Recueil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

IMPACT D'ERREURS D'APPARIEMENT SUR L'INFÉRENCE STATISTIQUE LORS D'ÉTUDES DE MORTALITÉ PAR COHORTES

D. Krewski[1, 2, 4], Y. Wang[3], S. Bartlett[3], J.M. Zielinski[3] et R. Mallick[1,2]

RÉSUMÉ

Grâce aux méthodes de couplage d'enregistrements, il est désormais plus facile d'effectuer des études de mortalité par cohortes où il y a couplage électronique des données d'exposition d'une base d'information et des données de mortalité d'une autre base. Le présent article est consacré à l'incidence des erreurs de couplage sur les estimations d'indicateurs épidémiologiques de risque comme les taux comparatifs de mortalité et les paramètres de modèles de régression de risques relatifs. Il démontre que ces indicateurs peuvent être entachés d'un biais et d'un surcroît de variabilité à cause d'erreurs de couplage, les faux liens et les non-liens introduisant un biais respectivement positif et négatif dans les estimations de taux normaliser de mortalité. Ces erreurs accroissent toujours l'incertitude des estimations, mais on peut réussir à éliminer le biais dans le cas particulier d'une égalité des faux positifs et des faux négatifs pour des états homogènes définis par un croisement des covariables d'intérêt.

MOTS CLÉS : Étude par cohortes; Couplage informatisé d'enregistrements; Erreurs de couplage; Valeurs seuils de pondération de couplage; Régression de Poisson; Régression de risques relatifs; Taux comparatif de mortalité

1. INTRODUCTION

Ces dernières années, on a effectué diverses études chronologiques par cohortes en épidémiologie environnementale en exploitant les bases de données administratives en place comme sources d'information (Howe et Spasoff, 1986; Carpenter et Fair, 1990). En général, il s'agit de coupler les enregistrements sur l'exposition humaine aux risques du milieu et les enregistrements sur l'état de santé, ce qui souvent se fait à l'aide de méthodes informatisées de mise en correspondance individuelle d'enregistrements appartenant à des bases de données différentes. Dans une étude de mortalité par cohortes, on détermine l'état civil de chaque membre d'une cohorte par couplage avec les dossiers de mortalité que tiennent les organismes publics. Un excès de mortalité dans une cohorte par rapport à l'ensemble de la population peut s'expliquer par le degré d'exposition de ses membres.

Plus précisément, il y a couplage d'enregistrements lorsqu'on met en correspondance deux éléments d'information ou plus pour une même entité (Bartlett et coll., 1993). Les méthodes de couplage informatique d'enregistrements (CIE) sont aujourd'hui des plus raffinées. Elles font appel à des algorithmes complexes pour évaluer les probabilités de juste appariement de deux enregistrements (Hill, 1988; Newcombe, 1988).

[1]Centre R. Samuel McLaughlin d'évaluation des risques pour la santé des populations, Université Carleton, Ottawa, Ontario, Canada, K1N 6N5.

[2]School of Mathematics and Statistics, Université Carleton, Ottawa, Ontario, Canada, K1S 5B6.

[3]Direction générale de la santé environnementale et sécurité des consommateurs, Ottawa, Ontario, Canada, K1A 0L2.

[4]C'est la personne avec qui on doit correspondre.

Statistique Canada a mis au point un système CIE appelé CANLINK qui peut traiter les mises en correspondance intrafichiers et interfichiers (Howe et Lindsay, 1981; Smith et Silins, 1981). Dans ce système, on attribue des valeurs de pondération à des paires d'enregistrements en fonction des probabilités d'appariement. On fixe deux seuils. Les appariements ayant une valeur de pondération plus grande que le seuil supérieur sont considérés comme de bons liens. Ceux dont la valeur de pondération se situe en deçà du seuil inférieure sont considérés comme des non-liens. On résout les cas d'appariement possible dont la valeur de pondération est comprise entre les seuils supérieure et inférieure en employant tout complément d'information si on en a, sinon on fixe une valeur seuil unique qui permet de distinguer les liens des non-liens.

On veille jalousement sur la confidentialité des enregistrements protégés par la *Loi sur la statistique* dans toute étude où il y a couplage d'enregistrements. Toutes les études prévoyant une mise en correspondance avec des bases de données protégées doivent satisfaire à de rigoureux critères d'examen et d'approbation avant de se réaliser. Tous les fichiers en couplage renfermant des données d'identification restent sous la garde de Statistique Canada (Labossière, 1986).

On a appliqué des méthodes informatisées de couplage d'enregistrements pour mettre en correspondance les données d'exposition au milieu et la Base canadienne de données sur la mortalité (BCDM). Ainsi, on a entrepris une étude sur les exploitants agricoles canadiens qui vise à dégager les éventuels rapports entre les causes de décès et diverses variables sociodémographiques et agricoles, et plus particulièrement l'utilisation de pesticides, chez plus de 326 000 de ces exploitants (Jordan-Simpson et coll., 1990). Dans cet exercice, on a mis en couplage la BCDM et les recensements de la population et de l'agriculture de 1971. Une autre étude permanente à grande échelle s'appuie sur le Registre national des doses (RND) du Canada (Ashmore et Grogan, 1985; Ashmore et Davies, 1989). Ce registre renferme des données sur l'exposition professionnelle aux rayonnements ionisants chez plus de 400 000 Canadiens depuis 1950. On a récemment effectué le couplage RND-BCDM afin d'examiner les liens entre l'excès de mortalité par cancer et l'exposition professionnelle à de faibles rayonnements ionisants (Ashmore et coll., 1997, 1998). Plus récemment encore, il y a eu raccordement du RND et de la Base canadienne de données sur le cancer (Sont et coll., 2001). Fair (1989) a dressé une liste complète des autres études sanitaires où il y a mise en correspondance de données d'exposition et de l'information de la BCDM.

