REVUE DE STATISTIQUE APPLIQUÉE

J.-P. NAKACHE
A. GUÉGUEN
M. ZINS
M. GOLDBERG

Analyse de données de survie groupées avec covariables dépendant du temps : application à l'étude de l'effet prédictif de l'état de santé perçu sur le décès, chez les hommes de la cohorte Gazel observés dans la période 1989-1999

Revue de statistique appliquée, tome 52, n° 2 (2004), p. 27-49
<http://www.numdam.org/item?id=RSA_2004__52_2_27_0>

Rev. Statistique Appliquée, 2004, **LII** (2), 27-49

ANALYSE DE DONNÉES DE SURVIE GROUPÉES AVEC COVARIABLES DÉPENDANT DU TEMPS : application à l'étude de l'effet prédictif de l'état de santé perçu sur le décès, chez les hommes de la cohorte Gazel observés dans la période 1989-1999

J-P. NAKACHE (*), A. GUÉGUEN (**), M. ZINS (**), M. GOLDBERG (**)

(*) CNRS/INSERM U88/IFR69
(**) INSERM U88/IFR69

RÉSUMÉ

Cet article concerne une étude de l'effet prédictif de l'état de santé perçu (ESP) sur la mortalité, dans le but de savoir pendant combien de temps avant la survenue de décès cet effet est significatif. L'évaluation a été faite jusqu'à 6 ans avant, pour les décès toutes causes confondues et aussi pour certaines causes spécifiques de décès.

Les analyses ont été effectuées à partir de données concernant 14 805 hommes de la cohorte Gazel âgés entre 41 et 50 ans et répondant, depuis 1989, à un auto-questionnaire annuel où l'ESP est évalué. Les dates de décès relevées jusqu'à fin 1999 ainsi que les causes de décès disponibles jusqu'à fin 1997 ont été prises en compte dans cette étude.

Le modèle logistique a été appliqué, après transformation des données initiales en données de survie groupées par année, pour expliquer le décès au cours d'une année par l'ESP déclaré 1 an, 2 ans, jusqu'à 6 ans avant, en ajustant sur l'âge en 1989, la catégorie socio-professionnelle et la période calendaire. Les résultats montrent que l'ESP est prédictif de la mortalité jusqu'à 6 ans avant la survenue de décès, et cet effet varie selon la cause de décès .

Mots-clés : *Données de survie groupées, Covariables dépendant du temps, Modèle discret de Cox, Risque relatif, Modèle logistique binaire et polytomique.*

ABSTRACT

This paper concerns a study of the predictive effect of self-rated health (SRH) on mortality, in order to know for how long before death this effect is significant. This effect was assessed until 6 years before, for all deaths and also for specific causes.

The analyses have been performed on a data set comprised of 14 805 men of the GAZEL cohort, aged 41-50 and followed-up since 1989 by a yearly self-administered questionnaire where SRH is evaluated. Records of deaths up to end of 1999 and death causes up to end of 1997 were used. Logistic regression model was performed after a modification of the data to account for grouped-survival data, using time-dependent SRH measurements and year as time unit, adjusting for age, socioeconomic status at baseline and calendar year.

SRH seems to have a predictive effect on mortality until 6 years before death occurrence, and this effect is different according to death causes.

Keywords : *Gouped survival data, Time-dependent covariates, Discrete Cox model, Relative risk, Binary and polytomous logistic regression.*

1. Introduction

De nombreux articles de la littérature traitent de l'état de santé perçu (ESP) et de son effet sur la mortalité. La durée du suivi de ces études de mortalité est cependant très variable. Ainsi dans une revue de la littérature, Idler et Benyamini (1997) retrouvent des suivis variant de 5 à 23 ans. Les auteurs analysent l'effet d'un mauvais état de santé mesuré à la date d'inclusion sur la survenue d'un décès quel que soit le délai entre la date d'inclusion et celle de la survenue du décès. Ils estiment donc un risque moyen de décéder sur toute la durée de suivi en faisant l'hypothèse que ce risque est le même si le décès survient un an, deux ans, voire 10 ans après la mesure de l'ESP.

Ces études soulèvent plusieurs questions : i) la mesure du lien entre l'ESP et la mortalité est-il constant dans le temps?, ii) cet effet est-il le même quelle que soit la cause de décès?

L'objectif de ce travail est donc d'étudier l'effet prédictif de l'ESP sur la mortalité en analysant pendant combien d'années avant la survenue de décès, l'ESP est prédictif. Dans ce travail, la mortalité est étudiée à la fois toutes causes confondues et par grande cause (maladies cardio-vasculaires, cancers, morts violentes, autres causes). Les auto-questionnaires annuels remplis par les volontaires de la cohorte Gazel de 1989 à 1999 permettent de recueillir les données nécessaires à cette analyse longitudinale. La méthode statistique utilisée pour analyser ces données de survie censurées est le modèle des risques proportionnels de Cox avec covariables dépendant du temps. Ce modèle avec temps continu requiert un temps de calcul trop important, qui augmente encore plus si plusieurs facteurs d'ajustement sont prévus dans le modèle. Comme nous avions à utiliser plusieurs fois ce modèle, nous avons opté pour son utilisation après regroupement des données de survie par intervalles de temps. Le modèle résultant, connu sous le nom de modèle discret de Cox avec covariables dépendant du temps, est un modèle logistique binaire spécifique qui requiert un temps calcul bien plus faible, même si plusieurs facteurs d'ajustement sont introduits dans le modèle. Nous avons utilisé le modèle logistique sous forme binaire pour étudier les décès toutes causes confondues, et sous forme multinomial pour prendre en compte différentes causes de décès simultanément.

2. Les données

En janvier 1989, 20 625 volontaires agents d'EDF-GDF ont accepté d'être l'objet d'un suivi épidémiologique prospectif sur une très longue période [Goldberg *et al.* (1994), (2001)]. Les informations sur ces sujets sont recueillies systématiquement pour toute la cohorte, auprès de différentes sources. Ce travail s'appuie sur des données issues des auto-questionnaires annuels permettant notamment de recueillir l'ESP, mais aussi sur des données fournies par des sources internes à l'entreprise et par l'INSERM. Cette étude concerne les 15 011 hommes de la cohorte.

2.1. Les variables de l'étude

Le statut vital est connu jusqu'à fin 1999. Les dates de décès sont recueillies auprès du service général des pensions d'EDF-GDF. Les causes de décès, recueillies auprès du Cépi-dc de l'INSERM, sont connues jusqu'à fin 1997.

L'état de santé perçu provient des réponses aux 10 questionnaires annuels successifs. La formulation de la question, identique d'une année à l'autre, est la suivante « Comment jugez vous votre état de santé ? » Les volontaires répondent au moyen d'une échelle visuelle numérique graduée de 1 (très bon état) à 8 (très mauvais état) sur laquelle il leur est demandé de se situer. La variable ESP a été utilisée dans les analyses sous la forme de variable binaire dont les deux modalités sont : bon état de santé (BS) si ESP = 1,2,3, 4 et mauvais état de santé (MS) si ESP = 5,6,7,8.

2.2. Facteurs d'ajustement

Les facteurs d'ajustement pris en compte sont l'âge en 1989 en deux classes : AGE89 = 1 (41-45 ans), AGE89 = 2 (46-50 ans) et la catégorie socio-professionnelle (PCS, donnée fournie par les services du personnel) en 1989 en trois classes : PCS = 3 (Ingénieurs et Cadres supérieurs), PCS = 4 (Professions intermédiaires) et PCS = 5 (Employés et Ouvriers).

