Nº 12-001-XIF au catalogue

Techniques d'enquête

Juin 2005

Statistique Canada
Statistics Canada
Canada

Comment obtenir d'autres renseignements

Toute demande de renseignements au sujet du présent produit ou au sujet de statistiques ou de services connexes doit être adressée à : Division des méthodes d'enquêtes auprès des entreprises, Statistique Canada, Ottawa, Ontario, K1A 0T6 (téléphone : 1 800 263-1136).

Pour obtenir des renseignements sur l'ensemble des données de Statistique Canada qui sont disponibles, veuillez composer l'un des numéros sans frais suivants. Vous pouvez également communiquer avec nous par courriel ou visiter notre site Web.

Service national de renseignements	1 800 263-1136
Service national d'appareils de télécommunications pour les malentendants	1 800 363-7629
Renseignements concernant le Programme des services de dépôt	1 800 700-1033
Télécopieur pour le Programme des services de dépôt	1 800 889-9734
Renseignements par courriel	infostats@statcan.ca
Site Web	www.statcan.ca

Renseignements pour accéder au produit

Le produit nº 12-001-XIF au catalogue est disponible gratuitement. Pour obtenir un exemplaire, il suffit de visiter notre site Web à www.statcan.ca et de choisir la rubrique Nos produits et services.

Normes de service à la clientèle

Statistique Canada s'engage à fournir à ses clients des services rapides, fiables et courtois, et ce, dans la langue officielle de leur choix. À cet égard, notre organisme s'est doté de normes de service à la clientèle qui doivent être observées par les employés lorsqu'ils offrent des services à la clientèle. Pour obtenir une copie de ces normes de service, veuillez communiquer avec Statistique Canada au numéro sans frais 1 800 263-1136. Les normes de service sont aussi publiées dans le site www.statcan.ca sous À propos de Statistique Canada > Offrir des services aux Canadiens.

Statistique Canada

Division des méthodes d'enquêtes auprès des entreprises

Techniques d'enquête

Juin 2005

Publication autorisée par le ministre responsable de Statistique Canada

Mai 2006

Nº 12-001-XIF au catalogue
ISSN 1712-5685

Périodicité : semestriel

Ottawa

This publication is available in English upon request (catalogue no. 12-001-XIE)

Note de reconnaissance

Le succès du système statistique du Canada repose sur un partenariat bien établi entre Statistique Canada et la population, les entreprises, les administrations canadiennes et les autres organismes. Sans cette collaboration et cette bonne volonté, il serait impossible de produire des statistiques précises et actuelles.

Vol. 31, N° 1, pp. 15-24
Statistique Canada, N° 12-001 au catalogue

L'effet des erreurs de couplage d'enregistrements sur les estimations du risque dans les études-cohorte de mortalité

D. Krewski, A. Dewanji, Y. Wang, S. Bartlett, J.M. Zielinski et R. Mallick [1]

Résumé

L'élaboration de la méthodologie de couplage informatisé d'enregistrements a facilité la réalisation d'études-cohorte de mortalité dans lesquelles les données sur l'exposition provenant d'une base de données sont couplées électroniquement à celles sur la mortalité provenant d'une autre base de données. Cependant, cette méthode donne lieu à des erreurs de couplage causées par l'appariement incorrect d'une personne figurant dans l'une des bases de données à une personne différente dans l'autre base de données. Dans le présent article, nous examinons l'effet des erreurs de couplage sur les estimations d'indicateurs épidémiologiques du risque, comme les ratios standardisés de mortalité et les paramètres des modèles de régression du risque relatif. Nous montrons que les effets sur les nombres observé et attendu de décès sont de sens opposé et que, par conséquent, ces indicateurs peuvent présenter un biais et une variabilité supplémentaire en présence d'erreurs de couplage.

Mots clés : Étude de cohorte; couplage informatisé d'enregistrements; erreurs de couplage; poids seuil de couplage; régression de Poisson; régression du risque relatif; ratio standardisé de mortalité.

1. Introduction

Ces dernières années, plusieurs études de cohorte historiques ont été réalisées en épidémiologie environnementale en se servant de bases de données administratives existantes comme sources d'information (Howe et Spasoff 1986; Carpenter et Fair 1990). En termes généraux, cette approche consiste à coupler des enregistrements de données sur l'exposition humaine à des risques environnementaux à des enregistrements de données sur l'état de santé, souvent au moyen de méthodes informatisées d'appariement d'enregistrements individuels provenant de bases de données différentes. Dans le cas d'une étude-cohorte de mortalité, le statut vital de chaque membre de la cohorte est déterminé par couplage aux enregistrements de décès des bases de données sur la mortalité tenues à jour par les organismes gouvernementaux. L'existence d'une surmortalité dans la cohorte comparativement à la population générale pourrait être due aux expositions subies par les membres de la cohorte.

En termes spécifiques, le couplage d'enregistrements est le processus consistant à regrouper deux ou plusieurs éléments d'information enregistrés distincts concernant une même entité (Bartlett, Krewski, Wang et Zielinski 1993). Les procédures de couplage informatisé d'enregistrements (CIE) sont devenues de plus en plus perfectionnées, grâce à l'utilisation d'algorithmes complexes pour évaluer la probabilité que l'appariement de deux enregistrements soit correct (Hill 1988; Newcombe 1988). Statistique Canada a mis au point un système de CIE appelé CANLINK capable de coupler les enregistrements d'un même fichier, ainsi que ceux de deux fichiers distincts (Howe et Lindsay 1981; Smith et Silins 1981). Ce système attribue à chaque paire d'enregistrements un poids reflétant la probabilité qu'il s'agisse d'un appariement. Deux seuils sont fixés : les appariements potentiels dont le poids de couplage est supérieur au seuil supérieur sont considérés comme des couplages, tandis que les appariements potentiels dont le poids de couplage est inférieur au seuil inférieur sont considérés comme des non-couplages. Les cas d'appariement possible dont le poids est compris entre les seuils inférieur et supérieur sont résolus à l'aide de renseignements supplémentaires, lorsqu'ils sont disponibles. Sinon, on choisit un seuil unique pour faire la distinction entre les couplages et les non-couplages.

Lors de toute étude comportant un couplage d'enregistrements, des mesures strictes sont prises pour assurer la non-divulgation des enregistrements protégés aux termes de la *Loi sur la statistique*. Toutes les études qui nécessitent le couplage d'enregistrements faisant partie de bases de données protégées doivent être soumises à un processus d'examen et d'approbation rigoureux avant d'être exécutées

1. D. Krewski, Centre McLaughlin d'évaluation du risque pour la santé des populations, Université d'Ottawa, Ottawa (Ontario), Canada K1N 6N5. School of Mathematics & Statistics, Carleton University, Ottawa (Ontario), Canada, K1S 5B6. La correspondance devrait être adressée à : A. Dewanji, Applied Statistics Unit, Indian Statistical Institute, Kolkata, India; Y. Wang, Santé environnementale et sécurité des consommateurs, Santé Canada, Ottawa (Ontario), Canada, K1A 0L2; S. Bartlett, Santé environnementale et sécurité des consommateurs, Santé Canada, Ottawa (Ontario), Canada, K1A 0L2; J. M. Zielinski, Santé environnementale et sécurité des consommateurs, Santé Canada, Ottawa (Ontario), Canada, K1A 0L2; R. Mallick, Centre McLaughlin d'évaluation du risque pour la santé des populations, Université d'Ottawa, Ottawa (Ontario), Canada K1N 6N5. School of Mathematics & Statistics, Carleton University, Ottawa (Ontario), Canada, K1S 5B6.

conformément à des procédures bien établies en vue d'assurer le respect de la confidentialité des données (Singh, Feder, Dunteman et Yu 2001). Tous les fichiers couplés contenant des renseignements permettant d'identifier des individus restent sous la garde de Statistique Canada (Labossière 1986).

