REVUE DE STATISTIQUE APPLIQUÉE

C. ABRAHAM
J.-P. DAURES
I. MOMAS

Utilisation de la proportion de cas attribuable : intérêts, limites et applications à une étude sur le cancer de la vessie dans l'Hérault

Revue de statistique appliquée, tome 45, n° 2 (1997), p. 5-20
<http://www.numdam.org/item?id=RSA_1997__45_2_5_0>

Rev. Statistique Appliquée, 1997, **XLV** (2), 5-20

UTILISATION DE LA PROPORTION DE CAS ATTRIBUABLE : INTÉRÊTS, LIMITES ET APPLICATIONS À UNE ÉTUDE SUR LE CANCER DE LA VESSIE DANS L'HÉRAULT

C. Abraham *,**, J.-P. Daures*, I. Momas***

* *Université Montpellier 1, Institut Universitaire de Recherche Clinique, 75, rue de la Cardonille, 34000 Montpellier*

** *Unité de Biométrie, ENSAM INRA, 2, Place P. Viala, 34060 Montpellier*

*** *Laboratoire d'hygiène, Faculté de Pharmacie, 4, Bd de l'observatoire 75006 Paris*

RÉSUMÉ

La proportion de cas attribuable (PCA) constitue une mesure importante en épidémiologie. Elle permet de quantifier la proportion de cas d'une maladie attribuable à l'exposition à un facteur de risque. Nous proposons une revue des différentes approches de la PCA ajustée sur des facteurs étiologiques en faisant le lien entre ces approches ainsi qu'une méthode pratique pour construire la PCA à plusieurs facteurs d'expositions. Nous donnons ensuite les résultats d'une application à une étude sur le cancer de la vessie dans l'Hérault et nous discutons de l'utilisation de la PCA (ajustée ou non) pour prédire les conséquences d'une suppression réelle de l'exposition à un facteur de risque dans la population totale.

Mots-clés : *Ajustement, Cancer de la vessie, Epidémiologie, Modèle logistique, Prévention, Proportion de cas attribuable.*

ABSTRACT

The attributable risk (AR) constitutes an important epidemiologic risk measure. It quantifies the proportion of cases of disease due to exposure factor. We propose a review of the different approaches of adjustment for aetiologic factors by connecting these approaches and a practical method for constructing a AR for several exposures. We give the results of an application to a study on the bladder cancer in Hérault and discuss the use of AR for forecasting the consequences of a real suppression of exposure to a risk factor.

Keywords : *Adjustment, Attributable Risk, Bladder cancer, Epidemiology, Logistic model, Prevention.*

1. Introduction

La connaissance de l'impact d'un facteur de risque sur la proportion des cas d'une maladie est de première importance en santé publique, en particulier en matière de prévention. Depuis son introduction sous le nom de risque attribuable par Levin (1953), ce concept a donné lieu à de nombreux travaux et plusieurs appellations (comme le risque attribuable en proportion, la fraction attribuable, la fraction étiologique, la proportion de cas attribuable) et interprétations ont rendu difficile son utilisation (voir Coste et Spira, 1991, pour une revue des appellations et Gefeller, 1992, pour une bibliographie détaillée et commentée). Nous adoptons l'apellation «proportion de cas attribuable», notée PCA, définie par

$$\mathrm{PCA} = 1 - \frac{P(M|E_0)}{P(M)} \tag{1}$$

où $P(M)$ est la probabilité de développer la maladie pendant une période donnée dans la population totale et $P(M|E_0)$ la probabilité de développer la maladie pendant la même période dans la population des individus non exposés au facteur de risque E considéré (on note E_0 et E_1 pour signifier respectivement la non exposition et l'exposition au facteur E). On obtient facilement des estimateurs de la PCA et de sa variance pour différents types d'études (Walter, 1978, Benichou, 1991).

Whittemore (1982) interprète le rapport $P(M|E_0)/P(M)$ comme étant la part du nombre de malades qui persisterait si la population totale se comportait comme la population des non exposés et par conséquent, la PCA représente la part du nombre de malades que l' on pourrait éviter si la population totale se comportait comme la population des non exposés. Cette interprétation a l'avantage d'être plus concrète que celles qui s'expriment en termes de part du nombre de malades «due» ou «attribuable» ou «liée» au facteur de risque et d'être plus prudente que l'interprétation «part du nombre de malades que l'on pourrait éviter si les effets associés au facteur de risque étaient absents». Nous reviendrons sur cette interprétation à la fin de cet article. Les interprétations des différentes définitions rencontrées dans la littérature sont analysées dans Greenland et Robins (1988) en distinguant parmi les sujets malades exposés ceux qui auraient contracté la maladie, avant une date t, même en l'absence d'exposition de ceux, appelés cas en excés, qui n'auraient pas contracté la maladie, avant la date t, en l'absence de l'exposition.

En plus du facteur de risque E considéré, plusieurs auteurs (entre autres Walter, 1980, Whittemore, 1982, Bruzzi *et al.*, 1985) s'accordent pour dire qu'il est nécessaire de prendre en compte d'autres facteurs d'exposition puisque la PCA n'est pas la même suivant que l'on se place dans les sous-populations déterminées par ces facteurs ou dans la population totale. Par exemple, si l'exposition est la consommation de tabac, la PCA au tabac parmi les consommateurs d'alcool peut être différente de la PCA au tabac dans la population totale. Ainsi, plusieurs approches de la PCA ajustée sur ces facteurs ont été proposées et des estimateurs ont été développés ainsi que leur comportement asymptotique.