Au total la population étudiée est constituée de 14 805 hommes à risque de décès au 01/01/90, sans données manquantes pour la PCS et l'ESP en 1989. En 1989, 58 % de ces sujets sont dans la classe d'âge 41-45 ans, et 42 % dans la classe 46-50 ans. 16 % sont des employés ou des ouvriers, 55 % sont dans la classe des professions intermédiaires et 29 % sont des ingénieurs ou des cadres. 394 décès surviennent entre le début 1990 et la fin 1999 . Le tableau 1 fournit la répartition de ces décès par année. 288 hommes sont décédés entre 1989 et fin 1997 : la cause de décès est connue pour 265 d'entre eux : 114 décès par cancer, 55 décès par maladie cardio-vasculaire, 56 décès par mort violente et 40 autres causes de décès.

TABLEAU 1
Répartition par année des décès au cours de la période d'observation 1990-1999

Pop à risque au début de l'année	**Nombre**	**Décédés dans l'année**
1990	14 805	16
1991	14 789	33
1992	14 756	31
1993	14 725	39
1994	14 686	40
1995	14 646	45
1996	14 601	35
1997	14 566	49
1998	14 517	55
1999	14 462	51
Total		394

2.3. Le schéma des données de l'étude

Les cohortistes sont observés entre le début 1989 et la fin 1999. Pour chacun d'entre eux, on dispose des dates de réponses aux auto-questionnaires (AQ) annuels, de la mesure de l'ESP, de la date d'un éventuel décès et de la cause de ce décès s'il survient avant fin 1997.

D'autre part, on observe des données manquantes pour l'ESP à partir de 1990. Or dans la plupart des procédures utilisées, les observations individuelles où la valeur de l'ESP est manquante sont éliminées, ce qui revient à ignorer les observations incomplètes et à les considérer donc comme manquantes complètement au hasard [Schafer, (2000)]. Pour cette étude, nous avons convenu de remplacer l'ESP non déclaré une année donnée par la dernière valeur connue, en cas d'observations incomplètes.

3. Méthode statistique

3.1. Méthode utilisée pour l'étude du décès toutes causes confondues

a) Modèle continu de Cox avec covariables dépendant du temps

Il s'agit du modèle de base de Cox [(1972, 1975)] étendu au cas de covariables dépendant du temps [Klein *et al*, (1997)] et qui s'écrit :

$$\lambda[t/x(t)] = Pr(t \leqslant T < t + \mathrm{d}t | T \geqslant t) = \lambda_0(t) \exp[\beta' x(t)]$$

où T est le temps (continu) où se produit le décès, $\lambda[t/x(t)]$ est le taux de survie au temps t pour un sujet dont le vecteur des covariables dépendant du temps est $x(t)$ et $\lambda_0(t)$ est une fonction arbitraire du temps t.

b) Modèle discret de Cox avec covariables dépendant du temps

L'écriture du modèle de Cox se simplifie quand il s'agit de données de survie groupées : valeurs de la variable décès Y (codée 0 ou 1) notées à différents temps discrets jusqu'à la survenue du décès et mesures des covariables $(X_1, X_2, \ldots, X_p)$ relevées à ces mêmes temps discrets. En effet, soit K le nombre d'intervalles de temps où l'on relève l'information concernant le décès et Y_k la variable binaire qui prend la valeur 1 si le sujet est décédé dans l'intervalle k et, 0 si le sujet est vivant dans l'intervalle k, pour k variant de 1 à K. Le risque $\lambda[t/x(t)]$ s'écrit pour le sujet « i » sous la forme :

$$\lambda_{ik} = P(Y_k = 1 | Y_{k'} = 0, k' < k; x_{ik})$$

qui représente la probabilité de décès dans l'intervalle k, sachant que le sujet « i », dont le vecteur des covariables associé à l'intervalle k est x_{ik}, était vivant dans l'intervalle $k-1$.

Remarque importante. – On cherche à modéliser les probabilités conditionnelles λ_{ik} par des covariables x_{ik} dépendant du temps. Les mesures de ces variables ne sont pas nécessairement relevées au cours de l'intervalle k. On peut chercher à

modéliser, par exemple, la probabilité de décéder dans l'intervalle k par une covariable mesurée à $k-3$. Dans notre étude, compte tenu du problème posé, on a successivement considéré six modèles : décès dans l'intervalle k en fonction de l'état de santé mesuré dans l'intervalle $(k-j)$ pour j allant de 1 à 6.

Pour définir la contribution V_i d'un sujet « i » à la vraisemblance de l'échantillon, il faut distinguer deux cas :

(1) Le sujet « i » décède dans l'intervalle k $(k = 1, , K)$

Dans ce cas, la contribution V_i du sujet « i » à la vraisemblance de l'échantillon est :

$$V_{id} = P\{Y_1 = 0, Y_2 = 0, \ldots, Y_{k-1} = 0, Y_k = 1/(x_{i1}x_{i2}\ldots x_{ik})\}$$

qui s'écrit sous la forme du produit de probabilités suivant :

$$V_{id} = P(Y_1 = 0; x_{i1}) \cdot P(Y_2 = 0|Y_1 = 0; x_{i2}) \cdot P(Y_3 = 0|Y2 = 0, Y_1 = 0; x_{i3}).$$
$$\ldots\ldots P(Y_{k-1} = 0|Y_{k'} = 0, k' < k-1; x_{ik-1}) \cdot P(Y_k = 1|Y_{k'} = 0, k' < k; x_{ik})$$
$$V_{id} = \lambda_{ik}(1-\lambda_{ik-1})(1-\lambda_{ik-2})\ldots(1-\lambda_{i2})(1-\lambda_{i1}) = \lambda_{ik}\prod_{k'=1}^{k-1}(1-\lambda_{ik'})$$

(2) Le sujet « i » est vivant jusqu'à la fin de l'étude.

Dans ce cas, la contribution V_{ic} du sujet x_i à la vraisemblance de l'échantillon est :

$$V_{ic} = P\{Y_1 = 0, Y_2 = 0, \ldots, Y_{K-1} = 0, Y_K = 0/(x_{i1}x_{i2}\ldots x_{iK})\}$$
$$= P(Y_1 = 0; x_{i1}) \cdot P(Y_2 = 0|Y_1 = 0; x_{i2}) \cdot P(Y_3 = 0|Y_2 = 0, Y_1 = 0; x_{i3}).$$
$$\ldots\ldots P(Y_{K-1} = 0|Y_{k'} = 0, k' < K-1; x_{ik-1}).P(Y_K = 0|Y_{k'} = 0, k' < K; x_{ik})$$
$$V_{ic} = (1-\lambda_{ik})(1-\lambda_{ik-1})(1-\lambda_{ik-2})\ldots(1-\lambda_{i2})(1-\lambda_{i1}) = \prod_{k'=1}^{K}(1-\lambda_{ik'})$$

Remarque. – Si le sujet « i » est censuré dans l'intervalle k, alors :

$$V_{ic} = \prod_{k'}^{k}(1-\lambda_{ik'})$$

La vraisemblance de l'échantillon a pour expression :

$$V = \prod_{i=1}^{n} V_{id}^{\delta_i} V_{ic}^{(1-\delta_i)}$$

où $\delta_i = 1$ si le sujet « i » décède et $\delta_i = 0$ si le sujet « i » est vivant à la fin de l'étude ou censuré avant la fin de l'étude.

V s'écrit aussi sous la forme :

$$V = \prod_{i=1}^{n} \prod_{k'=1}^{k_i} \left(\frac{\lambda_{ik'}}{1 - \lambda_{ik'}} \right)^{y_{ik'}} (1 - \lambda_{ik'}) \quad \text{où :}$$

$y_{ik'} = 0$ si $k' < k_i$ et $y_{ik'} = 1$ si $k' = k_i$ pour un sujet « i » qui décède dans l'intervalle k_i,

$k_i = K$ et $y_{ik'} = 0$ quel que soit l'intervalle k' pour un sujet « i » vivant à la fin de l'étude

et $y_{ik'} = 0$ pour tout $k' \leqslant k_i$ pour un sujet censuré dans l'intervalle k_i.