Des méthodes informatisées de couplage d'enregistrements ont été utilisées pour coupler des données sur l'exposition environnementale à celles de la Base canadiennes de données sur la mortalité (BCDM). Par exemple, une étude a été entreprise pour étudier les liens éventuels entre les causes de décès chez plus de 326 000 exploitants agricoles au Canada et diverses variables sociodémographiques et d'exploitation agricole, particulièrement l'utilisation de pesticides (Jordan-Simpson, Fair et Poliquin 1990). Cette étude comportait le couplage des données de la BCDM à celles du Recensement de la population de 1971 et du Recensement de l'agriculture de 1971. Une autre étude permanente de grande portée est fondée sur le Fichier dosimétrique national (FDN) du Canada (Ashmore et Grogan 1985; Ashmore et Davies 1989). Le FDN contient des renseignements remontant jusqu'à 1950 sur les expositions professionnelles aux rayonnement ionisants subies par plus de 400 000 Canadiens. Récemment, les enregistrements du FDN ont été couplés à ceux de la BCDM en vue d'étudier les associations entre la surmortalité due au cancer et l'exposition professionnelle à de faibles niveaux de rayonnements ionisants (Ashmore, Krewski et Zielinski 1997; Ashmore, Krewski, Zielinski, Jiang, Semenciw et Létourneau 1998). Plus récemment, les enregistrements du FDN ont été couplés à ceux de la Base canadienne de données sur l'incidence du cancer (Sont, Zielinski, Ashmore, Jiang, Krewski, Fair, Band et Létourneau 2001). La liste complète des autres études relatives à la santé fondées sur le couplage de données sur l'exposition à celles de la BCDM a été dressée par Fair (1989).

Le succès des études axées sur le couplage d'enregistrements dépend de la qualité des bases de données couplées (Roos, Soodeen et Jebamani 2001). À l'aide de données administratives longitudinales représentatives de la population, Roos et coll. ont examiné les problèmes de qualité de données dans les études sur l'état de santé et les soins de santé. Ardal et Ennis (2001) ont tenu compte des erreurs systématiques présentes dans les bases de données administratives intervenant dans l'analyse secondaire de l'information sur la santé. S'il est vrai que les études fondées sur le couplage d'enregistrements donnent de meilleurs résultats quand les données sont de haute qualité, les contraintes liées à la qualité des données sont compensées dans une certaine mesure par la grande taille des échantillons sur lesquels reposent de nombreuses bases de données administratives.

Les études par couplage d'enregistrements offrent plusieurs avantages par rapport aux études épidémiologiques classiques. L'utilisation des bases de données administratives existantes évite de devoir recueillir de nouvelles données pour les études sur la santé et permet d'obtenir des échantillons de grande taille moyennant assez peu d'efforts. Selon la nature des bases de données utilisées, le couplage d'enregistrements offre un moyen peu coûteux d'explorer de nombreuses associations éventuelles dans le cadre des études épidémiologiques. Le couplage d'enregistrements présente aussi certains inconvénients. Les chercheurs exercent généralement fort peu de contrôle sur l'information recueillie et le nombre de sujets perdus de vue lors des suivis peut être important. Les erreurs de couplage, qui sont le sujet du présent article, sont un autre inconvénient du couplage d'enregistrements. Inévitablement, certains enregistrements concordants ne seront pas couplés et certains enregistrements non concordants seront couplés incorrectement.

Assez peu de travaux ont été accomplis en vue de déterminer l'effet de ces erreurs de couplage sur les inférences statistiques. Neter, Maynes et Ramanathan (1965) ont utilisé un modèle de régression linéaire simple pour analyser l'effet des erreurs introduites durant le processus d'appariement. Selon leurs résultats, les erreurs de couplage font augmenter la variance résiduelle et introduisent un biais dans l'estimation de la pente de la droite de régression. Winkler et Scheuren (1991) établissent une expression du biais dû aux erreurs de couplage dans les estimations des coefficients de régression linéaire. Les progrès concernant l'estimation des taux d'erreurs de couplage réalisés par Belin et Rubin (1991) ont permis à Scheuren et Winkler (1993) de mettre en œuvre une méthode améliorée de correction du biais. L'application des méthodes de régression linéaire à l'analyse des fichiers de données appariées informatiquement est discutée plus en détail par Scheuren et Winkler (1997).

L'objet du présent article est d'étudier l'effet des erreurs de couplage sur les inférences statistiques dans les études-cohorte de la mortalité. À la section 2, nous décrivons les modèles de régression du risque relatif employés pour analyser les données provenant de ce genre d'études et nous élaborons des expressions pour les nombres observé et attendu de décès fondés sur ces modèles. À la section 3, nous discutons de l'effet des erreurs de couplage sur les nombres observé et attendu de décès et de personnes-années à risque. À la section 4, nous analysons l'effet des erreurs de couplage sur les estimations des ratios standardisés de mortalité (RSM) et sur les paramètres de régression du risque relatif. Les deux types d'erreurs peuvent introduire un biais et une variabilité supplémentaire dans les estimations

de ces paramètres. À la section 5, nous présentons nos conclusions.

2. Modèles de régression du risque relatif

Les méthodes statistiques d'analyse des données provenant d'études-cohorte de la mortalité sont bien établies (Breslow et Day 1987). L'objectif principal de ce genre d'analyse est de déterminer si l'exposition à l'agent d'intérêt augmente le taux de mortalité chez les membres de la cohorte. La mortalité est caractérisée par la fonction de risque, qui précise le taux de mortalité sous forme de fonction du temps. Si nous représentons par T le moment du décès, la fonction de risque au temps u se définit formellement comme suit

$$\lambda(u) = \lim_{\Delta u \downarrow 0} \frac{\Pr\{u \le T < u + \Delta u \,|\, T \ge u\}}{\Delta u}. \qquad (1)$$

Soit $\lambda_i(u)$ la fonction de risque pour une cause particulière de décès au temps u pour l'individu $i = 1, ..., N$ dans une cohorte de taille N, et soit $\mathbf{z}_i(u)$ un vecteur correspondant de covariables propres à cet individu. Nous supposons que ces covariables ont pour effet de modifier le risque de base $\lambda^*(u)$ conformément au modèle de régression du risque relatif

$$\lambda_i(u) = \lambda^*(u)\gamma\{\beta'\mathbf{z}_i(u)\}, \qquad (2)$$

où γ est une fonction positive des covariables et β est un vecteur de paramètres de régression.

Deux cas particuliers du modèle général de régression du risque relatif présentant un intérêt sont les modèles multiplicatif et additif de régression du risque. Définissons la fonction γ figurant dans (2) par

$$\log\gamma(z) = \frac{(1+z)^\rho - 1}{\rho}. \qquad (3)$$

Quand $\rho = 1$, le modèle général de régression du risque relatif se réduit au modèle multiplicatif de régression du risque

$$\lambda_i(u) = \lambda^*(u)\exp\{\beta'\mathbf{z}_i(u)\}, \qquad (4)$$

Ce modèle à risques proportionnels, qui a été introduit par Cox (1972), est d'usage très répandu en analyse des données sur la mortalité (Kalbfleish et Prentice 1980). Le modèle additif de régression du risque

$$\lambda_i(u) = \lambda^*(u) + \beta'\mathbf{z}_i(u) \qquad (5)$$

survient en tant que cas limite quand $\rho \to 0$.

