JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

PAUL DAMIANI

Approche méthodologique pour une étude des facteurs de risques cardiovasculaires, d'après une enquête effectuée auprès des patients du centre médical I. P. C. de 1970 à 1975

Journal de la société statistique de Paris, tome 120, n° 3 (1979), p. 198-202

<http://www.numdam.org/item?id=JSFS_1979__120_3_198_0>

APPROCHE MÉTHODOLOGIQUE POUR UNE ÉTUDE DES FACTEURS DE RISQUES CARDIOVASCULAIRES

d'après une enquête effectuée auprès des patients du Centre médical I. P. C. de 1970 à 1975

Paul DAMIANI

administrateur de l'I. N. S. E. E.

Les Investigations Pré-Cliniques (I. P. C.) (1) ont développé, grâce à la coopération d'universitaires et de médecins praticiens, un instrument de recherche en épidémiologie et en prévention médicale, notamment cardiovasculaire.

Le recueil informatisé de 600 variables cliniques et biologiques chez de nombreux sujets (100 000) commence à autoriser certaines déductions statistiques destinées à guider cette prévention médicale.

*
* *

Une analyse statistique des facteurs de risques cardiovasculaires a été réalisée à partir des données recueillies auprès des patients du Centre I. P. C., de 1970 à 1975. On a exclu de l'étude les personnes se sachant malades ou suivant un traitement au moment de l'examen et on n'a retenu que les hommes de 40 à 69 ans et les femmes de 50 à 69 ans, soit au total 29 400 personnes.

Les variables étudiées, au nombre de 16 (2), représentent les résultats de certaines investigations cliniques et biologiques, ainsi que les réponses à quelques interrogations du questionnaire médical.

Parmi les sujets, on a distingué un groupe de personnes présentant des symptômes ou des signes de pathologie cardiovasculaire et un groupe de personnes n'en présentant pas. Cette sélection a été faite d'après les résultats de l'électro-cardiogramme, de la radio-photographie cardio-aortique et d'après les réponses au questionnaire médical.

*
* *

Le Comité Scientifique des I. P. C. (3) n'ignore pas que le choix des patients et le choix des variables peuvent faire l'objet de critiques.

1. Représentées par MM. J. R. DEBRAY, J. CHRETIEN, M. GUENIOT, J. P. HARDOUIN, J. HIMBERT (†), P. PICHOT, G. RICHET, Cl. ROUSSEL, A. RYCKEWAERT, il y a dix ans, le 12 mai 1969, à la Société Médicale des Hôpitaux de Paris, sous le titre : « Une nouvelle conception de la médecine préventive utilisant le traitement automatique de l'information par ordinateur ». Ann. Méd. int., 1969, 120, 391 et 589.

2. Les variables retenues : antécédents familiaux, tabac, alcool, surmenage, cholestérol, glycémie, acide urique, urée, phosphatases alcalines, hémoglobine, capacité vitale, volume expiratoire maximal par seconde, surcharge pondérale, pression artérielle diastolique, pression artérielle systolique, pli cutané.

3. Composé de : Dr J. R. DEBRAY de l'Institut, Prof. Jacques CHRETIEN, Prof. Maurice GUENIOT, Prof. Ag. Louis GUIZE, Prof. Ag. J. P. HARDOUIN, Prof. Pierre PICHOT, Prof. Gabriel RICHET, Prof. A. RYCKEWAERT, Dr Claude ROUSSEL; assisté de Ariane BABOK, Fouad BENJELLOUN, Jean-Marie CHRETIEN, Robert MOYAL, André SALEM.

La définition des groupes de sujets présentant des symptômes ou signes de pathologie cardio-vasculaire demande à être nuancée et pourra, grâce à des études ultérieures, trouver de nouveaux fondements.

La publication de ce travail répond au souci d'une équipe pluridisciplinaire de faire connaître la méthodologie utilisée sous l'égide de Paul Damiani.

Le but de l'étude est d'analyser statistiquement, à partir des valeurs des différentes données les raisons pour lesquelles les individus sont classés dans l'un ou l'autre groupe définis ci-dessus.