Cette vraisemblance V est celle d'un modèle logistique binaire avec utilisation de la fonction de lien « complementary log-log » (CLL), appliqué à un fichier de données, où chaque sujet contribue pour k_i termes qui correspondent à k_i observations indépendantes (annexe 2). Cette propriété justifie la procédure logistique binaire classique. On montre d'autre part (annexe 2), que les fonctions de lien CLL et logit sont équivalentes quand les probabilités λ_{ik} sont petites ($\lambda_{ik} < 0.1$).

Le modèle logistique binaire (annexe A1.1) est donc utilisé dans le cas d'analyse de données de survie groupées, parce qu'il fournit la même vraisemblance que celle que l'on cherche à rendre maximum dans le modèle discret de Cox. De ce fait, les odds-ratio issus du modèle logistique sont les risques relatifs (RR) issus du modèle discret de Cox.

Remarque. – Le risque relatif de décès dans l'intervalle k, pour $x_{kj} = a$ *versus* $x_{kj} = b$, a pour expression : $\mathrm{RR}_k(x_{kj} = a / x_{kj} = b) = \dfrac{\lambda_k(x_{kj=a})}{\lambda_k(x_{kj=b})}$ où x_{kj} est la valeur prise par la variable j du vecteur à p composantes associé à l'intervalle k.

Le modèle logistique classique peut donc être utilisé dans ce cas, après avoir transformé les données de telle sorte que chaque intervalle k, pour chaque sujet, soit considéré comme une observation du fichier soumis au modèle logistique. La k-ième observation du sujet contient les valeurs des covariables relevées dans l'intervalle k et la valeur de la variable Y qui est égale à 1 (ou 0) suivant que le décès s'est produit (ou non) dans cet intervalle. Le nombre d'observations d'un sujet est donc égal à k si ce sujet décède ou est censuré dans l'intervalle k et à K si le sujet est vivant à la fin de la période d'observation considérée [Hosmer *et al.*, (1989)].

3.2. Méthode utilisée pour l'étude du décès par cause

Il s'agit de modéliser une variable réponse (statut vital) à plus de 2 modalités prenant les valeurs : 0 pour un sujet non décédé, 1 pour un décès par cancer, 2 pour un décès par maladie cardio-vasculaire, 3 pour une mort violente et 4 pour un décès d'une autre cause.

a) Méthode classique des « competing risks »

En pratique, dans ce cas d'une variable réponse à plus de 2 modalités, on est amené à faire l'hypothèse que les causes de décès sont indépendantes. On peut alors utiliser le modèle logistique binaire pour chaque cause de décès d_j, en considérant un sujet décédé d'une cause $d_{j'} \neq d_j = \{1, .., 4\}$ dans l'intervalle k, comme censuré dans l'intervalle $(k-1)$. Sa k-ième observation est donc tout simplement omise du fichier soumis à la procédure logistique binaire. C'est le principe des « competing risks » dont la théorie est plus complexe quand cette l'hypothèse d'indépendance entre causes de décès n'est pas respectée [Prentice *et al.*, 1978; David *et al.*, 1978].

b) Modèle logistique polytomique

Les modèles logistiques binaires définis en a) (un par cause de décès) fournissent des risques relatifs de décès pour chaque cause de décès comparée au groupe des vivants. Pour prendre en compte les différentes causes de décès simultanément dans un seul modèle, on peut utiliser le modèle logistique polytomique (annexe A1.2) qui permet de modéliser une variable réponse à plus de 2 modalités.

Le modèle logistique polytomique utilisé pour modéliser la variable réponse décès Y dont les modalités sont 0, 1, 2, 3, 4, où $Y = 0$ est le groupe des vivants pris comme référence et $Y = j$ le groupe des décédés (cause d_j), fournit 4 coefficients $\beta_1, \beta_2, \beta_3$ et β_4 associés à la covariable ESP. Les risques relatifs de décès (mauvaise santé *versus* bonne santé) $\mathrm{RR}_{j/0}$ $(j = 1, 2, 3, 4)$ qui se déduisent des coefficients β_j en prenant l'exponentielle, sont les risques relatifs de décès pour les différentes causes d_1, d_2, d_3 et d_4 comparées aux vivants.

Le modèle logistique polytomique est intéressant en ce sens qu'il permet, en plus, de comparer les risques relatifs de décès pour une cause comparée à une autre cause. Ainsi, pour obtenir les risques relatifs $\mathrm{RR}_{j/j'}$ correspondant à la cause d_j comparée à la cause $d_{j'}$, on calcule les coefficients $(\beta_j - \beta_{j'})$ pour $j \neq j' = 1, .., 4$ qui fournissent les risques relatifs respectifs $\mathrm{RR}_{j/j'} = \exp(\beta_j - \beta_{j'})$ et leurs intervalles de confiance IC95 % :

$$\exp\left[(\beta_j - \beta_{j'}) \pm 1.96 \times \sigma_{(\beta_j - \beta_{j'})}\right]$$

où $\sigma^2_{(\beta_j - \beta_{j'})}$ est la variance de $(\beta_j - \beta_{j'})$ égale à $\sigma^2_{\beta_j} + \sigma^2_{\beta'_j} - 2\mathrm{cov}(\beta_j, \beta'_j)$.

4. Résultats

4.1. Décès toutes causes confondues

a) Application du modèle continu de Cox avec covariables dépendant du temps

Nous avons utilisé la procédure PHREG du logiciel SAS [SAS/STAT (1997)] pour modéliser le délai de survie en fonction de l'ESP déclaré « j » années (j variant de 1 à 6) avant la survenue du décès et ce, en ajustant sur l'âge en 1989 et la PCS en

1989. Les résultats des six modèles sont résumés dans le tableau 2 où sont indiqués les temps de calcul (temps réel et temps CPU) requis pour le passage de chaque modèle.

TABLEAU 2
Modèle continu de Cox
Risques relatifs (RR) de* décès toutes causes confondues *au temps t en fonction de l'ESP déclaré* $(t-j)$ *années avant avec* $j = 1, .., 6$

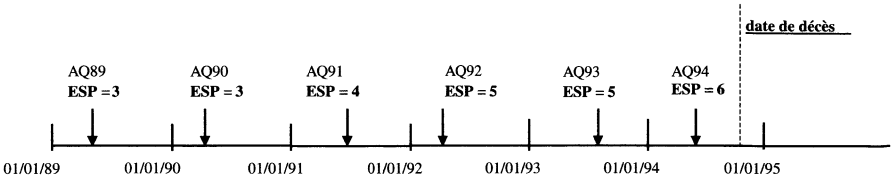
Santé à	Valeur	RR*	IC95 %		Temps réel	Temps CPU
$t-1$	1,2,3,4	1			23.10 min	1.14 min
	5,6,7,8	**2.826**	**2.303**	**3.467**		
$t-2$	1,2,3,4	1			21.58 min	1.11 min
	5,6,7,8	**2.341**	**1.888**	**2.903**		
$t-3$	1,2,3,4	1			20.43 min	1.04 min
	5,6,7,8	**1.910**	**1.511**	**2.414**		
$t-4$	1,2,3,4	1			18.18 min	58.46 sec
	5,6,7,8	**2.110**	**1.655**	**2.690**		
$t-5$	1,2,3,4	1			17.19 min	51.63 sec
	5,6,7,8	**1.716**	**1.305**	**2.259**		
$t-6$	1,2,3,4	1			17.12 min	44.12 sec
	5,6,7,8	**1.683**	**1.244**	**2.277**		

* ajusté sur l'âge en 89 et la PCS

b) Application du modèle discret de Cox avec covariables dépendant du temps

Avant d'appliquer la procédure logistique binaire classique, les données sont transformées comme suit : les décès sont regroupés par année, prise comme intervalle de temps. Le schéma des données fait l'objet de la figure 1.

