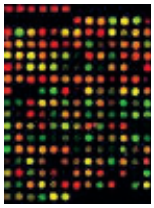
MEDECINE/SCIENCES 2008 ; 24 : 317-9

Défis statistiques posés par les biopuces

Autant d'espoir que de faux positifs ?

Stefan Michiels, Catherine Hill



Unité de Biostatistiques et d'Épidémiologie, Institut Gustave Roussy, 39, rue Camille Desmoulins, 94805 Villejuif Cedex, France.
michiels@igr.fr

La valeur pronostique des marqueurs

Nous disposons d'outils de plus en plus nombreux et performants pour étudier des marqueurs biologiques : les puces à ADN, qui mesurent simultanément l'expression de dizaines de milliers de gènes à la fois à partir d'un seul prélèvement, ou les *tissue-arrays,* qui permettent d'étudier l'expression d'un seul gène dans les prélèvements de centaines de patients [1]. Très souvent ces marqueurs biologiques sont mesurés pour étudier leur valeur pronostique c'est-à-dire leur capacité à prédire le devenir des malades [2, 3]. Dans le cas le plus simple, celui où un marqueur pronostique est soit positif, soit négatif, on comparera l'évolution dans les deux groupes de patients définis par la valeur du marqueur : positif *versus* négatif. Un marqueur n'est pas pronostique dans l'absolu, il est pronostique pour un risque défini. Par exemple, en cancérologie, un marqueur peut être lié au risque de développer des métastases sans être lié au risque de rechute locale, un autre sera lié à l'ensemble des risques d'évolution de la maladie : rechute locale, à distance ou décès. Dans le contexte d'un essai thérapeutique, on étudie aussi des marqueurs pour évaluer leur capacité à prédire l'effet du traitement étudié dans l'essai. Dans ce cas, on cherche à savoir si la différence d'efficacité entre le traitement étudié et le traitement de référence dépend de la valeur du marqueur (voir *Glossaire*). Dans le cas simple d'un marqueur positif ou négatif, sa valeur prédictive s'étudie en comparant la différence entre les deux traitements dans les deux sous-groupes de l'essai défini par le marqueur : marqueur positif et marqueur négatif. Malgré le grand nombre d'articles consacrés à leur étude, peu nombreux sont les facteurs dont la valeur prédictive est considérée comme établie. Beaucoup d'investigateurs pensent en effet qu'il suffit de comparer les deux traitements séparément dans chaque sous-groupe, ce qui est une source d'erreur considérable. Ils pensent aussi que si l'une des comparaisons est significative et l'autre non, cela établit la valeur prédictive du marqueur, ce qui est une autre erreur [4]. La démonstration de l'effet pronostique d'un marqueur nécessite la comparaison de deux groupes de patients définis par la valeur du marqueur, à traitement constant. La démonstration de l'effet prédictif d'un marqueur nécessite que soit réalisé un essai clinique « randomisé » comparant l'effet de deux traitements. Les données seront analysées en comparant la différence entre les traitements dans les deux sous-groupes définis par la valeur du marqueur.

La malédiction de la dimension

Nous constatons à l'heure actuelle une augmentation exponentielle du nombre de publications étudiant les marqueurs biologiques, et nombre d'entre elles affichent des résultats significatifs. Ce sont pour la plupart des études rétrospectives qui donnent lieu à des interprétations biologiques *a posteriori*. Les études actuelles portent toutes sur un petit nombre de patients, expliqué par le coût des puces. L'analyse de petits nombres de patients avec des milliers de variables pour chaque patient pose un problème méthodologique connu sous le nom de la « malédiction de la dimension ». Quelle que soit la technologie (Agilent, Affimetrix, etc.), une fois résolus certains détails techniques (normalisation, filtrages, contrôles de qualité, soustraction du bruit de fond, etc.), le problème se résume à l'exploitation et l'interprétation de l'information concernant l'expression des gènes [5]. Nous utiliserons les données publiées sur le cancer du sein pour illustrer notre propos. Dans une étude pionnière, à partir d'une population de 78 patientes sans

envahissement ganglionnaire (N-) étudiées au moment du diagnostic, van't Veer *et al.* [6] ont identifié, parmi 25 000 gènes, 70 gènes dont l'expression est différente chez les patientes ayant développé une métastase à distance dans les 5 ans après le diagnostic et chez les autres patientes.

