CHAPITRE IX

LES ETUDES DE STRATEGIES DIAGNOSTIQUES

Muriel Rabilloud, René Ecochard, Gilles Landrivon

Les examens cliniques ou paracliniques sont utilisés:

1 - Soit pour savoir si le patient est porteur d'une affection (par exemple, l'échographie pour la lithiase vésiculaire), afin, si le test est positif, d'envisager un traitement (ici, la chirurgie).

2 - Soit pour quantifier la valeur d'un paramètre (par exemple, la digitalinémie au cours d'un traitement digitalique) afin d'adapter une thérapeutique (ici, la posologie du tonicardiaque).

3 - Soit pour visualiser des structures normales ou pathologiques (par exemple, le réseau vasculaire avant une intervention chirurgicale intracrânienne, ou une lithiase du cholédoque au cours d'un cathétérisme rétrograde de la papille).

4 - Soit enfin pour déterminer l'ampleur d'une atteinte, afin d'établir un pronostic sans qu'il en découle de décision thérapeutique.

La plus grande partie de ce chapitre concerne la situation 1. Les situations 2 et 3 sont traitées dans le chapitre XIV sur les instruments de mesure. La situation 4 se rapproche de la situation 1 (notions de sensibilité et de spécificité, …) mais s'en écarte par l'absence de décision à l'issue du résultat.

Plan du chapitre

I - LE TEST DIAGNOSTIQUE REPLACE DANS SON CONTEXTE

Dans ce paragraphe, nous allons planter le décor, en introduisant les termes de probabilité pré-test, sensibilité, spécificité, probabilité post-test, seuil de traitement et utilité.

Tout examen trouve sa place dans une histoire, histoire qui commence par un symptôme, ou un examen de dépistage réalisé en l'absence de signe d'appel.

En amont du test diagnostique il y a un contexte clinique (âge, sexe, antécédents, symptômes déjà présents, éventuellement résultats d'autres examens) qui permet d'établir une probabilité d'existence de la maladie (avant de réaliser le test diagnostique étudié). Cette probabilité est appelée **probabilité pré-test**.

Un résultat positif du test va changer votre avis sur l'état du patient, la probabilité d'existence de la maladie étant plus élevée dans ce cas, moins élevée devant un résultat négatif. La probabilité que le patient soit porteur de la maladie connaissant le résultat du test est appelée **probabilité post-test**.

Parfois, la probabilité de la maladie sachant que le test est positif est de 100 % (égale à 1). C'est le cas si le test n'a aucun faux positif, par exemple l'histologie avec présence de tissu néoplasique sur une biopsie. Cet examen a une **spécificité** de 100 % (probabilité que le test soit négatif en l'absence de cancer).

Parfois, la probabilité de la maladie sachant que le test est négatif est nulle (égale à 0). Le test a complètement éliminé l'hypothèse d'atteinte par l'affection recherchée car il n'y a aucun faux négatif. C'est le cas par exemple du scanner abdominal pour le diagnostic de masse kystique rénale, ce dernier examen ayant une **sensibilité** de 100 % (probabilité que le test soit négatif en présence de la maladie).

Le plus souvent, la probabilité post-test est différente de 0 ou de 100%. Elle dépend du contexte clinique (probabilité pré-test) et de la qualité du test. Si le test est très spécifique et qu'il revient positif, l'information apportée par le test pour affirmer la présence de la maladie est importante et entraîne une augmentation importante de la probabilité d'avoir la maladie.

Il se peut que le test, qu'il soit positif ou négatif, ne change pas la décision thérapeutique. Quelle serait alors son **utilité** dans la situation 1 définie dans l'introduction du chapitre (savoir si le patient est porteur d'une affection, afin d'envisager un traitement si le test est positif) ?

En effet, pour chaque action thérapeutique ou diagnostique invasive, le clinicien a un seuil de probabilité en dessous duquel il s'abstient. On ne réalise pas une biopsie mammaire pour des images banales de mammographie en l'absence d'anomalie à la palpation. Ce seuil est le plus souvent appelé **seuil de traitement**. Il dépend du bénéfice global que l'on attend de l'intervention. Chaque geste thérapeutique ou diagnostique invasif a un bénéfice global, résultat de l'équilibre entre l'amélioration potentielle de l'état de santé et les effets secondaires possibles.

La mise sous anticoagulant en cas de suspicion d'embolie pulmonaire est une décision dans laquelle interviennent le risque d'hémorragie et le bénéfice apporté par le traitement. Le seuil de traitement est bas dans le cas de l'embolie pulmonaire. Dès qu'il y a une probabilité de 10 à 20 % d'embolie pulmonaire (voire moins) la décision est prise d'immobiliser le patient et de le décoaguler en attendant la réalisation des examens complémentaires. Ce seuil bas est dû d'une

part à la gravité des complications évitées par le traitement et d'autre part à la relative sécurité d'un traitement anticoagulant bien conduit.

Au contraire, lorsque le geste est lourd de conséquences, telle une gastrectomie totale ou une amputation, on attend un niveau élevé de probabilité d'existence de la maladie (par exemple, gastrite de Ménétrier, tumeur osseuse maligne, …) pour intervenir. Si l'incertitude demeure, on préfère, en effet, différer le geste et reprendre les examens complémentaires.

Si la probabilité de maladie dépasse largement le seuil de traitement, ce dernier est entrepris sans réaliser d'autres examens paracliniques, qui n'ont, en effet, aucune chance de faire changer l'indication du traitement. Un résultat négatif serait étiqueté faux négatif, le reste du contexte étant trop accusateur pour que cet examen puisse faire changer le diagnostic. Face à un cordon sous-cutané rouge et douloureux, la phlébite superficielle est affirmée et le traitement est mis en place. Une phlébographie normale aurait l'intérêt d'éliminer une atteinte profonde mais ne remettrait pas en cause l'atteinte superficielle qui serait donc traitée. Le test concerné (ici, la phlébographie) a donc comme intérêt de montrer l'extension de l'atteinte, non de faire évoluer la probabilité de phlébite superficielle. En effet, négative ou positive, elle n'a pas la possibilité d'infléchir le diagnostic suffisamment pour remettre en cause le traitement. Sa sensibilité est insuffisante pour qu'un résultat négatif abaisse la probabilité post-test en dessous du seuil de traitement.