Les données initiales sont regroupées par année. Chaque sujet se voit attribuer plusieurs enregistrements, un par année depuis 1990 jusqu'à l'année de décès, s'il survient pendant la période d'observation. Pour un sujet qui décède dans l'année t, la variable «décès» prend la valeur 0 (vivant) pour les années de 1990 à $t-1$ et la valeur 1 pour l'année t. Le tableau 3 fournit un exemple de transformation des données initiales pour un sujet (identifié A) dont le décès survient en 1994 et un autre sujet (identifié B) vivant dans toute la période d'observation [1989 , 1999].



Exemple : Décès dans l'année 1994

	1989	1990	1991	1992	1993	1994
Décès (toutes causes)	0	0	0	0	0	**1**
État de Santé Perçu (ESP)	**3**	**3**	**4**	**5**	**5**	**6**

FIGURE 1
Données groupées par année

TABLEAU 3
Exemple de transformation des données avant passage du modèle logistique classique sur données groupées

Sujet décédé en 1994								
ident	Année t	DCD à t	ESP à $t-1$	ESP à $t-2$	ESP à $t-3$	ESP à $t-4$	ESP à $t-5$	ESP à $t-6$
A1	90	0	ESP 89					
A2	91	0	ESP 90	ESP 89				
A3	92	0	ESP 91	ESP 90	ESP 89			
A4	93	0	ESP 92	ESP 91	ESP 90	ESP 89		
A5	94	1	ESP 93	ESP 92	ESP 91	ESP 90	ESP 89	

Sujet non décédé dans la période [1989,1999]								
ident	Année t	DCD à t	ESP à $t-1$	ESP à $t-2$	ESP à $t-3$	ESP à $t-4$	ESP à $t-5$	ESP à $t-6$
B1	90	0	ESP 89					
B2	91	0	ESP 90	ESP 89				
B3	92	0	ESP 91	ESP 90	ESP 89			
B4	93	0	ESP 92	ESP 91	ESP 90	ESP 89		
B5	94	0	ESP 93	ESP 92	ESP 91	ESP 90	ESP 89	
B6	95	0	ESP 94	ESP 93	ESP 92	ESP 91	ESP 90	ESP 89
B7	96	0	ESP 95	ESP 94	ESP 93	ESP 92	ESP 91	ESP 90
B8	97	0	ESP 96	ESP 95	ESP 94	ESP 93	ESP 92	ESP 91
B9	98	0	ESP 97	ESP 96	ESP 95	ESP 94	ESP 93	ESP 92
B10	99	0	ESP 98	ESP 97	ESP 96	ESP 95	ESP 94	ESP 93

Le sujet A est représenté dans le tableau 3 par 5 enregistrements A1,..., A5 correspondant aux années de 1990 à 1994, année de son décès. Le sujet B est lui représenté par 10 observations B1, B2, ..., B10 correspondant aux différentes années de la période d'observation (période calendaire). Les colonnes contiennent l'identification de l'observation annuelle, l'année d'observation, la valeur du statut vital à t et les valeurs de l'ESP déclarées l'année $t-1$, l'année $t-2,\ldots$, l'année $t-6$. À ces colonnes sont ajoutées des colonnes contenant les facteurs d'ajustement âge en 89 et PCS en 89 dont les valeurs sont les mêmes pour toutes les observations d'un même sujet.

Le tableau 4 présente les résultats de l'application de ce modèle logistique binaire : décès en fonction de l'ESP déclaré à une année avant, en ajustant sur l'âge en 1989, la catégorie socio-professionnelle et la période calendaire.

TABLEAU 4

Modèle discret de Cox (modèle logistique binaire)
Risques relatifs RR de* décès toutes causes confondues *et IC 95 % au temps t en fonction de l' ESP déclaré à (t − 1) (t variant de 90 à 99)*

Variable	**RR***	IC 95 %	
ESP déclaré à $t-1=(1,2,3,4)$	1		
ESP déclaré à $t-1=(5,6,7,8)$	3.12	2.55	3.82
Age en 89 41-45 (1)	**1**		
Age en 89 46-50 (2)	1.47	1.21	1.80
Ingénieurs. Cadres (3)	**1**		
Prof. intermédiaires (4)	1.36	1.05	1.76
Employés, Ouvriers (5)	2.02	1.50	2.72
Période 99	**1**		
Période **90**	**0.33**	**0.19**	**0.59**
Période 91	0.68	0.44	1.06
Période **92**	**0.64**	**0.41**	**1.00**
Période 93	0.78	0.52	1.19
Période 94	0.79	0.52	1.19
Période 95	0.88	0.59	1.32
Période 96	0.66	0.43	1.02
Période 97	0.94	0.64	1.40
Période 98	1.09	0.74	1.59

* ajusté sur l'âge en 89, la PCS et la période calendaire

Remarque. – L'ajustement sur la période calendaire est implicite dans le modèle continu de Cox : il est pris en compte dans le risque de décès de base $\lambda_0(t)$. Par contre dans le modèle discret de Cox, on doit le faire apparaître de manière explicite en mettant la période calendaire comme variable d'ajustement.

On observe, à la lecture du tableau 4, que le RR de décès (mauvaise santé versus bonne santé) est d'environ 3 et significatif, ce qui signifie que les sujets qui se déclarent en mauvaise santé une année ont un risque de décès 3 fois plus élevé l'année suivante. D'autre part, les sujets de la classe d'âge (46-50 ans) ont un risque de décès significativement plus élevé que ceux de la classe d'âge (41-45 ans). Les « employés et ouvriers » et les « professions intermédiaires » ont des risques significativement supérieurs à celui des « ingénieurs et cadres », Enfin le RR augmente avec la période calendaire, ce qui signifie tout simplement que le risque de décès augmente en vieillissant.

Ces RR ont été estimés jusqu'à $t - 6$ (en considérant l' ESP déclaré jusqu'à 6 ans avant la survenue du décès). Le tableau 5 et la figure 2 fournissent ces RR et leurs intervalles de confiance IC95 %, pour le décès en fonction de l'ESP déclaré 1 an avant, 2 ans avant, ..., jusqu'à 6 ans avant; ces RR sont ajustés sur l'âge en 89, la PCS et la période calendaire.

Les résultats des modèles du tableau 5 sont représentés sous forme de graphique dans la figure 2 avec en ordonnée, le risque relatif de décès (sujets en mauvaise santé versus sujets en bonne santé). Le RR de décès une année t est assez fort un an avant la survenue du décès (de l'ordre de 3), ce qui n'est pas surprenant. Deux ans avant il est plus petit (de l'ordre de 2). A partir de là et jusqu'à 6 ans avant la survenue de décès, il est pratiquement stable et significativement plus grand que 1.

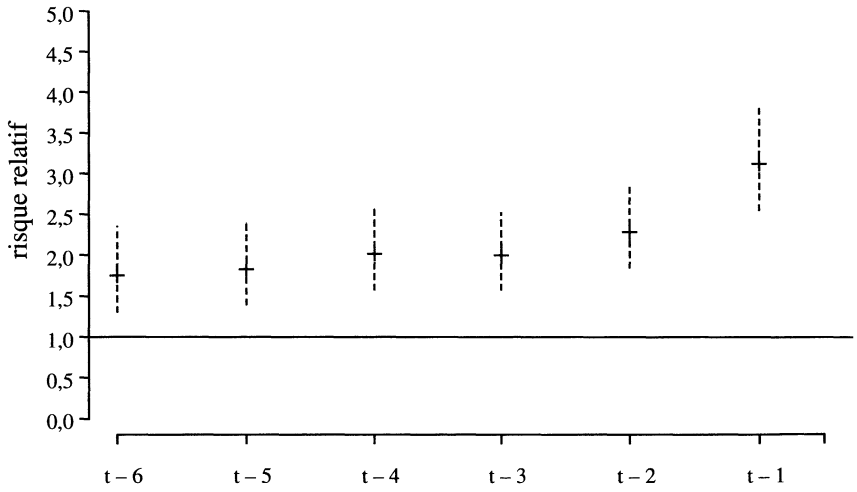


FIGURE 2
Modèle discret de Cox (Modèle logistique binaire)
Risques relatifs (RR) et intervalles de confiance IC95 % de* décès
(toutes causes confondues) *au temps t en fonction de l'ESP*
déclaré $(t - j)$ *années* $(j = 1, 2, \ldots, 6)$

