Le problème de stratification de populations dans les études cas-témoins : une approche théorique

Marie-Hélène CAZES, Myriam KHLAT
Institut national d'études démographiques (INED)
Emmanuelle GÉNIN, Marguerite GUIGUET
Institut national de la santé et de la recherche médicale (INSERM)

Introduction : position du problème

Parmi les interactions possibles entre phénomènes liés à la santé et processus démographique, il en est une que l'on peut rencontrer en épidémiologie génétique dans le cadre des études d'association entre un gène – ou un marqueur génétique – et une maladie.

Une méthode classique pour aborder ce type d'étude est la méthode cas-témoin qui consiste à comparer un échantillon de malades atteints de la maladie étudiée (pour laquelle on s'interroge et on recherche une éventuelle cause génétique) et un échantillon de personnes non malades comparables, ayant globalement les mêmes caractéristiques (d'âge, d'origine etc.).

Si l'on décèle qu'un marqueur génétique spécifique est beaucoup plus représenté (voire systématiquement présent) chez les individus malades et peu représenté (voire absent) chez les individus sains, la mesure statistique peut conclure à une association entre ce marqueur génétique et la maladie.

Ce type d'enquête et de méthodologie suppose implicitement que cas et témoins soient issus d'une population homogène. Cette méthode a soulevé une polémique par rapport au risque de biais qui lui était lié. En effet, si jamais il existait un effet de « structure » non connu de la population (ou effet de « stratification » selon le vocabulaire des épidémiologistes), la population supposée homogène d'où sont tirés les échantillons étant, en fait, un mélange de deux ou plusieurs populations aux caractéristiques très différentes, les résultats des tests statistiques d'association pourraient être biaisés et conclure à tort à une association, le facteur ethnique ou « sous-population » jouant dans ce cas comme un facteur de confusion dans l'étude de l'association gène-maladie.

Ce problème est important en pratique dans des pays où les populations sont composées de nombreuses communautés différentes, restées relativement fermées, et où les échantillons qui en sont issus sont susceptibles d'être très hétérogènes.

Dans le cadre de ces débats, de nombreux travaux ont été entrepris pour évaluer de façon théorique l'ampleur de ce biais en fonction de différents paramètres de population. Nous avons mené une étude de ce type pour apprécier dans quelles conditions concrètes d'observation ce biais risquait d'être observé. Ce travail a fait l'objet d'une publication dans *Cancer, Epidemiology, Biomarkers and Prevention* (1). Nous en reprenons ici les principaux résultats.

1. Analyse des facteurs impliqués dans le biais de stratification

En 2000, Wacholder et *al*. (2) avaient quantifié l'étendue du biais attaché à la stratification, en fonction d'un certain nombre de paramètres de population, en se basant sur un indice de rapport de risques, le CRR, indépendant de la taille de l'échantillon[1]. Il concluait que le biais restait faible dans la plupart des cas.

[1] Le CRR « confounding risk ratio » est le rapport du risque relatif brut de l'effet du génotype sur la maladie au risque relatif ajusté sur les groupes ethniques. Ce rapport reflète donc l'intensité du biais associé à la stratification.

Cependant, dans la mesure où c'est la *fiabilité* des associations positives qui compte dans ce type d'étude, une mesure plus sensée de l'impact du biais nous a semblé être l'erreur de type 1 (c'est-à-dire, dans la théorie classique des tests statistiques, la probabilité de conclure à tort à une association, prise classiquement à 5%). Pour une intensité donnée du biais, cette erreur de type 1 (ET1) dépend de la fréquence du génotype d'intérêt dans les sous-populations, mais aussi et surtout de la taille de l'échantillon. La question qu'on se pose est la suivante : même dans le cas d'un faible biais, l'erreur de type 1 ne pourrait-elle pas être assez importante ?

Pour répondre à cette question, des simulations ont été mises au point permettant de faire varier un grand nombre de paramètres de population. Elles ont été comparées aux résultats obtenus à partir de calculs théoriques, menés dans le cadre d'hypothèses asymptotiques, avec pour objectif de quantifier simultanément le biais (en termes de CRR) et l'ET1 en considérant différents scénarios de population. Nous avons volontairement choisi la situation la plus défavorable – celle où le risque de biais est le plus important – en nous limitant à la présence d'une seule sous-population cachée (baptisée « petite » sous-population) dans la population étudiée[2]. Nous avons pris en compte les variables suivantes : fréquence allélique du marqueur d'intérêt dans chacune des deux sous-populations, rapport de prévalence de la maladie entre la population cachée et l'autre population majoritaire, proportion relative de la population cachée dans la population globale et avons fait varier, de plus, les tailles d'échantillons.