Il est de coutume (et exact) de dire qu'un test diagnostique ne doit être réalisé que s'il a une chance de faire passer la probabilité de maladie de "l'autre côté" du seuil de traitement, c'est-à-dire de changer la décision. Devant un bilan hormonal évocateur d'insuffisance surrénale, on ne réalise pas de test à l'eau. Celui-ci n'a en effet aucune chance de faire changer de diagnostic car il a trop de faux négatifs. A l'inverse, dans le suivi d'un patient en neurologie, la percussion du tendon d'Achille sera utilisée, la perte du réflexe à ce niveau ayant une spécificité suffisante pour que sa positivité déclenche un bilan complémentaire.

II – L’EVALUATION D’UN TEST DIAGNOSTIQUE

A – Les études d’évaluation d’un test diagnostique

1 – Les différentes phases d’évaluation

Comme pour l’évaluation de l’efficacité d’un nouveau médicament, il est possible de définir 3 phases dans l’évaluation d’un nouveau test diagnostique.

La première phase, appelée aussi **phase exploratoire**, correspond à la phase précoce d’évaluation d’un nouveau test. L’objectif est de savoir si ce test peut avoir un intérêt diagnostique. Il s’agit par exemple de vérifier qu’un nouveau biomarqueur a une valeur en moyenne plus élevée chez les malades que chez les non-malades et qu’il fait mieux que le simple hasard pour prédire l’existence de la maladie. A ce stade, les études réalisées doivent permettre d’obtenir une réponse rapide pour décider de poursuivre l’évaluation ou de passer à autre chose.

La deuxième phase, appelée aussi **phase de challenge**, a pour objectif de mesurer les performances diagnostiques d’un nouveau test dans différents sous-groupes de malades et de non-malades. Les performances diagnostiques d’un test sont quantifiées par sa sensibilité et sa spécificité pour les tests ayant une réponse dichotomique, ou par les sensibilités et spécificités associées aux différents seuils de positivité pour un test ayant une réponse

ordinale ou continue. On dit classiquement que la sensibilité et la spécificité sont les qualités intrinsèques d'un test car elles ne font pas intervenir la prévalence de la maladie. La sensibilité est estimée chez les malades et la spécificité chez les non-malades. En revanche, elles dépendent souvent des caractéristiques des malades ou des non-malades. Par exemple, la sensibilité de la mammographie pour le diagnostic de cancer du sein dépend de la taille de la tumeur. Elle est plus faible dans une population de femmes dépistées que dans une population de femmes venant en consultation spécialisée à un stade plus avancé de la maladie. Au cours de cette phase d'évaluation, le nouveau test peut également être comparé aux autres tests existants.

La troisième phase, appelée aussi **phase clinique,** a pour objectif de mesurer les performances diagnostiques d'un nouveau test et de le comparer aux autres tests dans la population ciblée. Cela implique que l'étude porte sur un échantillon représentatif de la population dans laquelle le test va être utilisé. Pour les tests nécessitant une interprétation par un lecteur tels que les examens d'imagerie médicale, il est également nécessaire de réaliser l'étude avec un échantillon représentatif des médecins qui seront amenés à lire l'examen. C'est également à cette phase précédant l'utilisation du test en pratique clinique, que le seuil de positivité et l'impact du choix du seuil sur la sensibilité et sur la spécificité sont étudiés pour les tests avec une réponse ordinale ou continue.

2 – Les différents types d'étude

Les principaux types d'étude que l'on retrouve dans le domaine de l'évaluation des tests diagnostiques sont des études de type cas-témoins, des études de type cohorte et des essais cliniques randomisés.

Les études de type cas-témoins

Ces études sont appelées études de type cas-témoins car lorsque les sujets entrent dans l'étude, leur statut malade ou non malade est connu. Elles reposent sur la constitution d'un échantillon de sujets dont on sait qu'ils ont la maladie et de façon indépendante d'un échantillon de sujets dont on sait qu'ils n'ont pas la maladie. Le test à évaluer est ensuite mesuré dans le groupe des sujets malades et dans le groupe des sujets non malades. Ce type d'étude est utilisé à la phase exploratoire de l'évaluation d'un nouveau test. Les sujets inclus dans l'échantillon de malades sont souvent à un stade assez avancé de la maladie, alors que les sujets inclus dans l'échantillon de non-malades sont souvent des sujets sains qui n'ont aucune pathologie pouvant mimer la maladie que l'on cherche à diagnostiquer. Cela aboutit souvent à une surestimation des performances diagnostiques du test à évaluer. Ce type d'étude est également utilisé à la phase de challenge car il permet de constituer des groupes de sujets malades à différents stades de la maladie et des groupes de sujets non malades avec des caractéristiques différentes par exemple en termes d'âge ou de comorbidités.

Les études de type cohorte

Ces études sont appelées études de type cohorte car lorsque les sujets sont inclus dans l'étude, leur statut malade ou non-malade n'est pas connu. Un échantillon représentatif de la population dans laquelle le test va être utilisé est constitué. Les sujets inclus dans l'étude ont tous le test à évaluer et leur statut malade ou non-malade est déterminé de façon indépendante du résultat du test. La détermination du statut malade ou non-malade nécessite de disposer d'un test de référence parfait appelé *gold standard* (règle d'or). L'étude CASS [1] est un exemple d'étude de type cohorte. Dans cette étude, un échantillon de 1465 hommes pour

lesquels il existe une suspicion de coronaropathie a été constitué. L'objectif de l'étude était d'évaluer les performances de l'épreuve d'effort et de la douleur thoracique recherchée à l'interrogatoire pour faire le diagnostic de coronaropathie. Tous les sujets inclus dans l'étude ont eu, outre les 2 tests à évaluer, une coronarographie permettant de les classer dans le groupe des sujets ayant une coronaropathie ou dans le groupe de sujets n'ayant pas de coronaropathie. Ce type d'étude est surtout utilisé à la phase clinique de l'évaluation d'un test. A ce stade de l'évaluation, il est recommandé de privilégier les études multicentriques pour augmenter la représentativité de l'échantillon étudié.