* ajusté sur l'âge en 89, la PCS et la période calendaire

TABLEAU 5

Modèle discret de Cox (modèle logistique binaire)
Risques relatifs (RR) de* décès toutes causes confondues *et IC95 % au temps t en fonction de l'ESP déclaré à $(t-j)$ avec $j = 1, .., 6$*

Santé à	Valeur	RR*	IC95 %		Temps réel	Temps CPU
$t-1$	1,2,3,4	1			8.70 sec	7.29 sec
	5,6,7,8	**3.132**	**2.557**	**3.836**		
$t-2$	1,2,3,4	1			8.25 sec	6.61 sec
	5,6,7,8	**2.293**	**1.847**	**2.846**		
$t-3$	1,2,3,4	1			7.15 sec	5.71 sec
	5,6,7,8	**2.000**	**1.586**	**2.524**		
$t-4$	1,2,3,4	1			7.06 sec	5.09 sec
	5,6,7,8	**2.024**	**1.584**	**2.586**		
$t-5$	1,2,3,4	1			5.51 sec	4.06 sec
	5,6,7,8	**1.836**	**1.402**	**2.406**		
$t-6$	1,2,3,4	1			4.75 sec	3.25 sec
	5,6,7,8	**1.750**	**1.297**	**2.360**		

* ajusté sur l'âge en 89, la PCS et la période calendaire

La comparaison des résultats obtenus avec les modèles continu et discret de Cox met en évidence une différence très importante entre les temps requis pour les deux types de modèles : en moyenne 20 mn de temps réel et 1 mn de temps CPU pour les six modèles continus (tableau 2), alors que les temps calcul requis pour les six modèles discrets sont en moyenne de 7 sec et 5 sec respectivement (tableau 5). Par contre, les résultats concernant les risques relatifs sont assez semblables, d'où l'intérêt d'utiliser le modèle discret de Cox dans l'étude du décès par cause.

Remarque. – Le nombre de décès qui diminue d'année en année est insuffisant pour permettre l'analyse au-delà de $t-6$ car on se heurterait à un problème de puissance statistique et donc un manque de précision pour les estimations des risques relatifs.

4.2. Décès par cause

Les analyses ont été effectuées à partir du fichier des données précédemment transformées avec pour un sujet décédé, une valeur de 1 à 4 suivant sa cause de décès pour la variable réponse Y qui est donc codée ainsi : 0 = vivant, 1 = décès (tumeur), 2 = décès (maladie cardiovasculaire), 3 = décès (mort violente) et 4 = décès (autres causes).

Le tableau 6 contient les résultats des modèles logistiques binaires : un modèle par cause de décès d_j (1, 2, 3 ou 4), en censurant à $(k-1)$ les sujets qui décèdent dans l'intervalle k d'une autre cause que d_j.

Le tableau 7 fournit les résultats obtenus à l'aide du modèle logistique polytomique utilisé pour expliquer le décès en t (codé en 0, 1, 2, 3, 4) en fonction de l'ESP à $(t-1)$. Pour chaque modalité d'une variable explicative incluse dans le modèle (à l'exception de l'une d'entre elles choisie comme modalité de référence) on a autant de RR et IC95 % que de modalités de la variable réponse Y (causes de décès) moins la modalité de référence (vivant). Les risques relatifs des 4 premières lignes de ce tableau 7 (3.70, 2.01, 2.01 et 5.00) correspondent aux risques relatifs de décès d'un mauvais état de santé *versus* un bon état de santé pour chacune des causes de décès (décès par tumeur ; décès par maladie cardio-vasculaire ; décès par mort violente ; décès par autre cause).

Le tableau 8 contient les résultats du modèle logistique polytomique concernant les RR de décès par cause (et les IC95 %) au temps t, en fonction de l'état de santé déclaré $(t-j)$ années avant, pour j variant de 1 à 6.

Les résultats des deux types d'analyse – modèle logistique binaire par cause de décès avec censure des autres causes (tableau 6) et modèle logistique polytomique (tableau 8) – sont presque identiques. Mais l'analyse a nécessité, dans le premier cas, 24 (6×4) passages du modèle logistique binaire et dans le deuxième cas, six passages du modèle logistique polytomique avec un temps de calcul très raisonnable.

L'examen de la figure 3, qui représente les résultats du tableau 8 sous forme graphique, montre que pour les décès par cancers, le RR est relativement élevé un an avant ; au fur et à mesure que la santé perçue est considérée longtemps avant le décès, ce RR est plus faible et non significatif à partir de 4 ans avant. En ce qui concerne les décès par maladies cardio-vasculaires, le RR est plus constant. Il devient non significatif 4 ans avant, comme pour le décès par cancers. Pour les décès par morts violentes, on observe un RR constant et assez élevé (de l'ordre de 2 à 3). Enfin le RR est le plus fort pour les décès par autres causes (de l'ordre de 4), ce qui ne semble pas étonnant puisqu'il s'agit principalement de maladies chroniques (diabète, sida, cirrhose,...).

Remarque. – Si on s'intéresse à l'effet prédictif de l'ESP déclaré 4 ans avant la survenue de décès, on observe que, dans la comparaison au groupe des vivants, les RR sont les plus faibles et non significatifs pour les cancers et les maladies cardio-vasculaires, alors que pour les morts violentes ils sont plus élevés (de l'ordre de 2,5) et encore plus élevés pour les autres causes (de l'ordre de 5).

TABLEAU 6

*Modèles logistiques binaires***

Risques relatifs (RR) de* décès par cause *au temps t en fonction de l'ESP déclaré* ($t-j$) *années avant, avec* $j = 1, \ldots, 6$

	$t-1$		$t-2$		$t-3$		$t-4$		$t-5$		$t-6$	
Statut vital	RR*	IC95 %	RR*	IC95 %	RR*	IC95 %	RR*	IC95 %	RR*	IC95 %	RR*	IC95 %
Non décédés	1		1		1		1		1		1	
Tous cancers	3.70	2.55 5.37	2.38	1.59 3.55	1.64	1.04 2.59	1.56	0.96 2.56	1.63	0.98 2.70	1.20	0.62 2.31
Maladies card vasc	2.01	1.13 3.58	1.99	1.08 3.66	2.01	1.02 3.96	1.19	0.52 2.72	1.13	0.43 2.97	1.42	0.47 4.27
Morts violentes	2.01	1.12 3.60	1.79	0.95 3.57	1.92	0.96 3.82	2.42	1.15 5.09	2.49	1.09 5.68	3.07	1.16 8.12
Autres causes	5.00	2.67 9.35	3.54	1.88 6.67	4.91	2.56 9.41	5.09	2.50 10.36	5.67	2.61 12.31	3.46	1.44 8.30

* *ajusté sur l'âge en 89, la PCS et la période calendaire*

** pour la cause de décès d_j , les sujets décédés dans l'intervalle k d'une cause $d_{j'}$ différente de d_j sont censurés dans l'intervalle $(k-1)$

TABLEAU 7
Modèle logistique polytomique
Risques relatifs (RR) de* décès par cause *et IC 95 % au temps t en fonction de l' ESP(à t − 1) (t allant de 90 à 99)*