Les études par couplage d'enregistrements sont avantageuses à plusieurs égards par rapport aux études épidémiologiques classiques. Le recours aux bases de données administratives en place obvie à la nécessité de recueillir de nouvelles données aux fins des études sur la santé, et il est souvent possible d'obtenir assez aisément de grandes tailles d'échantillon. Selon la nature des bases d'information exploitées, le couplage peut être un moyen économique d'examen de nombreux rapports possibles dans des études épidémiologiques. Le couplage d'enregistrements présente aussi certains inconvénients. Ainsi, on est généralement peu maître des données recueillies et le suivi peut être impossible dans un nombre appréciable de cas. Un autre inconvénient est l'apparition d'erreurs de couplage, qui est ici notre propos même. Inévitablement, des enregistrements qui se correspondent ne seront pas couplés et d'autres qui ne se correspondent pas seront couplés par erreur.

On s'est relativement peu employé à établir l'incidence de ces erreurs sur l'inférence statistique. Neter et coll. (1965) se sont servis d'un simple modèle de régression linéaire pour analyser l'incidence des erreurs introduites par le couplage. Leurs résultats indiquent que ces erreurs viennent gonfler la variance résiduelle et introduire un biais dans le paramètre de pente estimé. Winkler et Scheuren (1991) ont exprimé le biais en question dans des estimations de coefficients de régression linéaire. Les progrès de l'estimation des taux d'erreur de couplage grâce aux travaux de Belin et Rubin (1991) ont permis à Scheuren et Winkler (1993) de mettre au point une méthode améliorée de correction de biais.

Notre propos sera de cerner l'incidence des erreurs de couplage sur l'inférence statistique dans des études de mortalité par cohortes. Nous décrirons à la section 2 les modèles de régression de risques relatifs ayant servi à l'analyse des données et établirons des expressions pour le nombre observé et prévu de décès par ces modèles. À la section 3, nous examinerons l'effet des erreurs de couplage sur le dénombrement observé et prévu de décès et d'années-personnes d'exposition. À la section 4, nous analyserons l'incidence de ces mêmes erreurs sur les

estimations de taux normalisés de mortalité (TNM) et de paramètres de régression de risques relatifs. L'un et l'autre de ces types d'erreurs peuvent introduire un biais et un surcroît de variabilité dans les estimations de paramètres. Enfin, nous livrerons nos conclusions à la section 5.

2. MODÈLES DE RÉGRESSION DE RISQUES RELATIFS

Les méthodes statistiques d'analyse sont bien établies dans le cadre des études de mortalité par cohortes (Breslow et Day, 1987). Une telle analyse vise principalement à déterminer si l'exposition à l'agent d'intérêt accroît le taux de mortalité chez les membres d'une cohorte. La mortalité est caractérisée par la fonction de risque, qui spécifie le taux de mortalité comme fonction du temps. Soit *T* le moment du décès. La fonction de risque au moment *u* est alors formellement définie de la manière suivante :

$$\lambda\ (u)=\lim_{\Delta u\downarrow 0}\ \Pr\left\{u\le T<u+\Delta u|T\ge u\right\}. \tag{1}$$

Soit $\lambda_i(u)$ la fonction de risque pour une cause particulière de décès au moment *u* du membre $i=1,...,N$ d'une cohorte de taille *N*. Soit $z_i(u)$ un vecteur correspondant de covariables particulières à ce membre. Nous supposons que ces covariables ont pour effet de modifier le risque de référence $\lambda^*(u)$ suivant le modèle de régression de risques relatifs

$$\lambda_i(u)=\lambda^*(u)\gamma\left\{\beta' z_i(u)\right\}, \tag{2}$$

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

Les modèles multiplicatif et additif sont deux cas d'espèce du modèle général de régression de risques relatifs. On peut définir la fonction γ dans (2) par

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

Lorsque $\rho = 1$, le modèle général se ramène au modèle multiplicatif

$$\lambda_i(u)=\lambda^*(u)\exp\left\{\beta' z_i(u)\right\}. \tag{4}$$

Le modèle des risques proportionnels, qui a été introduit par Cox (1972), sert largement à l'analyse des données de mortalité (Kalbfleisch et Prentice, 1980). Le modèle additif

$$\lambda_i(u)=\lambda^*(u)+\beta' z_i(u) \tag{5}$$

se présente comme un cas limite lorsque $\rho \to 0$.

Soit t_i^0 et t_i^1 les âges respectifs à l'entrée dans l'étude et au moment de l'impossibilité du suivi (par fin de l'étude ou encore retrait ou décès) du i^e membre de la cohorte. Soit δ_i = 1 ou 0 selon que le i^e membre était

décédé ou non au moment de l'impossibilité du suivi. On peut ainsi écrire la fonction de vraisemblance en expression logarithmique selon le modèle des risques relatifs (2) :

$$\log \mathrm{L} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \left\{ \delta_{\mathrm{i}} \log\left(\gamma\left\{\beta' \mathrm{z}_{\mathrm{i}}\left(\mathrm{t}_{\mathrm{i}}^{1}\right)\right\}\right) - \int_{\mathrm{t}_{\mathrm{i}}^{0}}^{\mathrm{t}_{\mathrm{i}}^{1}} \gamma\{\beta' \mathrm{z}_{\mathrm{i}}(\mathrm{u})\}\lambda^{*}(\mathrm{u})\mathrm{du} \right\}. \tag{6}$$

Si on a une seule covariable $z_i(u) \equiv 1$, l'estimation de maximum de vraisemblance de $\theta = \exp\{\beta\}$ se ramène au taux normalisé de mortalité TNM= *OBS/EXP*, où $\mathrm{OBS} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \delta_{\mathrm{i}}$ et $\mathrm{EXP} = \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{e}_{\mathrm{i}}$ sont respectivement le nombre observé et le nombre prévu de décès avec $\mathrm{e}_{\mathrm{i}} = \int_{\mathrm{t}_{\mathrm{i}}^{0}}^{\mathrm{t}_{\mathrm{i}}^{1}} \lambda^{*}(\mathrm{u})\mathrm{du}$.