Les essais cliniques randomisés

Lorsqu'il n'existe pas de *gold standard* parfait, l'évaluation d'un nouveau test peut se faire par un essai clinique randomisé avec un bras correspondant à la stratégie diagnostique et thérapeutique habituelle et un bras qui intègre le nouveau test dans la stratégie diagnostique et thérapeutique. Dans ce type d'étude, le critère de résultat est un critère clinique. Le test sera jugé performant si le résultat clinique est significativement meilleur dans le bras incluant le nouveau test. Ce type d'étude permet également d'évaluer l'impact de l'introduction du test dans la stratégie diagnostique et thérapeutique.

B – L'évaluation des performances d'un test diagnostique

1 – Sensibilité et spécificité

La sensibilité et la spécificité d'un test sont des probabilités conditionnelles. La sensibilité est la probabilité que le test soit positif (en faveur de la maladie) sachant que le sujet est malade. Il s'agit de la capacité du test à identifier les malades. La spécificité est la probabilité que le test soit négatif (pas en faveur de la maladie) sachant que le sujet n'a pas la maladie. Il s'agit de la capacité du test à identifier les non-malades.

Leur estimation peut être obtenue à partir des résultats d'une étude de type cas-témoins ou de type cohorte présentés sous forme d'un tableau 2×2 (Tableau 1). Il y a quatre résultats possibles en fonction du résultat du test et du statut vis-à-vis de la maladie. Le résultat du test est positif et le sujet est malade, il s'agit d'un vrai positif (VP). Le résultat du test est négatif et le sujet est malade, il s'agit d'un faux négatif (FN). Le résultat du test est négatif et le sujet est non-malade, il s'agit d'un vrai négatif (VN). Le résultat du test est positif et le sujet est non-malade, il s'agit d'un faux positif (FP).

La sensibilité est estimée chez les malades par la proportion de tests positifs : $\frac{VP}{VP + FN}$

La spécificité est estimée chez les non-malades par la proportion de tests négatifs : $\frac{VN}{VN + FP}$

Il s'agit des valeurs les plus probables de sensibilité et spécificité du test compte tenu des données observées (estimations du maximum de vraisemblance). Elles sont obtenues par une lecture verticale du tableau 2×2.

Les résultats de l'étude CASS ont permis d'estimer la sensibilité et la spécificité de la douleur thoracique pour faire le diagnostic de coronaropathie dans une population de sujets à risque (Tableau 2).

La sensibilité de la douleur thoracique était estimée à : $\frac{969}{1023} = 94{,}7\%$

La douleur thoracique était présente (positive) chez environ 95% des patients porteurs d'une coronaropathie.

La spécificité de la douleur thoracique était estimée à : $\frac{197}{442} = 44{,}6\%$

La douleur thoracique était absente (négative) chez environ 45% des sujets n'ayant pas de coronaropathie.

2 – La courbe ROC et l'aire sous la courbe ROC

On peut distinguer **trois types de réponse pour les tests diagnostiques**. La réponse peut être **dichotomique** comme pour la douleur thoracique. La réponse peut être **ordinale** ou **quantitative continue**. Un exemple de test avec une réponse ordinale est le score BIRADS développé par le collège américain de radiologie. Il s'agit d'un score à 5 niveaux qui permet de classer les mammographies en fonction du degré de suspicion de cancer. Les marqueurs biologiques tels que les PSA pour le diagnostic de cancer de la prostate, sont des exemples de tests avec une réponse quantitative continue.

A la phase précoce de l'évaluation des tests avec une réponse dichotomique la performance est mesurée par la sensibilité et la spécificité. Avec une réponse ordinale ou quantitative continue, il n'est pas possible de résumer la performance diagnostique par l'estimation d'une sensibilité et d'une spécificité. Il existe autant de sensibilités et de spécificités que de seuils de positivité possibles. La courbe ROC (de l'anglais Receiver Operator Characteristic) permet de représenter la relation entre la probabilité que le test soit positif chez les malades (sensibilité) et la probabilité que le test soit positif chez les non-malades (1-spécificité).

L'étude de Hall FM et al. [2] portait sur 400 femmes ayant eu une biopsie du sein pour suspicion de cancer à la mammographie et une palpation normale. Parmi ces femmes, 119 avaient un cancer du sein. Les auteurs ont relu les mammographies sans avoir connaissance du résultat de la biopsie et les ont classées selon le degré de suspicion de cancer (Tableau 3). Selon le seuil de positivité choisi pour classer les mammographies comme positives, la sensibilité et la spécificité évoluent en sens inverse. Si seules les mammographies avec un haut degré de suspicion de cancer sont classées comme positives, la sensibilité est faible car il y a beaucoup de faux négatifs. Au contraire, la spécificité est élevée car il y a peu de faux positifs. Plus le seuil de positivité choisi est bas, meilleure est la sensibilité et moins bonne est la spécificité. A partir des données observées (Figure 1), il est possible d'estimer la sensibilité (Se) et la spécificité (Sp) de la mammographie pour chaque seuil dans une population de femmes ayant une suspicion de cancer.

- Mammographie considérée comme positive pour les suspicions hautes de cancer et comme négative pour les suspicions moyennes, légères et minimes:

(a) $\mathrm{Se} = \frac{47}{119} = 0{,}39 \qquad 1 - \mathrm{Sp} = 1 - \left(\frac{281 - 6}{281}\right) = 1 - \frac{275}{281} = 1 - 0{,}979 \approx 0{,}02$

- Mammographie considérée comme positive pour les suspicions hautes ou moyennes

(b) $\mathrm{Se} = \frac{104}{119} = 0{,}87 \qquad 1\text{-}\mathrm{Sp} = 1 - \left(\frac{281-(6+117)}{281}\right) = 1\text{-}\frac{158}{281} = 1\text{-}0{,}56 = 0{,}44$

- Mammographie considérée comme positive pour les suspicions hautes, moyennes ou légères

(c) $\mathrm{Se} = \frac{113}{119} = 0{,}95 \qquad 1\text{-}\mathrm{Sp} = 1\text{-}\left(\frac{281\text{-}(6+117+37)}{281}\right) = 1\text{-}\frac{121}{281} = 1-0{,}43 = 0{,}57$

Si toutes les mammographies sont considérées comme positives, toutes les femmes ayant un cancer sont bien classées (Se=1), mais toutes les femmes n'ayant pas de cancer sont faussement positives (Sp=0). A l'autre extrême si toutes les mammographies sont considérées comme négatives, toutes les femmes n'ayant pas de cancer sont bien classées (Sp=1), mais toutes les femmes ayant un cancer sont faussement négatives (Se=0). A partir de chaque couple (sensibilité, 1- spécificité) estimé pour les différents seuils observés, il est possible en reliant les points de construire la courbe ROC empirique (Figure 1) permettant de représenter la performance diagnostique globale de la mammographie. Si l'on considère que la mesure du degré de suspicion de cancer est un continuum entre la normalité et le cancer certain, la courbe en pointillé représente la courbe ROC de la mesure latente quantitative continue d'où est issue la réponse ordinale observée.