				RR*	IC95 %	
ESP_1 MS vs BS	cancers	vs	non dcd	3.70	2.55	5.37
ESP_1 MS vs BS	cardio-vasc	vs	non dcd	2.01	1.13	3.58
ESP_1 MS vs BS	morts violentes	vs	non dcd	2.01	1.12	3.61
ESP_1 MS vs BS	autres causes	vs	non dcd	5.00	2.68	9.35
1990 vs 1997	cancers	vs	non dcd	0.26	0.11	0.65
1990 vs 1997	cardio-vasc	vs	non dcd	0.65	0.21	1.98
1990 vs 1997	morts violentes	vs	non dcd	0.87	0.26	2.85
1990 vs 1997	autres causes	vs	non dcd	...	...	...
1991 vs 1997	cancers	vs	non dcd	0.39	0.18	0.84
1991 vs 1997	cardio-vasc	vs	non dcd	1.16	0.45	3.00
1991 vs 1997	morts violentes	vs	non dcd	1.38	0.48	4.00
1991 vs 1997	autres causes	vs	non dcd	0.43	0.11	1.61
.....	.....			...	...	...
1996 vs 1997	cancers	vs	non dcd	0.70	0.38	1.29
1996 vs 1997	cardio-vasc	vs	non dcd	0.62	0.20	1.88
1996 vs 1997	morts violentes	vs	non dcd	0.49	0.12	1.98
1996 vs 1997	autres causes	vs	non dcd	0.98	0.37	2.60
AGE89 2 vs 1	cardio-vasc	vs	non dcd	0.64	0.38	1.09
AGE89 2 vs 1	cancers	vs	non dcd	0.61	0.42	0.88
AGE89 2 vs 1	morts violentes	vs	non dcd	0.89	0.53	1.52
AGE89 2 vs 1	autres causes	vs	non dcd	0.70	0.38	1.31
PCS 4 vs 3	cancers	vs	non dcd	1.63	0.98	2.69
PCS 4 vs 3	cardio-vasc	vs	non dcd	0.64	0.32	1.26
PCS 4 vs 3	morts violentes	vs	non dcd	1.51	0.78	2.91
PCS 4 vs 3	autres causes	vs	non dcd	1.80	0.78	4.17
PCS 5 vs 3	cancers	vs	non dcd	2.53	1.44	4.46
PCS 5 vs 3	cardio-vasc	vs	non dcd	2.58	1.33	5.01
PCS 5 vs 3	morts violentes	vs	non dcd	1.28	0.54	3.04
PCS 5 vs 3	autres causes	vs	non dcd	1.76	0.63	4.88

* ajusté sur l'âge en 89, la PCS et la période calendaire

TABLEAU 8

Modèle logistique polytomique

*Risques relatifs (RR**) de décès par cause et IC95 % au temps t en fonction de l'ESP déclaré ($t-j$) années avant avec $j = 1, .., 6$*

	$t-1$		$t-2$	p	$t-3$		$t-4$		$t-5$		$t-6$	
Statut vital	RR*	IC95 %	RR*	IC95 %	RR*	IC95 %	RR*	IC95 %	RR*	IC95 %	RR*	IC95 %
Non décédés	1		1		1		1		1		1	
Tous cancers	3.70	2.55 5.37	2.38	1.59 3.55	1.65	1.04 2.60	1.57	0.96 2.56	1.63	0.98 2.70	2.70	0.62 2.32
Maladies card vasc	2.01	1.13 3.58	1.99	1.08 3.68	2.01	1.02 3.97	1.19	0.52 2.73	1.13	0.43 2.98	1.43	0.48 4.28
Morts violentes	2.01	1.12 3.61	1.79	0.95 3.38	1.92	0.97 3.83	2.43	1.16 5.11	2.49	1.09 5.70	3.07	1.16 8.14
Autres causes	5.00	2.68 9.35	3.54	1.89 6.69	4.92	2.57 9.43	5.10	2.51 10.38	5.68	2.61 12.3	3.46	1.44 8.32

** ajusté sur l'âge en 89, la PCS et la période calendaire.

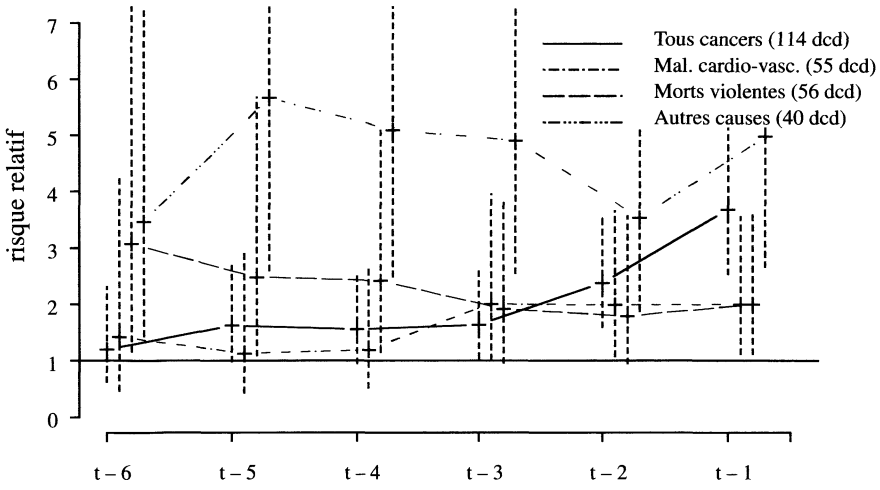


FIGURE 3

Modèle logistique polytomique

Risques relatifs (RR) et intervalles de confiance IC95 % de* décès par cause *au temps t en fonction de l'ESP déclaré* $(t-j)$ *années avant* $(j = 1, 2, \ldots, 6)$

* ajusté sur l'âge en 89, la PCS et la période calendaire.

D'autre part, les estimations des variances covariances des coefficients du modèle d'explication du décès à t en fonction de l'ESP 4 ans avant sont présentées dans le tableau 9. Elles permettent de calculer les risques relatifs (mauvaise santé/bonne santé) $\mathrm{RR}_{j/j'}$ de décès (cause d_j) comparés aux décès (cause $d_{j'}$ et leurs intervalles de confiance au seuil 95 %. L'examen de ces $\mathrm{RR}_{j/j'}$ qui font l'objet du tableau 10 montre que les seuls risques significatifs sont $\mathrm{RR}_{\text{autres causes/cancers}}$ et $\mathrm{RR}_{\text{autres causes/maladies cardio-vasculaires}}$.

TABLEAU 9

ESP déclaré 4 ans avant la survenue de décès

Variances covariances estimées des coefficients du modèle

Cov $(\beta_j, \beta_{j'})$	β_1 (cancers)	β_2 (card. vasculaires)	β_3 (morts violentes)	β_4 (autres causes)
β_1 (cancers)	0.063 248	0.000 134	0.000 110	0.000 116
β_1 (cardio-vasc.)	0.000 134	0.179 104	0.000 109	0.000 107
β_1 (morts violentes)	0.000 110	0.000 109	0.143 759	0.000 106
β_1 (autres causes)	0.000 116	0.000 107	0.000 106	0.131 403

TABLEAU 10
ESP déclaré 4 ans avant la survenue de décès
Risques relatifs $RR_{j/j'}$ et intervalles de confiance IC95 % de décès (cause d_j) comparés aux décès (cause $d_{j'}$)

Décès	Tous cancers		Maladies card vasc		Morts violentes	
	$RR_{j/j'}$	IC95 %	$RR_{j/j'}$	IC95 %	$RR_{j/j'}$	IC95 %
Tous cancers	–	–				
Maladies card vasc	0.76	0.29 1.99	–	–		
Morts violentes	1.55	0.64 3.78	2.04	0.67 6.21	–	–
Autres causes	3.26	**1.37 7.73**	3.54	**1.44 12.7**	2.1	0.75 5.87

5. Conclusion

Le modèle discret de Cox est très intéressant par rapport au modèle continu de Cox car il réduit de façon très importante les temps de calcul tout en fournissant de bons résultats, en permettant en plus l'introduction de plusieurs facteurs d'ajustement.