Dans le cas de grandes tailles d'échantillon, la maximisation de la fonction de vraisemblance (6) peut être d'un calcul laborieux. Breslow et coll. (1983) simplifient le calcul en posant que les covariables prennent des valeurs constantes dans les états où se trouve successivement un membre de la cohorte pendant que dure l'étude. Les états en question sont définis par classement recoupé des covariables d'intérêt. Plus précisément, on pose qu'il y a *J* états $\{\mathrm{S}_{\mathrm{j}};\ \mathrm{j} = 1,\ldots,\mathrm{J}\}$, de sorte que $\mathrm{z}_{\mathrm{i}}(\mathrm{u}) = \mathrm{z}_{\mathrm{j}}$ chaque fois que le i^{e} membre est dans l'état S_j au moment *u*. Ces états s'excluent les uns les autres et il n'y en a pas d'autres, aussi chaque membre de la cohorte se trouve-t-il dans un état et un seul à tout moment *u*. On peut alors écrire la fonction de vraisemblance en expression logarithmique (6) de la manière suivante :

$$\log \mathrm{L} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \left\{ \mathrm{d}_{\mathrm{j}} \log\left(\gamma\left\{\beta' \mathrm{z}_{\mathrm{j}}\right\}\right) - \gamma\left\{\beta' \mathrm{z}_{\mathrm{j}}\right\} \mathrm{e}_{\mathrm{j}} \right\}, \tag{7}$$

où

$$e_j = \sum_{i=1}^{N} \int_{[z_i(u) \in S_j]} \lambda^{*}(u)du \tag{8}$$

est la contribution au nombre prévu de décès de toutes les années-personnes d'observation dans l'état S_j, et où d_j désigne le nombre total de décès dans cet état. Pour $\Lambda_{\mathrm{j}}(\beta) = \log\left(\gamma\left\{\beta' \mathrm{z}_{\mathrm{j}}\right\}\right)$, l'estimation de maximum de vraisemblance $\hat{\beta}$ de β s'obtient comme la solution de l'équation de caractérisation

$$\frac{\partial \log \mathrm{L}}{\partial \beta} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \frac{\partial \Lambda_{\mathrm{j}}\left(\hat{\beta}\right)}{\partial \beta} \left\{ \mathrm{d}_{\mathrm{j}} - \exp\left\{\Lambda_{\mathrm{j}}\left(\hat{\beta}\right)\right\} \mathrm{e}_{\mathrm{j}} \right\} = 0 \tag{9}$$

3. INCIDENCE DES ERREURS DE COUPLAGE SUR LE NOMBRE OBSERVÉ ET PRÉVU DE DÉCÈS

Deux types d'erreurs peuvent se produire en cas de couplage informatisé d'enregistrements (CIE) de fichiers de données (Fellegi et Sunter, 1969). Il y a faux positif lorsqu'on considère à tort comme décédé un membre de la cohorte qui est vivant. Il y a faux négatif lorsqu'on considère comme toujours vivant un sujet décédé. Dans cette section, nous examinerons l'incidence des erreurs de couplage sur le nombre observé et prévu de décès respectivement. Pour ce faire, nous définissons d'abord des jeux d'indices intra-états qui représenteront des ensembles d'enregistrements qui ont été bien ou mal couplés.

3.1 Erreurs de couplage

Soit C_j l'ensemble de descripteurs des membres de la cohorte qui passent par l'état S_j. Dans l'hypothèse d'absence d'erreurs de couplage, soit A_j et D_j les ensembles respectifs de descripteurs des membres qui restent vivants tout au long de l'état S_j et qui décèdent dans ce même état. À noter que $A_j \cup D_j = C_j$ et $A_j \cap D_j = \{\text{ø}\}$.

Soit A_j^L et D_j^L les ensembles de descripteurs A_j et D_j dans l'hypothèse de présence d'erreurs de couplage. Nous définissons en outre D_j^P et A_j^N comme les ensembles d'indices de faux positifs et de faux négatifs dans le *j*e état S_j. Ces ensembles satisfont aux relations $A_j^L = (A_j - D_j^P) \cup A_j^N$ et $D_j^L = (D_j - A_j^N) \cup D_j^P$, où $A_j \cup D_j = C_j$ et $A_j^L \cap D_j^L = \{\text{ø}\}$.

Nous pouvons décrire l'incidence des erreurs de couplage sur la fonction de vraisemblance dans (7). Soit t_{ij}^0, t_{ij}^1 et t_{ij}^2 les moments où le *i*e membre entre, meurt (si la mort survient) et/ou sort du j$^{\text{ième}}$ état S_j. Par (8) et les décompositions de A_j^L et D_j^L, on peut formuler de la manière suivante le nombre prévu de décès e_j^L dans le *j*e état en cas d'erreurs de couplage :