Plus la courbe ROC se rapproche de l'angle supérieur gauche correspondant à une sensibilité de 1 et une spécificité de 1, meilleure est la performance globale du test. Au maximum un test quantitatif dont la courbe ROC passe par le point de sensibilité 1 et spécificité 1, est un *gold standard* parfait. Dans ce cas, les distributions des valeurs chez les malades et les non-malades ne se recouvrent pas et tous les sujets sont bien classés. Cet idéal rarement atteint est symbolisé par le soleil sur la figure 1.

Un test diagnostique dont la courbe ROC est sur la diagonale, est un test pour lequel la probabilité d'avoir une réponse positive chez les malades est égale à la probabilité d'avoir une réponse positive chez les non-malades quel que soit le seuil de positivité. Il ne fait pas mieux que le hasard. La pièce de monnaie symbolise la situation que l'on aurait en jetant une pièce et en décidant que le test est positif chaque fois que l'on tombe sur face et négatif chaque fois que l'on tombe sur pile (Se=0,5 et 1-Sp=0,5).

K symbolise les aptitudes diagnostiques du docteur Knock cher à Jules Romain. Le médecin affirmant que "tout bien portant est un malade qui s'ignore" a une sensibilité parfaite mais inquiète inutilement tous les bien-portants. Sa spécificité est nulle.

La performance diagnostique globale du test est d'autant meilleure que la courbe ROC s'éloigne de la diagonale. Elle se quantifie par l'estimation de l'aire sous la courbe. Un test dont la courbe ROC est sur la diagonale et qui n'a donc pas d'intérêt diagnostique, a une aire sous la courbe de 0,5. Elle peut s'interpréter comme la probabilité qu'un sujet malade ait une valeur du test supérieure à celle d'un sujet non malade, lorsqu'une valeur élevée du test est en faveur de la maladie. Le test est d'autant meilleur pour discriminer les malades des non-malades que son aire sous la courbe se rapproche de 1.

Une méthode non paramétrique d'estimation de l'aire sous la courbe consiste à calculer pour toutes les paires (malade, non malade), la proportion de paires pour lesquelles la valeur du test chez le sujet malade est supérieure à la valeur du test chez le sujet non malade, lorsqu'une valeur élevée du test est en faveur de la maladie. Il s'agit de la statistique de Mann et Whitney. L'aire sous la courbe ROC de la mammographie pour faire le diagnostic de cancer du sein dans l'étude de Hall FM et al. est estimée à 0,81 avec un intervalle de confiance à 95% compris entre 0,76 et 0,85. La mammographie fait significativement mieux que le hasard car la borne inférieure de l'intervalle de confiance est supérieure à 0,5.

3 - Choix du seuil de positivité d'un test

A la phase clinique de l'évaluation d'un test diagnostique ordinal ou continu, la détermination d'un seuil de positivité est nécessaire. Le seuil de positivité optimal est celui qui maximise l'utilité dans la population dans laquelle le test est utilisé. **L'utilité** est définie comme une mesure de l'état de santé ou de la préférence pour un état de santé ; il s'agit par exemple de l'espérance de vie pondérée par la qualité de vie. L'utilité moyenne dans la population dépend de l'utilité de chacune des situations (sujet malade et traité, sujet malade et non traité, sujet non malade et non traité, sujet non malade et traité) et de la fréquence de chacune de ces situations.

L'utilité moyenne pour le seuil c, notée U(c) s'écrit :

$$U(c) = Se \times p \times U_{VP} + (1 - Se) \times p \times U_{FN} + Sp \times (1 - p) \times U_{VN} + (1 - Sp) \times (1 - p) \times U_{FP}$$

Se = sensibilité du test

Sp = spécificité du test

p = prévalence de la maladie ou probabilité pré-test

U_{VP}, U_{FN}, U_{VN}, U_{FP} sont les utilités associées aux quatre situations : sujet malade et traité (vrai positif), sujet malade et non traité (faux négatif), sujet non malade et non traité (vrai négatif), sujet non malade et traité (faux positif).

L'utilité moyenne, U(c), peut être réécrite en fonction du bénéfice net en termes d'utilité à traiter à raison un sujet malade et du coût net en termes d'utilité à traiter à tort un sujet non malade

Les méthodes permettant d'estimer le seuil qui maximise l'utilité moyenne dépassent le cadre de cet ouvrage et ne sont donc pas présentées. Le lecteur intéressé pourra trouver ces méthodes dans les références données en fin de chapitre.

C - Evolution de la probabilité d'avoir la maladie à l'issue du test

1 – Les valeurs prédictives et le théorème de Bayes

L'estimation de la sensibilité et de la spécificité permet d'évaluer les performances diagnostiques d'un test, mais pour le clinicien qui va utiliser le test, ce qui compte ce sont les valeurs prédictives positive et négative.

La valeur prédictive positive (VPP) est la probabilité que le sujet ait la maladie sachant qu'il a un test positif. La valeur prédictive négative (VPN) est la probabilité que le sujet n'ait pas la

maladie sachant qu'il a un test négatif. Ces valeurs prédictives dépendent de la sensibilité et de la spécificité du test, mais également de la prévalence de la maladie ou probabilité pré-test. Dans une étude de type cohorte il est possible d'estimer les valeurs prédictives directement à partir du tableau 2×2 par une lecture horizontale du tableau. Reprenons les résultats de l'étude CASS présentés dans le tableau 2. L'échantillon constitué pour l'étude est *a priori* représentatif de la population des sujets adressés pour une coronarographie en raison d'une suspicion de coronaropathie. La prévalence de la maladie dans cette population peut être estimée à partir des données de l'étude à 70 % (1023/1465).

