Techniques d'enquête, juin 1993
Vol. 19, n° 1, pp. 3–13
Statistique Canada

Évaluation des taux d'erreur dans de grandes études par couplage d'enregistrements informatisé

S. BARTLETT, D. KREWSKI, Y. WANG et J.M. ZIELINSKI[1]

RÉSUMÉ

Les études épidémiologiques qui visent à étudier le rapport entre les risques environnementaux et l'état de santé comptent beaucoup sur l'appariement d'enregistrements de bases de données administratives différentes. Par des algorithmes complexes de couplage d'enregistrements appliqués à de grandes bases de données, on peut évaluer la possibilité d'un appariement de deux enregistrements particuliers en se fondant sur la comparaison d'une ou de plusieurs variables d'identification dans ces enregistrements. Puisque les erreurs d'appariement sont inévitables, il faut pouvoir tenir compte de leur effet sur les inférences statistiques faites à partir des fichiers couplés. Cet article donne un aperçu de la méthodologie utilisée pour le couplage d'enregistrements et traite les questions statistiques qui se rattachent aux erreurs de couplage.

MOTS CLÉS: Couplage d'enregistrements informatisé; étude des exploitants agricoles canadiens; étude de mortalité fondée sur le Fichier dosimétrique national; sélection de seuil.

1. INTRODUCTION

Depuis quelques années, les spécialistes de l'épidémiologie environnementale utilisent de plus en plus des bases de données administratives comme sources de renseignements pour les études sur la santé (Howe et Spasoff 1986; Carpenter et Fair 1990). Grosso modo, cette utilisation consiste à coupler des enregistrements sur l'exposition des personnes aux risques environnementaux avec des enregistrements sur l'état de santé en appariant, souvent par des méthodes informatiques, des enregistrements de bases de données différentes (Newcombe 1988).

On a utilisé récemment des méthodes de couplage d'enregistrements informatisé (CEI) pour faire une étude de mortalité sur plus de 326,000 exploitants agricoles du Canada au regard des pratiques agricoles (Jordan-Simpson *et al.* 1990). Dans cette étude, on a couplé la Base canadienne de données sur la mortalité (BCDM) au recensement de la population de 1971 et au recensement de l'agriculture de la même année. Selon des résultats provisoires basés sur un groupe de 70,000 exploitants agricoles de sexe masculin de la Saskatchewan, l'ensemble de la cohorte n'affichait pas de taux de mortalité excessif pour des causes de décès particulières, sauf qu'on notait l'existence d'une relation dose-réponse entre la mortalité due au lymphome non hodgkinien et la superficie traitée aux herbicides pour les exploitations de moins de 1,000 acres (Wigle *et al.* 1990).

Une autre grande étude en cours qui utilise le couplage d'enregistrements est basée sur le Fichier dosimétrique national (FDN) du Canada. Le FDN contient des données sur l'exposition professionnelle au rayonnement ionisant pour environ 255,000 Canadiens; ces données remontent aussi loin que 1950. On a couplé récemment le FDN à la BCDM afin d'étudier les liens possibles entre l'exposition au rayonnement ionisant et la mortalité par cancer (Ashmore *et al.* 1993).

Il s'est fait d'autres études sur la santé où on a couplé des données sur l'exposition professionnelle à la BCDM. Howe *et al.* (1987) ont observé un nombre significativement élevé de cas de cancer du poumon chez les travailleurs des mines d'uranium dans les Territoires du Nord-Ouest. De même, une étude de cohortes a permis d'établir un rapport significatif entre le cancer du poumon et les fumées des moteurs diesels ou la poussière de charbon chez des retraités de la Compagnie des chemins de fer nationaux du Canada (Howe *et al.* 1983). Shannon *et al.* (1984) ont couplé les fiches d'emploi de travailleurs du nickel de l'Ontario à la BCDM et ont observé chez ce groupe un taux de mortalité excessif pour le cancer du larynx et le cancer du poumon. Morrison *et al.* (1988) ont déterminé un risque significativement élevé de cancer du poumon, des glandes salivaires, de la cavité buccale et du pharynx chez les travailleurs des mines de fluorite de Terre-Neuve. Mao *et al.* (1988) ont eu recours au CEI pour coupler la BCDM au registre du cancer de l'Alberta afin de déterminer les taux de survie après diagnostic pour plusieurs types de cancer. Par ailleurs, on a couplé la base de données de l'Enquête sur la population active du Canada à la BCDM pour faire une étude de mortalité pour diverses professions (Howe et Lindsay 1983). Fair (1989) dresse une liste exhaustive des études sur la santé qui reposent sur le couplage de données sur l'exposition à la BCDM.

Le couplage d'enregistrements est l'opération par laquelle on associe deux ou plusieurs éléments d'information distincts qui se rapportent au même individu. Les méthodes de couplage d'enregistrements informatisé sont devenues très perfectionnées; elles utilisent en effet des

[1] S. Bartlett, D. Krewski, Y. Wang et J.M. Zielinski, Direction de l'hygiène du milieu; Direction générale de la protection de la santé, Santé et Bien-être social Canada, Ottawa (Ontario) Canada K1A 0L2.

algorithmes sophistiqués pour évaluer la probabilité d'une concordance entre deux enregistrements (Hill 1988; Newcombe 1988). Statistique Canada a mis au point un système de CEI appelé CANLINK qui peut effectuer des couplages à l'intérieur d'un même fichier – ou couplages internes – aussi bien que des couplages entre deux fichiers distincts (Howe et Lindsay 1981; Smith et Silins 1981).

La confidentialité des enregistrements protégés en vertu de Loi sur la statistique est rigoureusement respectée si ces enregistrements doivent servir dans des études qui nécessitent des couplages. Toutes les études qui nécessitent des couplages avec des bases de données protégées doivent être soumises à un processus d'examen et d'approbation minutieux avant d'être réalisées. Tous les fichiers couplés qui renferment des données personnelles demeurent sous la garde de Statistique Canada (Labossière 1986).

Les études par couplage d'enregistrements offrent plusieurs avantages par rapport aux études épidémiologiques classiques. En utilisant des bases de données administratives existantes, on n'a plus besoin de recueillir de nouvelles données pour des études sur la santé. En ayant accès à des bases de données existantes, on peut souvent obtenir de grands échantillons avec assez de facilité. Suivant la nature des bases de données utilisées, le couplage d'enregistrements est une façon peu coûteuse d'examiner de nombreuses associations possibles dans des études épidémiologiques.

Le couplage d'enregistrements a aussi des inconvénients. Des erreurs d'appariement peuvent se produire à cause de différences de codage ou de la non-unicité des identificateurs. On a généralement peu de contrôle sur les données recueillies et dans beaucoup de cas, les opérations de suivi peuvent être infructueuses. En outre, les études par couplage d'enregistrements présentent les mêmes lacunes que les études épidémiologiques classiques, par exemple possibilité de biais, confusion, et difficulté à reconnaître de faibles associations entre le milieu et la santé.