Soit t_i^0 et t_i^1 l'âge au moment de l'entrée dans l'étude et l'âge au moment de la perte de vue (due à l'abandon par le sujet, à l'interruption de l'étude ou au décès) du i^e sujet de la cohorte, respectivement. Soit $\delta_i = 1$ ou 0, selon que le i^e sujet est ou n'est pas décédé au moment de la perte de vue. La fonction de log-vraisemblance fondée sur le modèle du risque relatif (2) peut s'écrire

$$\log L = \sum_{i=1}^{N} \left\{ \begin{array}{l} \delta_i \log(\gamma\{\beta'\mathbf{z}_i(t_i^1)\}) \\ \qquad -\int_{t_i^0}^{t_i^1} \gamma\{\beta'\mathbf{z}_i(u)\}\lambda^*(u)\,du \end{array} \right\}. \qquad (6)$$

Lorsqu'il n'existe qu'une covariable $z_i(u) \equiv 1$, l'estimation du maximum de vraisemblance de $\theta = \exp\{\beta\}$ se réduit au ratio standardisé de mortalité RSM = OBS/ATT, où OBS = $\sum_{i=1}^N \delta_i$ et ATT = $\sum_{i=1}^N e_i$ sont les nombres observé et attendu de décès, respectivement, avec $e_i = \int_{t_i^0}^{t_i^1} \lambda^*(u)\,du$.

La maximisation de la fonction de vraisemblance (6) peut donner lieu à des calculs fastidieux dans le cas d'échantillons de grande taille. Breslow, Lubin et Langholz (1983) simplifient cette fonction en supposant que les covariables prennent des valeurs constantes dans les états par lesquels passe un sujet durant le cours de l'étude. Ces états sont définis par des classifications croisées des covariables d'intérêt. Plus précisément, supposons qu'il existe J états de ce genre $\{S_j; j = 1, ..., J\}$, tels que $\mathbf{z}_i(u) = \mathbf{z}_j$ chaque fois que le i^e sujet se trouve dans l'état S_j au temps u. Ces états sont mutuellement exclusifs et exhaustifs, si bien que, à tout temps u, chaque membre de la cohorte se trouvera dans un état, et uniquement un. La fonction de log-vraisemblance (6) peut s'écrire

$$\log L = \sum_{j=1}^{J} \{d_{jj}\log(\gamma\{\beta'\mathbf{z}_i\}) - \gamma\{\beta'\mathbf{z}_j\}e_j\}, \qquad (7)$$

où

$$e_j = \sum_{i=1}^{N} \int_{[\mathbf{z}_i(u)\in S_j]} \lambda^*(u)\,du \qquad (8)$$

est la contribution au nombre attendu de décès provenant de toutes les personnes-années d'observation dans l'état S_j, et d_{jj} est le nombre total de décès dans cet état. En posant que $\Lambda_j(\beta) = \log(\gamma\{\beta'\mathbf{z}_j\})$, nous obtenons l'estimation du maximum de vraisemblance $\hat\beta$ de β en tant que solution de l'équation de score

$$\frac{\partial \log L}{\partial \beta} = \sum_{j=1}^{J} \frac{\partial \Lambda_j(\hat\beta)}{\partial \beta}\{d_{jj} - \exp\{\Lambda_j(\hat\beta)\}e_j\} = 0. \qquad (9)$$

3. L'effet des erreurs de couplage sur les nombres observé et attendu de décès

Deux grands types d'erreurs peuvent se produire lors du couplage de fichiers de données dans le contexte du CIE (Fellegi et Sunter 1969). Un résultat faussement positif a lieu quand un membre de la cohorte encore en vie est

incorrectement désigné comme étant décédé et un résultat faussement négatif survient quand un membre décédé de la cohorte est considéré comme étant en vie. Plus précisément, pour le développement mathématique qui suit, un résultat faussement positif survient dans un état particulier quand un individu qui demeure en vie pendant tout le temps où il se trouve dans cet état est incorrectement étiqueté comme étant décédé dans cet état. Pareillement, un résultat faussement négatif survient dans un état particulier quand un membre de la cohorte qui est décédé avant d'atteindre cet état ou pendant qu'il se trouvait dans cet état est considéré comme étant en vie en étant dans cet état. Dans un état donné, les résultats faussement positifs et faussement négatifs représentent donc des cas particuliers de l'erreur de classification discutée par Anderson (1974, chapitre 6.2.1). À la présente section, nous examinons l'effet de ces deux types d'erreurs de couplage sur les nombres observé et attendu de décès, respectivement. À cet fin, nous commençons par définir des jeux d'indices dans les divers états que nous utiliserons pour représenter les ensembles d'enregistrements correctement appariés et incorrectement appariés.

3.1 Erreurs de couplage

Soit A_j et D_j l'ensemble d'étiquettes pour les membres de la cohorte qui demeurent en vie dans l'état S_j, et pour ceux qui sont décédés dans l'état S_j, respectivement. Soit D_{jj} le sous-ensemble de D_j correspondant aux personnes qui sont décédées dans l'état S_j. Soit A_j^L, D_j^L et D_{jj}^L les ensembles correspondants à la présence d'erreurs de couplage. Définissons en outre D_j^P comme étant l'ensemble d'étiquettes des individus en vie dans l'état S_j (c'est-à-dire dans A_j), mais étiquetés comme étant décédés dans l'état S_j, c'est-à-dire correspondant aux résultats faussement positifs dans S_j. De la même façon, A_j^N est l'ensemble d'individus décédés dans l'état S_j (c'est-à-dire dans D_j), mais étiquetés comme en étant en vie dans l'état S_j, c'est-à-dire correspondant aux résultats faussement positifs dans S_j. Représentons aussi par D_{jj}^P le sous-ensemble de D_j^P correspondant aux individus étiquetés comme étant décédés dans l'état S_j et, pareillement, par A_{jj}^N le sous-ensemble d'individus de A_j^N qui sont décédés dans l'état S_j (c'est-à-dire dans D_{jj}). Ces ensembles satisfont aux relations $A_j^L = (A_j - D_j^P) \cup A_j^N$, $D_j^L = (D_j - A_j^N) \cup D_j^P$ et $D_{jj}^L = (D_{jj} - A_{jj}^N) \cup D_{jj}^P$.

L'effet des erreurs de couplage sur la fonction de vraisemblance donnée par (7) peut être décrit comme suit. Soit t_{ij}^0 le temps auquel le i^{e} individu entre, réellement ou par erreur de couplage, dans le j^{e} état S_j. De même, t_{ij}^1 représente le moment du décès (s'il a lieu, réellement ou par erreur de couplage) du i^{e} individu dans l'état S_j et t_{ij}^2, le moment de la sortie de l'état S_j, réellement ou par erreur de couplage. Notons que, si t_{ij}^1 existe, il est inférieur ou égal à t_{ij}^2. Par souci de simplicité, supposons que t_{ij}^1, s'il existe, est égal à t_{ij}^0; autrement dit, tous les décès qui surviennent dans un état particulier le font au moment correspondant de l'entrée dans cet état. Bien que cette hypothèse produise une sous-estimation du nombre attendu de décès, aux fins de l'étude du biais, elle n'est peut-être pas si contestable. Le fait de supposer que tous les décès surviennent au moment de la sortie des états correspondants offre aussi une simplification comparable. Partant de (8) et de la décomposition de A_j^L, nous pouvons écrire le nombre attendu de décès e_j^L dans S_j en présence d'erreurs de couplage sous la forme

$$\begin{aligned} e_j^L &= \sum_{i \in A_j^L} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du \\ &= \sum_{i \in A_j} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du + \sum_{i \in A_j^N} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du \\ &\qquad - \sum_{i \in D_j^P} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du \\ &= e_j - \Delta e_j, \end{aligned} \tag{10}$$

où

$$e_j = \sum_{i \in A_j} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du, \text{ et } \Delta e_j = e_j^P - e_j^N \tag{11}$$

avec

$$e_j^P = \sum_{i \in D_j^P} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du \text{ et } e_j^N = \sum_{i \in A_j^N} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du. \tag{12}$$

Pour simplifier la notation, écrivons $T_\lambda(i, j)$ pour $\int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)\,du$ dans la suite. Le terme Δe_j représente le biais introduit par les erreurs de couplage dans le nombre attendu de décès dans le j^{e} état. Il découle de (10) et de (11) que les résultats faussement positifs ont tendance à réduire le nombre attendu de décès et que les résultats faussement négatifs ont tendance à l'augmenter.