La valeur prédictive positive de la douleur thoracique est estimée à : $\frac{969}{1214} = 80\%$

La valeur prédictive négative de la douleur thoracique est estimée à : $\frac{197}{251} = 78\%$

En revanche les études de type cas-témoins ne permettent pas d'estimer directement les valeurs prédictives, car elles ne reposent pas sur l'inclusion d'un échantillon représentatif d'une population, mais sur l'inclusion indépendante d'un échantillon de malades et d'un échantillon de non-malades dont les effectifs sont fixés par l'investigateur. Ayant une estimation de la sensibilité et de la spécificité du test à partir d'une étude de type cas-témoins, et par ailleurs une estimation de la prévalence de la maladie dans la population d'intérêt, il est possible d'estimer les valeurs prédictives positive et négative du test en utilisant le théorème de Bayes.

Le théorème de Bayes permet de façon générale d'inverser les probabilités conditionnelles et de passer par exemple de la probabilité que le test soit positif sachant que le sujet est malade (sensibilité) à la probabilité que le sujet ait la maladie sachant que le test est positif (VPP).

$$\begin{aligned}\text{VPP} = \text{P}(\text{M/Test}+) &= \frac{\text{P}(\text{M et Test}+)}{\text{P}(\text{Test}+)} = \frac{\text{P}(\text{Test}+/\text{M})\times\text{P}(\text{M})}{\text{P}(\text{Test}+\text{ et M})+\text{P}(\text{Test}+\text{ et NM})} \\ &= \frac{\text{P}(\text{Test}+/\text{M})\times\text{P}(\text{M})}{\text{P}(\text{Test}+/\text{M})\times\text{P}(\text{M})+\text{P}(\text{Test}+/\text{NM})\times\text{P}(\text{NM})} \\ &= \frac{\text{Se}\times\text{Prévalence}}{\text{Se}\times\text{Prévalence}+(1-\text{Sp})\times(1-\text{Prévalence})}\end{aligned}$$

Le théorème de Bayes permet également de passer de la probabilité que le test soit négatif sachant que le sujet est non-malade à la probabilité que le sujet soit non malade sachant que le test est négatif.

$$\begin{aligned}\text{VPN} = \text{P}(\text{NM/Test}-) &= \frac{\text{P}(\text{NM et Test}-)}{\text{P}(\text{Test}-)} = \frac{\text{P}(\text{Test}-/\text{NM})\times\text{P}(\text{NM})}{\text{P}(\text{Test}-\text{ et NM})+\text{P}(\text{Test}-\text{ et M})} \\ &= \frac{\text{P}(\text{Test}-/\text{NM})\times\text{P}(\text{NM})}{\text{P}(\text{Test}-/\text{NM})\times\text{P}(\text{NM})+\text{P}(\text{Test}-/\text{M})\times\text{P}(\text{M})} \\ &= \frac{\text{Sp}\times(1-\text{Prévalence})}{\text{Sp}\times(1-\text{Prévalence})+(1-\text{Se})\times\text{Pr}\textit{évalence}}\end{aligned}$$

Se = sensibilité, Sp = spécificité, M = malade, NM = non-malade

Pour illustrer le fait que les valeurs prédictives dépendent beaucoup de la prévalence, nous allons prendre l'exemple de l'utilisation de la mammographie pour faire le diagnostic de cancer du sein en situation de dépistage ou en consultation spécialisée. D'après les résultats de l'étude de Hall FM et al, et en considérant comme positives les mammographies pour lesquelles il y a une haute suspicion de cancer, la sensibilité est estimée à 39% et la spécificité à 98%.

Pour une prévalence de 4 pour mille dans la population des femmes dépistées entre 50 et 65 ans, la valeur prédictive positive est de : $\text{VPP} = \frac{0,39 \times 0,004}{0,39 \times 0,004 + (1 - 0,98) \times (1 - 0,004)} = 7,3\%$

Pour une prévalence de 30% dans la population venant en consultation spécialisée, la valeur prédictive est de : $\text{VPP} = \frac{0,39 \times 0,3}{0,39 \times 0,3 + (1 - 0,98) \times (1 - 0,3)} = 89\%$

L'information apportée par le test est la même dans les 2 cas, mais la probabilité pré-test de la maladie est très différente. La valeur prédictive positive est meilleure dans la population où la proportion de malades est plus importante. En revanche la valeur prédictive négative est meilleure dans la population où la proportion de non-malades est plus importante. Elle est estimée à 99,8% dans la population des femmes dépistées et à 78,9% dans la population des femmes qui viennent en consultation spécialisée.

2 – Les probabilités pré et post-test et les ratios de vraisemblance

L'information apportée par le test dépend de sa sensibilité et de sa spécificité et peut être quantifiée par les ratios de vraisemblance. On distingue le ratio de vraisemblance positif qui correspond à l'information apportée par le test lorsque le test est positif, et le ratio de vraisemblance négatif qui correspond à l'information apportée par le test lorsque le test est négatif.

Le ratio de vraisemblance positif d'un test (*positive likelihood* ratio en anglais, LR+) est le rapport de la vraisemblance d'un résultat positif chez les malades sur la vraisemblance d'un résultat positif chez les non-malades : $\text{RV+} = \frac{\text{P}(\text{Test} + /\text{M})}{\text{P}(\text{Test} + /\text{NM})} = \frac{Se}{1 - Sp}$

Un test qui ne fait pas mieux que le hasard pour discriminer les malades des non-malades, est un test pour lequel la vraisemblance d'un résultat positif chez les malades est égale à la vraisemblance d'un résultat positif chez les non-malades. Cette situation correspond à un RV+ égal à 1. Plus le ratio de vraisemblance positif est supérieur à 1, plus l'information apportée par un résultat positif du test est importante.