S'il n'y avait qu'un seul facteur de risque, il suffirait de comparer les valeurs moyennes de ce facteur dans chacun des groupes. Comme on dispose de plusieurs variables, on opère de la façon suivante.

On cherche une moyenne des valeurs des différentes variables qui explique le mieux la dispersion des sujets dans les deux groupes; dans cette moyenne, chaque variable est affectée d'un coefficient d'autant plus élevé que l'importance de la variable dans l'explication du phénomène est plus grande.

Pour déterminer cette moyenne, on applique la méthode d'analyse statistique appelée « régression pas à pas ». Avec cette méthode, on sélectionne les variables ayant une action significative dans la dispersion des sujets entre les deux groupes et on établit un modèle permettant de mesurer l'action des différents facteurs de risque retenus. La valeur de ce modèle est estimée par la part de la dispersion qu'il explique.

Une première sélection des variables, parmi les plusieurs centaines dont on dispose, avait auparavant été obtenue : 1° à partir de comparaisons univariées de variables entre le groupe pathologique et le groupe témoin (comparaison de moyennes entre variables quantitatives, comparaison de taux de réponses entre variables qualitatives), 2° à partir d'analyses factorielles des correspondances qui ont permis de sélectionner les variables indépendantes les plus représentatives parmi les variables redondantes et celles qui avaient la plus faible variabilité analytique.

L'étude a été réalisée par sexe et par groupe d'âge décennal. Les résultats obtenus sont les suivants :

Pour le sexe masculin

Le modèle explique entre 24 % et 28 % de la dispersion, ce qui correspond aux pourcentages trouvés en appliquant la méthode dans d'autres domaines des sciences humaines. Les facteurs de risque les plus significatifs sont, par groupe d'âge :

40-49 ans : la capacité vitale (1), la pression artérielle systolique, le surmenage,

50-59 ans : aux trois facteurs précédents (2) s'ajoutent : l'hémoglobine et la surcharge pondérale,

60-69 ans : la capacité vitale et le surmenage disparaissent, restent : la pression artérielle systolique, l'hémoglobine, la surcharge pondérale.

Autrement dit, il existe un facteur de risque cardio-vasculaire présent quel que soit l'âge, c'est la pression artérielle systolique; les facteurs de risque capacité vitale et surmenage

1. A elle seule la capacité vitale explique 18 % de la dispersion.
2. La capacité vitale est remplacée par le volume expiratoire maximal, variable en forte corrélation avec elle.

ont une importance qui décroît avec l'âge et disparaissent à 60 ans (1); les facteurs hémoglobine et surcharge pondérale ont une action qui augmente avec l'âge et n'apparaissent qu'après 50 ans.

Pour le sexe féminin

Le modèle est moins satisfaisant que pour le sexe masculin puisqu'il explique entre 11 % et 20 % de la dispersion; c'est un résultat que l'on rencontre souvent dans des études analogues (2).

Parmi les facteurs de risques retenus, le surmenage et l'hémoglobine sont présents dans les deux groupes d'âge.

On trouve en plus : le volume expiratoire maximal et la glycémie entre 50 et 59 ans; l'uricémie, la pression artérielle systolique, le pli cutané et le cholestérol, entre 60 et 69 ans.

On notera que les coefficients relatifs à certains facteurs (capacité vitale, surmenage, hémoglobine...) sont négatifs c'est-à-dire que plus un de ces facteurs a une valeur élevée et plus la probabilité d'être classé dans le groupe pathologique est faible.

Par ailleurs, on observe que, dans cette étude, deux facteurs de risque fréquemment retenus n'apparaissent pas discriminants : la cholesterolémie et la consommation tabagique dont nous possédons une meilleure analyse dans les questionnaires itératifs.

Il faut également savoir que lorsque deux variables sont fortement corrélées entre elles et ont une action comparable sur la distinction des deux groupes, ce modèle de régression ne fait apparaître qu'une seule de ces deux variables; c'est le cas, par exemple, de la pression artérielle diastolique face à la pression artérielle systolique.