Le problème des faux positifs dans l'identification des gènes

Si l'on étudie plusieurs marqueurs, et nous avons vu que plusieurs peut vouloir dire des dizaines de milliers, il faut avoir conscience du risque de faux positifs.

Cas où aucun marqueur n'est pronostique

Dans ce cas, tous les marqueurs qui apparaîtront comme pronostiques dans une étude seront de faux positifs. Avec une puce à ADN étudiant 10 000 gènes, le hasard seul fera ainsi apparaître à tort comme liés au pronostic environ 500 gènes (5 % x 10 000) si l'on prend un seuil de significativité de 5 % et environ 100 gènes si l'on prend un seuil de 1 %. Ces gènes, identifiés à tort comme pronostiques, sont des faux positifs. La caractéristique de ces associations apparentes mais fallacieuses est qu'elles ne sont pas reproductibles.

Cas où une proportion de marqueurs a une valeur pronostique

On peut estimer le nombre attendu de faux positifs en fonction de divers paramètres. Dans une étude comparant des groupes de patients à partir de données de biopuces, le nombre attendu de faux positifs qu'on identifiera à la suite de l'expérience dépend : (1) de cette proportion de gènes réellement différentiellement exprimés ; (2) de la distribution des vraies différences ; (3) de la variabilité de l'expression des gènes et (4) du nombre de sujets inclus dans l'étude. Ce dernier paramètre est le seul sous le contrôle de l'expérimentateur. Supposons que 100 des 10 000 gènes étudiés sont réellement deux fois plus ou deux fois moins exprimés dans l'un des groupes par rapport à l'autre et que nous connaissons les distributions statistiques des expressions des gènes (suivant une loi normale) [7]. Si l'on étudie 10 patients (5 par groupe), l'identification des 100 gènes les plus différentiellement exprimés sélectionnera 91 faux positifs et seulement 9 des 100 gènes vraiment différents. Le nombre attendu de faux positifs passe à 25 si l'effectif est de 30 patients par groupe et à 5 s'il atteint 56 patients par groupe. On remarque ainsi que la proportion de faux positifs est d'autant plus faible que le nombre de sujets est grand.

Analyse critique des études proposant une signature moléculaire

Il y a beaucoup de faux positifs parmi les gènes identifiés et donc les listes de gènes identifiés sont très peu reproductibles. Nous regardons les études qui ont pour but de déterminer une signature moléculaire (voir *Glossaire*) avec un œil critique. En l'absence d'un deuxième jeu de données indépendantes, la stratégie d'analyse de telles études pilotes comprend habituellement deux étapes. La première consiste à définir sur une partie des patients - échantillon d'apprentissage - une signature, c'est-à-dire un ensemble de gènes dont l'expression est la plus liée au pronostic. La seconde étape consiste à valider la classification sur les autres patients - constituant l'échantillon de validation -. L'efficacité de cette signature peut être évaluée par la proportion de mauvaises classifications sur l'échantillon de validation.

Nous avons repris les données de 7 études en cancérologie [8]. Ces études cherchaient à identifier des gènes prédictifs de la rechute ou du décès et étaient les seules études publiées entre janvier 1995 et avril 2003. Elles incluaient au moins 60 patients, la plus grande étude en comportait 240. Dans chaque étude, les gènes dont l'expression était différente chez les patients ayant rechuté et chez les patients sans rechute étaient identifiés à partir d'un échantillon d'apprentissage. La qualité de la prédiction était ensuite évaluée sur le reste de la population étudiée. Nous avons tiré au sort de façon répétée un échantillon d'apprentissage, afin de quantifier la reproductibilité de la liste des gènes les plus liés au pronostic et évalué les performances pronostiques des signatures moléculaires publiées en oncologie. La relation entre le nombre de patients dans l'échantillon d'apprentissage et la proportion moyenne de mauvaises classifications dans l'échantillon de validation était également étudiée.