Cet article a pour but d'analyser l'utilisation du couplage d'enregistrements informatisé dans des enquêtes épidémiologiques fondées sur des dossiers administratifs touchant la santé et l'environnement. Notre analyse s'intéresse particulièrement à l'effet des faux liens sur les inférences statistiques concernant les risques environnementaux. Dans la section 2, nous étudions des algorithmes pour le couplage d'enregistrements informatisé. La section 3 décrit l'application du couplage d'enregistrements dans des études sur l'exposition professionnelle au rayonnement ionisant et aux produits chimiques agricoles. Dans la section 4, nous nous penchons sur les questions statistiques que soulève l'analyse de bases de données constituées par le couplage d'enregistrements. Enfin, dans la section 5, nous présentons nos conclusions relativement à l'utilisation du couplage d'enregistrements en épidémiologie du milieu.

2. QUESTIONS RELATIVES AU COUPLAGE D'ENREGISTREMENTS

2.1 Position du problème

Considérons deux fichiers informatiques, **A** et **B**, qui renferment, dans le premier cas, des données sur la santé et, dans le second cas, des données sur l'exposition à des risques environnementaux, pour deux groupes de personnes. Chaque fichier est constitué d'un certain nombre d'enregistrements ou d'"observations" qui renferment chacun un certain nombre de zones ou "composantes". Chaque observation correspond habituellement à un membre de la population. Les zones représentent des attributs, comme le nom, l'adresse, l'âge et le sexe, qui caractérisent les observations. Le couplage d'enregistrements sert à déceler et à coupler les observations de chaque fichier qui correspondent au même individu (figure 1). Dans cet exemple,

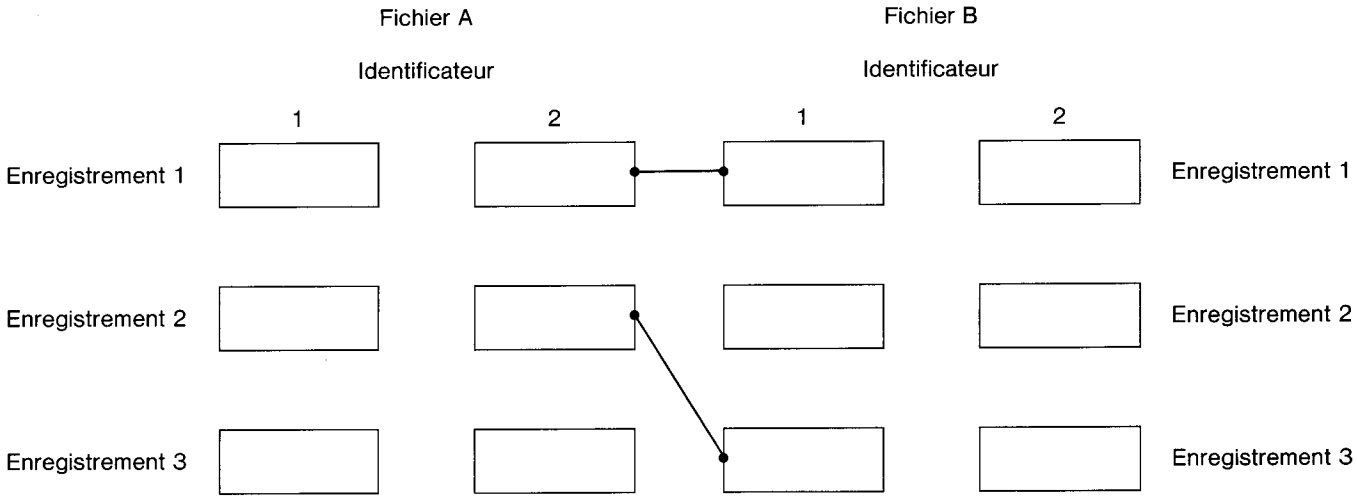

Figure 1. Schéma du couplage de deux fichiers

l'enregistrement 1 du fichier A concorde avec l'enregistrement 1 du fichier B et l'enregistrement 2 de la base de données A concorde avec l'enregistrement 3 du fichier B. L'enregistrement 3 du fichier A ne correspond à aucun des enregistrements du fichier B et l'enregistrement 2 du fichier B ne correspond non plus à aucun des enregistrements du fichier A.

Si les enregistrements renferment des identificateurs uniques qui ont été attribués avec précision, l'appariement sera une opération élémentaire. Le numéro d'assurance sociale est un exemple d'identificateur propre à un individu. Cependant, il n'existe pas toujours d'identificateurs uniques; dans ces circonstances, on ne peut effectuer de couplage "brut" et il faut donc envisager une forme quelconque de couplage probabiliste (voir section 2.3). Grâce à ce mode de couplage, on peut calculer la probabilité d'une concordance et utiliser un système de poids de couplage pour déterminer les liens et les non-liens.

2.2 Système de couplage d'enregistrements informatisé (CEI)

Dans un mode de couplage probabiliste, la décision préliminaire d'apparier ou non dépend d'un poids établi par suite de la comparaison de composantes de deux enregistrements (Newcombe 1988). Ce poids reflète la probabilité que cette paire d'enregistrements constitue un vrai lien: plus le poids est élevé, plus la paire est susceptible de constituer un vrai lien. Le poids dépend ordinairement de la probabilité d'appariement calculée au moment de la comparaison de deux enregistrements,

$$\frac{P(M|AB\ldots Z)}{P(\bar{M}|AB\ldots Z)} = \frac{P(A|M)P(B|M)\ldots P(Z|M)P(M)}{P(A|\bar{M})P(B|\bar{M})\ldots P(Z|\bar{M})P(\bar{M})} .$$

Dans l'expression ci-dessus, M désigne l'événement - c'est-à-dire la concordance de deux enregistrements - et $\{A, B, \ldots, Z\}$ sont les résultats de la comparaison d'identificateurs personnels. Le poids w est défini par le logit

$$w = log_2\left\{\frac{P(M|AB\ldots Z)}{P(\bar{M}|AB\ldots Z)}\right\}$$

$$= W_a + W_b + \ldots + W_z + W,$$

où

$$W_j = log_2\left\{\frac{P(J|M)}{P(J|\bar{M})}\right\}$$

pour tous $J \in \{A, B, \ldots, Z\}$, et

$$W = log_2\left\{\frac{P(M)}{P(\bar{M})}\right\} .$$

Notons que pour obtenir une valeur absolue, on doit connaître le nombre de concordances vraies et le nombre de non-concordances. Sinon, on ne peut déterminer que la probabilité relative. Le poids calculé par le système de CEI de Statistique Canada est la différence de logits.

On a élaboré des algorithmes visant à attribuer des poids qui reflètent la probabilité de couplage de deux enregistrements; à cette fin, on a posé l'hypothèse que les probabilités de concordance pour chacun des identificateurs pris individuellement sont statistiquement indépendantes (Howe et Lindsay 1981). Toutefois, certains identificateurs peuvent être corrélés, ce qui introduit un biais dans l'attribution du poids global.