Le modèle et la méthode utilisés, avec le détail des calculs, sont présentés en annexe : on se place sous l'hypothèse que le marqueur génétique n'a aucune influence sur le risque de maladie mais avec *n* cas et *n* témoins échantillonnés à partir d'une population divisée en deux sous-groupes. On a supposé que la « petite » sous-population cachée était celle qui présentait le plus fort risque de maladie.

Les scénarios considérés sont les suivants :

– rapport de prévalence de la maladie entre les deux sous-populations **de 2 ou de 10** (risque de maladie tantôt doublé dans la « petite » sous-population cachée, tantôt multiplié par 10) ;
– allèle marqueur de type **rare** (fréquence de l'allèle fixée à **0,1** dans la « grande » sous-population) ou **commun** (fréquence fixée à **0,5**),
– proportions successives de la « petite » sous-population cachée, ***f* = 1%, 5%, 10% ou 50%** (dans ce dernier cas, la population est composée de deux sous-populations équivalentes).

Pour chaque scénario, on a estimé à la fois le CRR et l'ET1. Pour calculer l'erreur de type 1, on a adopté des échantillons de 500 cas et 500 témoins. Mais on a aussi étudié l'impact de la taille de l'échantillon en la faisant varier de 100 à 1000.

2. Résultats

En partant d'un rapport de prévalence de la maladie de 2 et d'un échantillon de 500 cas et 500 témoins (Figure 1), on s'est attaché à rechercher l'influence sur le CRR et l'ET1 de :

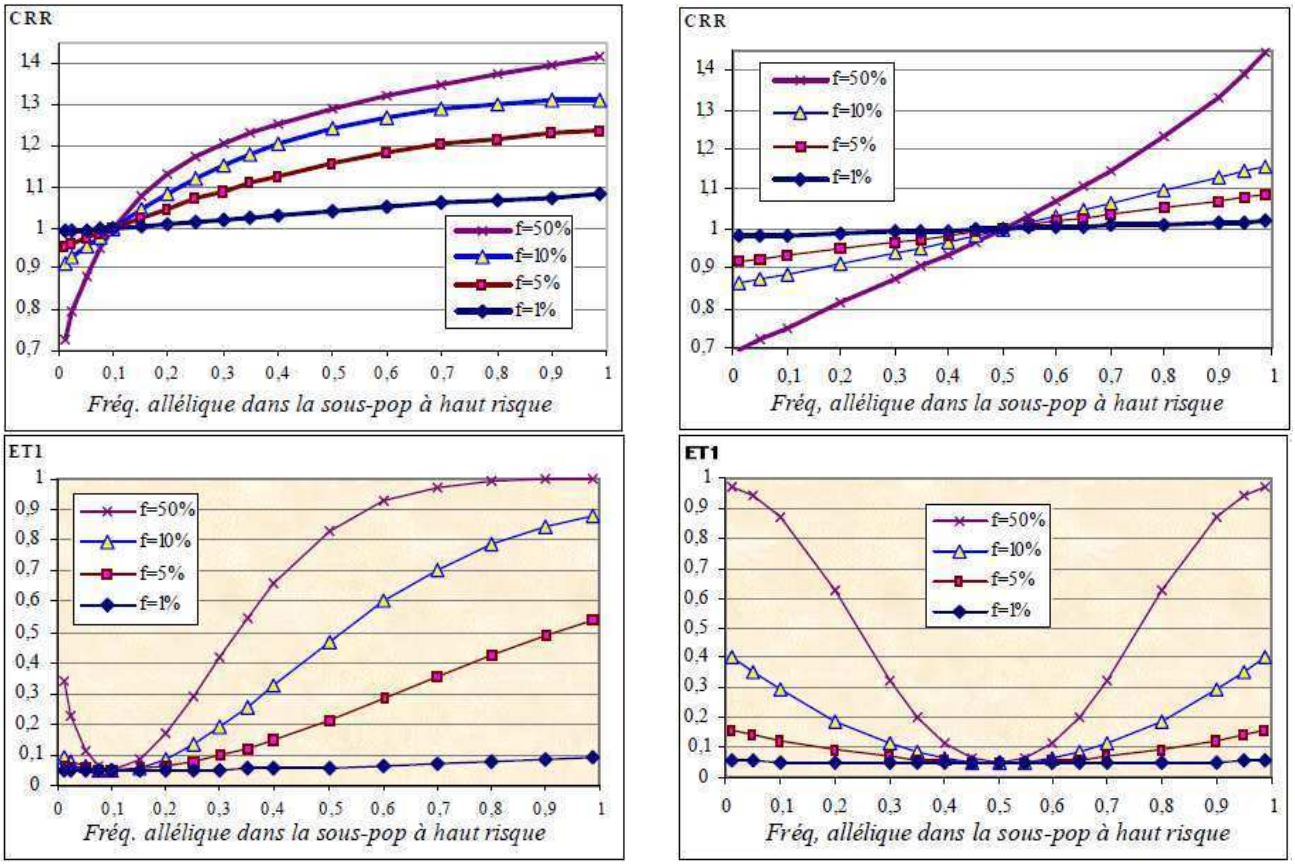
– la différence de fréquence allélique entre les deux sous-populations (la « petite » et la « grande »).
– l'importance relative de la « petite » sous-population à haut risque.
– le type de polymorphisme : allèle *rare* ou *commun*.