L'intérêt de la méthode est de remplacer les valeurs relevées pour différents facteurs de risque au cours d'un examen par un indice calculé sur un nombre réduit de facteurs les plus significatifs. A partir de cet indice, il est possible d'évaluer la probabilité pour un individu d'appartenir à une population présentant des signes de pathologie cardio-vasculaire et de mesurer la variation de probabilité due à une variation donnée d'un facteur. Il s'agit d'une probabilité calculée à partir de statistiques relevées sur une population nombreuse, elle ne doit être considérée que comme une indication pour un médecin.

Ces résultats ne représentent qu'une première approche et doivent nécessairement être vérifiés et complétés par les études fondées sur le suivi des sujets au moyen des examens itératifs.

ANNEXE

Données de base

Variable à expliquer

La variable à expliquer y représente le fait d'appartenir ou non à la sous-population présentant des symptômes ou des signes de pathologie cardiovasculaire. Elle peut prendre deux valeurs :

$y = 1$, si l'individu appartient à la sous-population pathologique,

$y = 0$, si l'individu appartient à la sous-population de référence.

1. La capacité vitale apparaît dans cette première étude plus discriminante que les réponses aux questions sur le tabagisme, mais ces deux variables sont très liées. De même le surmenage est fortement lié à la note d'hypocondrie qui n'a pas été retenue ici mais apparaîtra dans une étude ultérieure.

2. Il faut se rappeler que dans d'autres travaux de la littérature la définition de la pathologie cardio-vasculaire à partir d'un questionnaire et d'un électro-cardiogramme est moins performante chez la femme que chez l'homme.

Variables explicatives

On a retenu les 16 variables explicatives x_i suivantes ($i = 1, 2, \ldots, 16$) représentant les principaux facteurs de risque.

x_1 : antécédents familiaux; cette variable peut prendre 3 valeurs suivant l'importance des antécédents familiaux;
x_2 : tabac; cette variable peut prendre 4 valeurs suivant la quantité de tabac fumée;
x_3 : alcool; cette variable peut prendre 3 valeurs suivant la quantité d'alcool bue;
x_4 : surmenage; cette variable peut prendre 2 valeurs suivant que l'individu estime être ou n'être pas surmené;
x_5 : cholesterol, en g/l;
x_6 : glycémie, en g/l, dosage 45 minutes après absorption de 50 g de glucose;
x_7 : urée dans le sang, en mg/l;
x_8 : uricémie, en mg/l;
x_9 : phosphatases alcalines, en unités internationales;
x_{10} : hémoglobine, en g/100 ml;
x_{11} : pression artérielle systolique, en mm de Hg;
x_{12} : pression artérielle diastolique, en mm de Hg;
x_{13} : capacité vitale mesurée, en mm^3/l; c'est une mesure du volume d'air contenu dans les poumons;
x_{14} : volume expiratoire maximal par seconde, en mm^3/l; c'est une mesure du volume d'air expiré en un temps donné après inspiration;
x_{15} : pli cutané, en mm;
x_{16} : surcharge pondérale; cette variable représente le rapport du poids de l'individu considéré au poids moyen ajusté de la population étudiée de même taille, pour le même groupe d'âge et le même sexe.

MODÈLE

On suppose qu'il existe un modèle de régression linéaire entre y et les variables x_i, par groupe d'âge décennal et par sexe :

$$y = a + b_1x_1 + \ldots + b_ix_i + \ldots + b_{16}x_{16}$$

Les coefficients b_i et la constante a sont calculés par la méthode des moindres carrés.

On a appliqué la méthode régression pas à pas qui permet de ne retenir dans le modèle que les variables explicatives dont l'action sur y est jugée significative.

La valeur du modèle est mesurée par la proportion R^2 de la variance de y expliquée par le modèle.

RÉSULTATS

Les résultats obtenus figurent dans le tableau ci-après qui indique, par groupe d'âge et par sexe, les variables retenues, la proportion cumulée de variance expliquée par le modèle et le modèle sous deux formes. On a donné l'expression du modèle avec des variables brutes (en lettres minuscules), puis son expression avec des variables centrées réduites (en lettres majuscules) afin de permettre de comparer l'action des différents facteurs.