$$
\begin{aligned}
e_j^L &= \sum_{i \in A_j^L} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du + \sum_{i \in D_j^L} \int_{t_{ij}^0}^{t_{ij}^1} \lambda^*(u)du \\
&= \left(\sum_{i \in A_j} \int_{t_{ij}^0}^{t_{ij}^2} + \sum_{i \in A_j^N} \int_{t_{ij}^0}^{t_{ij}^2} - \sum_{i \in D_j^P} \int_{t_{ij}^0}^{t_{ij}^2} \right) \lambda^*(u)du + \\
&\quad \left(\sum_{i \in D_j} \int_{t_{ij}^0}^{t_{ij}^1} + \sum_{i \in D_j^P} \int_{t_{ij}^0}^{t_{ij}^1} - \sum_{i \in A_j^N} \int_{t_{ij}^0}^{t_{ij}^1} \right) \lambda^*(u)du \qquad (10) \\
&= \sum_{i \in A_j} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du + \sum_{i \in D_j} \int_{t_{ij}^0}^{t_{ij}^1} \lambda^*(u)du + \sum_{i \in A_j^N} \int_{t_{ij}^1}^{t_{ij}^2} \lambda^*(u)du - \sum_{i \in D_j^P} \int_{t_{ij}^1}^{t_{ij}^2} \lambda^*(u)du \\
&= e_j - \Delta e_j,
\end{aligned}
$$

où

$$
e_j = \sum_{i \in A_j} \int_{t_{ij}^0}^{t_{ij}^2} \lambda^*(u)du + \sum_{i \in D_j} \int_{t_{ij}^0}^{t_{ij}^1} \lambda^*(u)du \qquad (11)
$$

et

$$
\Delta e_j = e_j^P - e_j^N \qquad (12)
$$

avec

$$e_j^P = \sum_{i \in D_j^P} \int_{t_{ij}^1}^{t_{ij}^2} \lambda^*(u)du \tag{13}$$

et

$$e_j^N = \sum_{i \in A_j^N} \int_{t_{ij}^1}^{t_{ij}^2} \lambda^*(u)du. \tag{14}$$

Le terme Δe_j représente le biais dont est entaché le nombre prévu de décès dans le *j*e état à cause d'erreurs de couplage. Il découle de (10) et (12) que les faux positifs et les faux négatifs tendent respectivement à diminuer et à augmenter le nombre prévu de décès.

Par la décomposition de D_j^L, on peut formuler de la manière suivante le nombre observé de décès d_j^L en cas d'erreurs de couplage :

$$d_j^L = d_j + \Delta d_j, \tag{15}$$

où

$$\Delta d_j = d_j^P - d_j^N. \tag{16}$$

Ici, $d_j = \dim(D_j)$, $d_j^P = \dim(D_j^P)$ et $d_j^N = \dim(A_j^N)$, *dim* (*A*) désignant le nombre d'éléments de l'ensemble *A*. Le terme Δd_j représente le biais dont est entaché le nombre observé de décès dans le *j*e état à cause d'erreurs de couplage. Il découle de (15) et (16) que les faux positifs et les faux négatifs augmenteront et diminueront respectivement le nombre observé de décès.

On établit souvent l'état civil par couplage avec la BCDM, fonds d'information qui est généralement plus vaste que ce qui se rapporte directement à la cohorte d'intérêt. Si on relie à tort les données d'exposition d'une personne vivante à celles d'une personne décédée, cette dernière ne fera habituellement pas partie de la cohorte. Ainsi, les années-personnes d'exposition qu'apporte cette personne se termineront de façon prématurée dans l'année de son présumé décès; les années-personnes d'exposition perdues correspondent au temps s'écoulant entre le présumé décès et la fin du suivi. Par ailleurs, si on relie à tort les données d'exposition d'une personne décédée à celles d'une personne vivante, les années-personnes d'exposition qu'apporte cette dernière augmenteront du nombre d'années compris entre l'année de son décès effectif et la fin de la période de suivi. Ainsi, les faux positifs et les faux négatifs dégonflent et gonflent respectivement le nombre d'années-personnes d'exposition de la cohorte.

3.2 Espérances et variances des biais du nombre observé et prévu de décès

L'incidence des erreurs de couplage sur le nombre observé et prévu de décès dépend des taux de fausse positivité (falsipositivité) et de fausse négativité (falsinégativité). Soit D_{ij} $(\overline{D}_{ij})$ l'événement du décès (ou du non-décès) du *i*e membre de la cohorte dans le *j*e état et L_{ij} $(\overline{L}_{ij})$ l'événement de caractérisation de ce membre comme étant décédé (n'étant pas décédé) à la suite du couplage avec la BCDM. Comme Fellegi et Sunter (1969), nous définissons le taux de fausse positivité comme la probabilité conditionnelle $p_{ij}^I = \Pr\{L_{ij} | \overline{D}_{ij}\}$.

De même, le taux de fausse négativité est $p_{ij}^{II} = \Pr\left\{\overline{L}_{ij} \middle| D_{ij}\right\}$. Par souci de simplicité, nous supposons que les taux de fausse positivité et de fausse négativité sont constants dans chaque état $j = 1,\ldots,J$. Plus précisément, $p_{ij}^{I} = p_j^{I}$ et $p_{ij}^{II} = p_j^{II}$ pour $i \in C_j$. L'hypothèse conviendra chaque fois que les sujets se trouvant dans le même état sont hautement homogènes, plus particulièrement pour des attributs comme la qualité des descripteurs individuels qui influent sur les taux d'erreur de couplage. Il est improbable que l'on satisfasse pleinement dans la pratique à cette hypothèse idéale, mais celle-ci permet de considérablement simplifier l'évaluation subséquente des effets des erreurs de couplage.