Pour chaque étude, nous avons fait varier la taille de l'échantillon d'apprentissage et pour une taille donnée, nous avons tiré 500 fois au sort l'échantillon d'apprentissage sur l'ensemble de la population d'étude, en calculant à chaque fois la signature, c'est-à-dire l'ensemble des 50 gènes les plus corrélés au pronostic. Nous avons pu montrer que la liste des gènes les plus pronostiques était très instable. Par exemple, dans l'étude de van't Veer, la confirmation a été faite initialement sur un petit jeu de validation de 19 patientes [6]. En considérant la même taille de jeu d'apprentissage (n = 78) parmi les 97 (78+19) patientes et en analysant 500 signatures avec des échantillons de 78 patientes différents, seuls 14 des 70 gènes identifiés par van't Veer ont été retrouvés dans plus de la moitié des signatures. Et à l'inverse, 10 autres gènes qui n'apparaissaient pas dans l'étude initiale étaient identifiés dans plus de la moitié des signatures. Au total, 564 gènes étaient inclus dans au moins une signature. En outre, la même signature n'a jamais été obtenue deux fois. Récemment, Eindor *et al.* [9] ont proposé une méthode de calcul permettant d'estimer le nombre de patients nécessaire pour obtenir une reproductibilité élevée de la liste des gènes déclarés pronostiques. En reprenant le même jeu de données dans le cancer du sein, ils ont montré que plusieurs milliers de patientes sont nécessaires pour que deux listes de gènes pronostiques se recouvrent d'au moins 50 %. Nous avons aussi montré que la qualité de la prédiction, évaluée par la proportion de mauvaises classifications sur l'échantillon de validation, n'était pas très bonne et d'autant plus mauvaise que l'échantillon d'apprentissage était petit.

Reproductibilité des résultats sur d'autres données

Une étude de validation est une étude conçue pour confirmer les résultats d'une étude précédente, dont l'objectif est de réduire la place laissée au hasard et aux éventuels biais [10]. Par définition, cette étude est indépendante des données ayant servi à générer l'hypothèse à vérifier. Ainsi les patients de la première étude ne doivent pas faire partie de la seconde

étude, cela peut sembler évident cependant cette règle élémentaire n'est pas toujours respectée. Ainsi dans l'étude de van't Veer, les 70 gènes ont été identifiés dans une population de 78 patientes N- dont 34 patientes ayant eu une métastase à distance, et la validation de cette signature [6] a été faite sur une population incluant 31 des 34 patientes de la première étude [11]. Si l'on considère les 180 patientes qui n'étaient pas incluses dans l'étude pilote et pour lesquelles nous connaissons le statut métastatique à 5 ans, nous pouvons calculer les indices de sensibilité et spécificité, correspondant aux proportions de patientes correctement diagnostiquées. Nous constatons que la signature de van't Veer est un outil très sensible : elle prédit correctement 93 % des cas qui vont métastaser à 5 ans (intervalle de confiance à 95 % : 81 à 99 %) mais peu spécifique : elle identifie seulement 53 % (44 à 61 %) des patientes qui ne vont pas métastaser. Très récemment, une deuxième validation [12] a été réalisée sur 307 patientes atteintes d'un cancer du sein sans envahissement ganglionnaire et elle a conduit à des résultats similaires : une bonne sensibilité atteignant 90 % (78 à 95 %) mais une mauvaise spécificité de 42 % (36 à 48 %).

En outre, pour que la validation ait un sens, le marqueur étudié doit être mesuré de la même manière : s'il a été identifié à partir d'une puce à ADN, il faut valider sa valeur pronostique en l'étudiant avec le même type de puce sans changer de technique de mesure [5]. La règle de classement des patients en bon et mauvais pronostic en fonction des résultats du marqueur doit être exactement la même pour la validation que celle qui a été déterminée dans la première étude. Adapter la règle aux nouvelles données est une façon très efficace de piper les dés, c'est pourtant une erreur assez commune. Enfin, il faut montrer que la combinaison de marqueurs ajoute de la valeur pronostique aux variables classiques déjà utilisées en routine clinique. Dans l'exemple du cancer du sein que nous décrivons, à l'heure actuelle, il n'est pas démontré que la signature moléculaire de van't Veer apporte une nouvelle information par rapport aux facteurs histologiques connus qui sont l'âge de la patiente, le grade de la tumeur, le nombre de ganglions envahis et la présence de récepteurs hormonaux dans la tumeur [5].