Le RV+ permet de passer de la probabilité pré-test à la probabilité post-test lorsque le test est positif. Il multiplie l'Odds pré-test de la maladie. Reprenons l'exemple de la mammographie avec un seuil de positivité correspondant à une haute suspicion de cancer.

$$\text{RV+} = \frac{0,39}{1 - 0,98} = 19,5$$

L'Odds de cancer du sein en situation de dépistage est égal à :

$$\text{Odds pré test} = \frac{\text{prévalence}}{1 - \text{prévalence}} = \frac{0.004}{1 - 0.004} \approx 0{,}004$$

L'Odds post-test lorsque la mammographie est positive :

$$\text{Odds post test} = \text{Odds pré test} \times \text{RV} + = 0{,}004 \times 19{,}5 = 0{,}078$$

La probabilité post-test est égale à : $\frac{\text{Odds post test}}{1 + \text{Odds post test}} = 7{,}2\%$

On retrouve la valeur prédictive positive ou probabilité d'avoir la maladie sachant que le test est positif. Il s'agit d'une autre façon d'appliquer le théorème de Bayes.

Le ratio de vraisemblance négatif (*negative likelihood* ratio en anglais, LR-) est le rapport de la vraisemblance d'un résultat négatif chez les malades sur la vraisemblance d'un résultat négatif chez les non-malades : $\text{RV} - = \frac{\text{P}(\text{Test} - / \text{M})}{\text{P}(\text{Test} - / \text{NM})} = \frac{1 - Se}{Sp}$

Plus le ratio de vraisemblance négatif se rapproche de 0, plus l'information apportée par un résultat négatif du test est importante.

Le RV- de la mammographie est égal à : $\text{RV} - = \frac{1 - 0{,}39}{0{,}98} = 0{,}62$

Si la mammographie est négative l'Odds de la maladie est divisé par 1,6.

$$\text{Odds post test} = \text{Odds pré test} \times \text{RV} - = 0{,}004 \times 0{,}62 \approx 0{,}0025$$

La probabilité post-test est égale à : $\frac{\text{Odds post test}}{1 + \text{Odds post test}} \approx 2{,}5$ pour mille

La probabilité post-test correspond à la probabilité d'avoir la maladie sachant que le test est négatif. Il s'agit de 1 moins la valeur prédictive négative.

Le RV+ dépend surtout de la spécificité du test. Meilleure est la spécificité du test, meilleur est le test pour affirmer la présence de la maladie lorsqu'il est positif. Le RV- dépend surtout de la sensibilité. Meilleure est la sensibilité, meilleur est le test pour éliminer la maladie lorsqu'il est négatif. Prenons l'exemple de 3 tests : la gazométrie dans le sang artériel pour faire le diagnostic d'embolie pulmonaire, la culture de liquide pleural pour faire le diagnostic de tuberculose et le scanner pour faire le diagnostic de masse rénale kystique (tableau 4).

La gazométrie est sensible mais peu spécifique pour le diagnostic d'embolie pulmonaire. Ce test permet de réduire de façon importante la probabilité d'avoir la maladie s'il est négatif en divisant l'Odds pré-test par 10. En revanche il multiplie l'Odds pré-test seulement par 2 lorsqu'il est positif.

A l'inverse la culture de liquide pleural est très spécifique pour le diagnostic de tuberculose mais très peu sensible. Ce test permet d'augmenter de façon importante la probabilité d'avoir la maladie s'il est positif en multipliant l'Odds pré-test par 24. En revanche, il divise l'Odds pré-test seulement par 1,3 s'il est négatif.

Le scanner est un test qui a à la fois une sensibilité de 100% et une spécificité élevée. C'est un test qui n'a pas de faux négatifs. Il permet d'éliminer la maladie lorsqu'il est négatif. Un test positif multiplie par 50 l'Odds pré-test.

III - QUELQUES ASPECTS PARTICULIERS AUX ETUDES D'EVALUATION D'UN TEST DIAGNOSTIQUE

A – Quelques biais spécifiques aux études d'évaluation des tests diagnostiques

1 – Biais de vérification

Dans les études d'évaluation des stratégies diagnostiques, il y a un risque d'obtenir des estimations biaisées chaque fois que le statut malade/non-malade n'est pas mesuré de façon indépendante du test à évaluer ou l'inverse. Par exemple le biais d'incorporation survient lorsque la détermination du statut malade, non-malade repose au moins en partie sur le résultat du test à évaluer. Cela entraîne une surestimation de la sensibilité et de la spécificité.

Dans cette catégorie de biais, on trouve le biais de vérification qui survient lorsque la probabilité d'avoir le *gold standard* dépend du résultat du test à évaluer. Cette situation se présente classiquement lorsque le *gold standard* est invasif ou coûteux et ne peut pas être réalisé chez tout le monde. Dans ce cas, il est plus souvent réalisé chez les sujets qui ont un test positif que chez ceux qui ont un test négatif.

Une étude mise en place pour évaluer les performances diagnostiques de l'électrocardiogramme d'effort a porté sur 414 sujets à risque de coronaropathie. Tous les sujets ont eu un électrocardiogramme d'effort. Tous les sujets ayant un électrocardiogramme d'effort positif ont eu une coronarographie. Pour les sujets ayant un électrocardiogramme d'effort négatif, seuls 40 % pris au hasard ont eu une coronarographie. Les résultats sont présentés dans le tableau 5. La sensibilité et la spécificité estimées chez les sujets qui ont eu une coronarographie sont respectivement de : $Se = \frac{92}{92+46} = 67\%$ et $Sp = \frac{72}{72+27} = 73\%$.