Soit ξ_{ij} une variable indicatrice, $\xi_{ij} = 1$ s'il y a faux positif pour le *i*e sujet dans le *j*e état et $\xi_{ij} = 0$ dans les autres cas. De même, soit $\psi_{ij} = 1$ dans le cas d'un faux négatif pour le même sujet et $\psi_{ij} = 0$ dans les autres cas. À supposer que les erreurs de couplage relatives aux divers sujets soient indépendantes, l'espérance et la variance du biais du nombre observé de décès dans le *j*e état Δd_j sont

$$E\left\{d_j^P - d_j^N\right\} = E\left\{\sum_{i \in C_j} \xi_{ij} - \sum_{i \in C_j} \psi_{ij}\right\} = \sum_{i \in C_j} \Delta p_{ij} = \Delta p_j N_j \tag{17}$$

et

$$\begin{aligned} \mathrm{Var}\left\{d_j^P - d_j^N\right\} &= \sum_{i \in C_j} \mathrm{Var}\left\{\xi_{ij}\right\} + \sum_{i \in C_j} \mathrm{Var}\left\{\psi_{ij}\right\} - \sum_{i \in C_j} \mathrm{Cov}\left\{\xi_{ij}, \psi_{ij}\right\} \\ &= \sum_{i \in C_j} \left\{p_{ij}^I\left(1 - p_{ij}^I\right) + p_{ij}^{II}\left(1 - p_{ij}^{II}\right)\right\} \\ &= \left\{p_j^I\left(1 - p_j^I\right) + p_j^{II}\left(1 - p_j^{II}\right)\right\} N_j, \end{aligned} \tag{18}$$

où $\Delta p_{ij} = p_{ij}^I - p_{ij}^{II}$ et $\Delta p_j = p_j^I - p_j^{II}$. À noter que, comme il ne peut y avoir simultanément de faux positif et de faux négatif,

$$\mathrm{Cov}\left\{\xi_{ij}, \psi_{ij}\right\} = E\left\{\left(\xi_{ij}\ \psi_{ij}\right)^2\right\} - \left(E\left\{\xi_{ij}\ \psi_{ij}\right\}\right)^2 = 0. \tag{19}$$

De même, l'espérance et la variance du biais du nombre prévu de décès dans le *j*e état Δe_j peuvent s'exprimer de la manière suivante :

$$E\left\{e_j^P - e_j^N\right\} = \Delta p_j \sum_{i \in C_j} \int_{\min\left(t_{ij}^1, t_{ij}^2\right)}^{t_{ij}^{'}} \lambda^*(u)\,du \tag{20}$$

et

$$\mathrm{Var}\left\{e_j^P - e_j^N\right\} = \left\{p_j^I\left(1 - p_j^I\right) + p_j^{II}\left(1 - p_j^{II}\right)\right\} \sum_{i \in C_j} \left\{\int_{\min\left(t_{ij}^1, t_{ij}^2\right)}^{t_{ij}^2} \lambda^*(u)\,du\right\}^2, \tag{21}$$

Les résultats indiquent que, si les taux de fausse positivité et de fausse négativité sont égaux dans chaque état, le nombre observé de décès sera égal en moyenne au nombre effectif de décès, auquel cas le biais dont est

entaché le nombre prévu de décès par erreurs de couplage disparaîtra. Il reste que les deux types d'erreurs de couplage introduiront un surcroît de variation dans le nombre observé et prévu de décès.

Il est difficile de minimiser la variance du biais dans (21), puisque p_j^I et p_j^{II} ne sont pas fonctionnellement indépendants. En règle générale, décroître p_j^I, c'est accroître p_j^{II} et vice versa. On pourrait néanmoins minimiser le facteur multiplicatif $\left\{p_j^I\left(1-p_j^I\right)+p_j^{II}\left(1-p_j^{II}\right)\right\}$ dans (21) compte tenu de la relation fonctionnelle entre p_j^I et p_j^{II}. Bien que ce facteur soit indépendant du modèle sous-jacent de régression de risques relatifs γ dans (2), l'erreur quadratique moyenne obtenue par une combinaison de (20) et (21) ne saurait être minimisée sans une spécification du risque de référence $\lambda^*(u)$.

4. INCIDENCE DES ERREURS DE COUPLAGE SUR LES ESTIMATIONS DES TAUX NORMALISÉS DE MORTALITÉ ET DES COEFFICIENTS DE RÉGRESSION

4.1 Taux normalisés de mortalité

Pour déterminer l'incidence des erreurs de couplage sur les taux normalisés de mortalité (TNM), nous remplaçons les nombres et observé et prévu de décès d_j et e_j dans une hypothèse d'absence d'erreurs de couplage par les nombres et observé et prévu de décès d_j^L et e_j^L dans une hypothèse de présence de telles erreurs dans l'expression $TNM = \sum d_j / \sum e_j$. Soit TNM_L les taux normalisés de mortalité en cas d'erreurs de couplage. Nous avons alors

$$TNM_L = TNM\left[1+\frac{\sum \Delta d_j}{\sum d_j}\right]/\left[1-\frac{\sum \Delta e_j}{\sum e_j}\right]. \qquad (22)$$

Il s'ensuit que les faux positifs et les faux négatifs augmenteront et diminueront respectivement les TNM.