Nous nous intéressons dans cet article à l'utilisation de la PCA en santé publique. Aussi, il nous est apparu utile, en premier lieu, de donner une présentation reliant entre elles les différentes approches des PCA ajustées. Nous proposons, ensuite, une

méthode simple pour construire la PCA pour plusieurs facteurs de risque dans le but, d'une part, d'isoler un petit nombre de facteurs de risque dont l'exposition pourrait être diminuée par une campagne de prévention et, d'autre part, de savoir si tous les principaux facteurs de risque sont connus. Cette méthode est appliquée, dans une troisième partie, à une étude sur le cancer de la vessie dans l'Hérault et nous donnons plusieurs estimations de la PCA ajustée sur différents facteurs. Enfin, dans une dernière partie, nous discutons les résultats de cette étude (ils mettent en relief la dépendance de la PCA aux choix des niveaux de base, le niveau de base pour un facteur d'exposition étant le seuil en dessous duquel on considère qu'il n'y a pas d'exposition) et nous nous interrogeons sur la valeur prédictive de la PCA dans l'hypothèse d'une suppression réelle des facteurs de risque dans la population totale.

2. Les différentes approches des PCA ajustées

Nous rappellons la construction de la PCA ajustée de Whittemore (1982) à partir de laquelle nous présentons les autres approches. Nous insistons sur les estimateurs valides dans le cadre d'une étude cas-témoins (dans laquelle deux échantillons sont tirés séparément, un dans la population des malades, les cas, l'autre dans la population des non malades, les témoins) car ces études, moins couteuses que les études de cohorte (dans lesquelles on s'intéresse à l'apparition de la maladie dans un unique échantillon que l'on suit pendant plusieurs années) sont les plus fréquentes dans la pratique. La particularité d'une étude cas-témoins provient qu'elle ne fournit pas d'information sur la distribution de la maladie. Nous supposerons, par la suite, que toutes les variables considérées ne prennent qu'un nombre fini de valeurs bien qu'il soit possible d'étendre certains résultats à des variables continues (Benichou et Gail, 1990).

Si le facteur d'exposition E considéré est corrélé avec un autre facteur étiologique C, l'élimination du facteur d'exposition dans la population totale peut ne pas conduire à la probabilité de la maladie $P(M|E_0)$ (puisque la répartition du facteur C dans la sous-population des non exposés n'est pas la même que dans la population totale). Si le facteur C est décrit par les classes $C_1, C_2, ..., C_K$, Whittemore (1982) propose pour la probabilité de la maladie en l'absence de l'exposition, la quantité $\sum_k P(C_k)P(M|E_0, C_k)$ où $P(C_k)$ est la probabilité de la classe C_k et $P(M|E_0, C_k)$ la probabilité de développer la maladie (pendant une période donnée) dans la sous-population des sujets non exposés de la classe C_k. On peut donner une interprétation simple de cette quantité en la multipliant par le nombre N d'individus de la population totale. $N \times P(C_k)$ représente le nombre attendu d'individus de la classe C_k et $N \times P(C_k) \times P(M|E_0, C_k)$ le nombre attendu de malades de la classe C_k si tous les individus de cette classe se comportaient comme les non exposés de cette classe. Ainsi, la modélisation de l'absence d'exposition respecte la distribution du facteur C dans la sous-population des exposés. On obtient alors la PCA ajustée, notée PCAA,

$$PCAA = 1 - \frac{\sum_k P(C_k)P(M|E_0, C_k)}{P(M)} \qquad (2)$$

(cette définition correspond à celle de Walter, 1980).

Les classes (ou strates) C_1,C_2,...,C_K peuvent aussi provenir des croisements de plusieurs facteurs. Par exemple, si on considère les facteurs dichotomiques fumeur-non fumeur, noté F_1-F_0, et consommateur d'alcool-non consommateur d'alcool, noté A_1-A_0, le facteur C est constitué des 4 strates (F_0, A_0), (F_0, A_1), (F_1, A_0) et (F_1, A_1). La formule (2) correspond alors à la PCAA sur plusieurs facteurs. Elle est égale à formule (1) de la PCA sans ajustement lorsque les facteurs E et C sont indépendants.

• La formule (2) peut aussi s'écrire (appendice 1) :

$$PCAA = \sum_k P(C_k|M)PCA_k \tag{3}$$

où $PCA_k = 1 - P(M|E_0, C_k)/P(M|C_k)$ est la PCA restreinte à la strate C_k et $P(C_k|M)$ la probabilité ayant contracté la maladie d'être dans la classe C_k. Ainsi, la PCAA s'exprime comme une somme pondérée des PCA_k. Si la maladie est suffisamment rare de sorte que $P(E_0|C_k) \simeq P(E_0|C_k, \bar{M})$, $P(E_0|C_k, \bar{M})$ étant la probabilité d'être non exposé parmi les individus non malades de la strate C_k, la PCAA peut être estimée dans une étude cas-témoins en remarquant que

$$PCA_k = 1 - P(E_0|C_k, M)/P(E_0|C_k) \simeq 1 - P(E_0|C_k, M)/P(E_0|C_k, \bar{M}).$$

Whittemore (1982) donne des estimateurs de la PCAA et leur variance asymptotique pour un échantillonnage simple et un échantillonnage apparié pour une étude cas-témoins (voir le paragraphe 4.1 pour la description d'un échantillonnage simple dans un étude cas-témoins).