Fellegi et Sunter (1969) ont proposé un modèle mathématique dans le but de définir un cadre théorique pour le couplage d'enregistrements. Dans ce modèle, le poids tient compte des probabilités d'erreur pour chaque zone en utilisant un rapport de vraisemblance; le poids w est défini

$$w = \sum_{i \in \{\text{zones}\}} w_i ,$$

où

$$w_i = \begin{cases} log_2\{m_i/u_i\} & \text{si les zones } i \text{ de deux enregistrements concordent} \\ log_2\{(1-m_i)/(1-u_i)\} & \text{si les zones } i \text{ de deux enregistrements ne concordent pas,} \end{cases}$$

avec

$$m_i = \Pr\{\text{concordance des zones } i \mid \text{paire d'enregistrements} \in \boldsymbol{M}\} \quad (1)$$

et

$$u_i = \Pr\{\text{concordance des zones } i \mid \text{paire d'enregistrements} \in \boldsymbol{U}\} . \quad (2)$$

Dans les équations ci-dessus, $\boldsymbol{M}$ est un ensemble de paires d'enregistrements qui sont des concordances et $\boldsymbol{U}$ est un ensemble de paires d'enregistrements qui sont des non-concordances. Les résultats de chaque comparaison de zones sont aussi supposés statistiquement indépendants (Jaro 1989).

Newcombe (1988), Fellegi et Sunter (1969), Tepping (1968) et Copas et Hilton (1990) ont élaboré diverses méthodes - probabilistes ou basées sur un modèle - pour attribuer des poids à des composantes (zones) d'enregistrements. Un système probabiliste comme celui utilisé à Statistique Canada détermine les poids de couplage en calculant le logarithme de la probabilité empirique d'une concordance; d'autres systèmes, basés sur un modèle, utilisent l'algorithme EM (Dempster *et al.* 1977) pour estimer les poids de couplage (Jaro 1989; Belin 1989; Winkler 1988).

2.3 Sources d'erreur

Dans le couplage d'enregistrements, il existe un certain nombre de sources d'erreur qui peuvent causer un mauvais appariement d'enregistrements. Des erreurs de codage (par ex. date de naissance erronée) peuvent se produire lorsque les enregistrements sont entrés dans les bases de données. Il peut aussi y avoir des différences de code, par exemple des versions différentes du prénom ou du nom de famille.

Outre les erreurs de codage et les différences de code, les données manquantes, spécialement celles qui ont trait à des identificateurs importants, feront s'accroître sensiblement le taux d'erreur pour le couplage d'enregistrements (Fair et Lalonde 1981). L'existence d'enregistrements en double, c'est-à-dire qu'un enregistrement d'un fichier est couplé à plus d'un enregistrement de l'autre fichier, peut aussi causer des erreurs de couplage (Jabine et Scheuren 1986). C'est pourquoi les systèmes de CEI doivent inclure des règles qui permettent les appariements multiples.

Une façon d'accroître la fiabilité de l'identificateur du nom de famille est d'utiliser un système de codage phonétique. Par exemple, deux observations d'un identificateur, **ANDERSON** et **ANDERSEN**, seront recodées **ANDAR** au moyen du New York State Intelligence and Identification System (NYSIIS) (Newcombe 1988). Cela aura pour conséquence d'amenuiser l'effet des variations de nom sur le couplage des enregistrements. Cependant, en comprimant le nom, on risque de réduire le pouvoir de différenciation puisque deux noms différents peuvent avoir le même code NYSIIS. Du même coup, la probabilité d'effectuer des couplages incorrects s'accroît (Newcombe 1988).

Le prénom peut se présenter sous diverses formes, selon les bases de données. Par exemple, Joseph et Joe, Cynthia et Cindy, David et Dave. Newcombe *et al.* (1992) étudient des façons de mettre à profit les connaissances acquises sur les variations de prénoms afin d'accroître la probabilité d'effectuer des couplages exacts.

Parfois, les identificateurs existants peuvent avoir un pouvoir de différenciation trop faible. Il arrive aussi que l'algorithme de couplage exploite l'information contenue dans les zones d'identification qui servent au couplage. Les deux situations peuvent être la cause d'erreurs d'appariement.

Lorsqu'il s'agit de grands fichiers, il n'est plus commode de comparer tous les couples d'enregistrements possibles. Afin de réduire le nombre de comparaisons, on peut répartir les enregistrements des deux fichiers à coupler dans des groupes exhaustifs s'excluant mutuellement et faire les comparaisons à l'intérieur des groupes. Le groupage s'effectue généralement par un tri appliqué aux deux fichiers à l'aide d'une ou de plusieurs variables d'identification. L'inconvénient de cette méthode est que les éléments d'une paire qui sont classés dans deux groupes différents ne seront pas comparés l'un à l'autre et seront par conséquent assimilés à des non-concordances. Les paires d'éléments à rapprocher proviendront uniquement des enregistrements pour lesquels il y concordance des variables de tri. Par conséquent, le nombre de faux liens négatifs augmentera (Newcombe 1987; Jaro 1989). Une bonne variable de groupage est une variable qui repose sur des groupes qui comptent à peu près le même nombre d'enregistrements (Jaro 1989).

Dans la plupart des applications de la méthode de Fellegi-Sunter, les résultats des comparaisons faites pour différentes zones d'appariement sont supposés indépendants. Kelley (1986) a fait des études de simulation pour analyser la robustesse du système de couplage du Bureau of the Census des É.-U. à l'égard du non-respect de l'hypothèse d'indépendance. Pour certaines populations et certaines variables de couplage, on a constaté que le non-respect de l'hypothèse d'indépendance peut avoir une incidence notable sur les taux d'erreur de couplage.

Newcombe *et al.* (1983) ont comparé la précision de l'appariement informatisé à celle de la recherche manuelle dans une étude de suivi épidémiologique. Ils ont constaté que l'appariement informatisé était plus avantageux que la recherche manuelle et moins susceptible de favoriser un couplage (nécessairement faux) avec des enregistrements qui n'ont aucun rapport avec la population à l'étude. Dans les deux cas, la précision dépendait largement de la quantité de renseignements personnels contenus dans les enregistrements qui faisaient l'objet du couplage. Fair et Lalonde (1987) arrivent aux mêmes conclusions après avoir étudié l'effet de la présence (ou de l'absence) de divers identificateurs sur les taux d'erreur de couplage.

Schnatter *et al.* (1990) ont vérifié si le système de CEI utilisé à Statistique Canada permet d'évaluer avec précision un nombre de décès. On a comparé le nombre de décès survenus parmi une cohorte de 17,446 travailleurs de l'industrie du raffinage du pétrole avec le nombre de décès établi au moyen d'un couplage à la BCDM. Le système de CEI de Statistique Canada avait signalé 98% des décès survenus au Canada.

2.4 Choix du seuil et estimation du taux d'erreur

Après que des poids ont été attribués à toutes les concordances probables, une décision est prise concernant la probabilité que la concordance soit une concordance désignée (c.-à-d. un lien). Suivant la méthode de Fellegi-Sunter, on compare chaque poids à des limites supérieure et inférieure et on prend une décision selon la règle suivante:

$$\text{concordance probable} = \begin{cases} \text{un lien} & \text{si} \quad w \geq w_u \\ \text{un cas indéterminé} & \text{si} \quad w_l < w < w_u \\ \text{un non-lien} & \text{si} \quad w \leq w_l. \end{cases}$$

Dans la règle ci-dessus, w_l et w_u sont les limites inférieure et supérieure pour les poids de couplage; idéalement, elles sont choisies de manière à réduire au maximum le nombre de cas indéterminés, ce qui a pour effet de maintenir, voire d'abaisser le taux d'erreur de classification pour les deux types d'erreur (lien authentique classé comme non-lien, et non-lien authentique classé comme lien).