On peut ainsi exprimer la différence $\Delta TNM = TNM_L - TNM$:

$$\frac{\Delta TNM}{TNM} = \frac{\sum \Delta d_j}{\sum d_j} + \frac{\sum \Delta e_j}{\sum e_j} + o\left(\frac{\sum \Delta e_j}{\sum e_j}\right). \qquad (23)$$

Dans la plupart des études par cohortes, le nombre de décès est bien moindre que la taille de l'ensemble de la cohorte, auquel cas le terme d'erreur est petit dans (23). On peut approcher la moyenne et la variance du biais relatif du TNM par

$$E\left\{\frac{\Delta TNM}{TNM}\right\} \approx \sum \Delta p_j \left\{\left(\sum d_j\right)^{-1} N_j + \left(\sum e_j\right)^{-1} \sum_{i \in C_j} \int_{\min\left(t_{ij}^1, t_{ij}^2\right)}^{t_{ij}^2} \lambda^*(u)du\right\} \qquad (24)$$

et

$$Var\left\{\frac{\Delta TNM}{TNM}\right\} \approx \sum\left\{p_j^I\ \left(1-p_j^I\right)+ p_j^{II}\ \left(1-p_j^{II}\right)\right\}\left\{\left(\sum d_j\right)^{-2}\ N_j\ + \right.$$

$$\left. + 2\left(\sum d_j\right)^{-1}\left(\sum e_j\right)^{-1}\ \sum_{i\in C_j}\int_{\min\left(t_{ij}^1,t_{ij}^2\right)}^{t_{ij}^2}\ \lambda^*\left(u\right)du + \left(\sum e_j\right)^{-2}\ \sum_{i\in C_j}\left(\int_{\min\left(t_{ij}^1,t_{ij}^2\right)}^{t_{ij}^2}\ \lambda^*\left(u\right)du\right)^2\ \right\} \tag{25}$$

respectivement.

On peut tirer trois conclusions de (24) et (25). Premièrement, si le taux de fausse positivité est supérieur (inférieur) au taux de fausse négativité, il y aura en moyenne surestimation (sous-estimation) du TNM. Deuxièmement, si les taux de fausse positivité et de fausse négativité sont égaux dans chaque état, les biais introduits dans le TNM par les erreurs de couplage s'annuleront en réalité. Troisièmement, les deux types d'erreurs de couplage introduisent un surcroît de variation dans les estimations du TNM.

4.2 Paramètres de régression de risques relatifs

Pour établir l'incidence des erreurs de couplage sur les estimations de paramètres de régression, considérons d'abord le modèle général de régression de risques relatifs (2). Si nous remplaçons les nombres et observé et prévu de décès $\mathrm{d_j}$ et $\mathrm{e_j}$ dans la fonction de vraisemblance en expression logarithmique (7) par les nombres et observé et prévu de décès en cas d'erreurs de couplage $\mathrm{d_j^L}$ et $\mathrm{e_j^L}$, nous obtenons

$$\log \mathrm{L} = \sum_{\mathrm{j}=1}^{\mathrm{J}}\left\{\mathrm{d_j^L}\ \log\left(\gamma\left\{\beta'\,\mathbf{z}_\mathrm{j}\right\}\right) - \gamma\left\{\beta'\,\mathbf{z}_\mathrm{j}\right\}\mathrm{e_j^L}\right\}. \tag{26}$$

Soit $\hat{\beta}$ et $\tilde{\beta}$ les estimateurs de maximum de vraisemblance (EMV) de β en fonction de $\left\{\mathrm{d_j}, \mathrm{e_j}\right\}$ et $\left\{\mathrm{d_j^L}, \mathrm{e_j^L}\right\}$ respectivement. Si $\hat{\Lambda}_\mathrm{j} = \Lambda_\mathrm{j}\left(\hat{\beta}\right)$, la fonction de caractérisation (9) peut s'écrire de la manière suivante :

$$\sum_{\mathrm{j}=1}^{\mathrm{J}}\frac{\partial\Lambda_\mathrm{j}\left(\tilde{\beta}\right)}{\partial\beta}\left\{\mathrm{d_j} + \Delta\,\mathrm{d_j} - \exp\left\{\hat{\Lambda}_\mathrm{j} + \Delta\,\Lambda_\mathrm{j}\right\}\left(\mathrm{e_j} - \Delta\,\mathrm{e_j}\right)\right\} = 0 \tag{27}$$

où $\Delta\,\Lambda_\mathrm{j} = \Lambda\left(\tilde{\beta}\right) - \Lambda\left(\hat{\beta}\right)$. Si nous supposons que $\Delta\,\beta = \tilde{\beta} - \hat{\beta}$ est petit, un développement du premier ordre de $\exp\left\{\hat{\Lambda}_\mathrm{j} + \Delta\,\Lambda_\mathrm{j}\right\}$ autour de $\hat{\beta}$ donne

$$\exp\left\{\hat{\Lambda}_\mathrm{j} + \Delta\,\Lambda_\mathrm{j}\right\} \approx \exp\left\{\hat{\Lambda}_\mathrm{j}\right\} + \exp\left\{\hat{\Lambda}_\mathrm{j}\right\}\frac{\partial\,\hat{\Lambda}_\mathrm{j}}{\partial\beta}\,\Delta\,\beta \tag{28}$$

Si on substitue (28) dans (27), on obtient :

$$\sum_{j=1}^{J} \frac{\partial \Lambda_j(\hat{\beta})}{\partial \beta}\left[d_j - \exp\{\hat{\Lambda}_j\} e_j\right] + \sum_{j=1}^{J} \frac{\partial \Lambda_j(\hat{\beta})}{\partial \beta}\left[\Delta d_j + \gamma\{\hat{\beta}' z_j\}\Delta e_j - \right.$$
$$\left. - \gamma\{\hat{\beta}' z_j\} e_j \frac{\partial \Lambda_j'(\hat{\beta})}{\partial \beta}\Delta\beta + \gamma\{\hat{\beta}' \mathbf{z}_j\}\Delta e_j \frac{\partial \Lambda_j'(\hat{\beta})}{\partial \beta}\Delta\beta\right] \approx 0. \tag{29}$$

Si on utilise (9), la première sommation dans (29) est nulle. Ainsi, comme $\Delta e_j \Delta\beta$ est petit, $\Delta\beta$ peut être approximé par

$$\Delta\beta \approx \left\{\sum \frac{\partial \hat{\Lambda}_j}{\partial \beta}\gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \hat{\Lambda}_j'}{\partial \beta}\right\}^{-1} \sum \frac{\partial \hat{\Lambda}_j}{\partial \beta}\{\Delta d_j + \gamma\{\hat{\beta}' \mathbf{z}_j\}\Delta e_j\} \tag{30}$$