En utilisant la décomposition de D_{jj}^L, nous pouvons écrire le nombre observé de décès d_{jj}^L en présence d'erreurs de couplage comme suit

$$d_{jj}^L = d_{jj} + \Delta d_{jj}, \tag{13}$$

où

$$\Delta d_{jj} = d_{jj}^P - a_{jj}^N, \tag{14}$$

avec d_{jj}, d_{jj}^P et a_{jj}^N le nombre d'individus dans les ensembles D_{jj}, D_{jj}^P et A_{jj}^N, respectivement. Le terme Δd_{jj} représente la variation du nombre observé de décès dans le j^{e} état due aux erreurs de couplage. Il découle de (13) et de (14) que les résultats faussement positifs font augmenter le nombre observé de décès et que les résultats faussement négatifs le réduisent.

Le statut vital est souvent déterminé par couplage des données sur la cohorte étudiée à celles de la BCDM, dont l'effectif est généralement beaucoup plus grand que celui de

la cohorte d'intérêt. Lorsque les enregistrements sur l'exposition d'une personne en vie sont associés incorrectement à ceux d'une personne décédée, cette dernière n'appartient habituellement pas à la cohorte. Donc, la contribution de personnes-années à risque de la personne qui demeure en vie cessera prématurément dans l'année du décès présumé; les personnes-années à risque perdues correspondent à la période écoulée de l'année du décès présumé jusqu'à la fin du suivi. Par ailleurs, si les enregistrements sur l'exposition d'un individu décédé sont associés incorrectement à ceux d'une personne en vie, la contribution de personnes-années à risque de cet individu inclura une période supplémentaire s'étendant de l'année réelle du décès jusqu'à la fin du suivi. Par conséquent, les résultats faussement positifs réduiront le nombre de personnes-années à risque dans la cohorte et les résultats faussement négatifs l'augmenteront.

3.2 Espérances et variances des différences dans les nombres observé et attendu de décès

L'effet des erreurs de couplage sur les nombres observé et attendu de décès dépend des taux de résultats faussement positifs et faussement négatifs. Soit p_j^P et p_j^N les taux de résultats faussement positifs et de résultats faussement négatif, respectivement, dans l'état S_j, pour $j = 1, \ldots, J$, que l'on suppose être constants dans S_j et les mêmes pour tous les individus dans A_j et D_j, respectivement. Cette hypothèse est raisonnable si les individus qui se trouvent dans le même état sont très homogènes, particulièrement en ce qui concerne des attributs tels que la qualité des identificateurs personnels, qui influent sur les taux d'erreurs de couplage. Bien que cette hypothèse idéaliste soit peu susceptible d'être entièrement satisfaite en pratique, elle simplifie considérablement l'évaluation subséquente des effets des erreurs de couplage. Formellement, p_j^P (p_j^N) est la probabilité conditionnelle qu'un individu compris dans $A_j (D_j)$ soit étiqueté comme étant décédé (en vie) dans l'état S_j. Autrement dit, $p_j^P = P[i \in D_j^P \mid i \in A_j]$ et $p_j^N = P[i \in A_j^N \mid i \in D_j]$.

Soit a_j, d_j, a_j^N et d_j^P le nombre d'individus dans A_j, D_j, A_j^N et D_j^P, respectivement. Alors, notons que d_j^P suit une loi binomiale(a_j, p_j^P) et que a_j^N suit une loi binomiale(d_j, p_j^N). En outre, d_{jj}^P suit une loi binomiale(a_j, p_{jj}^P), où p_{jj}^P est la probabilité conditionnelle qu'un individu compris dans A_j soit étiqueté comme étant décédé dans l'état S_j. Autrement dit, $p_{jj}^P = P[i \in D_{jj}^P \mid i \in A_j]$. De toute évidence, $p_{jj}^P \le p_j^P$. De la même façon, a_{jj}^N suit une loi binomiale(d_{jj}, p_{jj}^N), où p_{jj}^N est la probabilité conditionnelle qu'un individu compris dans D_{jj} soit étiqueté comme étant en vie dans l'état S_j. Autrement dit, $p_{jj}^N = P[i \in A_{jj}^N \mid i \in D_{jj}]$. Bien qu'il n'existe pas de relation sans importance entre p_j^N et p_{jj}^N en général, il est raisonnable de supposer que $p_j^N = p_{jj}^N$ dans le contexte des erreurs couplage.

En supposant que les erreurs de couplage associées à divers individus sont indépendantes, l'espérance et la variance de la différence dans le nombre observé de décès dans l'état S_j, donnée par Δd_{jj} dans (14), sont

$$E[\Delta d_{jj}] = E[d_{jj}^P] - E[a_{jj}^N] = a_j\, p_{jj}^P - d_{jj}\, p_j^N \qquad (15)$$

et

$$\begin{aligned} V[\Delta d_{jj}] &= V[d_{jj}^P] + V[a_{jj}^N] \\ &= a_j\, p_{jj}^P (1 - p_{jj}^P) + d_{jj}\, p_j^N (1 - p_j^N). \end{aligned} \qquad (16)$$

Puisque A_j et D_{jj} sont constitués d'ensembles différents d'individus, d_{jj}^P et a_{jj}^N sont indépendants.

De la même façon, l'espérance et la variance de la différence dans le nombre attendu de décès dans l'état S_j, donnée par Δe_j dans (11), peuvent être calculées comme suit. À cette fin, il est commode d'écrire e_j^P et e_j^N en fonction des variables indicatrices qui suivent. Pour $i \in A_j$, définissons $\xi_{ij} = I\{i \in D_j^P\}$ et $\xi_{ijj} = I\{i \in D_{jj}^P\}$. En outre, pour $i \in D_j$, définissons $\psi_{ij} = I\{i \in A_j^N\}$. Alors, il découle de (12) et des définitions de D_j^P et A_j^N que

$$e_j^P = \sum_{i \in A_j} \xi_{ij}\, T_\lambda(i,\ j) \qquad (17)$$

et

$$e_j^N = \sum_{i \in D_j} \psi_{ij}\, T_\lambda(i,\ j). \qquad (18)$$

En particulier, nous pouvons écrire $d_{jj}^P = \sum_{i \in A_j} \xi_{ijj}$ et $a_{jj}^N = \sum_{i \in D_{jj}} \psi_{ij}$, qui sont utiles pour établir (15) et (16). D'après (17) et (18), nous obtenons

$$\begin{aligned} E[\Delta e_j] &= E[e_j^P] - E[e_j^N] \\ &= p_j^P \sum_{i \in A_j} T_\lambda(i,\ j) \ - \ p_j^N \sum_{i \in D_j} T_\lambda(i,\ j), \end{aligned} \qquad (19)$$

et

$$\begin{aligned} V[\Delta e_j] &= V[e_j^P] + V[e_j^N] \\ &= p_j^P (1 - p_j^P) \sum_{i \in A_j} T_\lambda^2(i,\ j) \\ &\quad + p_j^N (1 - p_j^N) \sum_{i \in D_j} T_\lambda^2(i,\ j), \end{aligned} \qquad (20)$$

puisque A_j et D_j sont constitués d'ensembles différents d'individus.