Lorsque les circonstances le permettent, on recourt à une solution manuelle pour les paires classées parmi les cas indéterminés. On peut s'appuyer sur des données supplémentaires pour prendre une décision sur les cas indéterminés. Dans beaucoup d'applications toutefois, la solution manuelle n'est pas pratique, surtout lorsqu'il y a un grand nombre de cas indéterminés. Dans ces circonstances, on peut déterminer un seuil unique, $w_t = w_l = w_u$, de sorte que deux possibilités seulement existent: les paires dont le poids est plus grand que w_t sont reconnues comme des liens; celles dont le poids est plus petit que w_t sont reconnues comme des non-liens.

Le choix de la limite w_t n'est pas sans difficultés. Les méthodes existantes reposent sur une connaissance des taux d'erreur de couplage, qui sont estimés soit par la résolution manuelle d'un échantillon (ou d'une population) de cas indéterminés, soit par l'application d'une méthode analytique. La première possibilité représente une méthode fondée sur un échantillon puisqu'elle implique une collecte de données pour estimer le taux d'erreur de couplage.

Les taux d'erreur de couplage dépendent des seuils fixés. Plus l'écart est grand entre les limites inférieure et supérieure, plus le nombre de cas indéterminés est élevé. Si nous avons un seuil unique, le nombre de "faux négatifs" augmente et le nombre de "faux positifs" diminue à mesure que le seuil s'accroît.

L'emploi d'une méthode simple fondée sur un échantillon pour le choix du seuil nécessite l'application d'une phase pilote. Premièrement, on tire un échantillon d'enregistrements dans le plus petit des deux fichiers qui doivent faire l'objet d'un couplage. Deuxièmement, on établit des concordances par une technique manuelle et au moyen d'un système de couplage probabiliste informatisé. Troisièmement, en supposant que les concordances établies manuellement sont des concordances vraies, le seuil choisi correspond au poids pour lequel est minimisée la somme des faux positifs et des faux négatifs. Malgré tout, des erreurs de couplage pourraient encore se produire dans les opérations manuelles à cause d'erreurs de codage, du pouvoir de différenciation insuffisant des identificateurs utilisés ou d'autres problèmes de couplage.

Pour estimer les taux d'erreur d'un système de CEI, on peut construire un tableau de contingence comme celui-ci:

CEI	Système manuel	
	Couplage	Non-couplage
Couplage	n_{11}	n_{12}
Non-couplage	n_{21}	n_{22}

On peut alors estimer le taux de faux positifs (FP) et le taux de faux négatifs (FN) au moyen des expressions

$$\mathbf{FP} = \frac{n_{12}}{n_{11} + n_{12}}$$

et

$$\mathbf{FN} = \frac{n_{21}}{n_{21} + n_{22}}.$$

Fellegi et Sunter (1969) soulignent que les taux d'erreur rattachés à des seuils donnés sont une fonction des probabilités de concordance pour les liens et les non-liens. Les estimations des probabilités de concordance peuvent donc servir à déterminer les seuils. Cette approche est aussi examinée par Jaro (1989).

En ce qui concerne les systèmes de couplage d'enregistrements basés sur un modèle, les règles de couplage d'enregistrements et la méthode de détermination des seuils ressemblent, à quelques différences près, à celles appliquées dans la méthode à échantillon décrite plus haut. L'essentiel de la méthode basée sur un modèle consiste à ajuster des modèles en vue d'estimer les probabilités conditionnelles définies en (1) et (2) et d'estimer le taux d'erreur au moyen du logit de deux probabilités conditionnelles estimées. Cette méthode prévoit l'utilisation de l'algorithme EM pour estimer les probabilités conditionnelles m_i et u_i définies en (1) et (2) pour la zone i de l'enregistrement en supposant l'indépendance des comparaisons entre les zones,

$$Pr(\gamma^j|M) = \Pi_{i=1}^{n} m_i^{\gamma_i^j}(1 - m_i)^{1-\gamma_i^j},$$

$$Pr(\gamma^j|U) = \Pi_{i=1}^{n} u_i^{\gamma_i^j}(1 - u_i)^{1-\gamma_i^j},$$

($i = 1, \ldots, n$ et $j = 1, \ldots, N$), où n est le nombre de zones, N est le nombre total de paires de comparaison, et

$$\gamma_i^j = \begin{cases} 1 & \text{s'il y a concordance des zones } i \text{ pour la paire d'enregistrements } j \\ 0 & \text{s'il y a non-cordance des zones } i \text{ pour la paire d'enregistrements } j. \end{cases}$$

En répétant alternativement l'étape de l'espérance (étape E) et l'étape de la maximisation (étape M) de l'algorithme EM, on obtient les estimations des probabilités conditionnelles $\hat{m}_i$ et $\hat{u}_i$. Ces dernières peuvent ensuite servir à estimer la probabilité globale d'un appariement exact et à déterminer par conséquent les poids (Jaro 1989). Ces poids servent ensuite à estimer la probabilité globale d'un appariement exact.

Belin et Rubin (1991) proposent une méthode qui utilise les résultats d'opérations d'appariement par ordinateur pour ajuster un modèle mixte destiné à l'estimation du taux d'erreur de couplage. Une transformation de Box-Cox (Box et Cox 1964) est appliquée aux poids pour les concordances et pour les non-concordances, de sorte que les poids transformés w_i^* forment des distributions gaussiennes, φ_T et φ_F, de moyennes μ_T et μ_F, et de variances σ_T^2 et σ_F^2 respectivement. Tous les poids transformés sont donc supposés provenir d'une distribution mixte

$$\lambda\varphi_T\left(\frac{w_i^* - \mu_T}{\sigma_T}\right) + (1 - \lambda)\varphi_F\left(\frac{w_i^* - \mu_F}{\sigma_F}\right).$$

Après avoir estimé le coefficient de mixité, λ, à l'aide de résultats d'opérations d'appariement antérieures, on peut ajuster le modèle ci-dessus au moyen de poids issus de la procédure de couplage. On peut ensuite se servir du modèle ajusté pour estimer le taux d'erreur de l'algorithme de couplage d'enregistrements, étant donné un seuil particulier. L'erreur type correspondante est aussi estimée au moyen de l'algorithme SEM, par lequel est estimée la covariance des estimations de paramètres produites par l'algorithme EM (Meng et Rubin 1991).

3. EXEMPLES DE GRANDES ÉTUDES PAR COUPLAGE D'ENREGISTREMENTS

3.1 Étude des exploitants agricoles canadiens

L'étude des exploitants agricoles canadiens a été créée dans le but d'analyser les liens possibles entre des causes de décès chez les exploitants agricoles et diverses variables socio-démographiques et agricoles, en particulier le rapport entre l'emploi de pesticides et la mortalité. Les données sur la mortalité provenaient de la BCDM, tandis que celles sur les variables socio-démographiques et agricoles étaient tirées du recensement de la population et du recensement de l'agriculture. Comme les bases de données des recensements ne contenaient pas d'informations sur l'exposition aux pesticides proprement dite, on s'est servi de variables substitutives comme le nombre d'acres traités aux insecticides ou aux herbicides et le coût des produits chimiques agricoles. Le fichier d'analyse contenant les données pertinentes a été construit à l'aide d'une méthode de couplage probabiliste.