La probabilité d'avoir une coronarographie étant plus élevée chez les sujets ayant un test positif que chez ceux ayant un test négatif, il y a une surreprésentation des tests positifs. La sensibilité du test est surestimée et la spécificité sous-estimée. La probabilité d'avoir une coronarographie ne dépendant que du résultat du test, il est possible d'obtenir les estimations non biaisées de la sensibilité et de la spécificité en utilisant le théorème de Bayes. A partir des résultats présentés dans le tableau 5, nous avons une estimation de :

- la probabilité que le test soit positif dans la population des sujets à risque de coronaropathie : $\frac{119}{119+295} = 28{,}7\%$

- la probabilité d'avoir la maladie sachant le test positif : $\frac{92}{92+27} = 77{,}3\%$

- la probabilité de ne pas avoir la maladie sachant le test négatif : $\frac{72}{46+72} = 61\%$

Estimation de la sensibilité :

$$Se = P(Test+ / M) = \frac{\mathrm{P(M/Test+)} \times \mathrm{P(Test+)}}{\mathrm{P(M/Test+)} \times \mathrm{P(Test+)} + \mathrm{P(M/Test-)} \times \mathrm{P(Test-)}}$$
$$= \frac{0{,}773 \times 0{,}287}{0{,}773 \times 0{,}287 + (1-0{,}61) \times (1-0{,}287)} = 44\%$$

Estimation de la spécificité :

$$Sp = P(Test- / NM) = \frac{\mathrm{P(NM/Test-)} \times \mathrm{P(Test-)}}{\mathrm{P(NM/Test-)} \times \mathrm{P(Test-)} + \mathrm{P(NM/Test+)} \times \mathrm{P(Test+)}}$$
$$= \frac{0{,}61 \times (1-0{,}287)}{0{,}61 \times (1-0{,}287) + (1-0{,}773) \times 0{,}287} = 87\%$$

Lorsqu'il existe un *gold standard* mais qu'il ne peut pas être utilisé chez tous les sujets inclus dans l'étude, il est possible d'estimer les performances du test à évaluer en utilisant le *gold standard* sur un échantillon de sujets positifs et un échantillon de sujets négatifs pris au hasard.

2 – Biais lié à l'utilisation d'un *gold standard* imparfait

Il est très fréquent que le test utilisé comme référence ne soit pas parfait. Si l'on estime la sensibilité et la spécificité du test à évaluer en faisant comme si le test de référence était parfait, ces estimations sont biaisées. En particulier, il est impossible de montrer la supériorité du nouveau test par rapport au test de référence.

Prenons l'exemple d'un nouveau test parfait, dont on évalue les performances par rapport à un test de référence qui a une sensibilité de 90 % et une spécificité de 90 %. Dans une étude portant sur 100 sujets malades et 100 sujets non malades, 10 sujets malades seront classés comme négatifs par le test de référence et 10 sujets non-malades seront classés comme positifs par le test de référence (tableau 6). La sensibilité et la spécificité du nouveau test seront sous-estimées à 90 %.

Dans le cas où les 2 tests sont indépendants conditionnellement au statut vis-à-vis de la maladie, un défaut de sensibilité du test de référence entraîne une sous-estimation de la spécificité du nouveau test. A l'inverse un défaut de spécificité du test de référence entraîne une sous-estimation de la sensibilité du nouveau test.

Il est possible d'estimer les performances diagnostiques d'un test dans la situation où le test de référence n'est pas parfait. Le statut malade, non-malade des sujets inclus dans l'étude n'est pas directement observé, c'est une variable latente. Les résultats positif ou négatif du test à évaluer et du test de référence apportent de l'information sur le statut des sujets.

Dans une étude portant sur un échantillon d'une population dans laquelle les sujets inclus ont eu le test à évaluer et le test de référence, il y a 5 paramètres à estimer : la sensibilité et la spécificité du nouveau test, la sensibilité et la spécificité du test de référence et la prévalence de la maladie. Le tableau 2×2 présentant les résultats croisés des 2 tests permet d'estimer 3 paramètres. Si l'on connaît la sensibilité et la spécificité du test de référence, alors il est possible d'estimer la sensibilité et la spécificité du nouveau test et la prévalence de la maladie. Les données observées apportent 3 degrés de liberté.

Si aucun des paramètres n'est connu avec certitude, alors il est nécessaire d'augmenter l'information apportée par les données. Hui et Walter [3] ont proposé d'utiliser des échantillons de sujets provenant de 2 populations ayant des prévalences de la maladie très différentes. Ils ont pris l'exemple de l'évaluation d'un nouveau test cutané pour faire le diagnostic de tuberculose, le test de Tine. Le test de référence est le test cutané de Mantoux. Ils ont repris les données de 2 études : l'une dans laquelle les 2 tests ont été appliqués à un échantillon provenant de la population d'un district scolaire ayant une faible prévalence de la maladie, et l'autre dans laquelle les 2 tests ont été appliqués à un échantillon d'une population d'un sanatorium ayant une prévalence élevée de la maladie.

Sous l'hypothèse d'indépendance conditionnelle des 2 tests et de performances diagnostiques identiques dans les 2 populations, il y a 6 paramètres à estimer : la sensibilité et la spécificité de chacun des tests et la prévalence de la maladie dans chacune des populations. Les tableaux 2×2 présentant les résultats croisés des 2 tests dans chacune des populations nous apportent chacun 3 degrés de liberté. Le nombre de degrés de liberté est de 6, l'information apportée par les données est donc suffisante pour estimer tous les paramètres. Cette approche peut être généralisée à plus de 2 tests ou plus de 2 populations.

La présentation des méthodes d'estimation dépasse le cadre de cet ouvrage. Le lecteur intéressé pourra trouver ces méthodes dans les références données en fin de chapitre.

B – L'intervalle de confiance et le calcul du nombre de sujets nécessaire

1 – Intervalle de confiance de la sensibilité et de la spécificité

La sensibilité et la spécificité sont estimées respectivement par la proportion de résultats positifs chez les malades et la proportion de résultats négatifs chez les non-malades. Leur variance et erreur standard estimées correspondent à la variance et à l'erreur standard d'une proportion.

Pour la sensibilité : $\text{Variance} = \dfrac{\text{Se}\times(1-\text{Se})}{\text{M}}$ $\qquad$ $\text{Erreur standard} = \sqrt{\dfrac{\text{Se}\times(1-\text{Se})}{\text{M}}}$

M = effectif de malades

Pour la spécificité : $\text{Variance} = \dfrac{\text{Sp}\times(1-\text{Sp})}{\text{NM}}$ $\qquad$ $\text{Erreur standard} = \sqrt{\dfrac{\text{Sp}\times(1-\text{Sp})}{\text{NM}}}$

NM = effectif de non-malades

Si les effectifs de malades et non-malades sont suffisamment grands et si la sensibilité et la spécificité ne sont pas trop proches de 100 %, l'intervalle de confiance peut alors être construit en utilisant la méthode basée sur l'approximation de la distribution binomiale par une distribution de Gauss.