• Bruzzi *et al.* (1985) exprime la PCAA en fonction des risques relatifs $RR_k = P(M|E_1, C_k)/P(M|E_0, C_k)$ restreints aux strates C_k :

$$PCAA = 1 - \sum_k \{P(E_0, C_k|M) + P(E_1, C_k|M)RR_k^{-1}\} \tag{4}$$

(on montre, en appendice 1, l'équivalence des formules (2) et (4)) et introduit le modèle logistique pour décrire la probabilité de la maladie conditionnellement aux facteurs E et C. Plus précisément, supposons que C soit constitué par le croisement des facteurs F^1,F^2,...,F^I ($C = (F^1, F^2, ..., F^I)$), notons G^1,G^2,...,G^J des éventuels produits des variables E,F^1,F^2,...,F^I et $x_1, x_2, ..., x_L$ les valeurs de la variable $X = (1, E, F^1, F^2, ..., F^I, G^1, G^2, ..., G^J)$; ainsi, il existe une bijection naturelle de l'ensemble des strates $\{(E_m, C_k), m \in \{0, 1\},\ k \in \{1, 2, ..., K\}\}$ sur l'ensemble $\{x_l,\ l \in \{1, 2, ..., L\}\}$. $P(M|x_l)$ est alors modélisée par $P(M|x_l) = exp\ {}^t\theta x_l/\{1 + exp\ {}^t\theta x_l\}$ où θ est un vecteur de $\mathbf{R}^{2+I+J}$. De plus, si z_l représente le vecteur identique à x_l sauf pour la variable d'exposition E qui est maintenue à son niveau le plus faible E_0, le quotient $P(M|x_l)/P(M|z_l)$, noté $\widetilde{RR}_l$, est égal soit à un risque relatif (lorsque la valeur de l'exposition de x_l vaut E_1), soit à 1 (lorsque la valeur de l'exposition de x_l vaut E_0) et la formule (4) devient $PCAA = 1 - \sum_l P(x_l|M)\widetilde{RR}_l^{-1}$. Dans le cadre

d'une étude cas- témoins, si la maladie est rare, on peut effectuer l'approximation suivante

$$\widetilde{RR}_l = \frac{P(M|x_l)}{P(M|z_l)} \simeq \frac{P(M|x_l)/(1-P(M|x_l))}{P(M|z_l)/(1-P(M|z_l))} = exp\ {}^t\theta(x_l - z_l).$$

On obtient un estimateur de $\widetilde{RR}_l$ en remplaçant θ par son estimateur du maximum de vraisemblance (emv) (voir Prentice et Pyke, 1979 et Drescher et Schill, 1991, pour l'emv de θ dans des études cas-témoins ou des études de cohorte) et on obtient alors un estimateur de la PCAA en estimant les probabilités $P(x_l|M)$ par les proportions observées.

L'emploi du modèle logistique permet de considérer d'éventuelles intéractions entre les facteurs par la construction des variables $G^1, G^2, ..., G^J$ et conduit à des estimations robustes des risques relatifs même pour des strates ayant peu d'individus puisque θ est estimé à partir de l'ensemble des individus (alors que, sans le modèle logistique, chaque risque relatif RR_k serait estimé à partir des individus de la strate C_k uniquement).

Cependant, l'estimateur ci-dessus (que nous appellerons estimateur de Bruzzi) n'est pas l'emv basé sur le modèle logistique puisque le modèle logistique n'est utilisé que pour construire un estimateur des risques relatifs. C'est pourquoi Greenland et Drescher (1993) proposent une variante de cet estimateur dans laquelle les probabilités $P(x_l|M)$ sont estimées à partir du modèle logistique. Greenland et Drescher (1993) donnent la normalité et la variance asymptotique de leur estimateur pour une étude cas-témoins et une étude de cohorte.

• Enfin, une dernière approche qui permet d'obtenir un estimateur de la PCAA est l'approche Mantel-Haenszel. On exprime la PCAA (2) en fonction des risques relatifs de la façon suivante (appendice 1) :

$$PCAA = \sum_k P(E_1, C_k|M)\frac{RR_k - 1}{RR_k}. \tag{5}$$

De même que dans l'approche de Bruzzi, si la maladie est rare, les risques relatifs peuvent être estimés par les odds ratio OR_k

$$RR_k \simeq OR_k = \frac{P(M|E_1, C_k)/P(\bar{M}|E_1, C_k)}{P(M|E_0, C_k)/P(\bar{M}|E_0, C_k)} = \frac{P(E_1|C_k, M)/P(E_1|C_k, \bar{M},)}{P(E_0|C_k, M)/P(E_0|C_k, \bar{M})}.$$

Sous l'hypothèse d'homogénéité des odds ratio ($OR_k = OR$ pour tout k), on obtient

$$PCAA \simeq P(E_1|M)\frac{OR - 1}{OR}. \tag{6}$$

Bien que l'indice k ne figure pas dans l'expression ci-dessus, il s'agit bien d'une quantité ajustée puisque OR est l'odd ratio calculé dans chaque strate C_k (OR est

différent de l'odd ratio global

$$OR^* = \{P(M|E_1)/P(\bar{M}|E_1)\}/\{P(M|E_0)/P(\bar{M}|E_0)\}).$$

OR peut être estimé par l'estimateur de Mantel-Haenszel $\widehat{OR}_{MH}$ (voir, par exemple, Kleinbaum, 1982) et $P(E_1|M)$ par la proportion d'exposés parmi les cas.

L'avantage de cette méthode repose sur les bonnes propriétés de l'estimateur de la variance de $\widehat{OR}_{MH}$, à savoir son bon comportement même pour des données éparses. Son inconvénient majeur provient que son utilisation suppose l'homogénéité des odds ratio. On peut trouver l'expression de l'estimateur de la variance pour différents types d'échantillonnages d'une étude cas-témoins ainsi que des comparaisons entre les estimateurs de la PCAA des sommes pondérées, de Bruzzi et de Mantel-Haenszel dans Bénichou (1991).

3. Méthode de recherche des facteurs de risque

Il est nécessaire, dans le cadre d'une campagne de prévention par exemple, de connaître le plus possible de facteurs de risque causaux liés à la maladie et de pouvoir isoler un petit nombre de facteurs prépondérants sur lesquels on pourrait agir. Ainsi, au lieu de ne s'intéresser qu'à un seul facteur, on peut s'interroger sur le gain que l'on pourrait attendre, dans des conditions idéales (la valeur prédictive de la PCA sera discutée dans le dernier paragraphe), si plusieurs facteurs de risque étaient supprimés simultanément. L'objet de ce paragraphe est de proposer une méthode simple pour détecter les facteurs de risque qui permettraient d'atteindre un gain maximum.