3.1.1 Définition de la cohorte

La cohorte se compose de tous les agriculteurs de sexe masculin qui répondaient à la définition d'exploitant agricole dans le recensement de 1971. L'exploitant agricole est la personne responsable des décisions quotidiennes prises pour la bonne marche de l'exploitation agricole. Il n'est pas nécessairement propriétaire de l'exploitation; il peut en être locataire ou agir à titre de gérant engagé. Un seul exploitant est désigné pour chaque exploitation agricole. Selon la définition que l'on en donnait dans le recensement de 1971, une ferme est une exploitation agricole qui a une superficie d'un acre ou plus et dont les ventes de produits agricoles sont de $50 ou plus. La cohorte des exploitants agricoles comptait 326,000 individus de sexe masculin (Jordan-Simpson *et al.* 1990). On en a fait le suivi (au point de vue de la mortalité) jusqu'en 1987.

3.1.2 Méthode de couplage d'enregistrements

Le fichier d'analyse de l'étude des exploitants agricoles canadiens a été constitué par suite de trois couplages indépendants, le dernier d'entre eux étant le couplage du fichier de la cohorte des exploitants agricoles à la BCDM. Mais avant d'effectuer ce troisième couplage, il a fallu construire le fichier de la cohorte des exploitants agricoles.

Les données socio-démographiques ont été tirées du recensement de la population de 1971 tandis que les données sur les pratiques agricoles provenaient du recensement de l'agriculture de 1971. La base de données du recensement de la population contient des enregistrements pour chaque personne au Canada et elle est constituée de données recueillies au moyen de deux types de questionnaires: un questionnaire abrégé et un questionnaire complet. Ce dernier exigeait plus de renseignements que le premier et était distribué aléatoirement au tiers des ménages. Le questionnaire du recensement de l'agriculture était envoyé à toutes les exploitations agricoles.

Le nom des exploitants agricoles n'apparaît pas comme tel dans le fichier du recensement de l'agriculture ni dans celui du recensement de la population. Leurs nom et adresse figurent dans le registre central des fermes, créé pour servir de liste d'envoi pour les questionnaires sur l'agriculture.

La BCDM fait état de tous les décès enregistrés par les provinces et les territoires depuis 1950 et est stockée sous une forme standard dans un ordinateur à Statistique Canada (Smith et Newcombe 1982). Entre 1950 et 1987, 5.9 millions de décès ont été enregistrés dans la BCDM. La base contient des renseignements personnels, plus la date et l'endroit du décès de même que la cause du décès, codée selon la Classification internationale des maladies (CIM).

La Loi sur la statistique garantit la confidentialité de tous les enregistrements contenus dans la BCDM et les bases de données des recensements de la population et de l'agriculture. Comme nous l'avons mentionné plus tôt, toutes les études qui nécessitent des couplages avec ces bases de données doivent être soumises à un processus d'examen et d'approbation minutieux avant d'être réalisées, et les fichiers couplés qui contiennent des renseignements personnels demeurent sous la garde de Statistique Canada.

Pour constituer le fichier d'analyse, il a fallu coupler en trois étapes tous les fichiers décrits plus haut.

a) **Suivi**. Le registre central des fermes de 1971 a été couplé à celui de 1981 au moyen du CEI pour déterminer si les agriculteurs qui figuraient sur la liste en 1971 étaient encore vivants en 1981. Cette information était ajoutée dans le registre central des fermes de 1971 afin d'accroître la probabilité d'un couplage fructueux avec la BCDM.

b) **Base de données de la cohorte des exploitants agricoles.** Le registre central des fermes de 1971, enrichi des données de l'étape précédente, a été fusionné au fichier du recensement de l'agriculture afin d'inclure des noms dans la base de données de la cohorte. Cette opération était nécessaire pour effectuer le couplage avec la base de données sur la mortalité. Le fichier ainsi obtenu était ensuite couplé au fichier du recensement de la population à l'aide du CEI pour constituer la base de données de la cohorte des exploitants agricoles.

c) Fichier d'analyse. La base de données de la cohorte des exploitants agricoles a été couplée avec la BCDM au moyen du CEI. Le fichier ainsi obtenu contenait des données socio-démographiques de même que des données sur l'exposition aux risques environnementaux et sur la mortalité, et pouvait donc se prêter à l'analyse.

3.1.3 Choix des seuils

Il fallait déterminer des valeurs de seuil pour chacun des trois couplages effectués en vue de constituer le fichier d'analyse et pour les couplages basés sur le questionnaire abrégé et le questionnaire complet. En ce qui a trait au couplage avec des bases de données sur la mortalité, Statistique Canada a fixé les seuils à l'aide d'une méthode basée sur un échantillon. Cette méthode est illustrée pour le couplage de la base de données de la cohorte des exploitants agricoles avec la base de données sur la mortalité (en ce qui concerne ceux qui ont rempli le questionnaire de recensement abrégé).

On a prélevé un échantillon d'environ 10% parmi les questionnaires abrégés qui avaient été remplis par la cohorte des exploitants agricoles (Statistique Canada 1991a). On a ensuite déterminé les liens de deux manières: par le système de CEI de Statistique Canada et par une technique manuelle qui utilise les données de registres des décès. Après quoi on a comparé les résultats des couplages en supposant que les couplages réalisés par la technique manuelle produisaient des concordances vraies.

La figure 2 représente le nombre de faux positifs et de faux négatifs qui correspondent à une série de seuils comme poids de couplage. Le seuil a été choisi de manière à minimiser le nombre total de faux positifs et de faux négatifs. La valeur-seuil en question est w_t, point d'intersection des deux courbes. Le taux d'erreur pour les faux positifs est estimé à $(36/453) \times 100 = 7.9\%$, tandis que celui pour les faux négatifs est estimé à $(38/20{,}847) \times 100 = 0.2\%$, pour un taux d'erreur global de 8.1% (tableau 1).