Les résultats (15)–(16) et (19)–(20) indiquent que les erreurs de couplage d'enregistrements introduisent un biais et une variation supplémentaire dans les nombres observé et attendu de décès. Minimiser les termes de variance dans (16) et (20) est difficile, puisque les deux taux d'erreurs p_j^P et p_j^N ne sont pas fonctionnellement indépendants. En

général, la diminution de p_j^P donnera lieu à une augmentation de p_j^N, et inversement (voir la section 5 pour une discussion plus approfondie de ce point). Bien que ces taux d'erreurs soient indépendants du modèle de régression du risque relatif sous-jacent γ donné par (2), l'erreur quadratique moyenne obtenue par combinaison des termes d'espérance et de variance ne peut être minimisée sans qu'on spécifie le risque de base $\lambda^*(u)$, qui figure dans T_λ.

4. L'effet des erreurs de couplage sur les estimations des RSM et des coefficients de régression

4.1 Ratios standardisés de mortalité

Pour déterminer l'effet des erreurs de couplage sur les RSM, nous remplaçons les nombres observé et attendu réels de décès d_{jj} et e_j par les nombres observé et attendu de décès en présence d'erreurs de couplage d_{jj}^L et e_j^L dans l'expression $\text{RSM} = \sum d_{jj} / \sum e_j$. En représentant par RSM_L les ratios standardisés de mortalité en présence d'erreurs de couplage, nous obtenons

$$\text{RSM}_L = \text{RSM}\left[1+\frac{\sum \Delta d_{jj}}{\sum d_{jj}}\right] \Big/ \left[1-\frac{\sum \Delta e_j}{\sum e_j}\right]. \qquad (21)$$

Il découle des équations (10) à (14) que les résultats faussement positifs feront augmenter le RSM, tandis que les résultats faussement négatifs le feront diminuer.

En utilisant un développement en série de premier ordre de Taylor comme approximation de RSM_L autour de RSM, la différence $\Delta\text{RSM} = \text{RSM}_L - \text{RSM}$ peut s'exprimer sous la forme

$$\frac{\Delta\text{RSM}}{\text{RSM}} = \frac{\sum_j \Delta d_{jj}}{\sum_j d_{jj}} + \frac{\sum_j \Delta e_j}{\sum_j e_j}. \qquad (22)$$

Alors, la moyenne et la variance de la différence relative de RSM peuvent être approximées par

$$E\left[\frac{\Delta\text{RSM}}{\text{RSM}}\right] \simeq \frac{\sum_j E[\Delta d_{jj}]}{\sum_j d_{jj}} + \frac{\sum_j E[\Delta e_j]}{\sum_j e_j} \qquad (23)$$

et

$$V\left[\frac{\Delta\text{RSM}}{\text{RSM}}\right] \simeq \left(\sum_j d_{jj}\right)^{-2} V\left[\sum_j \Delta d_{jj}\right] + \left(\sum_j e_j\right)^{-2} V\left[\sum_j \Delta e_j\right] + 2\left(\sum_j d_{jj}\right)^{-1}\left(\sum_j e_j\right)^{-1} \text{Cov}\left[\sum_j \Delta d_{jj}, \sum_j \Delta e_j\right], \qquad (24)$$

respectivement. Il est facile de calculer le deuxième membre de (23) en utilisant (15) et (19). Pour calculer le deuxième membre de (24), notons que

$$V\left[\sum_j \Delta d_{jj}\right] = \sum_j V[\Delta d_{jj}] + 2\sum_{j<j'} \text{Cov}[\Delta d_{jj},\ \Delta d_{j'j'}], \qquad (25)$$

$$V\left[\sum_j \Delta e_j\right] = \sum_j V[\Delta e_j] + 2\sum_{j<j'} \text{Cov}[\Delta e_j,\ \Delta e_{j'}], \qquad (26)$$

et

$$\text{Cov}\left[\sum_j \Delta d_{jj},\ \sum_j \Delta e_j\right] = \sum_j \text{Cov}[\Delta d_{jj},\ \Delta e_j] + \sum_{j\neq j'} \text{Cov}[\Delta d_{jj},\ \Delta e_{j'}]. \qquad (27)$$

Sans perte de généralité, supposons, pour $j < j'$, que $t_{ij}^0 \le t_{ij'}^0$ pour le même individu i (en vie ou décédé) dans S_j et $S_{j'}$; autrement dit, le moment de l'entrée dans S_j est identique ou antérieur à celui de l'entrée dans $S_{j'}$. Nous avons alors, pour $j < j'$,

$$\text{Cov}[\Delta d_{jj}, \Delta d_{j'j'}] = -\left(\sum_{i\in A_j\cap A_{j'}} p_{jj}^P\, p_{j'j'}^P + \sum_{i\in A_j\cap D_{j'j'}} p_{jj}^P\, p_{j'}^N\right), \qquad (28)$$

$$\begin{aligned}\text{Cov}[\Delta e_j,\ \Delta e_{j'}] &= \sum_{i\in A_j\cap A_{j'}} p_j^P (1-p_{j'}^P)\, T_\lambda(i,\ j)\, T_\lambda(i,\ j') \\ &+ \sum_{i\in A_j\cap D_{j'}} p_j^P\ p_{j'}^N\, T_\lambda(i,\ j)\, T_\lambda(i,\ j') \\ &+ \sum_{i\in D_j\cap D_{j'}} p_{j'}^N (1-p_j^N)\, T_\lambda(i,\ j)\, T_\lambda(i,\ j'), \qquad (29)\end{aligned}$$

$$\begin{aligned}\text{Cov}[\Delta d_{jj},\ \Delta e_j] &= \sum_{i\in A_j} p_{jj}^P (1-p_j^P)\, T_\lambda(i,\ j) \\ &+ \sum_{i\in D_{jj}} p_j^N (1-p_j^N)\, T_\lambda(i,\ j), \qquad (30)\end{aligned}$$

$$\begin{aligned}\text{Cov}[\Delta d_{jj},\ \Delta e_{j'}] &= \sum_{i\in A_j\cap A_{j'}} p_{jj}^P (1-p_{j'}^P)\, T_\lambda(i,\ j') \\ &+ \sum_{i\in A_j\cap D_{j'}} p_{jj}^P\ p_{j'}^N\, T_\lambda(i,\ j') \\ &+ \sum_{i\in D_{jj}\cap D_{j'}} p_{j'}^N (1-p_j^N)\, T_\lambda(i,\ j'), \text{ et } (31)\end{aligned}$$

$$\begin{aligned}\text{Cov}[\Delta d_{j'j'},\ \Delta e_j] &= -\sum_{i\in A_j\cap A_{j'}} p_j^P\ p_{j'j'}^P\, T_\lambda(i,\ j) \\ &+ \sum_{i\in A_j\cap D_{j'j'}} p_j^P\ p_{j'}^N\, T_\lambda(i,\ j). \qquad (32)\end{aligned}$$

En utilisant les équations (25) à (32), nous pouvons approximer la variance de la différence relative ΔRSM/RSM au moyen du deuxième membre de (24). Nous pouvons tirer deux conclusions des équations (23) et (24). En premier lieu, les erreurs de couplage peuvent introduire un biais dans l'estimation du RSM. En deuxième lieu, les deux types d'erreurs de couplage introduisent une variation supplémentaire dans les estimations du RSM. Notons que le premier terme de (32) est dominé par le premier terme de (29) pour $p_{j'}^{P} < 0{,}5$, et que le terme de covariance négatif (28) est dominé dans le calcul de la variance dans (25). Par conséquent, la variance supplémentaire (24) est strictement positive, puisque les taux de résultats faussement positifs et de résultats faussement négatifs sont tous deux positifs.