Une réponse possible consiste à considérer qu'un individu est exposé s'il est exposé à au moins un des facteurs de risque pris en compte et qu'il est non exposé s'il n'est exposé à aucun de ces facteurs. La suppression de l'exposition correspond, par exemple, à l'arrêt de la consommation chez une personne qui consomme de l'alcool et à l'arrêt simultané de la consommation d'alcool et de tabac pour un individu qui fume et consomme de l'alcool. Soient $E^1, E^2, ..., E^P$ les facteurs dichotomiques d'exposition considérés. On note E_1^p (ou $E^p = 1$) pour un individu exposé à E^p et E_0^p (ou $E^p = 0$) pour un individu non exposé à E^p. On définit « E^i ou E^j » par E^i ou $E^j = 1$ (resp. E^i ou $E^j = 0$) si $E^i = 1$ ou $E^j = 1$ (resp. si $E^i = 0$ et $E^j = 0$). L'idée est de calculer successivement les PCA aux expositions $E^{*1} = E^{\pi(1)}$, $E^{*2} = E^{*1}$ ou $E^{\pi(2)}$, $E^{*3} = E^{*2}$ ou $E^{\pi(3)}$,..., $E^{*p} = E^{*(p-1)}$ ou $E^{\pi(p)}$, $p \in \{1, 2, ..., P\}$ où $E^{\pi(i)}$ est le facteur ajouté en ième position, choisi selon une règle π définie plus loin.

Dans le cas sans ajustement, par définition

$$PCA(E^{*p}) = 1 - \frac{P(M|E_0^{\pi(1)}, E_0^{\pi(2)}, ..., E_0^{\pi(p)})}{P(M)}.$$

La PCA à plusieurs facteurs peut être utile en pratique. En effet, supposons que suivant l'avis des experts, une opération préventive, pour être efficace, ne puisse avoir pour cible qu'un petit nombre p de facteurs de risque. Il semble alors naturel de choisir les facteurs qui réalisent la PCA maximale pour p facteurs $PCA^p_{\max}$. De plus,

la quantité $PCA(E^{*p})$ permet de savoir, en pratique, si tous les principaux facteurs de risques sont connus puisque dans cette situation $PCA(E^{*p}) \simeq 1$.

On propose, pour construire $PCA(E^{*p})$, d'ajouter à chaque pas, le facteur qui augmente le plus la PCA. La démarche «pas à pas» pose principalement deux questions :

– existe-t-il un ordre privilégié pour les adjonctions successives des facteurs?

– obtient-on nécessairement, à chaque pas, une augmentation de la PCA, c'est-à-dire, a-t-on $PCA(E^{*p}) \leqslant PCA(E^{*p}$ ou $E^{\pi(p+1)})$?

On montre, en appendice 2, que l'adjonction d'une variable supplémentaire n'implique pas nécessairement une augmentation de la PCA. Cependant, pour un facteur d'exposition $E^{\pi(p+1)}$ tel que $P(E_0^{\pi(p+1)}|M) \leqslant P(E_0^{\pi(p+1)})$, on obtient une condition suffisante à cet accroîssement. En particulier, on montre que moins E^{*p} et $E^{\pi(p+1)}$ sont corrélés et corrélés conditionnellement à la population des malades, plus l'adjonction de $E^{\pi(p+1)}$ a tendance à augmenter la PCA. Quant à l'ordre d'adjonction, on montre en appendice 3, qu'il n'existe pas d'ordre privilégié. Plus précisement, si PCA_π^p désigne la PCA pour p variables ajoutées successivement suivant une méthode π, alors, on n'a pas nécessairement $PCA_\pi^p = PCA_{\max}^p$ dès que $p \geqslant 2$ et ceci quelle que soit la méthode π utilisée.

4. Application à une étude sur le cancer de la vessie

4.1. Présentation de l'étude

Les données ont été recueillies suivant un échantillonnage simple dans une étude cas-témoins : un échantillon est prélevé dans la population des malades (les cas) et un autre dans la population des non malades (les témoins); ces échantillons sont prélevés indépendamment des facteurs de risques considérés. L'étude porte sur le cancer de la vessie dans l'Hérault. Les 1015 individus de l'étude (219 cas et 796 témoins), tous de sexe masculin, ont été soumis à un questionnaire rétrospectif portant principalement sur les lieux de naissance et de résidence (les cancers de la vessie étant plus fréquents au bord de la Méditerranée), les antécédents médicaux, la consommation de divers produits (café, alcool, thé, lait, légumes frais, fruits, graisses, huiles, épices, édulcorants, médicaments,...), la consommation de tabac, la profession et les loisirs.

Les travaux antérieurs (Momas *et al.*, 1994a et 1994b) ont permis de retenir les facteurs de risque suivants : la consommation de tabac, d'épices, d'alcool, de café, l'âge et le métier; et les facteurs protecteurs suivants : la consommation de vitamines (vita) et le lieux de résidence (n-sud, un individu étant exposé au facteur protecteur n-sud s'il a vécu hors du pourtour Méditerranéen).

Ces variables ont été rendues dichotomiques de la façon suivante :

– pour le tabac, le seuil a été choisi comme étant la consommation cumulée à 365 cigarettes (soit une cigarette par jour pendant un an), ce qui correspond au premier quartile de la distribution observée

– pour la consommation cumulée d'alcool, le seuil été choisi à 1 073 864 grammes (soit de l'ordre de 10 unités par jour de boissons alcoolisées pendant 30 ans), ce qui correspond au troisième quartile de la distribution observée

– pour la consommation cumulée de café, trois seuils (correspondant au premier quartile, à la médiane et au troisième quartile) ont été considérés : 18 200 (café1), 34 944 (cafém) et 52 416 (café3) grammes

– pour le métier, nous avons séparé les métiers à risques (teinturier, cuisinier, plombier, mécanicien, chauffeur et métiers liés au pétrole) des autres métiers.