Tableau 1

Comparaison d'enregistrements couplés et non couplés au moyen du CEI et de la technique manuelle; selon un échantillon d'enregistrements du recensement pour l'étude des exploitants agricoles

Couplage d'enregistrements informatisé (CEI)	Technique manuelle		Total
	Couplés	Non couplés	
	Questionnaire abrégé		
Couplés	417	36	453
Non couplés	38	20,809	20,847
Total	455	20,845	21,300
	Questionnaire complet		
Couplés	286	13	299
Non couplés	15	18,498	18,513
Total	301	18,511	18,812

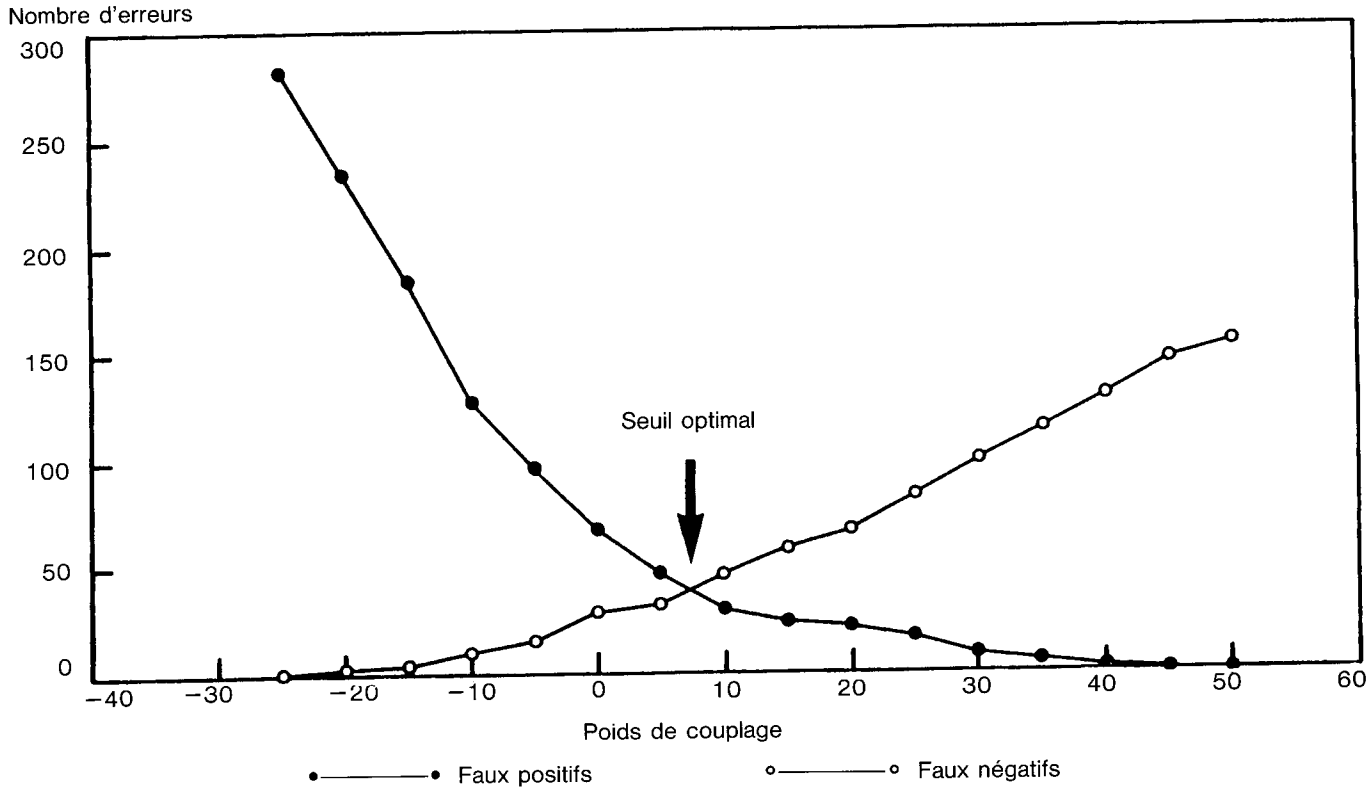

Figure 2. Faux liens positifs et faux liens négatifs, étude de mortalité des exploitants agricoles canadiens: questionnaire abrégé

Pour illustrer l'effet de l'addition de renseignements, on peut construire un tableau semblable pour les questionnaires complets (tableau 1). Les taux d'erreur pour les faux positifs et les faux négatifs sont 4.3% ((13/299) × 100) et 0.1% ((15/18,511) × 100) respectivement, pour un taux global de 4.4%. On peut donc réduire les taux d'erreur en augmentant la quantité de renseignements personnels.

3.2 Étude de mortalité fondée sur le Fichier dosimétrique national

Le Fichier dosimétrique national (FDN) du Canada contient des données sur l'exposition professionnelle au rayonnement pour environ 255,000 Canadiens; ces données remontent aussi loin que 1951. Le FDN a été couplé récemment à la BCDM. L'étude de mortalité fondée sur le Fichier dosimétrique national a pour but d'établir un lien entre la surmortalité attribuable au cancer et à d'autres affections et l'exposition professionnelle à de faibles niveaux de rayonnement ionisant (Ashmore *et al.* 1993).

3.2.1 Définition de la cohorte

La cohorte se compose de tous les travailleurs qui ont fait l'objet d'une surveillance pour rayonnement ionisant (y compris les produits de filiation du radon et du tritium) et qui avaient un dossier dans le Fichier dosimétrique national au 31 décembre 1983. Le FDN contient les fiches de presque tous les travailleurs exposés à des rayonnements ionisants au Canada qui font l'objet d'une surveillance; certaines fiches contiennent des données qui remontent à 37 ans. Le FDN comprend en outre 80 catégories d'emploi, qui vont du travailleur de centrale nucléaire au dentiste en passant par le radiologiste d'hôpital. En tout, 248,940 personnes formaient la cohorte à l'étude.

Selon le genre de rayonnement et le degré d'exposition prévu pour des catégories d'emploi particulières, des données sont recueillies annuellement, trimestriellement, mensuellement ou à la quinzaine. À chaque année, un indice global de la dose absorbée annuellement par chaque individu est inscrit dans le Lifetime Dose History System (LDHS). C'est à partir des données contenues dans le LDHS que l'on pourra tenter d'établir des liens entre l'exposition professionnelle aux rayonnements et l'état de santé.

Les données individuelles contenues dans le LDHS permettent aussi de calculer la dose reçue cumulée pour chaque personne. Bien que le degré d'exposition des personnes ne sera pas le même à chaque année, on peut calculer une dose annuelle moyenne pour les individus en divisant la dose reçue cumulée par le nombre d'années qui se sont écoulées depuis le moment de la première exposition. Les analyses statistiques peuvent se faire en fonction de la dose reçue cumulée, de la dose annuelle moyenne ou des doses annuelles inscrites dans le LDHS. Jusqu'en 1986, les renseignements personnels tels que le nom de famille, le prénom, le sexe, l'année de naissance et les numéros d'identification attribués aux fichiers des individus étaient stockés dans un fichier indépendant appelé fichier d'identification central (FIC) (Ashmore et Grogan 1985).

3.2.2 Méthode de couplage d'enregistrements

Les variables d'identification ont changé plusieurs fois depuis la création du FDN, ce qui complique par moments la reconstitution du passé dosimétrique des individus. À cause de cela et d'autres difficultés du même genre, on se sert du numéro d'assurance sociale comme variable d'identification clé depuis 1977.