Modèles de régression par sexe et groupe d'âge

Groupe d'âge et facteurs	Proportion cumulée de variance expliquée R^2	Modèles [1]
		Sexe masculin
40-49 ans		
x_{13} : capacité vitale	0,184	$y = 0{,}49985 - 0{,}0003\ x_{13} + 0{,}0082\ x_{11} + 0{,}0163\ x_4$ (0,0000) (0,0023) (0,0079)
x_{11} : pression artérielle systolique	0,260	$Y = -0{,}4370\ X_{13} + 0{,}2550\ X_{11} + 0{,}1468\ X_4$ (0,0706) (0,0707) (0,0712)
x_4 : surmenage	0,280	
50-59 ans		
x_{11} : pression artérielle systolique	0,112	
x_{14} : volume expiratoire maximal	0,164	$y = 0{,}09705 + 0{,}0051\ x_{11} - 0{,}0001\ x_{14} + 0{,}0246\ x_4$ (0,0013) (0,0000) (0,0073) $-0{,}00030\ x_9 + 0{,}0035\ x_{16}$ (0,0012) (0,0021)
x_4 : surmenage	0,205	
x_{10} : hémoglobine	0,225	
x_{16} : surcharge pondérale	0,236	$Y = 0{,}2528\ X_{14} - 0{,}1803\ X_{11} + 0{,}2040\ X_4 - 0{,}1522\ X_{10} + 0{,}1058\ X_{16}$ (0,0633) (0,0633) (0,0606) (0,0626) (0,0625)
60-69 ans		
x_{10} : hémoglobine	0,137	$y = -0{,}1095 - 0{,}0064\ x_{10} + 0{,}0059\ x_{11} + 0{,}0049\ x_{16}$ (0,0012) (0,0014) (0,0023)
x_{11} : pression artérielle systolique	0,222	$Y = -0{,}3297\ X_{10} + 0{,}2678\ X_{11} + 0{,}1335\ X_{16}$ (0,0628) (0,0640) (0,0636)
x_{16} : surcharge pondérale	0,239	
		Sexe féminin
50-59 ans		
x_{14} : volume expiratoire maximal	0,048	$y = 1{,}0598 - 0{,}0002\ x_{14} + 0{,}0242\ x_4 - 0{,}0038\ x_{10} + 0{,}0015\ x_6$ (0,0001) (0,0101) (0,0025) (0,0011)
x_4 : surmenage	0,079	$Y = -0{,}1685\ X_{14} + 0{,}2115\ X_4 - 0{,}1345\ X_{10} + 0{,}1201\ X_6$ (0,0914) (0,0885) (0,0912) (0,0893)
x_{10} : hémoglobine	0,092	
x_6 : glycémie	0,106	
60-69 ans		
x_8 : uricémie	0,076	
x_{11} : pression artérielle systolique	0,136	$y = -0{,}6069 + 0{,}0095\ x_8 + 0{,}0053\ x_{11} - 0{,}0037\ x_{15} - 0{,}0306\ x_{10}$ (0,0028) (0,0013) (0,0023) (0,0021) $+ 0{,}0173\ x_4 + 0{,}0016\ x_5$ (0,0091) (0,0008)
x_{15} : pli cutané	0,153	
x_{14} : hémoglobine	0,169	
x_4 : surmenage	0,182	$Y = 0{,}2323\ x_8 + 0{,}2768\ X_{11} - 0{,}1133\ X_{15} - 0{,}1180\ X_{10} + 0{,}1315\ X_4 + 0{,}1302\ X_5$ (0,0694) (0,0703) (0,0692) (0,0693) (0,0692) (0,0701)
x_5 : cholestérol	0,198	

1. Deux modèles sont indiqués : avec variables brutes (lettres minuscules), avec variables centrées réduites (lettres majuscules). Sous chaque coefficient du modèle figure l'erreur-type correspondante.