La médiane de l'âge des patients de l'étude est de 65 ans et a été choisie comme seuil. La consommation d'épices a été rendue booléenne ainsi que la consommation de vitamines (consommation au moins un fois par semaine de carottes ou de courgettes ou d'épinards) et que le lieu de résidence (sujet ayant vécu au moins un an au dehors du pourtour de la Méditerranée).

4.2. Présentation des résultats

Notre choix s'est porté sur l'estimateur développé par Greenland et Drescher (1993), noté $\widehat{PCAA}$, parce qu'il bénéficie des avantages du modèle logistique et ne nécessite aucune hypothèse particulière pour être applicable.

Le tableau 1 donne des estimations de la PCAA. Chacun des six facteurs est considéré, à tour de rôle, comme étant le facteur d'exposition et les cinq autres comme étant les facteurs sur lesquels on ajuste auquels on rajoute vita et n-sud (les variables vita et n-sud sont, d'après les travaux antérieurs, des facteurs protecteurs importants). Par exemple, la PCA au tabac ajustée sur la consommation d'épices, de café, d'alcool, de vitamines, sur le lieu de résidence, sur le métier et sur l'âge est estimée à 76,27%.

TABLEAU 1
Estimation de la PCAA par l'estimateur de Greenland et Drescher (1993) et intervalle de confiance à 95%

Exposition	$\widehat{PCAA}$	Exposition	$\widehat{PCAA}$
Tabac	0.7627 (0.6378,0.8876)	Métier	0.1457 (0.0728,0.2187)
Epice	0.1563 (0.0836,0.2290)	Age	0.2668 (0.1269,0.4066)
Cafém	0.2653 (0.1126,0.4179)	Alcool	0.1160 (0.0096,0.2135)

Chaque facteur est considéré, à tour de rôle, comme étant le facteur d'exposition et les autres comme étant les facteurs sur lesquels on ajuste auquels on rajoute vita et n-sud.

Dans le tableau 2 (où l'alcool a été retiré puisque sa $\widehat{PCAA}$ est faible) chacun des cinq facteurs de risque restants est considéré, à tour de rôle, comme étant le facteur d'exposition et les quatre autres comme étant les facteurs sur lesquels on ajuste; trois niveaux de base pour la consommation de café ont été considérés. Les variables vita et n-sud ne sont plus prises en compte dans les tableaux 2 à 5. On construit dans le tableau 3 une PCA maximale (91,6%) pour plusieurs facteurs afin de savoir quelle part de la maladie est expliquée par les facteurs considérés.

TABLEAU 2
Estimation de la PCAA par l'estimateur de Greenland et Drescher (1993) et intervalle de confiance à 95% pour 3 niveaux de base différents pour la consommation de café

Exposition	$\widehat{PCAA}$	Exposition	$\widehat{PCAA}$
Tabac	0.7431 (0.6175,0.8687)	Tabac	0.7336 (0.6030,0.8642)
Epice	0.1494 (0.0781,0.2206)	Epice	0.1503 (0.0790,0.2216)
Café1	0.2074 (-0.0578,0.4726)	Cafém	0.2915 (0.1488,0.4343)
Métier	0.1420 (0.0722,0.2118)	Métier	0.1385 (0.0679,0.2091)
Age	0.2977 (0.1703,0.4251)	Age	0.2850 (0.1545,0.4155)

Exposition	$\widehat{PCAA}$
Tabac	0.7335 (0.6029,0.8641)
Epice	0.1530 (0.0821,0.2238)
Café3	0.1921 (0.0981,0.2860)
Métier	0.1411 (0.0708,0.2113)
Age	0.2916 (0.1628,0.4204)

Chaque facteur est considéré, à tour de rôle, comme étant le facteur d'exposition et les autres comme étant les facteurs sur lesquels on ajuste.

TABLEAU 3
Estimation de la PCA sans ajustement (intervalle de confiance à 95%) pour plusieurs facteurs d'exposition suivant la méthode pas à pas exposée

Exposition	$\widehat{PCA}$
Tabac	0.7515 (0.6315,0.8715)
Tabac/Age	0.8661 (0.7341,0.9981)
Tabac/Age/Café3	0.8884 (0.7613,1.0155)
Tabac/Age/Café3/Epice	0.9163 (0.7996,1.0330)
Tabac/Age/Café3/Epice/Métier	0.9004 (0.7614,1.0394)
Tabac/Age/Café3/Epice/Alcool	0.9050 (0.7724,1.0376)
Tabac/Age/Café3/Epice/AlcoolM	0.9149 (0.7964,1.0334)

Pour des raisons de place, on remplace la notation «E^1 ou E^2» par «E^1/E^2». AlcoolM représente la variable dichotomique pour la consommation cumulée d'alcool pour un seuil égal à $2,5 \times 10^6$ g, 3×10^6 g ou $3,5 \times 10^6$ g, les estimations étant égales pour ces 3 seuils.

Enfin, nous avons tenté d'isoler un petit nombre de facteurs de risque en vue de prédire, dans le cadre d'une campagne de prévention, quel serait le gain que l'on pourrait optenir (sous certaines conditions discutées dans le dernier paragraphe) lors de la suppression des expositions (tableau 4). Les variables «âge» et «métier» ont été enlevées puisqu'une campagne serait sans effet sur la première et d'un effet limité

sur la seconde. On obtient une PCA de 84,7% en considérant la consommation de tabac, d'épices ou de café (dont le seuil est situé à la médiane). On calcule ensuite, pour cette même exposition, la PCA ajustée sur l'âge et le métier puis, uniquement sur l'âge (tableau 5).