Plusieurs couplages ont été nécessaires pour rassembler les identificateurs personnels, les antécédents dosimétriques et les données sur les décès appropriés.

a) **Couplage à l'intérieur des fichiers dosimétriques.** Depuis 1984, Statistique Canada exécute des fusions dynamiques avec son fichier LDHS afin de regrouper les enregistrements dosimétriques; ces fusions ont pour effet de réduire le nombre d'enregistrements fragmentaires et de réunir les enregistrements dans un dossier dosimétrique complet pour chaque individu étudié. Le fichier dérivé des couplages internes indique quels enregistrements du FDN semblent se rapporter au même individu.

b) **Couplage avec la BCDM.** La cohorte du FIC, qui avait déjà fait l'objet d'un couplage interne, a été couplée avec les données sur la mortalité (couplage à deux fichiers). En couplant ces deux fichiers, il est possible d'évaluer le risque ultérieur de décès pour les membres de la cohorte. Dans cette étude, la BCDM permettait de connaître la cause du décès, l'année et le lieu du décès, de même que le lieu et l'année de naissance.

c) **Fichier d'analyse.** On a apparié des données de trois fichiers - FIC, BCDM et LDHS - pour créer un enregistrement complet pour chaque membre de la cohorte étudiée. Chaque enregistrement indique, si possible, le mois et l'année de naissance, le sexe, les données sur le décès énumérées ci-dessus, le poids de couplage avec les données sur le décès, et les antécédents dosimétriques. Les enregistrements du FIC ou du fichier dosimétrique qui ne pouvaient être appariés étaient soumis à un examen minutieux.

3.2.3 Choix des seuils

En ce qui a trait au couplage du fichier de cohorte avec la BCDM, le choix du seuil s'est effectué de la même manière que dans l'étude des exploitants agricoles canadiens. Premièrement, on a déterminé les poids des cas indéterminés. Tous les cas indéterminés qui avaient un poids inférieur à −30 étaient classés comme des non-liens. On comptait 4,429 femmes et 8,686 membres de la cohorte avec un poids de couplage supérieur à cette valeur. On a tiré un échantillon dans ce groupe de personnes et on l'a traité manuellement en examinant les certificats de décès pour déterminer si les liens observés étaient des liens authentiques ou des non-liens. Le seuil a été choisi de manière à correspondre au poids de couplage pour lequel le nombre de faux liens positifs est égal au nombre de faux liens négatifs pour les femmes et pour les hommes, pris séparément (figure non insérer).

4. QUESTIONS TOUCHANT L'ANALYSE DES ENSEMBLES DE DONNÉES COUPLÉES

Peu d'études ont été faites pour déterminer l'effet du couplage d'enregistrements sur les résultats de l'analyse de régression. Neter *et al.* (1965) ont reconnu que les erreurs introduites durant l'appariement pouvaient nuire aux analyses qui se font à partir des fichiers qui viennent d'être couplés. Supposons que les valeurs vraies d'une variable aléatoire étudiée sont inscrites dans un fichier constitué de N enregistrements. Désignons par Y_i la valeur vraie pour l'enregistrement $i = 1, \ldots, N$. Ce fichier est couplé à un autre fichier, qui contient des renseignements personnels et selon lequel une valeur Z_i est attribuée à l'enregistrement $i = 1, \ldots, N$. En supposant que toutes les erreurs d'appariement ont la même probabilité, nous avons

$$Z_i = \begin{cases} Y_i & \text{avec une probabilité } p \\ Y_j & \text{avec une probabilité } q (j \neq i), \end{cases}$$

où $p + (N - 1)q = 1$.

Neter *et al.* (1965) se sont servis de ce modèle pour étudier l'incidence des erreurs d'appariement sur la moyenne et la variance empiriques de la variable Z. Ils ont aussi étudié l'effet des erreurs d'appariement sur la corrélation entre Z et une autre variable aléatoire, X, incluse dans le même fichier et sur les estimations des paramètres de la régression de Z_i par rapport à X_i. Leurs conclusions étaient les suivantes: 1) l'estimation de la moyenne de Z est non biaisée pour la moyenne des Y; 2) si W est corrélé positivement avec Y, la variance résiduelle d'une régression de Z par rapport à X sera plus grande que la variance d'une régression de Y par rapport à W; et 3) la pente de la droite de régression sera sous-estimée si Z est utilisée au lieu de Y.

Belin et Rubin (1991) et Winkler et Thibaudeau (1991) étudient des cadres théoriques, des algorithmes de calcul et des logiciels pour estimer les probabilités d'appariement. Ces recherches ont incité Scheuren et Winkler (1991) à mettre à jour les travaux de Neter *et al.* (1965). Ils ont utilisé le modèle

$$Z_i = \begin{cases} Y_i & \text{avec probabilité } p_i \\ Y_j & \text{avec probabilité } q_{ij} (j \neq i), \end{cases}$$

où $p_i + \sum_{j \neq i} q_{ij} = 1$, pour étudier l'incidence des erreurs d'appariement sur les estimations des coefficients β dans le modèle de régression linéaire

$$Y = X\beta + \epsilon.$$

L'effet des erreurs d'appariement sur le modèle de régression ci-dessus peut être exprimé par l'équation

$$E(Z_i) = Y_i + B_i,$$

où le terme de biais est défini

$$B_i = (p_i - 1) Y_i + \sum_{j \neq i} q_{ij} Y_i.$$

Au lieu d'utiliser la paire formée de la variable indépendante et de la variable dépendante, (X_i, Y_i), pour ajuster le modèle, on se sert de la paire formée de la variable indépendante et de la variable dépendante couplée, (X_i, Z_i). Notons que la variable dépendante couplée peut être exprimée sous la forme $Z = Y + B$, les coefficients étant estimés par l'expression

$$\hat{C} = (X^T X)^{-1} X^T Z = \hat{\beta} + (X^T X)^{-1} XB,$$

où $\hat{\beta} = (X^T X)^{-1} X^T Y$, de sorte que le facteur de correction pour le biais est $(X^T X)^{-1} XB$.

Scheuren et Winkler (1991) se sont servis de ces modèles pour exécuter des études de simulation basées sur des données réelles. Leur méthode consistait à prendre un fichier d'enregistrements couplés et d'enregistrements non couplés et à soumettre de nouveau ces enregistrements à une opération de couplage en utilisant différentes variables d'appariement. Cette simulation démontre que le fait de pouvoir estimer avec précision les probabilités d'appariement influe fondamentalement sur la précision des estimations des coefficients. Si les probabilités d'appariement peuvent être estimées avec justesse, la procédure de correction donnera des résultats convenables.

5. ANALYSE

Le couplage d'enregistrements est une méthode intéressante pour examiner les liens entre l'exposition professionnelle et l'état de santé au moyen de bases de données existantes. Cependant, des erreurs de couplage peuvent se produire pour diverses raisons: erreurs de codage, changement dans les identificateurs, données manquantes, pouvoir de différenciation insuffisant des identificateurs.

Les taux d'erreur dépendent de la quantité de renseignements personnels, comme le montre l'étude des exploitants agricoles. Dans cette étude, le taux d'erreur était moins élevé avec le couplage du questionnaire de recensement complet, qui contenait plus de renseignements personnels que le questionnaire abrégé. Il est donc important de disposer de renseignements personnels de qualité pour effectuer un couplage d'enregistrements.

On a porté assez peu d'attention à l'incidence des erreurs de couplage sur les inférences statistiques basées sur des études par couplage d'enregistrements. Ce genre d'erreur peut introduire un biais dans les estimations des mesures d'association des variables de santé et des variables environnementales, comme les coefficients de régression. Des recherches sont en cours dans le but d'analyser l'effet de ces erreurs sur les résultats des études épidémiologiques présentées dans cet article.