[2] Wacholder et *al* (2) ont montré en effet que la diversité ethnique réduisait le biais de stratification de population. Le biais potentiel est le plus fort avec 2 ou 3 ethnies et il tend à diminuer à mesure que le nombre d'ethnies augmente, chacun des biais s'annulant les uns les autres.

FIGURE 1 : BIAIS ET ERREUR DE TYPE 1 DANS LES ÉTUDES CAS-TÉMOINS EN PRÉSENCE DE STRATIFICATION DE POPULATION ;
RAPPORT DE PRÉVALENCE DE LA MALADIE ENTRE LA "PETITE" SOUS-POPULATION À HAUT RISQUE ET LA "GRANDE" SOUS-POPULATION À FAIBLE RISQUE = **2** ;
500 CAS – 500 TÉMOINS

Allèle rare
Fréquence de l'allèle dans la « grande » sous-population à faible risque :
$p_2 = 0{,}1$

Allèle commun
Fréquence de l'allèle dans la « grande » sous-population à faible risque :
$p_2 = 0{,}5$



2.1. Analyse du biais

Quand l'allèle étudié est moins fréquent dans la « petite » sous-population ayant la prévalence la plus forte de maladie que dans l'autre (partie gauche des graphes), le CRR a des valeurs inférieures à 1 comme si le génotype d'intérêt jouait un rôle protecteur par rapport à la maladie (association faussement négative).

Inversement, si l'allèle est plus fréquent dans la « petite » sous-population à haut risque que dans l'autre, le CRR est supérieur à 1, reflétant ainsi une fausse association positive.

Le schéma de variation du CRR dans chacune des 4 courbes indique que :

– plus la fréquence allélique dans la « petite » sous-population à haut risque diffère de la fréquence dans l'autre sous-population, plus la valeur du CRR s'éloigne de 1 (*i.e.* plus le biais augmente),

– plus la taille relative *f* de la « petite » sous-population cachée à haut risque augmente, plus le biais augmente (biais maximum quand les deux sous-populations sont en proportions égales, $f = 50\%$).

On constate, de plus, que la situation avec un allèle rare (graphe de gauche) est plus exposée au biais, (hormis les cas extrêmes) : ainsi, pour une proportion *f* de la « petite »

population de 10% et une différence de fréquence allélique de 0,5, le CRR monte à 1,27. Dans le cas d'un allèle commun (graphe de droite), la valeur équivalente atteinte par le CRR est de 1,16.

En considérant maintenant des scénarios beaucoup plus réalistes dans lesquels la « petite » sous-population ne représente qu'une part de la population étudiée (1%, 5%, ou 10%) et la différence de fréquence allélique reste inférieure ou égale à 0,2 (voir tableau 1 pour les chiffres détaillés), on constate que, dans ces cas, le biais est contenu dans des limites beaucoup plus raisonnables (1,15 au plus pour des allèles rares ; 1,07 pour des allèles communs).