TABLEAU 4

Estimation de la PCA sans ajustement (intervalle de confiance à 95%) pour plusieurs facteurs d'exposition, sans tenir compte du métier et de l'âge, suivant la méthode pas à pas exposée

Exposition	$\widehat{PCA}$
Tabac	0.7515 (0.6315,0.8715)
Tabac/Epice	0.79831 (0.6852,0.9114)
Tabac/Epice/Cafém	0.8475 (0.7252,0.9898)
Tabac/Epice/Cafém/Alcool	0.8091 (0.6553,0.9629)

Pour des raisons de place, on remplace la notation « E^1 ou E^2 » par « E^1/E^2 ».

TABLEAU 5

Estimation de la PCAA par l'estimateur de Greenland et Drescher (1993) et intervalle de confiance à 95% pour la variable d'exposition «Tabac ou Epice ou Cafém»

Variable(s) d'ajustement	$\widehat{PCAA}$
Métier Age	0.8447 (0.7202,09693)
Age	0.8503 (0.7304,0.9702)

Discussion

Les résultats

On remarque tout d'abord, autant pour la PCA que pour la PCAA, que la consommation de tabac est le facteur de risque prépondérant pour le cancer de la vessie puisque la proportion de cas attribuable à une consommation cummulée de tabac supérieure à 365 cigarettes se situe autour de 75%. En plus de l'importance de l'âge, du métier et de la consommation de café, la consommation d'épices semble être un facteur de risque spécifique au cancer de la vessie, ce qui pourrait expliquer la forte proportion de cancers de la vessie au bord de la Méditérranée. Le tableau 3 indique que les facteurs «tabac», «âge», «café» et «épice» expliquent 91,6% du nombre des malades. On peut donc considérer qu'ils constituent les principaux facteurs de risque.

On remarque, comme on pouvait s'y attendre, que l'adjonction d'un facteur de risque à l'exposition n'augmente pas nécéssairement la PCA. Cette augmentation

étant d'autant plus probable que le facteur ajouté est indépendant et indépendant conditionnellement à la maladie des autres facteurs, on a été conduit à augmenter les niveaux de base des derniers facteurs ajoutés (afin de ne faire intervenir que la partie indépendante de ces facteurs). Cependant, même en augmentant le seuil de la variable alcool à $2,5 \times 10^6$, 3×10^6 ou $3,5 \times 10^6$ grammes, on n'obtient pas d'accroissement de la PCA (il semble que, même pour un seuil très élevé, la variable «alcool» reste corrélée aux autres variables).

Les tableaux 4 et 5 indiquent que le gain maximal d'une campagne ayant pour but de réduire la consommation de tabac, d'épices et de café pourrait être de l'ordre de 84%. L'ajustement sur l'âge et le métier changent peu le résultat (tableau 5). De façon générale, les résultats pour la PCA et la PCAA sont très proches. Cela pourrait s'expliquer par l'absence de corrélation entre les variables de même que la faible sensibilité de la PCAA aux seuils des variables d'ajustement (tableau 2). Par contre, les résultats sont sensibles aux choix des niveaux de base des variables d'exposition. En effet, dans le tableau 2, la PCAA pour la consommation de café dont le seuil est fixé à la médiane est de 29% alors que pour un seuil fixé au troisième quartile, elle est de 19%. Un choix possible pour le niveau de base de la variable d'exposition est celui pour lequel la PCA (ou la PCAA) est maximale. Si on interprète la PCA comme la proportion des cas évitable (ceci sera discuté par la suite), ce choix correspond au seuil en dessous duquel la population devrait se situer pour réduire au maximun l'apparition de la maladie. On montre en appendice 4 que les seuils pour lesquels la PCA est égale à 1 sont ceux qui permettent de n'avoir aucun non exposé parmi les cas.

La PCA (ou PCAA) comme outil de prévision

On peut s'interroger sur le crédit que l'on peut apporter à la PCA (ou la PCAA) en matière de prévision dans le cas d'une suppression réelle d'un facteur de risque. Peut-on raisonnablement penser que l'on pourrait réduire de 75% le nombre des cancers de la vessie si toute la population cessait de fumer?

Coste et Spira (1991) rappellent que cette interprétation en terme de proportion de cas évitable n'est valable que si le facteur d'exposition est un facteur causal distribué dans la population indépendamment des autres facteurs de risque et si la suppression de l'exposition ne modifie pas l'exposition aux autres facteurs de risque.

En effet, utiliser la PCA (non ajustée) pour prévoir le gain induit par la suppression de l'exposition revient à identifier la probabilité de la maladie dans la population totale après la suppression de l'exposition à la probabilité de la maladie dans la population des non exposés (avant la suppression de l'exposition). Cette identification est source de biais dès que les distributions de certains facteurs de risque autre que l'exposition considérée ne sont pas identiques dans les deux populations. Par exemple, on peut penser que les gens qui ne fument pas sont peu consommateurs de café alors que ceux qui s'arrêtent de fumer continuent d'être de gros consommateurs de café. Ainsi, la probabilité d'être malade pour un non fumeur peut être inférieure à celle d'être malade pour un individu ayant cessé de fumer. On peut alors penser que la PCAA se prête mieux à la prédiction puisque la suppression de l'exposition pour la PCAA s'opère en respectant les distributions des autres facteurs de risque. Utiliser la PCAA revient alors à supposer que la suppression de l'exposition n'entraîne pas

de modification des distributions des autres facteurs de risque. Or, il se peut qu'un individu qui arrête de fumer ait tendance à accroître sa consommation d'alcool ou de café pour compenser le manque lié au tabac, ou au contraire, se mette à faire du sport, à mener une vie plus saine...

Ainsi, il apparaît que la PCA, ajustée ou non, ne constitue pas toujours un bon modèle pour la prévision puisque la suppression d'une exposition à un facteur de risque chez un individu peut entraîner un changement de comportement de cet individu qui n'est pas pris en compte par la PCA ou la PCAA.

