

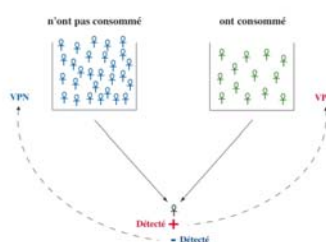
Comment valider une analyse ?

Docteur Yves JACOMET
Responsable de l'APMD de Nice

Avec l'attribution d'un Prix Nobel en 2002 aux travaux de Tversky et Kahneman, une reconnaissance de grande ampleur a été accordée à l'interprétation bayésienne des méthodes d'analyse scientifiques. Dans le droit fil de ces travaux, Gigerenzer et Hoffrage ont expliqué que le cerveau humain n'avait pas l'intuition des probabilités puis ils ont démontré qu'il valait mieux les remplacer par les fréquences correspondantes pour se faire comprendre. Ils les ont appelées "fréquences naturelles" et leur usage est maintenant largement répandu pour l'étude des enjeux de santé publique.

Plus aucune démarche diagnostique ne peut échapper à la démarche bayésienne car ce qui compte n'est pas le résultat d'un examen, fût-il obtenu avec les appareils "les plus puissants ou les plus sensibles" du marché, mais son interprétation. Cette interprétation dite diagnostique se fait avec le théorème de Bayes et calcule des valeurs prédictives positives et négatives (VPP et VPN) qui sont très éloignées des définitions de la sensibilité et de la spécificité. La sensibilité et la spécificité sont indispensables au calcul des VPP et VPN mais ne peuvent en aucun cas se substituer à une interprétation diagnostique en bonne et due forme comportant obligatoirement le calcul des VPP et des VPN.

Théorème de Bayes



Aucune exception dans toutes les méthodes de mesure (médecine, physique, chimie, astronomie, téléphonie etc.)

Dans la lutte contre le dopage, ce qui est intéressant n'est pas le résultat de l'analyse de laboratoire mais son interprétation résumée dans l'alternative: "consommation ou pas consommation". L'analyse du laboratoire doit bien sûr être irréprochable sinon elle n'est pas interprétable.

En médecine, l'interprétation diagnostique est quotidienne. Et, pour des enjeux de santé publique, les critères d'interprétation des examens de radiologie ou de laboratoire sont

étudiés sur des échantillons de population immenses ou cohortes. C'est le cas du cancer du côlon, du cancer de la prostate, de la sérologie du VIH, du cancer du sein et des empreintes génétiques. Mais, en fait, toutes les maladies sont passées progressivement au crible de l'interprétation diagnostique bayésienne. Le but est, par exemple, d'opérer les femmes porteuses d'un cancer du sein au lieu de celles porteuses d'une image radiologique anormale.

Ce raisonnement est applicable à tous les signaux et donc tous les résultats fournis par toutes les machines scientifiques aussi sophistiquées soient-elles. Autrement dit, l'interprétation diagnostique bayésienne s'applique naturellement à la physique et à la chimie mais aussi à l'astronomie, la téléphonie, l'électronique, l'aviation etc. Aucun domaine ne peut y échapper dès lors que la mesure d'un signal a retourné un résultat. En téléphonie, l'activation d'une borne donne un signal et la distance qui sépare le téléphone de cette borne relève de l'interprétation diagnostique bayésienne. Cette distance peut être petite mais aussi très grande en fonction des contextes. Il n'y a pas d'interprétation automatique ou binaire à la grande surprise des tribunaux parfois quand il faut retracer un itinéraire lors d'une expertise médico-légale.

En médecine, l'interprétation diagnostique intuitive ou spontanée donne actuellement des résultats surprenants. On dit "actuellement" parce que la formation des médecins fera disparaître peu à peu cet écart entre une démarche irrationnelle due à l'intuition et une démarche rationnelle due à l'interprétation diagnostique bayésienne.