TABLEAU 1 : ERREUR DE TYPE 1 ET CRR DANS DES SCÉNARIOS PLUS COURAMMENT RENCONTRÉS
RAPPORT DU TAUX DE PRÉVALENCE = 2 ; 500 CAS - 500 TÉMOINS

Tableau 1a (Allèle rare)

$\Delta p=p_1-p_2$	f = 1%		f = 5%		f = 10%	
	ET1	CRR	ET1	CRR	ET1	CRR
-0,090	0,05	0,99	0,06	0,95	0,09	0,91
-0,075	0,05	0,99	0,06	0,96	0,08	0,93
-0,050	0,05	0,99	0,05	0,97	0,06	0,95
-0,025	0,05	1,00	0,05	0,99	0,05	0,98
0,000	0,05	1,00	0,05	1,00	0,05	1,00
0,050	0,05	1,01	0,05	1,02	0,06	1,04
0,100	0,05	1,01	0,06	1,05	0,09	1,08
0,150	0,05	1,02	0,08	1,07	0,13	1,12
0,200	0,05	1,02	0,10	1,09	0,19	1,15

f = proportion de la « petite » sous-population à haut risque
p_2 = fréquence allélique dans la « grande » sous-population fixée à 0,1

Tableau 1b (Allèle commun)

$\Delta p=p_1-p_2$	f = 1%		f = 5%		f = 10%	
	ET1	CRR	ET1	CRR	ET1	CRR
0,20	0,05	0,99	0,07	0,96	0,11	0,94
-0,15	0,05	0,99	0,06	0,97	0,08	0,95
-0,10	0,05	1,00	0,05	0,98	0,07	0,97
-0,05	0,05	1,00	0,05	0,99	0,05	0,98
0,00	0,05	1,00	0,05	1,00	0,05	1,00
0,05	0,05	1,00	0,05	1,01	0,05	1,02
0,10	0,05	1,00	0,05	1,02	0,07	1,03
0,15	0,05	1,01	0,06	1,03	0,08	1,05
0,20	0,05	1,01	0,07	1,04	0,11	1,07

f = proportion de la « petite » sous-population à haut risque
p_2 = fréquence allélique dans la « grande » sous-population fixée à 0,5

2.2. Analyse de l'erreur de type 1

Si on regarde la relation entre le CRR et l'ET1 (Figure 1), on trouve, comme on pouvait s'y attendre, que plus le CRR est élevé, plus l'ET1 est grande et dans le cas des scénarios extrêmes fournissant les valeurs maximales du CRR, l'erreur de type 1 est de 100% (*i.e.* on a alors la certitude de conclure faussement à une association entre l'allèle d'étude et la maladie).

Mais là encore, si on se limite aux scénarios plus réalistes définis plus haut, les ET1 restent dans une limite acceptable dans le cas d'échantillons de 500 cas et 500 témoins : par exemple, dans le cas d'allèles rares (tableau 1a), nous obtenons au pire une ET1 de 19% (avec un CRR de 1,15) chaque fois que la différence de fréquence allélique est de 0,2 et ceci dans l'hypothèse d'un rapport de prévalence de maladie doublé. Dans le cas d'allèles communs (tableau 1b), cette ET1 plafonne à 11% (CRR de 1,07).