On pourrait donc s'interroger sur le comportement d'un individu après la suppression d'un facteur de risque. Par exemple, l'individu qui cesse de fumer va-t-il se comporter, vis-à-vis des autres facteurs de risque,

– comme quelqu'un qui n'a jamais fumé et dans ce cas l'emploi de la PCA sans ajustement peut convenir à la prévision,

– comme il se comportait avant l'arrêt de tabac et dans ce cas l'emploi de la PCAA semble alors indiquée,

– ou bien d'une autre façon?

Par conséquent, plusieurs questions se posent naturellement :

– peut-on connaître le comportement de quelqu'un qui cesse de fumer et quel serait le modèle comportemental?

– comment utiliser cette information pour prédire le gain d'une campagne de prévention, c'est-à-dire, comment tenir compte à la fois du modèle actuel et du modèle comportemental?

Appendices

Appendice 1. On note

$$PCAA^a = 1 - \frac{\sum P(C_k)P(M|E_0, C_k)}{P(M)}$$

$$PCAA^b = \sum P(C_k|M)\left(1 - \frac{P(M|E_0, C_k)}{P(M|C_k)}\right)$$

$$PCAA^c = 1 - \sum\{P(E_0, C_k|M) + P(E_1, C_k|M)RR_k^{-1}\}$$

$$PCAA^d = \sum P(E_1, C_k|M)\frac{RR_k - 1}{RR_k}$$

Montrons que $PCAA^a = PCAA^b = PCAA^c = PCAA^d$.

$$PCAA^a = \frac{\sum P(M|C_k)P(C_k) - \sum P(M|E_0,C_k)P(C_k)}{P(M)}$$

$$= \sum \frac{P(C_k)}{P(M)} \left(P(M|C_k) - P(M|E_0,C_k)\right) = PCAA^b$$

puisque $\dfrac{P(C_k)}{P(M)} = \dfrac{P(C_k|M)}{P(M|C_k)}$.

$$PCAA^a = 1 - \frac{\sum \{P(C_k,E_0)P(M|E_0,C_k) + P(C_k,E_1)P(M|E_0,C_k)\}}{P(M)}$$

$$= 1 - P(E_0|M) - \frac{\sum P(C_k,E_1)P(M|E_0,C_k)}{P(M)}$$

$$= 1 - P(E_0|M) - \frac{\sum P(C_k,E_1)P(M|E_1,C_k)RR_k^{-1}}{P(M)} = PCAA^c$$

puisque $RR_k^{-1} = P(M|E_0,C_k)/P(M|E_1,C_k)$.

$$PCAA^c = \sum \{P(E_0,C_k|M) + P(E_1,C_k|M)\} - \sum \{P(E_0,C_k|M) + P(E_1,C_k|M)RR_k^{-1}\}$$

$$= \sum P(E_1,C_k|M)(1 - RR_k^{-1}) = PCAA^d.$$

Appendice 2. Les facteurs considérés sont des facteurs de risque.

$$PCA(E^{*p}) \leqslant PCA(E^{*p} \text{ ou } E^{\pi(p+1)})$$

$$\Leftrightarrow P(M|E_0^{*p}) \geqslant P(M|E_0^{*p}, E_0^{\pi(p+1)})$$

$$\Leftrightarrow \frac{P(E_0^{*p}, E_0^{\pi(p+1)}|M)}{P(E_0^{*p}|M)} \leqslant \frac{P(E_0^{*p}, E_0^{\pi(p+1)})}{P(E_0^{*p})}$$

Soit X_1 (resp. X_2) la variable qui vaut 1 si l'individu n'est pas exposé à E^{*p} (resp. $E^{\pi(p+1)}$) et 0 sinon, alors $\Gamma(X_1,X_2) = \mathrm{E}(X_1X_2) - \mathrm{E}(X_1)\mathrm{E}(X_2) = P(E_0^{*p}, E_0^{\pi(p+1)}) - P(E_0^{*p})P(E_0^{\pi(p+1)})$ et si on note Γ_M la covariance conditionellement à la population des malades, on obtient de même $\Gamma_M(X_1,X_2) = P(E_0^{*p}, E_0^{\pi(p+1)}|M) - P(E_0^{*p}|M)P(E_0^{\pi(p+1)}|M)$.

Sous l'hypothèse $\Gamma(X_1,X_2) \geqslant 0$ et $\Gamma_M(X_1,X_2) \leqslant 0$ on a, si $P(E_0^{\pi(p+1)}|M) \leqslant P(E_0^{\pi(p+1)})$

$$\frac{P(E_0^{*p}, E_0^{\pi(p+1)}|M)}{P(E_0^{*p}|M)} \leqslant P(E_0^{\pi(p+1)}|M) \leqslant P(E_0^{\pi(p+1)}) \leqslant \frac{P(E_0^{*p}, E_0^{\pi(p+1)})}{P(E_0^{*p})}$$

soit $PCA(E^{*p}) \leqslant PCA(E^{*p}$ ou $E^{\pi(p+1)})$.

Cependant, il est rare dans la pratique de trouver des variables corrélées positivement et corrélées négativement conditionellement à la maladie. Par contre, il est vraisemblable que les variables soient suffisamment peu corrélées de sorte que

$$\frac{P(E_0^{*p}, E_0^{\pi(p+1)}|M)}{P(E_0^{*p}|M)} \simeq P(E_0^{\pi(p+1)}|M) \leqslant P(E_0^{\pi(p+1)}) \simeq \frac{P(E_0^{*p}, E_0^{\pi(p+1)})}{P(E_0^{*p})}.$$

Appendice 3. On reprend les notations du paragraphe 3.

On suppose que les distributions théoriques sont données par le tableau suivant.