Dans le cas du cancer du sein, une mammographie déclarée positive sans plus de précision chez une femme de 40 à 50 ans fait poser spontanément à 75% des médecins un diagnostic de cancer du sein alors que le pourcentage attendu de cancers du sein dans cette tranche d'âge lors d'une mammographie anormale n'est que de 7,5% en tenant compte d'une prévalence de 1%, d'une sensibilité de la mammographie de 80% et d'une spécificité de 90%. Quand on explique aux médecins testés la démarche diagnostique, un grand nombre d'entre eux se rendent compte de leur erreur de raisonnement. Dans tous les cas, comme l'ont démontré Gigerenzer et Hoffrage, la justesse diagnostique ne peut pas être intuitive ni spontanée; elle doit être enseignée.

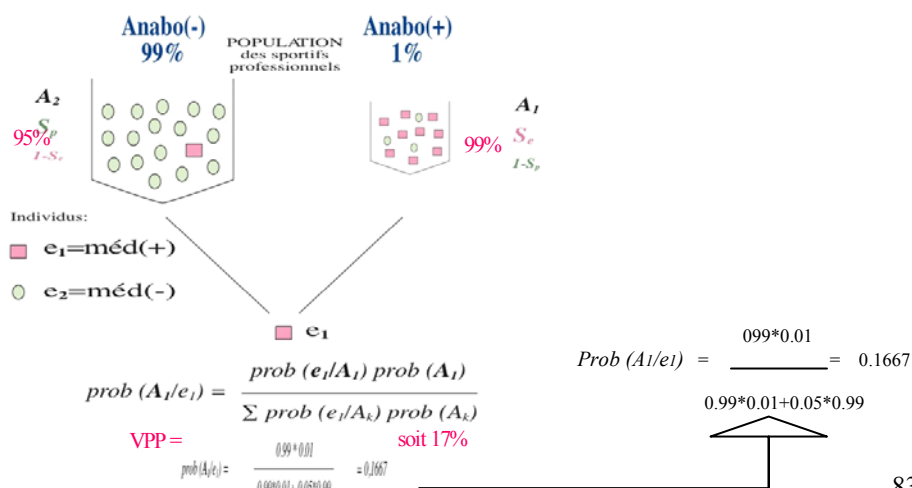
Dans le cas des empreintes génétiques, la valeur prédictive positive d'une empreinte génétique est proche de zéro et donc quasiment nulle si les empreintes génétiques sont menées sur le tout venant. Il faut faire des empreintes génétiques uniquement sur les personnes qui ont un rapport avec la question posée. Pour la recherche d'un criminel, il faut que les personnes prélevées aient un lien même ténu avec l'affaire d'où l'importance des autres indices et donc de l'enquête. C'est à cette condition préalable que l'empreinte génétique a une valeur prédictive positive très élevée et permet de faire un diagnostic de culpabilité très proche de la certitude. Dans tous les autres cas, l'erreur judiciaire n'est plus marginale. Il est d'ailleurs possible de mesurer un taux d'erreurs judiciaires prévisible en faisant analyser des échantillons anonymes et non signalés au laboratoire. Le résultat a été édifiant puisqu'il atteignait 2% des échantillons dans certains cas aux Etats-Unis. On est donc très loin de l'affirmation péremptoire que les empreintes génétiques sont infaillibles. Et, après leur avoir

fourni les valeurs de la sensibilité et de la spécificité de l'analyse par empreintes génétiques, Gigerenzer et Hoffrage ont montré que 32% des élèves magistrats testés en fin de formation et 56% des étudiants en droit persistaient à fournir un diagnostic erroné de culpabilité. Ce qui montre la force du penchant pour l'intuition au détriment de la démonstration.