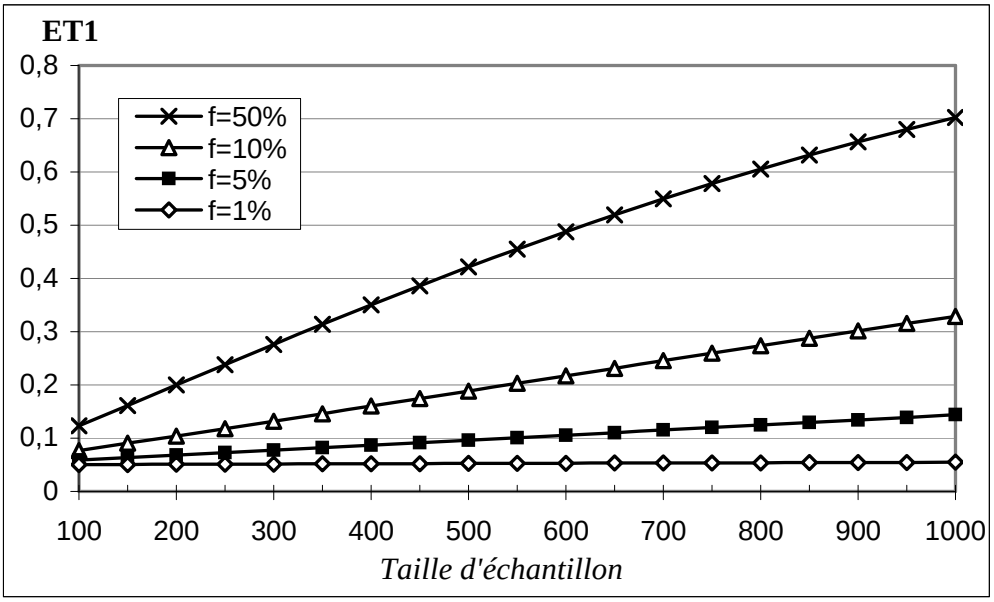
On retrouve les mêmes schémas de variation quand *la prévalence de la maladie est 10 fois plus élevée* dans la « petite » sous-population à haut risque que dans l'autre (Figure non montrée) mais dans ce cas, l'intensité du biais, comme l'ET1, sont tous deux considérablement augmentés. Dans le scénario le plus extrême (sous-populations égales (f=50%) et différence de fréquence allélique maximale), le biais est de 4 pour des allèles communs et de 3,2 pour des allèles rares, tandis que l'ET1 est de 100% pour des échantillons de 500 cas - 500 témoins.

En terme d'erreur de type 1, l'avantage des allèles communs sur les allèles rares est beaucoup moins évident dans ces conditions extrêmes sauf dans le cas d'une sous-population à haut risque proportionnellement très petite (f=1%). Pour f=5% ou 10%, on est presque certain de trouver des fausses associations pour des allèles rares, en partant d'une différence de fréquence allélique de 0,2.

2.3. Influence de la taille des échantillons

Nous avons étudié l'influence de la taille de l'échantillon sur l'ET1 en nous basant sur un scénario avec un rapport des taux de prévalence de la maladie de 2, et une fréquence allélique de 0,3 dans la « petite » sous-population à haut risque contre 0,1 dans l'autre, correspondant à des CRR de 1,02, 1,09, 1,15 ou 1,20 pour les valeurs de f de 1%, 5%, 10% ou 50% respectivement (Figure 3).

FIGURE 3 : INFLUENCE DE LA TAILLE D'ÉCHANTILLON (N CAS – N TÉMOINS) SUR L'ERREUR DE TYPE 1 (ET1) DANS LES ÉTUDES CAS – TÉMOINS EN PRÉSENCE DE STRATIFICATION DE POPULATION, RAPPORT DE PRÉVALENCE DE LA MALADIE ENTRE LA « PETITE » ET LA « GRANDE » SOUS-POPULATION = 2
FRÉQUENCE ALLÉLIQUE DE 0,1 DANS LA « GRANDE » SOUS-POPULATION À FAIBLE RISQUE ET DE 0,3 DANS LA « PETITE » SOUS-POPULATION À HAUT RISQUE



Quand la « petite » sous-population cachée à haut risque ne représente qu'une très petite portion de la population totale (1%), la probabilité de fausse association reste clairement proche des 5% attendus, même pour des échantillons allant jusqu'à 1000 cas – 1000 témoins.

Quand la proportion de la « petite » sous-population à haut risque est plus grande, l'ET1 croît avec la taille d'échantillon et plus la proportion est élevée, plus l'augmentation est rapide.

Pour une « petite » sous-population représentant 10% du total, l'ET1 atteint 10% pour des échantillons de taille 200, alors qu'il faut des échantillons de taille 550 (550cas-550 témoins) pour atteindre le même niveau d'erreur quand la proportion est de 5%.

Ainsi contrairement à ce qu'on pourrait penser un peu rapidement *a priori*, augmenter la taille de l'échantillon augmente du même coup le biais et la probabilité de conclure à tort à une association.

