodologiques rencontrées avec l'évaluation des tests diagnostiques, il paraît raisonnable de recommander plusieurs études de phase III pour la validation d'un test.

– *Phase IV : La pratique de ce test diagnostique améliore-t-elle l'état de santé des patients ?*
Le rôle de la mesure du BNP dans l'amélioration de l'état de santé des sujets avec DVG n'a pas été évalué par une étude phase IV. Nous citerons donc l'exemple du dépistage du cancer du sein par mammographie. Une méta-analyse récente [7-9] ayant inclus uniquement les études avec une randomisation imprévisible n'a pas montré de bénéfice sur le critère décès, toutes causes confondues, chez les femmes dépistées par mammographie (risque relatif 1, IC : 0,98-1,05).
Alors que le calcul du nombre de sujets nécessaires est systématique pour les essais cliniques, très peu d'évaluations de tests diagnostiques ou de biomarqueurs justifient ce calcul et la définition des hypothèses *a priori*. Nous disposons, actuellement, de techniques statistiques permettant de bien définir le nombre de sujets nécessaires selon la phase du développement pour atteindre l'objectif de l'étude avec la précision [10] et la rigueur scientifique nécessaires.

Méta-analyse et test diagnostique

La méta-analyse des tests diagnostiques comme celle des essais thérapeutiques peut être très utile pour l'évaluation de la précision de mesure des critères intermédiaires. La méta-analyse permet de présenter des résultats quantifiés d'une revue systématique de l'ensemble des études effectuées avec un test diagnostique. Ces résultats quantifient la sensibilité, la spécificité et le *Diagnostic Odds Ratio*. Ils permettent d'affiner la précision de l'estimation des performances diagnostiques, d'expliquer l'hétérogénéité, et d'étudier la reproductibilité des résultats [11]. Cependant, la méta-analyse des études de tests diagnostiques n'est pas une pratique courante et pose des problèmes méthodologiques spécifiques non résolus [12, 13]. Voici l'exemple de l'ultrasonographie et du fibrinogène à l'iode radioactif.

Première étape : recherche exhaustive de la littérature sur les études comparant l'ultrasonographie à la phlébographie

Plus de 50 études ont été recueillies. Les données ont été extraites par deux médecins de l'équipe en insu [14]. Nous avons évalué grâce aux nouvelles techniques méta-analytiques la performance de l'ultrasonographie en tenant compte des variables qui peuvent expliquer leur hétérogénéité [15]. Le rôle de chacune des covariables (position du patient, expérience du technicien par exemple) a aussi été analysé à l'aide des techniques de métarégression. Comme pour les essais thérapeutiques la méta-analyse des tests diagnostiques a permis de mettre en évidence l'importante lacune méthodologique des études réalisées [16].

Problèmes méthodologiques d'évaluation du test

La *figure 1* montre que seulement 6 parmi les 13 études ayant évalué le fibrinogène radioactif par rapport à la phlébographie dans le diagnostic des thromboses veineuses asymptomatiques ont minimisé les possibilités de **biais de sélection** en incluant des patients de façon consécutive ; de **biais d'interprétation**, en évaluant les résultats des deux examens en insu par deux observateurs différents ; de **biais de vérification** *(« work up bias »)* en effectuant les deux examens de façon systématique chez tous les patients [17].

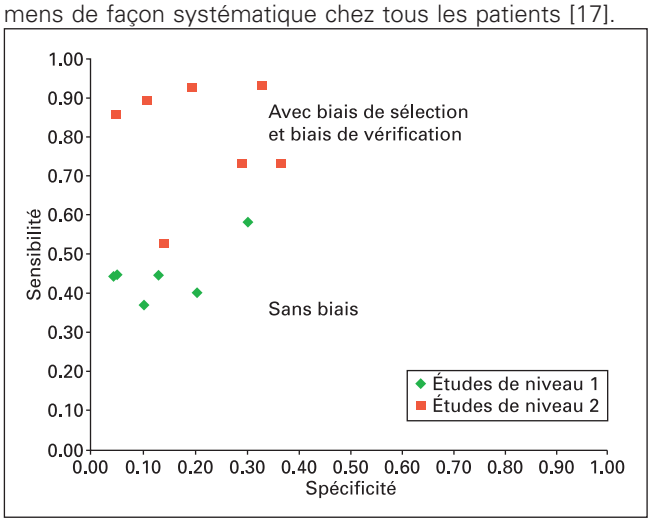

Figure 1. Résultats de 13 études d'évaluation d'I-Fibrinogène *vs.* phlébographie.

Rappelons que le *Diagnostic Odds Ratio* (DOR) est supérieur à 1 quand le test est plus souvent positif chez les sujets malades que chez les sujets sains, égal à 1 lorsque le test n'a aucune valeur discriminante. Avec un DOR égal à 21,9 (7,6-63), les études potentiellement biaisées surestiment nettement la précision du fibrinogène radioactif lorsqu'on les compare aux études possiblement non biaisées avec un DOR de 1,51 (0,54-4,2). D'autre part, la grande hétérogénéité des résultats de ces études rend difficile leur synthèse quantifiée. Cependant, la méta-analyse garde tout son intérêt dans l'évaluation de l'hétérogénéité des résultats et les facteurs qui l'influencent [11, 16, 18].