Pour résumer, ce n'est pas la sensibilité d'un appareil qui importe mais le couple sensibilité-spécificité. Plus la sensibilité augmente plus la spécificité diminue et inversement. Il n'y a pas d'exception. Ensuite, il faut connaître la prévalence de ce qu'on recherche c'est à dire la fréquence a priori du diagnostic qu'on veut porter et ce dans la sous-population concernée. A partir de ce moment-là, on peut calculer une VPP et une VPN. La VPP est la probabilité d'avoir consommé quand l'examen est positif et la VPN est la probabilité de ne pas avoir consommé quand l'examen est négatif. La probabilité d'avoir consommé n'est jamais certaine et la probabilité de ne pas avoir consommé non plus. En revanche, il est possible d'augmenter la VPP et donc le diagnostic de consommation en relevant le seuil de détection ou de décision. On parle alors de seuil discriminant. Plus la concentration est élevée plus le diagnostic de consommation est probable. De même, plus la concentration du produit est élevée plus son identification est probable. Elle devient même possible en spectrométrie de masse dès lors que la concentration est suffisamment élevée pour permettre une acquisition des données en full scan et la recherche du même produit en bibliothèque de spectres. Dans un grand nombre de cas, l'identification du produit est alors proche de la certitude. En revanche, quand le seuil discriminant est trop bas, l'interprétation diagnostique est aléatoire et n'a guère plus d'efficacité qu'un tirage au sort à pile ou face comme le montre tous les calculs effectués avec le théorème de Bayes.

A ma connaissance, l'Agence Mondiale Antidopage est informée de ces questions et ses recommandations en tiennent compte au fil des ans.

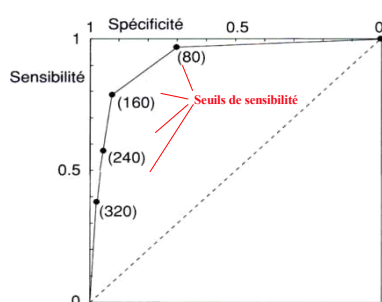
Dans le cas d'un produit dopant anabolisant comme la nandrolone ou de toute substance qui réunit les mêmes conditions supposées à savoir une prévalence de consommation de 1% et une méthode de détection sensible à 99% et spécifique à 95%, la probabilité d'une consommation dans un délai compatible avec le moment de prélèvement avant la détection en laboratoire n'est que de 83% soit 17% d'erreurs judiciaires ou disciplinaires prévisibles. En vertu de la loi des grands nombres, il y aura assurément 17 déclarations erronées de culpabilité sur 100 échantillons examinés si on ne dispose d'aucun autre indice. Et, pour ne pas arranger les choses, plus la prévalence de la consommation est faible plus le nombre de déclarations de culpabilité erronées augmente.



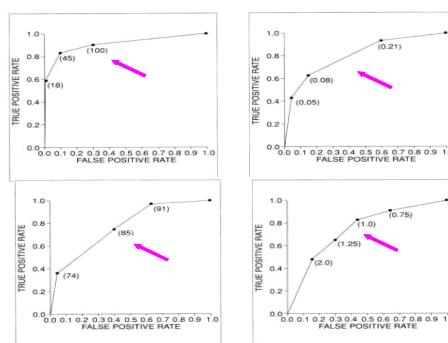
Progressivement, on ne devrait plus faire un diagnostic de consommation sur la foi d'un examen rendu sous la forme négatif ou positif. Il faut impérativement y ajouter la mention du seuil discriminant si le produit n'a pas pu être clairement identifié en spectrométrie de masse. La quantification et l'identification sont deux opérations distinctes. Et, dans un nombre grandissant de cas, il faudrait remplacer totalement l'alternative négatif-positif par l'alternative VPP-VPN. Ce qui veut dire en clair que les autorités doivent être informées de la possibilité d'une erreur judiciaire jamais totalement écartée et surtout de son ordre de grandeur. Cet ordre de grandeur s'exprime presque toujours en cas pour cent et quasiment jamais en cas par million. C'est la confusion très répandue entre la sensibilité et la VPP, qui ont des ordres de grandeur très différents, qui a conduit à des déclarations imprudentes y compris de très hauts responsables comme le président de la société allemande de médecine légale en 1998 qui affirmait dans la grande presse que les empreintes génétiques ne comportaient aucun faux positif. On sait à présent sans aucun doute possible que c'est totalement faux.