Conclusion

Il est important de comprendre le rôle des facteurs qui sous-tendent la stratification de population pour interpréter correctement les résultats d'études cas-témoins. Cela nécessite une connaissance approfondie des caractéristiques de la structure de la population analysée.

Nos résultats permettent de clarifier l'impact séparé de chacune des différentes composantes de structure de population et de taille d'échantillon. Ainsi :

– plus grande est la différence entre la population étudiée et la sous-population cachée, en termes soit de fréquence allélique, soit de prévalence de maladie, et plus grande est l'ET1.
– plus la taille relative de la sous-population cachée est grande relativement à la population étudiée, plus grande est l'ET1.
– l'ET1 est plus élevée pour des allèles rares que pour des allèles communs[3].

À partir de ces résultats, on peut cibler dans quel scénario la stratification de population peut risquer d'être vraiment préoccupante et dans quel cas, elle ne le sera pas ; en résumé, la situation la moins favorable correspond à celle qui met en jeu un allèle rare, un fort rapport des prévalences de la maladie, une grande différence dans les fréquences alléliques, et une sous-population relativement importante. La situation la plus favorable concerne un polymorphisme courant, un faible rapport des taux de prévalence de la maladie, une petite différence de fréquences alléliques, et une sous-population relativement petite.

Alors que l'impact potentiel du rapport de prévalence de la maladie et celui de la différence des fréquences alléliques semblaient intuitivement simples, celui trouvé concernant le type de polymorphisme (rare contre commun) et la proportion relative de la sous-population cachée étaient plus difficiles à anticiper.

En ce qui concerne la taille d'échantillon, il est intéressant de constater que plus cette taille augmente, plus le biais se répand plutôt que de diminuer et ceci est tout à fait compréhensible puisque les grands échantillons sont associés à des puissances statistiques plus grandes pour détecter des associations, que celles-ci soient réelles ou artificielles ((Pritchard et Donnelly (3)).

Dans les populations naturelles, de grandes différences de fréquence alléliques entre sous-populations sont susceptibles d'arriver surtout entre groupes ethniques identifiables et l'on peut donc aisément les traiter par des méthodes d'appariement, d'ajustement ou autres méthodes standard. Garte et *al* (4,5) rapportent par exemple une différence de fréquence allélique significative entre Caucasiens, Asiatiques, Africains et Afro-américains, mais une différence

[3] Par exemple, un allèle relativement commun comme *GSTM1*0* (fréquence autour de 50%) sera beaucoup moins exposé au biais de stratification qu'un allèle rare comme *CYP1A1*2A* (fréquence autour de 5 à 6%).

seulement de 0,04 entre des populations d'origine différentes en Europe pour des gènes métaboliques communément étudiés.

En appariant donc soigneusement des échantillons cas-témoins de taille modérée dans les populations cosmopolites des États-Unis et d'Europe, on a peu de chances d'obtenir des niveaux de structure qui entraînent de fortes associations positives artificielles.

En conclusion, cette étude éclaire les schémas de variation de l'erreur de type 1 en fonction des différentes composantes de structuration de la population et montre que le biais, comme l'erreur de type 1, résultant de la stratification de populations est vraisemblablement limité dans les études cas-témoins de taille modérées, méthodologiquement bien menées, sauf dans des cas de scénarios assez irréalistes.

Si l'on respecte le principe d'ajuster l'analyse sur l'origine géographique ou ethnique des individus, le risque de biais lié à l'utilisation de la méthode cas-témoins est quasiment insignifiant.

Toutefois, chaque fois qu'une association statistique est susceptible de résulter d'une stratification, on devrait utiliser d'autres approches pour confirmer l'association.

BIBLIOGRAPHIE

(1) KHLAT M., CAZES M.H., GÉNIN E., GUIGUET M., « Robutness of Case-Control studies of genetic factors to population stratification : magnitude of biais and type 1 error », *Cancer Epidemiol Biomarkers Prev.*, 13 (10) : 1660-1164, 2004.

(2) WACHOLDER S., ROTHMAN N., CAPORASO N., « Population stratification in epidemiologic studies of common genetic variants and cancer : quantification of bias », *J. Natl. Cancer Inst.*, 92 : 1151-8, 2000.