4.2 Paramètres de régression du risque relatif

Pour déterminer l'effet des erreurs de couplage sur les estimations des paramètres de régression, considérons d'abord le modèle général de régression du risque relatif (2). En remplaçant dans la fonction de log-vraisemblance (7) les nombres observé et attendu de décès d_{jj} et e_j par les nombres observé et attendu de décès en présence d'erreurs de couplage d_{jj}^{L} et e_j^{L}, nous obtenons

$$\log L = \sum_{j=1}^{J} \{ d_{jj}^{L} \log(\gamma\{\beta' \mathbf{z}_j\}) - \gamma\{\beta' \mathbf{z}_j\} e_j^{L} \}. \qquad (33)$$

Soit $\hat{\beta}$ et $\tilde{\beta}$ les estimations du maximum de vraisemblance de β fondées sur $\{d_{jj}, e_j\}$ et $\{d_{jj}^{L}, e_j^{L}\}$, respectivement. L'équation de score (9) peut s'écrire sous la forme

$$\sum_{j=1}^{J} \frac{\partial \Lambda_j(\tilde{\beta})}{\partial \beta} [d_{jj} + \Delta d_{jj} - \exp\{\Lambda_j(\tilde{\beta})\}(e_j - \Delta e_j)] = 0. \qquad (34)$$

En supposant que $\Delta\beta = \tilde{\beta} - \hat{\beta}$ est faible, un développement en série de premier ordre de $\exp\{\Lambda_j(\tilde{\beta})\}$ autour de $\hat{\beta}$ donne

$$\exp\{\Lambda_j(\tilde{\beta})\} \simeq \exp\{\hat{\Lambda}_j\} + \exp\{\hat{\Lambda}_j\} \frac{\partial \hat{\Lambda}_j}{\partial \beta} \Delta\beta, \qquad (35)$$

où $\hat{\Lambda}_j = \Lambda_j(\hat{\beta})$ et $\partial\hat{\Lambda}_j / \partial\beta$ est $\partial\Lambda_j / \partial\beta$ évalué à $\beta = \hat{\beta}$. En introduisant (35) par substitution dans (34), nous obtenons

$$\sum_{j=1}^{J} \frac{\partial \hat{\Lambda}_j}{\partial \beta} [d_{jj} - \exp\{\hat{\Lambda}_j\} e_j] + \sum_{j=1}^{J} \frac{\partial \hat{\Lambda}_j}{\partial \beta} \begin{bmatrix} \Delta d_{jj} + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j \\ - \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \hat{\Lambda}_j'}{\partial \beta} \Delta\beta \\ + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j \frac{\partial \hat{\Lambda}_j'}{\partial \beta} \Delta\beta \end{bmatrix} \simeq 0. \qquad (36)$$

En utilisant (9), la première somme dans (36) est nulle. Par conséquent, puisque $\Delta e_j \Delta\beta$ est faible, $\Delta\beta$ peut être approximé par

$$\Delta\beta \simeq \left(\sum_j \frac{\partial \hat{\Lambda}_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \hat{\Lambda}_j'}{\partial \beta} \right)^{-1} \sum_j \frac{\partial \hat{\Lambda}_j}{\partial \beta} \{ \Delta d_{jj} + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j \}. \qquad (37)$$

Il découle de (37) que

$$E[\Delta\beta] \simeq \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right)^{-1} \sum_j \frac{\partial \Lambda_j}{\partial \beta} \alpha_j, \qquad (38)$$

où $\alpha_j = E[\Delta d_{jj}] + \gamma\{\hat{\beta}' \mathbf{z}_j\} E[\Delta e_j]$, qui peut être calculé d'après (15) et (19). En outre,

$$V[\Delta\beta] \simeq \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right)^{-1} \left(\sum_j \sum_{j'} \frac{\partial \Lambda_j}{\partial \beta} \Theta_{jj'} \frac{\partial \Lambda_{j'}'}{\partial \beta} \right) \left(\sum_j \frac{\partial \Lambda_j}{\partial \beta} \gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta} \right)^{-1} \qquad (39)$$

avec

$$\Theta_{jj'} = \mathrm{Cov}[\Delta d_{jj} + \gamma\{\hat{\beta}' \mathbf{z}_j\} \Delta e_j,\ \Delta d_{j'j'} + \gamma\{\hat{\beta}' \mathbf{z}_{j'}\} \Delta e_{j'}],$$

qui peut aussi être obtenu facilement en utilisant (16), (20) et (28) à (32).

Dans le cas particulier du modèle multiplicatif de risque (4), la différence $\Delta\beta$ due aux erreurs de couplage peut être approximée par

$$\Delta\beta \simeq (X' W X)^{-1} X' (\Delta D + \Delta W), \qquad (40)$$

où $X' = (\mathbf{z}_1', \ldots, \mathbf{z}_J')$, $\Delta D' = (\Delta d_{11}, \ldots, \Delta d_{JJ})$, $W = \mathrm{diag}(\exp(\mathbf{z}_1' \hat{\beta}) e_1, \ldots, \exp(\mathbf{z}_J' \hat{\beta}) e_J)$, et $\Delta W' = (\exp(\mathbf{z}_1' \hat{\beta}) \Delta e_1, \ldots, \exp(\mathbf{z}_J' \hat{\beta}) \Delta e_J)$. Notons que la matrice de poids W est la matrice d'information de Fisher pour $\hat{\beta}$. Il découle de (38) que

$$E[\Delta\beta] \simeq (X' W X)^{-1} X' \Pi, \qquad (41)$$

où $\Pi' = (\pi_1, \ldots, \pi_J)$ avec π_j identique à α_j, mais $\gamma\{\hat{\beta}' \mathbf{z}_j\}$ remplacé par $\exp(\mathbf{z}_j' \hat{\beta})$.

En outre,

$$V[\Delta\beta] \simeq (X' W X)^{-1} X' \Psi X (X' W X)^{-1}, \qquad (42)$$

où Ψ est la matrice des $\Theta_{jj'}$ avec $\gamma\{\hat{\beta}' \mathbf{z}_j\}$ remplacé par $\exp(\mathbf{z}_j' \hat{\beta})$. Notons que les expressions (40) à (42) sont des cas particuliers des expressions (37) à (39), respectivement, écrites en notation matricielle.

Avec une seule covariable $z_i = 1$, $X' W X = e^{\hat{\beta}} \sum_j e_j$, $X' \Delta D = \sum_j d_{jj}$ et $X' \Delta W = e^{\hat{\beta}} \sum_j \Delta e_j$. Dans ce cas,

$$\Delta\beta \simeq \frac{\sum_j \Delta d_{jj} + e^{\hat{\beta}} \sum_j \Delta e_j}{e^{\hat{\beta}} \sum_j e_j}. \qquad (43)$$

Puisque le $\text{RSM} = e^{\hat{\beta}} = \sum_j d_{jj} / \sum_j e_j$, avec $\Delta\beta = \Delta\text{RSM/RSM}$ ici, nous obtenons

$$\Delta\beta \simeq \frac{\sum_j \Delta d_{jj}}{\sum_j d_{jj}} + \frac{\sum_j \Delta e_j}{\sum_j e_j}. \qquad (44)$$

Donc, l'expression (44) peut être considérée comme un cas particulier de (22).