Il découle de (30) que

$$E\{\Delta\beta\} \approx \left\{\sum \frac{\partial \Lambda_j}{\partial \beta}\gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta}\right\}^{-1} \sum \frac{\partial \Lambda_j}{\partial \beta}\Delta p_j \alpha_j, \tag{31}$$

où

$$\alpha_j = N_j + \gamma\{\hat{\beta}' \mathbf{z}_j\} \sum_{i \in C_j} \int_{\min(t_{ij}^1, t_{ij}^2)}^{t_{ij}^2} \lambda^*(u) du. \tag{32}$$

De plus,

$$\mathrm{Var}\{\Delta\beta\} \approx \left\{\sum \frac{\partial \Lambda_j}{\partial \beta}\gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta}\right\}^{-1} \left[\sum \frac{\partial \Lambda_j}{\partial \beta}\Theta_j \frac{\partial \Lambda_j'}{\partial \beta}\right] \left\{\sum \frac{\partial \Lambda_j}{\partial \beta}\gamma\{\hat{\beta}' \mathbf{z}_j\} e_j \frac{\partial \Lambda_j'}{\partial \beta}\right\}^{-1}, \tag{33}$$

avec

$$\Theta_j = \{p_j^I(1-p_j^I) + p_j^{II}(1-p_j^{II})\} \sum_{i \in C_j} \left\{1 + \gamma\{\hat{\beta}' z_j\} \int_{\min(t_{ij}^1, t_{ij}^2)}^{t_{ij}^2} \lambda^*(u) du\right\}^2. \tag{34}$$

Dans le cas d'espèce du modèle multiplicatif (4), le biais $\Delta\beta$ par erreurs de couplage peut être approché par

$$\Delta\beta \approx (X' W X)^{-1} X'(\Delta D + \Delta W), \tag{35}$$

où $X' = (\mathbf{z}_1', ..., \mathbf{z}_J')$, $\Delta D' = (\Delta d_1, ..., \Delta d_J)$, $W = \mathrm{diag}\left(\exp(\mathbf{z}_1'\hat{\beta}) e_1, ..., \exp(\mathbf{z}_J'\hat{\beta}) e_J\right)$, and $\Delta W' = \left(\exp(\mathbf{z}_1'\hat{\beta})\Delta e_1, ..., \exp(\mathbf{z}_J'\hat{\beta})\Delta e_J\right)$. À noter que la matrice de pondération *W* est la matrice de pondération de Fisher pour $\hat{\beta}$. De (35), il s'ensuit que

$$E\{\Delta\beta\} \approx (X' W X)^{-1} X'\Pi, \tag{36}$$

où $\Pi' = (\pi_1, ..., \pi_J)$ avec

$$\pi_j = \Delta\, p_j \left\{ N_j + \exp\left(\mathbf{z}'_j\, \hat{\beta} \right) \sum_{i \in C_j} \int_{\min\left(t_{ij}^1, t_{ij}^2\right)}^{t_{ij}^2} \lambda^{*}(u) du \right\}. \tag{37}$$

De plus,

$$\mathrm{Var}\left\{\Delta\beta\right\} \approx \left(X'\, W\, X\right)^{-1}\, X'\, \Psi\, X\, \left(X'\, W\, X\right)^{-1}, \tag{38}$$

où $\Psi = \mathrm{diag}(\psi_1, ..., \psi_J)$ avec éléments diagonaux

$$\psi_j = \left\{ p_j^{I}\left(1 - p_j^{I}\right) + p_j^{II}\left(1 - p_j^{II}\right) \right\} \sum_{i \in C_j} \left\{ 1 + \exp\left(\mathbf{z}'_j\, \hat{\beta}\right) \int_{\min\left(t_{ij}^1, t_{ij}^2\right)}^{t_{ij}^2} \lambda^{*}(u)\, du \right\}^2. \tag{39}$$

Avec une covariable unique $z_i = 1$, $X'\, W\, X = e^{\hat{\beta}}\, \Sigma e_j$, $X'\, \Delta\, D = \Sigma d_j$ et $X'\, \Delta\, W = e^{\hat{\beta}}\, \Sigma\, \Delta e_j$. Dans ce cas,

$$\Delta\beta \approx \left(\sum \Delta d_j + e^{\hat{\beta}} \sum \Delta e_j \right) / \left(e^{\hat{\beta}} \sum e_j \right). \tag{40}$$

Comme

$$TNM = e^{\hat{\beta}} = \frac{\Sigma\, d_j}{\Sigma\, e_j}, \tag{41}$$

avec $\Delta\, \beta = \Delta TNM\, /\, TNM$ dans ce cas, nous obtenons

$$\Delta\, \beta \approx \frac{\Sigma\, \Delta d_j}{\Sigma d_j} + \frac{\Sigma\, \Delta e_j}{\Sigma e_j}. \tag{42}$$

Ainsi, (42) peut être considéré comme un cas d'espèce de (23).

Les résultats qui précèdent indiquent que, en cas d'égalité des taux de fausse positivité et de fausse négativité dans chaque état, les biais des estimations des paramètres de régression par erreurs de couplage seront presque éliminés. Il faut cependant préciser que et les faux positifs et les faux négatifs introduiront un surcroît de variation dans les estimations des paramètres de régression de risques relatifs.