(3) PRITCHARD J.K., DONNELLY P., « Case-control studies of association in structured or admixed populations ». *Theor. Popul. Biol.*, 60 : 227-237, 2001.

(4) GARTE S., GASPARI L., ALEXANDRIE A.K., et *al.*, « Metabolic gene polymorphism frequencies in control populations ». *Cancer Epidemiol Biomarkers Prev*, 10 : 1239-48, 2001.

(5) GARTE S., « The role of ethnicity in cancer susceptibility gene polymorphisms : the example of CYP1A1 », *Carcinogenesis*, 19 : 1329-32, 1998.

Annexe

Méthode

On considère une étude cas-témoins comprenant *Nc* cas et *Nt* témoins, échantillonnés à partir d'une population qui est divisée en 2 sous-groupes Pop1 et Pop2.

Soit f la proportion de Pop1. On suppose que Pop1 est la sous-population ayant le risque de maladie le plus élevé (appelée « petite » population). On considère un gène à 2 allèles (A, a) et *p1* et *p2* les fréquences respectives de l'allèle variant dans les 2 sous-populations.

On veut déduire la distribution génotypique attendue pour le gène marqueur chez les cas et les témoins sous l'hypothèse nulle H_0, c'est-à-dire sous l'hypothèse que le marqueur n'a aucune influence sur le risque de maladie. Comme la maladie est suffisamment rare, on peut faire l'hypothèse que la distribution génotypique chez les témoins est quasiment la même que dans la population toute entière. On peut écrire :

$$P_{2t} = P(AA/témoins) = f \quad p_1^2 + (1-f) \quad p_2^2$$
$$P_{1t} = P(Aa/témoins) = 2f \quad p_1(1-p_1) + 2 \quad (1-f) \quad p_2 \quad (1-p_2)$$
$$P_{0t} = P(aa/témoins) = f \quad (1-p_1)^2 + (1-f) \quad (1-p_2)^2$$

Soit K le rapport de la prévalence de la maladie dans la Pop1 à la prévalence de la maladie dans la Pop2. La distribution génotypique attendue chez les cas peut s'écrire :

$$P_{2c} = P(AA/cas) = \frac{P(AA/Pop1) \quad P(cas/Pop1) \quad P(Pop1) + P(AA/Pop2) \quad P(cas/Pop2) \quad P(Pop2)}{P(cas/Pop1) \quad P(Pop1) + P(cas/Pop2) \quad P(Pop2)}$$

$$= \frac{f \quad p_1^2 \quad K + (1-f) \quad p_2^2}{f \quad K + (1-f)}$$

$$P_{1c} = P(Aa/cas) = \frac{2f \quad p_1 \quad (1-p_1)K + 2 \quad (1-f) \quad p_2 \quad (1-p_2)}{f \quad K + (1-f)}$$

$$P_{0c} = p(aa/cas) = \frac{f \quad (1-p_1)^2 K + (1-f) \quad (1-p_2)^2}{f \quad K + (1-f)}$$

Pour tester l'effet du génotype sur le risque de maladie, on a fait une régression logistique dans laquelle on incluait le génotype comme une variable quantitative codée 0 (aa), 1 (Aa), ou 2(AA) sous l'hypothèse d'un modèle additif d'hérédité.

Dans l'hypothèse d'indépendance entre le génotype et la maladie, le CRR se réduit au risque relatif brut de l'effet du génotype sur la maladie[4].

L'erreur de type 1 du test du rapport de vraisemblance de l'association entre le marqueur et la maladie a été évalué en utilisant une distribution du Chi2 non centrée à 1 ddl et un paramètre λ de non centralité qui dépend des différentes valeurs des variables :

$$\lambda = \sum_{i=1}^{3} \quad \frac{N_c N_t \quad (P_{ic} - P_{it})^2}{N_c P_{ic} + N_t P_{it}}$$

Ce procédé nous fournit une approximation asymptotique de l'erreur de type 1 qui s'est révélée très proche de l'erreur de type 1 observée au travers de simulations (non montrées).

---

[4] Dans notre modèle, les calculs étant faits sous l'hypothèse que le marqueur génétique est indépendant de la maladie, un CRR égal à 1 signifie qu'il n'y a pas de stratification et correspond à un biais nul.