Conclusion

Nous disposons maintenant d'une méthodologie mieux standardisée pour évaluer un test en évitant les biais potentiels. Cette évaluation comprend un plan de développement complet avec différentes phases (de I à IV). Parce que la méthode expérimentale utilisée pour évaluer la précision d'un test n'exclut pas un biais, il est indispensable de répéter les études de validations de phase III et d'utiliser les techniques méta-analytiques pour synthétiser les résultats et étudier l'hétérogénéité clinique et statistique observées.

Références :

1. Sackett DL, Haynes RB. The architecture of diagnostic research. BMJ. 2002;324:539-41.
2. Knottnerus JA, Muris J. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, editor. The evidence base of clinical diagnosis. London: BMJ Publishing Group; 2002.
3. Talwar S, Siebenhofer A, Williams B, Ng L. Influence of hypertension, left ventricular hypertrophy, and left ventricular systolic dysfunction on plasma N terminal proBNP. Heart. 2000;83:278-82.
4. Selvais PL, Donckier JE, Robert A, Laloux O, van Linden F, Ahn S, et al. Cardiac natriuretic peptides for diagnosis and risk stratification in heart failure: influences of left ventricular dysfunction and coronary artery disease on cardiac hormonal activation. Eur J Clin Invest. 1998;28:636-42.
5. Landray MJ, Lehman R, Arnold I. Measuring brain natriuretic peptide in suspected left ventricular systolic dysfunction in general practice: cross-sectional study. BMJ. 2000;320:985-6.
6. Van den Bruel A, Aertgeerts B, Buntinx F. Results of diagnostic accuracy studies are not always validated. J Clin Epidemiol. 2006;59:559-66.
7. Gotzsche PC. Update on effects of screening mammography. Lancet. 2002;360:338-9; author reply 339-40.
8. Gotzsche PC. Mammography service screening and mortality. Lancet. 2003;362:329-30; author reply 330.
9. Gotzsche PC, Olsen O. Is screening for breast cancer with mammography justifiable? Lancet. 2000;355:129-34.
10. Flahault A, Cadilhac M, Thomas G. Sample size calculation should be performed for design accuracy in diagnostic test studies. J Clin Epidemiol. 2005;58:859-62.
11. Lijmer JG, Bossuyt PM, Heisterkamp SH. Exploring sources of heterogeneity in systematic reviews of diagnostic tests. Stat Med. 2002;21:1525-37.
12. Irwig L, Macaskill P, Glasziou P, Fahey M. Meta-analytic methods for diagnostic test accuracy. J Clin Epidemiol. 1995;48:119-30.
13. Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology. 1997;8:12-7.
14. Kassai B, Sonie S, Shah NR, Boissel FH. Literature search parameters marginally improved the pooled estimate accuracy for ultrasound in detecting venous thrombosis. J Clin Epidemiol. 2006;In Press.
15. Kassai B, Boissel JP, Cucherat M, Sonie S, Shah NR, Leizorovicz A. A systematic review of the accuracy of ultrasound in the diagnosis of deep venous thrombosis in asymptomatic patients. Thromb Haemost. 2004;91:655-66.
16. Lijmer JG, Mol BW, Heisterkamp S, Bonsel GJ, Prins MH, van der Meulen JH, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA. 1999;282:1061-6.
17. Lensing AW, Hirsh J. 125I-fibrinogen leg scanning: reassessment of its role for the diagnosis of venous thrombosis in post-operative patients. Thromb Haemost. 1993;69:2-7.
18. Whiting P, Rutjes AW, Reitsma JB, Glas AS, Bossuyt PM, Kleijnen J. Sources of variation and bias in studies of diagnostic accuracy: a systematic review. Ann Intern Med. 2004;140:189-202.

En résumé : l'évaluation des tests diagnostiques

La méthodologie de l'évaluation des tests diagnostiques doit répondre à des questions précises à chaque phase de développement clinique du test.

- Phase I : Le test a-t-il des résultats différents chez les patients et chez les sujets sains (études cas-témoins) ?
- Phase II : La maladie est-elle plus vraisemblable chez les sujets aux résultats positifs (études cas-témoins) ?
- Phase III : Parmi les sujets à risque, quelle est la précision du test (étude transversale, prospective) ?
- Phase IV : L'examen diagnostique améliore-t-il l'état de santé des patients (essai clinique randomisé en groupe parallèle avec une randomisation imprévisible) ?

La méta-analyse des tests diagnostiques comme celle des essais thérapeutiques peut être très utile pour l'évaluation de la précision de mesure des critères intermédiaires. Cependant, elle n'est pas une pratique courante et pose des problèmes méthodologiques spécifiques non résolus.