REMERCIEMENTS

Nous tenons à remercier Martha Fair, de Statistique Canada, et le docteur Howard Morrison, de Santé et Bien-être social Canada, pour les précieux commentaires que ces personnes ont exprimés sur cet article. Nous remercions également les deux réviseurs anonymes pour leurs nombreux conseils pratiques.

BIBLIOGRAPHIE

ASHMORE, J.P., et GROGAN, D. (1985). The National Dose Registry of Canada. *Radiation Protection Dosimetry*, 11, 95-100.

ASHMORE, J.P., KREWSKI, D., et ZIELINSKI, J.M. (1993). National Dose Registry Study. *European Journal of Cancer*, soumis.

BELIN, T.R. (1989). Results from evaluation of computer matching. Note de service, Statistical Research Division, U.S. Bureau of the Census, Washington, D.C.

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.

BOX, G., et COX, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society*, Séries B, 26, 211-246.

CARPENTER, M., et FAIR, M.E. (Éds.) (1990). *Canadian Epidemiology Research Conference* - 1989: *Proceedings of Record Linkage Sessions and Workshop*. Ottawa, Ontario: Ottawa Select Printing.

COPAS, J.B., et HILTON, F.J. (1990). Record linkage: Statistical models for matching computer records. *Journal of the Royal Statistical Society*, Séries A, 153, 287-320.

DEMPSTER, A.D., LAIRD, N.M., et RUBIN, D.B. (1977). Maximum likelihood estimation from incomplete data via EM algorithm, (avec discussion). *Journal of the Royal Statistical Society*, Séries B, 39, 1-38.

FAIR, M.E. (1989). Studies and references relating to uses of the Canadian Mortality Data Base. Rapport de la section de la recherche sur l'hygiène du travail et de l'environnement, division de santé, Statistique Canada, Ottawa.

FAIR, M.E., et LALONDE, P. (1988). Identificateurs manquants et justesse de l'observation suivie. *Recueil: Symposium sur les utilisations statistiques des données administratives*, Statistique Canada, Ottawa, 111-125.

FELLEGI, I.P., 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. Release 2.7. Rapport de la sous-division de la recherche et des systèmes généraux, division des services et développement informatiques, 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 LINDSAY, J. (1983). A follow-up study of a ten-percent sample of the Canadian Labor Force. I. Cancer mortality in males, 1965-73. *Journal of the National Cancer Institute*, 70, 37-44.

HOWE, G.R., FRASER, D., LINDSAY, J., PRESNAL, B., et YU, S.Z. (1983). Cancer mortality (1965-77) in relation to diesel fume and coal exposure in a cohort of retired railway workers. *Journal of the National Cancer Institute*, 70, 1015-1019.

HOWE, G.R., NAIR, R.C., NEWCOMBE, H.B., MILLER, A.B., BURCH, J.D., et ABBATT, J.D. (1987). Lung cancer mortality (1950-80) in relation to radon daughter exposure in a cohort of workers at the Eldorado Port radium uranium mine: Possible modification of risk by exposure rate. *Journal of the National Cancer Institute*, 79, 1255-1260.

HOWE, G.R., et SPASOFF, R.A. (Éds.) (1986). *Proceeding of the Workshop on Computerized Linkage in Health Research*. Toronto: University of Toronto Press.

JARO, M.A., (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 414-420.

JABINE, T.B., et SCHEUREN, F.J. (1986). Record linkages for statistical purposes: methodological issues. *Journal of Official Statistical*, 2, 255-277.

JORDAN-SIMPSON, D.A., FAIR, M.E., et POLIQUIN, C. (1990). Étude des exploitants agricoles canadiens: Méthodologie. *Rapports sur la santé*. N° 83-003 au catalogue, Statistique Canada, 2, 141-155.

KELLEY, R.P. (1986). Robustness of the Census Bureau's record linkage system. Rapport présenté à la réunion de l'American Statistical Association, août 1986.

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

MAO, Y., SEMENCIW, R., MORRISON, H., KOCH, M., HILL, G., FAIR, M., et WIGLE, D. (1988). Survival rates among patients with cancer in Alberta in 1974-78. *Canadian Medical Association Journal*, 138, 1107-1113.

MENG, L., et RUBIN, D.B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86, 899-911.

MORRISON, H.I., SEMENCIW, R.W., MAO, Y., et WIGLE, D.T. (1988). Cancer mortality among a group of fluorspar miners exposed to radon progeny. *American Journal of Epidemiology*, 128, 1266-1275.

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., SMITH, M.E., HOWE, G.R., MINGAY, J., STRUGNELL, A., et ABBATT, J.D. (1983). Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. *Computers in Biology and Medicine*, 13, 157-169.

NEWCOMBE, H.B. (1988). *Handbook of Record Linkage: methods for health and statistical studies, administration and business*. Oxford: Oxford Medical Publications.

NEWCOMBE, H.B., FAIR, M.E., et LALONDE, P. (1992). The use of names for linking personal records. *Journal of the American Statistical Association*, 87, 1193-1204.

SCHEUREN, F., et WINKLER, W.E. (1991). An error model for regression analysis of data files that are computer matched. *Proceedings of the 1991 Annual Research Conference*, U.S. Bureau of the Census, Washington, D.C.

SCHNATTER, A.R., ACQUIVELLA, J.F., THOMPSON, F.S., DONALESKI, D., et THERIAULT, G. (1990). An analysis of death ascertainment and follow-up through Statistics Canada's Mortality Data Base system. *Canadian Journal of Public Health*, 81, 60-65.

SHANNON, H.S., JULIAN, J.A., et ROBERTS, R.S. (1984). A mortality study of 11,500 nickel workers. *Journal of the National Cancer Institute*, 73, 1251-1258.

SMITH, M.E., et NEWCOMBE, H.B. (1982). Use of the Canadian Mortality Data Base for epidemiological follow-up. *Canadian Journal of Public Health*, 73, 39-46.

SMITH, M.E., et SILINS, J. (1981). Generalized Iterative Record Linkage System. *Proceedings of the Social Statistics Section, American Statistical Association*, 128-137.

STATISTIQUE CANADA, (1991a). Canadian Farm Operators' mortality study general work plan. Rapport interne de la section de la recherche sur l'hygiène du travail et de l'environnement. Statistique Canada, Ottawa.

STATISTIQUE CANADA, (1991b). Données non-publiées.

TEPPING, B.J. (1968). A model for optimum linkage of records. *Journal of the American Statistical Association*, 63, 1321-1332.

WIGLE, D.T., SEMENCIW, R.M., WILKINS, K., RIEDEL, D., RITTER, L., MORRISON, H.I., et MAO, Y. (1990). Mortality study of Canadian male farm operators: Non-Hodgkin's lymphoma mortality and agriculture practices in Saskatchewan. *Journal of the National Cancer Institute*, 82, 575-581.

WINKLER, W.E. (1988). Using the E.M. algorithm for weight computation in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 667-671.

WINKLER, W.E., et THIBAUDEAU, Y. (1991). An application of the Fellegi-Sunter model of record linkage to the 1990 U.S. Census. U.S. Bureau of the Census, Statistical Research Division, rapport technique.