5. CONCLUSIONS

On utilise les méthodes informatisées de couplage d'enregistrements depuis un certain temps dans les études de mortalité par cohortes, mais on n'a pas encore soumis à un examen détaillé la question de l'incidence des

erreurs de couplage sur la fiabilité des inférences statistiques qui en sont tirées. Les résultats théoriques que nous avons présentés vont en ce sens.

Ils montrent que les faux positifs non seulement gonflent le nombre observé de décès, mais tendent aussi à dégonfler le nombre prévu de décès. À l'inverse, les faux négatifs gonflent le nombre prévu et dégonflent le nombre observé. On a également pu voir que les erreurs de couplage introduisent un biais dans les estimations de TNM, les faux liens et les faux non-liens étant généralement source d'un biais positif et d'un biais négatif. L'estimation des coefficients de régression de risques relatifs est également susceptible d'être entachée d'un biais, mais le sens de ce biais dépendra de la nature des coefficients. Il n'y a pas que les biais que nous venons d'énumérer, puisque les erreurs de couplage ajoutent de l'incertitude aux estimations tant des TNM que des coefficients de régression.

Il est possible de presque éliminer les biais des estimations de TNM si les taux de fausse positivité et de fausse négativité sont égaux dans chaque état défini par un classement recoupé des covariables d'intérêt, à condition que toutes ces covariables prennent des valeurs constantes pour chacun des états en question. Il s'agit de conditions exigentes auxquelles il est improbable que l'on satisfasse dans la pratique. Il reste que, d'après nos résultats, il serait sans doute souhaitable d'établir des valeurs seuils de couplage qui équilibrent séparément les taux de fausse positivité et de fausse négativité dans des états homogènes.

Les résultats analytiques que nous avons livrés répandent un éclairage considérable sur les effets que peuvent avoir les erreurs de couplage dans les études de mortalité par cohortes, mais il importe d'examiner de tels effets dans des conditions qui soient le plus proches possible de celles qui pourraient se présenter dans la pratique. Voilà pourquoi nous effectuons actuellement, à l'aide de données réelles du Registre national des doses du Canada, une étude par simulation informatique où l'introduction de faux liens et de faux non-liens aux probabilités connues sert à mieux cerner l'incidence des erreurs de couplage sur les estimations de risque de cancer.

Nos résultats peuvent aider à clarifier l'incidence des erreurs de couplage sur l'inférence statistique, mais il reste à développer les méthodes par lesquelles ces erreurs sont prises en compte dans les analyses statistiques. Il peut s'agir de modèles d'erreurs de réponse employés en échantillonnage d'enquête combinées avec les méthodes statistiques classiques d'analyse des données de mortalité par cohortes. Des recherches sont aussi en cours dans ce domaine.

REMERCIEMENTS

Ces travaux de recherche ont été en partie financés par une subvention accordée par le Conseil national de recherches en sciences naturelles et en génie du Canada à D. Krewski, titulaire actuel de la chaire CRSNG-CRSH- McLaughlin en évaluation des risques pour la santé des populations à l'Université d'Ottawa. De premières versions du présent document ont été présentées au congrès annuel mixte de l'American Statistical Association à San Francisco du 8 au 12 août 1993 et au congrès annuel de la Société statistique du Canada à Montréal du 10 au 16 juillet 1995. On a eu droit à la version définitive à la séance où on a honoré J.N.K. Rao (18 octobre 2001 à Ottawa) au symposium 2001 de Statistique Canada. Le premier de ses auteurs (D. Krewski) est particulièrement reconnaissant d'avoir été invité à prendre la parole à cette occasion. J.N.K. Rao a été son directeur de thèse de doctorat à une certaine époque.

BIBLIOGRAPHIE

Ashmore, J.P. et Grogan, D. (1985), AThe National Dose Registry of Canada@, *Radiation Protection Dosimetry*, 11, 95-100.

Ashmore, J.P. et Davies, B.D. (1989), AThe National Dose Registry: A Centralized Record Keeping System for Radiation Workers in Canada@, in *Applications of Computer Technology to Radiation Protection. IAEA-SR-136/58, J. Stephan Institute,* Ljublyua, 505-520.

Ashmore, J.P. Krewski, D. et Zielinski, J.M. (1997), AProtocol 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), AFirst 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), AEvaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisé@, *Technique d'enquête,* 19, 3-13.

Belin, T.R. et Rubin, D.B. (1991), ARecent 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), AMultiplicative 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), ARegression Models and Life Tables (with discussion)@, *Journal of Royal Statistical Society*, B34, 187-220.

Efron, B. (1982), AThe Jackknife, the Bootstrap and Other Resampling Plans@, *Society for Industral and Applied Mathematics,* Philadelphia, Pennsylvania, 13-19.

Fair, M.E. (1989), AStudies and References Relating to Uses of the Canadian Mortality Data Base@, Rapport de la Occupational and Environmental Health Research Unit, Division de la santé, Statistique Canada, Ottawa.

Fellegi, I. et Sunter, A. (1969), AA Theory for Record Linkage@, *Journal of the American Statistical Association,* 64, 1183-1210.

Hill, T. (1988), AGeneralized Iterative Record Linkage System: GIRLS Strategy (Release 2.7)@, Rapport du Research and General System, Division du développement de systèmes informatiques, Statistique Canada, Ottawa.

Howe, G.R. et Lindsay, J. (1981), AA 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), ACanadian Farm Operator Study: Methodology@, *Health Reports.* 2, 141-155.

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

Labossiére, G. (1986), AConfidentiality 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.

Neter, J., Maynes, E.S., et Ramanathan, R. (1965), AThe 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.

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

Smith, M.E. et Silins, J. (1981), AGeneralized 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. and Létourneau, E. (2001), AFirst 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), AHow Computer Matching Error effect Regression Analysis: Exploratory and Confirmatory Analysis@, Technical Report, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.