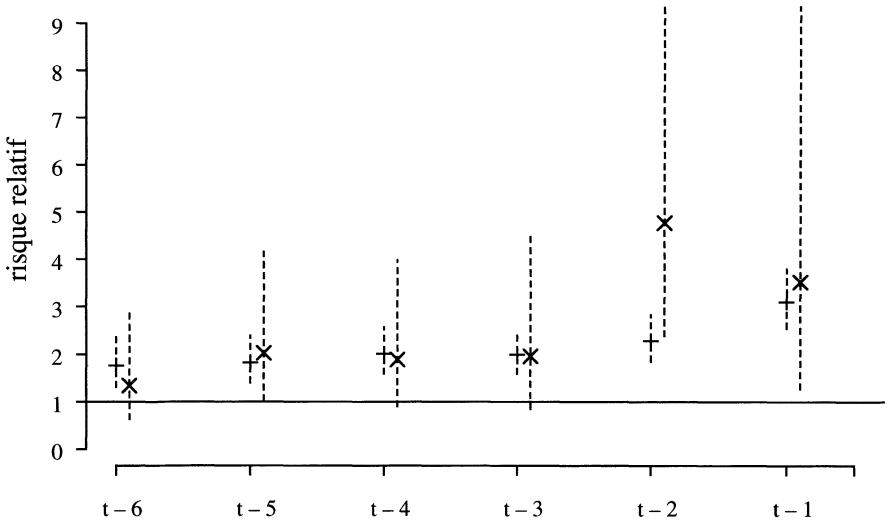
Le modèle logistique polytomique permet d'étudier plusieurs causes de décès simultanément, mais son utilisation est surtout justifiée par le fait qu'il permet de comparer les risques relatifs de décès d'une cause par rapport à une autre cause.

D'autre part, la variable ESP a été transformée dans notre étude en variable binaire en prenant le codage suivant : bon état de santé (ESP = 1,2,3,4) et mauvais état de santé (ESP = 5,6,7,8) : d'autres codages en 2 ou 3 classes pourraient être testés.

En ce qui concerne les données manquantes, nous avons choisi dans cette étude la solution simple qui consiste à estimer l'état de santé perçu non déclaré une année par la dernière valeur connue, mais on pourrait utiliser dans une étude ultérieure une méthode récente connue sous le nom de méthode des imputations multiples [Schafer J.L. (2000)], permettant de garder tous les sujets en fournissant des estimations des risques relatifs non biaisées.

Dans la plupart des études (études courantes), on ne dispose que d'une mesure de chaque variable explicative – mesure relevée en général au début de l'étude – et non de mesures de ces variables relevées à différents temps au cours de l'étude. Il était donc intéressant de regarder les résultats obtenus en n'utilisant que la mesure de l'ESP à l'inclusion des sujets en 1989, ce qui revient à considérer plusieurs modèles : décès en t par rapport à l'ESP déclaré en 1989 pour t variant de 1 à 6, et ce dans le but de répondre à la question : l'ESP prédit-il le décès t années plus tard ? Les résultats correspondants sont représentés sous forme graphique dans la figure 4 qui montre que les RR obtenus ont des intervalles de confiance très larges, comparés à ceux obtenus en utilisant les données longitudinales de l'ESP.

L'originalité des résultats obtenus dans cette étude tient à la spécificité de la cohorte Gazel et plus particulièrement au fait que, chaque année les cohortistes remplissent un auto-questionnaire qui contient la mesure de l'ESP.



\+ Modèle discret de Cox avec ESP dépendant du temps : décès au temps t en fonction de l'ESP $(t-j)$ années avant $(j = 1,2,.,6)$

× Modèle classique de Cox avec ESP fixe : décès au temps $t-j$ en fonction de l'ESP 1989 $(j = 1,2,.,6)$

FIGURE 4
Risques relatifs (RR) et intervalles de confiance IC95 % de* décès toutes causes confondues

* ajusté sur l'âge en 89, la PCS et la période calendaire.

Références

COX D.R. (1972), *Regression models and life-tables (with discussion)*, Journal of the Royal Statistical Society, Series B, 34 :187-220.

COX D.R. (1975), *Partial likelihood*, Biometrika, 62 : 269-276.

DAVID H.A., MOESCHBERGER M.L. (1978), *The theory of Competing Risks*, Griffin's Statistical Monographs, 39, MacMillan, New York.

GOLDBERG M., LECLERC A. (1994), *Cohorte GAZEL, 20 000 volontaires d'EDF-GDF pour la recherche médicale. Bilan 1989-1993*, Les Éditions INSERM, Paris.

GOLDBERG M., CHASTANG J.F., LECLERC A., ZINS M., BONENFANT S., BUGEL I., KANIEWSKI N., SCHMAUS A., NIEDHAMMER I., PICIOTTI

M., CHEVALIER A., GODARD C., IMBERNON E. (2001), *Socio-economic, demographic, occupational and health factors associated with participation in a long-term epidemiologic survey. A prospective study of the French Gazel Cohort and its target population*, Am. J. Epidemiol. 154; 373-84.

HOSMER D.W., LEMESHOW S. (1989), *Applied logistic regression*, New York : John Wiley & Sons, Inc.

IDLER E.L., BENYAMINI Y. (1997), Self-Rated Health and Mortality : A Rewiew of Twenty-Seven Community Studies Journal of Health and Social Behaviour.

KLEIN J.P., MOESCHBERGER M.L. (1997), *Survival Analysis : Techniques for censored and truncated data*, Springer (Statistics for Biology and Health); 271-282.

PRENTICE R.L., KALBFLEICH J.D., PETERSON A.V., FLOURNOY N., FAREWELL V.T., BRESLOW N.E. (1978), *The Analysis of Failure Time Data in the presence of Competing Risks*, Biometrics, 34, 541-554.

PRENTICE R.L., GLOECKLER L.A. (1978), *Regression analysis of grouped survival data with applications to breast cancer data*, Biometrics, 34; 57-67.

SAS/STAT Software (1997), *Changes and Enhancements through Release 6.12*, SAS Institute Inc., Cary, NC (USA).

SCHAFER J.L. (2000), *Analysis of Incomplete Multivariate Data*, Chapman & Hall/ CRC.

WU M., WARE J.H. (1979), *On the use of repeated measurements in regression analysis with dichotomous responses*, Biometrics; 35 : 513-521.

Annexe 1

A1.1 Modèle logistique : variable réponse binaire (k = 2 modalités)

Il s'agit de modéliser une variable réponse binaire Y prenant les valeurs 0 ou 1 par p variables explicatives $X_1, X_2, \ldots, X_p$ en utilisant les informations d'un échantillon de n réalisations $x_1, x_2, \ldots, x_p$ formant le vecteur x :

$$\pi(x) = P(Y = 1|x)$$

On utilise, dans le modèle logistique binaire, le logit $g(x)$ comme fonction de lien entre $\pi(x)$ et les p variables explicatives, dont l'expression est :

$$g(x) = \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \ldots + \beta_j x_j + \ldots + \beta_p x_p$$

On obtient des estimations $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_j, \ldots, \hat{\beta}_p$, des paramètres $\beta_0, \beta_1, \ldots, \beta_j, \ldots, \beta_p$, en rendant maximum la vraisemblance de l'échantillon $\{(Y_i, x_i); i = 1, 2, \ldots n\}$.

La contribution du couple (Y_i, x_i) à la fonction de vraisemblance de l'échantillon est $V_i(\beta) = \pi(x_i)^{\delta_i}[1 - \pi(x_i)]^{1-\delta_i}$ qui vaut $\pi(x_i)$ si $Y_i = 1$ et $1 - \pi(x_i)$ si $Y_i = 0$. Les individus de l'échantillon étant indépendants, la vraisemblance s'écrit comme le produit des contributions des n sujets de l'échantillon.