Les résultats qui précédent indiquent que les résultats faussement positifs ainsi que faussement négatifs introduisent un biais et une variation supplémentaire dans les estimations des paramètres de régression du risque relatif. La seule contribution négative à cette variance supplémentaire (39) a lieu par la voie de $\text{Cov}[\Delta d_{jj}, \Delta d_{j'j'}]$, donné par (28), et du premier terme de (32) (voir $\Theta_{jj'}$). En utilisant le même argument qu'à la section 4.1, il s'ensuit que cette variance supplémentaire est strictement positive.

5. Conclusion

Le couplage d'enregistrements est maintenant une technique bien établie dans le contexte des études épidémiologiques des risques pour la santé des populations. En couplant l'information sur les expositions des individus provenant d'une base de données à celles sur les résultats en ce qui concerne la santé provenant d'une autre base de données, il est possible de construire de grandes bases de données informatives sur les risques que courent les populations et les sous-groupes de population. Le succès de ce genre d'études dépend, en grande partie, de la qualité des deux bases de données que l'on couple, y compris la quantité d'information sur les identificateurs personnels utilisés pour coupler les individus représentés dans les deux bases de données. Dans la plupart des études, l'exactitude du couplage est examinée en estimant les taux de faux couplages (résultats faussement positifs) et de faux non-couplages (résultats faussement négatifs) associés au processus de couplage. En pratique, on procède habituellement au tirage d'un échantillon d'enregistrements couplés et non couplés, puis on détermine l'exactitude des couplages dans l'échantillon en se servant de données auxiliaires provenant d'autres sources.

Bien que le CIE soit utilisé depuis un certain temps dans les études-cohorte de mortalité, l'effet des erreurs de couplage sur la fiabilité des inférences statistiques faites d'après ce genre d'études n'a pas fait l'objet d'un examen détaillé. Les résultats théoriques présentés dans le présent article visent à combler cette lacune. Ces résultats montrent qu'en plus d'accroître le nombre observé de décès, les résultats faussement positifs ont tendance à réduire le nombre attendu de décès. Inversement, les résultats faussement négatifs accroissent le nombre attendu de décès et réduisent le nombre observé de décès. Nous avons montré que les erreurs de couplage introduisent un biais dans les estimations des RSM. Les estimations des coefficients de régression du risque relatif sont également entachées d'un biais, dont la direction dépend de la nature du coefficient de régression. En plus de ces biais, les erreurs de couplage introduisent une incertitude additionnelle dans les estimations des RSM, ainsi que des coefficients de régression.

Bien que nous émettions l'hypothèse simplificatrice que $t_{ij}^1 = t_{ij}^0$, il est possible d'établir les expressions pertinentes du biais et de la variabilité supplémentaire sans le faire; cependant, les expressions sont trop complexes pour fournir des éclaircissements supplémentaires sur les effets des erreurs de couplage. Il en est également ainsi de l'hypothèse selon laquelle $p_{jj}^N = p_j^N$. La définition de A_j pour le ou les états correspondant à la dernière tranche d'âge, qui est habituellement ouverte jusqu'à l'infini du côté droit, pose un problème technique. Dans ces états, l'hypothèse que $t_{ij}^1 = t_{ij}^0$ est problématique si la probabilité de mourir dans cette dernière tranche d'âge est appréciable. On peut contourner le problème en supposant que la durée de vie humaine a une limite supérieure finie.

Comme nous en discutons à la section 3.1, les résultats faussement positifs surviennent principalement lorsqu'un individu en vie à la fin de la période de suivi est couplé incorrectement à une personne décédée. Cependant, une personne décédée dans l'un des états S_j peut être couplée incorrectement à une autre personne décédée à une période antérieure, ce qui donne un résultat faussement positif qui persiste jusqu'au moment réel du décès; l'analyse de la section 3 tient compte de ce genre d'erreur. De même, une personne décédée peut être couplée incorrectement à une autre personne décédée à une date ultérieure, qui n'est pas en vie à la fin de la période de suivi. Ce cas est traité comme un résultat faussement négatif uniquement jusqu'au moment incorrect du décès. À ce moment-là, il y aura une contribution incorrecte au nombre de décès, erreur qui n'a pas été prise en compte à la section 3. Toutefois, ce genre d'erreur ne serait normalement pas décelé dans les études par couplage d'enregistrements habituelles dans lesquelles on procède à une vérification manuelle simplifiée pour repérer les résultats faussement positifs et faussement négatifs. Puisque ce genre d'erreurs est vraisemblablement rare, nous nous attendons à ce que son effet soit faible.

Afin d'étudier plus en détail l'effet éventuel des erreurs de couplage d'enregistrements, supposons que τ_j est la limite d'âge supérieure pour le j^{e} état S_j. (Notons que certains τ_j peuvent être égaux.) Alors, si nous représentons

par α la probabilité d'une erreur de couplage (de l'un ou l'autre type), nous pouvons écrire les taux de résultats faussement positifs et de résultats faussement négatifs, p_j^P et p_j^N, sous la forme $\alpha P[T \le \tau_j]$ et $\alpha P[T > \tau_j]$, respectivement. En particulier, $p_{jj}^P = \alpha P[\tau_{j-1} < T \le \tau_j]$, où τ_{j-1} est la limite inférieure d'âge pour le j^{e} état, et $p_{jj}^N = p_j^N$. Par conséquent, les taux de résultats faussement positifs peuvent être supérieurs aux taux de résultats faussement négatifs pour les groupes d'âge avancé, l'inverse se produisant pour les groupes d'âge plus jeune. Si l'on suppose que le profil de taille est le même pour les D_j et A_j, certains termes s'annulent dans le calcul de $E[\Delta e_j]$ dans (19) et dans celui de $E[\Delta d_{jj}]$ dans (15). Cet effet d'annulation réduira les biais attendus dans le RSM et dans les paramètres de régression du risque donnés par (23) et (38), respectivement.

Bien que nous ayons considéré uniquement la mortalité toutes causes confondues dans le présent article, la mortalité par cause peut être étudiée en apportant des modifications simples aux définitions de D_{jj}, D_{jj}^L et D_{jj}^P. Ces ensembles devraient alors ne tenir compte que des décès dus à la cause particulière étudiée. Par conséquent, d_{jj} et e_j devraient représenter, respectivement, les nombres observé et attendu de décès du type spécifié dans S_j. Dans (1) et (2), la fonction de risque devrait avoir trait au type spécifique de décès, avec $\lambda^*(u)$ le taux de risque par cause de base correspondant. Enfin, à la section 2, l'indicateur δ_i devrait indiquer le type spécifique de décès.

Les résultats analytiques qui précèdent fournissent d'importants éclaircissements sur les effets des erreurs de couplage dans les études-cohorte de la mortalité, mais il est important d'examiner ce genre d'effets dans des conditions aussi proches que possible de celles rencontrées en pratique. À cette fin, nous avons réalisé une étude en simulation informatisée fondée sur des données réelles provenant du Fichier dosimétrique national du Canada, dans laquelle nous avons introduit des couplages incorrects et des non-couplages incorrects avec probabilités connues pour évaluer plus en profondeur l'effet des erreurs de couplage sur les estimations du risque de cancer (Mallick, Krewski, Dewanji et Zielinski 2002). Les résultats de cette simulation corroborent les résultats théoriques exposés dans l'article.