Conclusion

Il faut considérer avec prudence les annonces d'identification de signatures moléculaires de bon ou mauvais pronostic déterminées sur quelques dizaines de patients. L'étude de milliers de gènes sur quelques dizaines de patients conduit à des résultats instables et peu reproductibles. Nous attendons beaucoup des nouveaux outils de la génomique qui sont très puissants, mais il faut les évaluer avec rigueur et avancer sur des bases solidement établies. ◊

The statistical challenge of false positives with microarrays

RÉFÉRENCES

1. Jordan B. Chroniques génomiques. Des puces ADN en clinique ? *Med Sci (Paris)* 2007 ; 23 : 210-4.
2. Figarella-Branger D, Colin C, Chinot O, *et al.* Tumorothèque de l'AP-HM : cartographie moléculaire des gliomes. *Med Sci (Paris)* 2006 ; 22 (suppl 1) : 54-9.
3. Jardin F, Tilly H. Un saut (de puce) vers une application clinique des profils d'expression génique dans les lymphomes. *Med Sci (Paris)* 2004 ; 20 : 848-50.
4. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005 ; 365 : 176-86.
5. Michiels S, Koscielny S, Hill C. Interpretation of microarray data in cancer. *Br J Cancer* 2007 ; 96 : 1155-8.
6. Van 't Veer LJ, Dai H, van de Vijver MJ, *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 2002 ; 415 : 530-6.
7. Pawitan Y, Michiels S, Koscielny S, *et al.* False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* 2005 ; 21 : 3017-24.
8. Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet* 2005 ; 365 : 488-92.
9. Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci USA* 2006 ; 103 : 5923-8.
10. Ransohoff DF. Bias as a threat to the validity of cancer molecular-marker research. *Nat Rev Cancer* 2005 ; 5 : 142-9.
11. Van de Vijver MJ, He YD, van't Veer LJ, *et al.* A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002 ; 347 : 1999-2009.
12. Buyse M, Loi S, van't Veer L, *et al.* Transbig consortium. Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer. *J Natl Cancer Inst* 2006 ; 98 : 1183-92.

GLOSSAIRE DES BIOPUCES

Signature : information obtenue à partir d'un nombre limité de gènes permettant de différencier plusieurs groupes de malades ou de tissus. Rechercher la signature prédictive du risque de métastase à distance dans les 5 ans après le diagnostic sous-entend l'existence d'une empreinte génétique unique pour ce risque, il s'agit d'une hypothèse très forte.

Marqueur pronostique : marqueur biologique associé à un événement spécifique, par exemple un gène qui est surexprimé dans les tumeurs des patients qui développeront une métastase par comparaison avec les tumeurs des patients qui resteront sans métastases. La mesure de l'expression de ce gène permet de prédire le risque de métastases.

Marqueur prédictif : terme employé tantôt pour désigner un marqueur pronostic, tantôt pour désigner un marqueur prédictif de l'utilité d'un traitement donné. Dans ce second cas, le bénéfice du traitement est par exemple plus grand chez les patients ayant des valeurs positives pour le marqueur. Pour établir le caractère prédictif d'un marqueur, il faut qu'une différence d'effet soit mise en évidence, de préférence dans le cadre d'un essai clinique randomisé. Si l'on veut sélectionner un traitement pour un groupe de patients sur la base de marqueurs génétiques, il doit avoir été démontré préalablement que ceux-ci sont prédictifs de l'effet du traitement.

TROIS CARACTÉRISTIQUES SOUHAITABLES D'UN FACTEUR PRONOSTIQUE OU PRÉDICTIF

- Une méthode de mesure reproductible, précise et applicable à grande échelle.
- Une fréquence de marqueur positif (ou négatif) intermédiaire, c'est-à-dire ne pas avoir un découpage de la population en 1 % *versus* 99 % mais plutôt en 25 % *versus* 75 %.
- Une faible corrélation avec des facteurs connus, c'est-à-dire apporter une nouvelle information.

TIRÉS À PART
S. Michiels