Relation entre odds-ratio et paramètres du modèle logistique

L'estimation du odds-ratio du groupe $Y = 1$ *versus* le groupe de référence $Y = 0$, pour la valeur $X_j = a$ *versus* la valeur $X_j = b$ de la variable explicative X_j est liée au paramètre $\hat{\beta}_j$ associé à la variable X_j par la formule :

$$\hat{\Psi}(X_j = a/X_j = b) = \mathrm{e}^{\hat{\beta}_j(a-b)}$$

Le odds-ratio exprime l'intensité de la liaison entre la variable réponse Y et la variable explicative auquel il est associé. Cette liaison n'est pas significative (au risque $\alpha = 5\,\%$) si l'intervalle de confiance du odds-ratio $\mathrm{e}^{(a-b)(\hat{\beta}_j \pm 1.96\hat{\sigma}_{\hat{\beta}_j})}$ (où $\hat{\sigma}_{\hat{\beta}_j}$ est l'estimation de l'écart-type du coefficient $\hat{\beta}_j$) contient la valeur 1.

Remarque. – Si X_j est une variable binaire (codée en 1 ou 0), $\hat{\Psi}(1/0) = \hat{\Psi} = \mathrm{e}^{\hat{\beta}_j}$.

A1.2 Modèle logistique polytomique : variable réponse nominale à k modalités (k > 2)

Les modalités de la variable réponse Y étant codées en $0, 1, 2, \ldots, k$ avec $Y = 0$ prise comme modalité de référence et le vecteur des covariables de longueur $p + 1$ étant noté $x' = (x_0, x_1, \ldots, x_p)$ avec $x_0 = 1$ associé au coefficient β_0 du modèle, on définit les k logits suivants :

$$g_r(x) = \ln\left[\frac{P(Y = r|x)}{P(Y = 0|x)}\right] = \beta_{r0} + \beta_{r1}x_1 + \ldots + \beta_{rj}x_j + \ldots + \beta_{rp}x_p \quad (r = 1, 2, \ldots, k)$$

Probabilités conditionnelles des différentes modalités de Y étant donné x

$$\pi_0(x) = P(Y = 0|x) = \frac{1}{1 + \sum_{r=1}^{k} \exp[g_r(x)]}$$

$$\pi_r(x) = P(Y = r|x) = \frac{\exp[g_r(x)]}{1 + \sum_{r=1}^{k} \exp[g_r(x)]} \quad (r = 1, 2, \ldots, k)$$

La vraisemblance d'un échantillon de n observations indépendantes a pour expression :

$$V(\beta) = \prod_{i=1}^{n} V_i(\beta) = \prod_{i=1}^{n} \left[\prod_{r=0}^{n} \pi_r(x_i)^{y_{ir}} \right]$$

où $y_{ir} = 1$ si $Y = r$ pour le sujet « i » et $y_{ir} = 0$ si $Y \neq r$ pour le sujet « i ».

Odds-ratios

Si $Y = 0$ est prise comme modalité de référence de la variable Y, on définit pour la variable explicative X_j du modèle, les k odds-ratios suivants :

$$\Psi_r(X_j = a/X_j = b) = \mathrm{e}^{\beta_{rj}(a-b)} \quad (r = 1, 2, \ldots, k)$$

qui représentent les odds ratio du groupe $Y = r$ *versus* le groupe de référence $Y = 0$, pour la valeur $X_j = a$ *versus* la valeur $X_j = b$ de la variable explicative X_j, pour r variant de 1 à k.

Remarque. – Si X_j est une variable binaire (codée en 1 ou 0) :

$$\Psi_r(1/0) = \Psi_r = \mathrm{e}^{\beta_{rj}} \quad (r = 1, 2, \ldots, k)$$

Annexe 2

Vraisemblance de l'échantillon : données de survie groupées par année

$$V = \prod_{i=1}^{n} V_{id}^{\delta_i} V_{ic}^{(1-\delta_i)}$$

où $\delta_i = 1$ si le sujet décède et $\delta_i = 0$ si le sujet est vivant à la fin de l'étude ou censuré avant la fin de l'étude avec :

$V_{id} = \lambda_{ik} \prod_{k'=1}^{k_i - 1}(1 - \lambda_{ik'})$ si le sujet « i » décède dans l'intervalle k_i et,

$V_{ic} = \prod_{k'=1}^{k_i}(1 - \lambda_{ik'})$ si le sujet est censuré dans l'intervalle $k_i (1 \leqslant k_i < K)$ ou s'il est vivant à la fin de l'étude, et dans ce dernier cas $k_i = K$.

Cette vraisemblance s'écrit aussi :

$$V = \prod_{i=1}^{n} \prod_{k'=1}^{k_i} \left(\frac{\lambda_{ik'}}{1 - \lambda_{ik'}} \right)^{y_{ik'}} (1 - \lambda_{ik'})$$

V est la vraisemblance d'un échantillon de $\sum_{i=1}^{n} k_i$ observations (correspondant à k_i observations par sujet avec : $Y_{ik'} = 0$ pour $k' < k_i$ et $Y_{ik'} = 1$ pour $k' = k_i$ si le sujet décède en k_i, $Y_{ik'} = 0$ pour $k' \leqslant k_i$ si le sujet est censuré en k_i et $Y_{ik'} = 0$ pour $k' \leqslant K$ si le sujet est vivant à la fin de l'étude, auquel cas $k_i = K$) dans la modélisation d'une variable réponse binaire $Y_{ik'}$ avec $Pr(Y_{ik'} = 1) = \lambda_{ik'}$.

Il s'agit donc de modéliser les probabilités $\lambda_{ik'}$ en fonction des $x_{ik'}$. On peut, pour ce faire, utiliser différentes fonctions de lien dont les plus courantes sont la transformation « complementary log-log » (CLL) et la transformation « logit ».

Par ailleurs, Prentice et Gloeckler [1978] ont montré que si le modèle des risques proportionnels de Cox avec temps continu est vérifié alors :

$$\lambda_{ik'} = 1 - \exp[-\exp(\alpha_{k'} + \beta' x_i)]$$

où x_i est le vecteur des covariables qui ne dépendent pas du temps, β est le vecteur des paramètres identique à celui du modèle continu de Cox et $\alpha_{k'}$ est une constante correspondant à la probabilité de « survie » conditionnelle dans l'intervalle k' pour $x_i = 0$.

Or : $1 - \lambda_{ik'} = \exp[-\exp(\alpha_{k'} + \beta' x_i)]$, d'où $\log[-log(1 - \lambda_{ik'})] = \alpha_{k'} + \beta' x_i$ qui est la fonction de lien (CLL) entre les probabilités $\lambda_{ik'}$ et la combinaison linéaire des covariables connue sous le nom de (CLL).

Le modèle de Cox pour données de survie groupées est donc équivalent à un modèle d'explication d'une variable réponse binaire avec la transformation CLL comme fonction de lien entre les probabilités $\lambda_{ik'}$ et la combinaison linéaire des covariables indépendantes du temps.

Wu et Ware [1979] ont appliqué le modèle logistique binaire avec la fonction de lien CLL pour des covariables dépendant du temps. Dans ce cas, la k-ième observation d'un sujet contient les valeurs des covariables relevées dans l'intervalle k et la valeur de la variable Y qui est égale à 1 (ou 0) suivant que le décès s'est produit (ou non) dans cet intervalle. Le nombre d'observations d'un sujet est donc égal à k si ce sujet décède dans l'intervalle k , à k_i si le sujet est censuré dans l'intervalle et à K si le sujet est vivant à la fin de l'étude.

D'autre part si les probabilités $\lambda_{ik'}$ sont petites pour chaque intervalle ($\lambda_{ik'} < 0.1$), la modélisation d'une variable réponse binaire avec la fonction de lien CLL est équivalente à la modélisation de cette variable réponse binaire avec la fonction de lien logit : en effet, quand $\lambda_{ik'}$ est petit $-\ln(1 - \lambda_{ik'}) = -\ln[1 + (-\lambda_{ik'})]$ est équivalent à $\lambda_{ik'}$ et le rapport $\lambda_{ik'}/(1 - \lambda_{ik'})$ est équivalent à $\lambda_{ik'}$; donc $-\ln(1 - \lambda_{ik'})$ est équivalent à $\lambda_{ik'}/(1 - \lambda_{ik'})$.