Individu	M	E^1	E^2	E^3	Individu	M	E^1	E^2	E^3
1	0	0	1	1	11	1	1	1	0
2	0	0	1	1	12	1	0	1	0
3	0	0	1	1	13	0	0	1	0
4	1	0	0	1	14	0	0	0	1
5	1	1	0	1	15	1	1	0	0
6	1	1	0	1	16	0	0	0	1
7	0	1	0	0	17	0	1	0	0
8	0	1	0	0	18	0	1	0	0
9	0	1	0	0	19	0	0	0	1
10	1	1	1	0					

On a

$$P(M|E_0^1) = 2/9,\ P(M|E_0^2) = 1/3,$$
$$P(M|E_0^3) = 2/5,\ P(M|E_0^1, E_0^2) = 1/4,$$
$$P(M|E_0^1, E_0^3) = 1/2,\ P(M|E_0^2, E_0^3) = 1/6,$$

d'où $PCA^1_{\max} = PCA(E^1)$ et $PCA^2_{\max} = PCA(E^2$ ou $E^3)$.

Cet exemple montre donc qu'il n'existe pas de méthode pas à pas π telle que $PCA^p_\pi = PCA^p_{\max}$ dès que $p \geqslant 2$. De plus, si π est la méthode proposée dans le paragraphe 3, $PCA^1_\pi = PCA^1_{\max}$ et $PCA^2_\pi = PCA(E^1\ ou\ E^2) = 0.32$ alors que $PCA^2_{\max} = 0.55$.

Appendice 4. On suppose que $P(M) > 0$.

Sans ajustement, on suppose que $P(E_0) > 0$.

$$PCA = 1 \Leftrightarrow P(M|E_0) = 0 \Leftrightarrow P(E_0|M)P(M)/P(E_0) = 0$$
$$\Leftrightarrow P(E_0|M) = 0.$$

Avec ajustement, on suppose que $P(E_0, C_k)P(C_k) > 0 \; \forall k$.

$$PCA = 1 \Leftrightarrow \sum P(C_k)P(M|E_0, C_k) = 0 \Leftrightarrow P(M|E_0, C_k) = 0 \; \forall k$$
$$\Leftrightarrow P(M, E_0, C_k) = 0 \; \forall k \Leftrightarrow P(E_0, C_k|M) = 0 \; \forall k \Leftrightarrow P(E_0|M) = 0.$$

Remerciements

Nous remercions les rapporteurs de cet article pour leurs nombreuses remarques qui ont permis d'améliorer ce travail.

Nous remercions aussi, très sincèrement, l'ensemble du personnel du laboratoire de Biométrie de l'INRA de Montpellier pour son accueil et la mise à notre disposition de moyens informatiques qui ont permis la rédaction de cet article.

Bibliographie

BENICHOU J. (1991), Methods of ajustment for estimating the attributable risk in case-control studies : a review, *Statistics in Medicine*, **10**, 1753-1773.

BENICHOU J., GAIL M.H. (1990), Variance calculations and confidence intervals for estimates of the attributable risk based on logistic models, *Biometrics*, **46**, 991-1003.

BRUZZI P., GREEN S.B., BYAR D.P., BRINTON L.A., SCHAIRER C. (1985), Estimating the population attributable risk for multiple risk factors using case-control data, *American Journal of Epidemiology*, **122**, 904-914.

COSTE J., SPIRA A. (1991), La proportion de cas attribuable en santé publique : définitions(s), estimation(s) et interprétation, *Revue Epidémiologique et Santé Publique*, **39**, 339-411.

DRESCHER K., SCHILL W. (1991), Attributable risk estimation from case-control data via logistic regression, *Biometrics*, **47**, 1247-1256.

FAREWELL V.T. (1979), Some results on the estimation of logistic models based on retrospective data, *Biometrika*, **66**, 1, 27-32.

GEFELLER O. (1992), An annotated bibliography on the attributable risk, *Biometrical Journal*, **34**, 1007-1012.

GREENLAND S., DRESCHER K. (1993), Maximun likelihood estimation of the attributable fraction from logistic models, *Biometrics*, **49**, 865-872.

GREENLAND S., ROBINS J.M. (1988), Conceptual problems in the definition and interpretation of attributable fractions, *American Journal of Epidemiology*, **128**, 1185-1197.

LEVIN M.L. (1953), The occurence of lung cancer in man, *Acta Unio Internationalis contra Cancrum*, **9**, 531-541.

KLEINBAUM D.G., KUPPER L.L., MORGENSTERN H. (1982), *Epidemiologic Research. Principles and Quantitative Methods*, Van Nostrand Reinhold, 115 Fifth Avenue, New York.

MIETTINEN O.S. (1974), Proportion of disease caused or prevented by a given exposure, trait or intervention, *American Journal of Epidemiology*, **99**, 5, 322-325.

MOMAS I., DAURES J.P., FESTY B., BONTOUX J., GREMY F. (1994a), Bladder cancer and black cigarette smoking : some new results from a french case-control study, *European Journal of Epidemiology*, **10**, 599-604.

MOMAS I., DAURES J.P., FESTY B. (1994b), Relative importance of risk factors in lader carcinogenesis : some crem results about mediteranean habits, *Cancer Causes and Control*, **5**, 326-332.

PRENTICE R.L., PYKE K. (1979), Logistic disease incidence models and case-control studies, *Biometrika*, **66**, 3, 403-411.

WALTER S.D. (1976), Estimation and interpretation of attributable risk in health research, *Biometrics*, **32**, 829-849.

WALTER S.D. (1978), Calculation of attributable risks from epidemiological data, *International Journal of Epidemiology*, **7**, 2, 175-182.

WALTER S.D. (1980), Prevention for multifactorial diseases, *American Journal of Epidemiology*, **112**, 3, 409-416.

WALTER S.D. (1983), Effects of interaction, confounding and observational error on attributable estimation, *American Journal of Epidemiology*, **117**, 5, 598-604.

WHITTEMORE A.S. (1982), Statistical methods for estimating attributable risk from retrospective data, *Statistics in Medicine*, **1**, 229-243.