Les militaires savent depuis des dizaines d'années que les radars les plus chers n'empêchent pas les erreurs et conduisent assez fréquemment à tirer sur des objectifs non identifiés à temps comme amis. Car, c'est l'identification ou spécificité qui compte pour faire un diagnostic et pas la sensibilité. Un radar est immensément sensible mais il n'est aucunement spécifique. Une méthode de détection perd toujours en spécificité ce qu'elle gagne en sensibilité. De la même manière, on peut dire en médecine qu'il n'existe pas d'analyse de laboratoire pathognomonique c'est à dire sensible à 100% et simultanément spécifique à 100%. Pour s'en rendre compte, il suffit d'établir la courbe ROC d'un appareil ou d'une méthode de mesure. Cette courbe ROC, qui n'était jamais signalée en médecine, commence à y être connue. L'établissement complet d'une telle courbe coûterait extrêmement cher mais rien n'empêche de déterminer le point précis de cette courbe auquel on travaille c'est à dire les valeurs mesurées de la sensibilité et de la spécificité pour un problème posé. En y ajoutant la connaissance de la prévalence, le calcul exact des VPP et VPN devient facile.

Courbe expérimentale ROC



Diversité des courbes ROC



vertical : True Positive rate
horizontal : False Positive Rate

En résumé, seul le meilleur couple sensibilité-spécificité permet de faire le diagnostic le plus probable. La sensibilité seule ne mène à rien. En médecine, ce problème est maintenant bien cerné avec la comparaison des examens radiologiques de différentes sensibilités comme la radiographie simple, le scanner, l'IRM et le PET-scan. Ce n'est pas l'appareil ou la méthode la plus sensible qui permet de faire le meilleur diagnostic. Tout dépend de ce qu'on cherche en fonction du couple sensibilité-spécificité pour le problème posé. Il en va de même en biologie et tout particulièrement avec la spectrométrie de masse dans la lutte contre le dopage avec les méthodes d'acquisition des données de sensibilité différentes comme le full scan et la SIM en SM et la MRM en SM-SM. Tout dépend de la molécule qu'on cherche et, en tout état de cause, le résultat final dépendra du couple sensibilité-spécificité et pas de la sensibilité seule.

Quant à l'interprétation diagnostique de consommation d'une substance quelconque, elle ne peut pas s'exprimer autrement qu'en VPP et VPN si l'on veut en minimiser la part subjective et objectiver le risque d'erreur.

Questions de la salle

Martine PREVOST, médecin au CROS Limousin

Je souhaite savoir si l'analyse de l'échantillon B dans un autre laboratoire produirait des résultats plus fiables.

Yves JACOMET

L'analyse de l'échantillon B doit rester possible à tout moment. Quand une carrière de sportif en dépend, il est nécessaire de respecter le principe du contradictoire et de favoriser la discussion. Un dosage rassemble des appareils de mesure, des méthodes de travail et des individus. Le fonctionnement de cet ensemble est forcément différent d'un endroit à l'autre. Il faut favoriser la discussion entre les équipes pour confronter les résultats quand ils sont différents. J'ai participé à une discussion sur un cas présumé de dopage porté en appel devant les instances internationales du football où Martial Saugy avait conclu dans le même sens que nous mais dans un sens différent du laboratoire qui avait procédé aux analyses. Il ne fait pas de doute que c'est au jury disciplinaire de prendre une décision de synthèse qui tiendra compte d'indices et de preuves rassemblés qui dépassent largement les seuls résultats de biologie et dont le laboratoire n'a pas connaissance.

Patrick MAGALOFF

Compte tenu du faible nombre de laboratoires agréés, une telle proposition serait irréalisable dans la situation actuelle.

Yves JACOMET

Il est plus facile pour un laboratoire d'identifier une molécule qu'il voit souvent qu'une molécule qu'il voit rarement. L'hyperspécialisation des LNDD, qui voient très peu souvent chaque molécule détectée, peut s'avérer être un handicap quand il faut faire un diagnostic de consommation à partir d'une concentration très faible. Je crois que les spécialistes de L'AMA sont très au courant de ce type de problème qui, au demeurant, n'est pas facilement surmontable. A titre de comparaison, il n'est plus accepté en médecine de faire un diagnostic grave sur la foi d'un seul résultat sans l'avoir répété. De même dans l'aviation civile ou militaire, il est impensable de faire un diagnostic sur la nature d'un avion sur un écran radar sans l'avoir formellement identifié séparément.