Alors que les résultats présentés ici permettent de mieux comprendre l'effet des erreurs de couplage sur l'inférence statistique, des méthodes tenant compte de ce genre d'erreurs dans les analyses statistiques n'ont pas encore été élaborées. Ces méthodes pourraient s'inspirer des modèles d'erreur de réponse utilisés dans le domaine du sondage, conjugués aux méthodes statistiques classiques d'analyse des données sur la mortalité des cohortes. Des travaux de recherche dans ce domaine sont en cours.

6. Remerciements

La présente étude a été financée en partie par une bourse du Conseil national de recherches en sciences et en génie du Canada octroyée à D. Krewski, qui est titulaire à l'heure actuelle de la chaire CRSNG-CRHS-McLaughlin d'évaluation du risque pour la santé des populations à l'Université d'Ottawa. Des versions préliminaires du présent article ont été présentées à l'Annual Joint Meeting de l'American Statistical Association qui s'est tenue à San Francisco du 8 au 12 août 1993 et à l'Assemblée annuelle de la Société statistique du Canada qui s'est tenue à Montréal du 10 au 16 juillet 1995. La version finale a été présentée à la session dédiée à J.N.K. Rao du Symposium 2001 de Statistique Canada qui a eu lieu à Ottawa le 18 octobre 2001. L'auteur principal (D. Krewski) est particulièrement reconnaissant d'avoir été invité à prendre la parole à la session en l'honneur de J.N.K. Rao, qui avait été son directeur de thèse de doctorat il y a de nombreuses années. L'étude a été achevée pendant les séjours de A. Dewanji au Centre McLaughlin d'évaluation du risque pour la santé des populations à titre de chercheur invité durant les étés de 2002 et de 2003.

Bibliographie

Anderson, T.W. (1974). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons, Inc.

Ardal, S., et Ennis, S. (2001). Enquêtes sur les données : Mise en évidence d'erreurs systématiques dans les bases de données administratives. *Recueil : Symposium 2001, La qualité des données d'un organisme statistique : Une perspective méthodologique*, Statistique Canada, Ottawa.

Ashmore, J.-P., et Grogan, D. (1985). The national dose registry of Canada. *Radiation Protection Dosimetry*, 11, 95-100.

Ashmore, J.-P., et Davies, B.D. (1989). The national dose registry: A centralized record keeping system for radiation workers in Canada. Dans *Applications of Computer Technology to Radiation Protection*, IAEA-SR-136/58, J. Stephan Institute, Ljublyua, 505-520.

Ashmore, J.-P., Krewski, D. and Zielinski, J.M. (1997). Protocol for a cohort mortality study of occupational radiation exposure based on the national dose registry of Canada. *European Journal of Cancer*, 33, S10-S21.

Ashmore, J.-P., Krewski, D., Zielinski, J.M., Jiang, H., Semenciw, R. et Létourneau, E. (1998). First analysis of occupational radiation mortality based on the national dose registry of Canada. *American Journal of Epidemiology*, 148, 564-574.

Bartlett, S., Krewski, D., Wang, Y. et Zielinski, J.M. (1993). Évaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisé. *Techniques d'enquête*, 19, 3-13.

Belin, T.R., et Rubin, D.B. (1991). Recent developments in calibrating error rates for computer matching. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, 657-668.

Breslow, N.E., Lubin, J.H. et Langholz, B. (1983). Multiplicative models and cohort analysis. *Journal of the American Statistical Association*, 78, 1-12.

Breslow, N.E., et Day, N.E. (1987). *Statistical Methods in Cancer Research*, Vol. 2 : *The Design and Analysis of Cohort Studies*. IARC scientific publication No. 82, international agency for research on cancer, Lyon, France.

Carpenter, M., et Fair, M.E. (Eds.) (1990). *Canadian Epidemiology Research Conference – 1989: Proceedings of Record Linkage Sessions & Workshop*. Ottawa Select Printing, Ottawa.

Cox, D.R. (1972). Regression models and life tables (avec discussion). *Journal of Royal Statistical Society*, B, 34, 187-220.

Fair, M.E. (1989). Studies and References Relating to Uses of the Canadian Mortality Data Base. Report from the occupational and environmental health research unit, Division de la Santé, Statistique Canada, Ottawa.

Fellegi, I., et Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Hill, T. (1988). Generalized Iterative Record Linkage System: GIRLS Strategy (Relâcher 2.7). Report from research and general system, informatics services and development division, Statistique Canada, Ottawa.

Howe, G.R., et Lindsay, J. (1981). A generalized iterative record linkage computer system for use in medical follow-up studies. *Computers and Biomedical Research*, 14, 327-340.

Howe, G.R., et Spasoff, R.A. (Eds.) (1986). *Proceeding of the Workshop on Computerized Linkage in Health Research*. University of Toronto Press, Toronto.

Jordan-Simpson, D.A., Fair, M.E. et Poliquin, C. (1990). Étude des exploitants agricoles canadians : Méthodologie. *Rapports sur la santé*, 2, 141-155.

Kalbfleish, J.D., et Prentice, R.L. (1980). *The Statistical Analysis of Failure Time Data*. New York: John Wiley & Sons, Inc.

Labossière, G. (1986). Confidentiality and access to data: The practice at Statistics Canada. *Proceedings of the Workshop on Computerized Record Linkage in Health Research*, University of Toronto Press, Toronto.

Mallick, R., Krewski, D., Dewanji, A. et Zielinski, J.M. (2002). A simulation study of the effect of record linkage errors in cohort mortality data. *Proceedings of International Conference in Recent Advances in Survey Sampling*. Carleton University, Ottawa, à paraître.

Neter, J., Maynes, E.S. et Ramanathan, R. (1965). The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60, 1005-1027.

Newcombe, H.B. (1988). *Handbook of Record Linkage*: *Methods for Health and Statistical Studies*, *Administration*, *and Business. Oxford Medical Publications*. Oxford.

Roos, L.L., Soodeen, R. et Jebamani, L. (2001). Un environnement riche en information : La qualité des données des systèmes d'appariement de dossiers au Canada. *Recueil* : *Symposium 2001*, *La qualité des données d'un organisme statistique* : *Une perspective méthodologique*, Statistique Canada, Ottawa.

Scheuren, F., et Winkler, W.E. (1993). Analyse de regression de fichiers de données couplés par ordinateur. *Techniques d'enquête*, 19, 45-65.

Scheuren, F., et Winkler, W.E. (1997). Analyse de régression des fichiers de données appariés par ordinateur – Partie II. *Techniques d'enquête*, 23, 171-180.

Singh, A.C., Feder, M., Dunteman, G. et Yu, F. (2001). Protection de la confidentialité et maintien de la qualité des microdonnées à grande diffusion. *Recueil* : *Symposium 2001*, *La qualité des données d'un organisme statistique* : *Une perspective méthodologique*, Statistique Canada, Ottawa.

Smith, M.E., et Silins, J. (1981). Generalized iterative record linkage system. *Social Statistics Section*, *Proceedings of the American Statistical Association*, 128-137.

Sont, W.N., Zielinski, J.M., Ashmore, J.P., Jiang, H., Krewski, D., Fair, M.E., Band, P. et Létourneau, E. (2001). First analysis of cancer incidence and occupational radiation exposure based on the national dose registry of Canada. *American Journal of Epidemiology*, 153, 309-318.

Winkler, W.E., et Scheuren, F. (1991). How computer matching error effect regression analysis: Exploratory and confirmatory analysis. Rapport technique, Statistical research division, U.S. Bureau of the Census, Washington, D.C.