Intervalle de confiance à 95% de la sensibilité ou spécificité estimées : $\text{P} \pm 1{,}96 \times \sqrt{\dfrac{\text{P}\times(1-\text{P})}{\text{N}}}$

P correspond à la sensibilité ou à la spécificité estimées

N correspond à l'effectif de malades ou non-malades

Cette méthode de construction de l'intervalle de confiance fondée sur l'approximation gaussienne est en général applicable quand NP>= 5 et N(1-P)>=5.

Quand les effectifs sont trop petits ou les estimations trop proches de 100 %, il convient de construire l'intervalle de confiance exact fondé sur la loi binomiale.

2 – Calcul du nombre de sujets nécessaire

Quand la sensibilité et la spécificité attendues ne sont pas trop proches de 100 %, la méthode de calcul du nombre de sujets nécessaire pour estimer une sensibilité et une spécificité est la même que pour estimer une proportion de malades dans une population (prévalence de la maladie).

Si l'étude mise en place est une étude de type cas-témoins, le nombre de malades à inclure pour estimer la sensibilité et le nombre de non-malades à inclure pour estimer la spécificité sont déterminés séparément. Il convient alors de fixer la sensibilité et la spécificité attendues, la largeur souhaitée de l'intervalle de confiance et sa probabilité de couverture qui est en général de 95 %.

Si l'étude mise en place est une étude de type cohorte, il est nécessaire de tenir compte de la prévalence de la maladie dans la population d'où sera tiré l'échantillon de l'étude. Dans la plupart des cas, la prévalence de la maladie est inférieure à 50 %. La stratégie à suivre est alors la suivante. On calcule le nombre de malades à inclure pour estimer la sensibilité puis on calcule le nombre de sujets à inclure pour avoir le nombre de malades nécessaire, compte tenu de la prévalence.

$$N_{Total} = \frac{M}{Prévalence}$$

N_{Total} est l'effectif total de sujets à inclure dans l'étude

M est l'effectif de malades

Si la prévalence de la maladie est supérieure à 50 %. La même stratégie est appliquée mais c'est le nombre de non-malades à inclure pour estimer la spécificité qui est calculé en premier.

Dans le cas où la sensibilité et la spécificité attendue sont proches de 100 %, il convient d'utiliser une méthode exacte de calcul du nombre de sujets reposant sur la distribution binomiale.

IV – CONCLUSION

L'objectif de ce chapitre est de donner au lecteur les outils méthodologiques nécessaires pour la mise en place d'une étude visant à estimer les capacités diagnostiques d'un nouveau test. L'accent a été mis sur les études visant à estimer la sensibilité et la spécificité, ce qui correspond à la phase précoce de l'évaluation d'un nouveau test. Les deux ouvrages qui sont donnés en référence permettront aux lecteurs qui le désirent d'aller plus loin [4 ; 5]

Références

1. Weiner DA, Ryan TJ, McCabe CH, Kennedy JW, Schloss M, Tritani F, Chaitman BR, Fisher LD. Correlations among history of angina, ST-segment response and prevalence of coronary artery disease in the coronary artery surgery study (CASS). N Engl J Med 1979; 301: 230-5.

2. Hall FM, Storella JM, Silverstone DZ, Wyshak G. Non palpable breast lesions: recommendations for biopsy based on suspicion of carcinoma at mammography. Radiology 1988; 167: 353-8.

3. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. Biometrics 1980; 36: 167-71.

4. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Ed Oxford University Press, New York, 2003.

5. Zhou XH, Obuchowsky NA, McClish DK. Statistical methods in diagnostic medicine. Ed John Wiley & Sons, New York, 2002.

Tableau 1 - Les quatre situations possibles selon le résultat du test diagnostique et le statut malade ou non malade

	Maladie présente	Maladie absente	
Test positif	Vrai Positif (VP)	Faux Positif (FP)	VP+FP
Test négatif	Faux Négatif (FN)	Vrai Négatif (VN)	FN+VN
	VP+FN	FP+VN	N

Tableau 2 – Existence de douleurs thoraciques en fonction de la présence ou non d'une coronaropathie chez des sujets à risque (étude CASS)

	Coronaropathie présente	Coronaropathie absente	
Douleur thoracique	969	245	1214
Pas de douleur thoracique	54	197	251
	1023	442	1465

Tableau 3 – Classement de 119 femmes ayant un cancer du sein et de 281 femmes n'ayant pas de cancer du sein selon le degré de suspicion de cancer à la mammographie

Résultat mammographie Degré de suspicion de cancer	Cancer du sein	Pas de cancer du sein
Haute	47	6
Moyenne	57	117
Légère	9	37
Minime	6	121
Total	119	281

Tableau 4 – Sensibilité, spécificité et ratios de vraisemblance positif et négatif de trois tests différents

Test	Sensibilité	Spécificité	RV+	RV-
Gazométrie pour diagnostic d'embolie pulmonaire	0,95	0,5	1,9	1/10 = 0,1
Culture de liquide pleural pour le diagnostic de tuberculose	0,24	0,99	24	1/1,3 = 0,77
Scanner pour le diagnostic de masse kystique rénale	1	0,98	50	0

Tableau 5 – Résultats de l'électrocardiogramme (ECG) d'effort chez 414 sujets à risque de coronaropathie

	Résultat de la coronarographie			
	Maladie coronarienne	Pas de maladie coronarienne	Pas de coronarographie	Total
ECG d'effort positif	92	27	0	119
ECG d'effort négatif	46	72	177	295

Tableau 6 – Résultats d'une étude visant à évaluer les performances d'un nouveau test parfait par rapport à un test de référence qui a une sensibilité de 90% et une spécificité de 90%. L'étude porte sur un échantillon de 200 sujets comprenant 100 malades et 100 non-malades

	Test de référence positif	Test de référence négatif
Test à évaluer positif	90 VP	10 FP
Test à évaluer négatif	10 FN	90 VN
	100 = 90 VP + 10 FN	100 = 90 VN + 10 FP

Figure 1 – Courbe ROC de la mammographie pour le diagnostic de cancer du sein chez des femmes ayant une biopsie positive (étude de Hall FM et al)