

No 11-522-XIF au catalogue

**La série des symposiums internationaux
de Statistique Canada - Recueil**

**Symposium 2006 : Enjeux
méthodologiques reliés à la
mesure de la santé des
populations**



2006



Statistics
Canada

Statistique
Canada

Canada

Couplage d'enregistrements déterministe et probabiliste

Claude Nadeau, Marie P. Beaudet et Jocelyne Marion¹

Résumé

Les efforts des programmes d'enquête en vue de réduire les coûts et le fardeau de réponse se sont traduits par une utilisation croissante de l'information disponible dans les bases de données administratives. Le couplage des données provenant de ces deux sources est un moyen d'exploiter leur nature complémentaire et de maximiser leur utilité respective. Le présent article décrit diverses méthodes suivies pour procéder au couplage des enregistrements des bases de données de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) et de l'Information-santé orientée vers la personne (ISOP). Les fichiers résultant de certaines de ces méthodes de couplage sont utilisés pour analyser les facteurs de risque d'hospitalisation pour une maladie cardiaque. La sensibilité de l'analyse est évaluée en fonction des diverses approches de couplage.

MOTS-CLÉS : maladie cardiaque; analyse de sensibilité.

1. Introduction

Le couplage interne des enregistrements d'une base de données et le couplage des enregistrements d'une base de données à ceux d'une autre peuvent accroître la quantité et la qualité de l'information disponible pour l'analyse. À l'intérieur d'une base de données, il permet de repérer les enregistrements en double. Entre bases de données, il peut servir à accroître l'étendue des mesures ou à vérifier le degré de concordance entre diverses versions d'un concept. La combinaison de sources de données rend possibles des analyses qui, autrement, ne le seraient pas. En outre, le couplage d'enregistrements peut permettre d'inclure des personnes qui ont tendance à ne pas participer aux enquêtes à cause de contraintes de temps, parce qu'ils sont malades ou parce qu'ils refusent simplement, et contribuer ainsi à la réduction des coûts et du fardeau de réponse.

La présente étude, dans laquelle des fichiers de données ont été couplés selon une approche déterministe ainsi qu'une approche probabiliste, illustre la sensibilité de l'analyse en fonction des méthodes de couplage utilisées. Fondée sur le couplage des bases de données de l'Enquête sur la santé dans les collectivités canadiennes (ESCC) de 2001 et de l'Information-santé orientée vers la personne, l'analyse porte sur l'association entre les caractéristiques personnelles et les facteurs de risque modifiables, d'une part, et la probabilité d'être hospitalisé pour une maladie cardiaque, d'autre part.

La section 2 donne une brève description des sources de données et du nombre de répondants à l'enquête qui ont accepté que l'information qu'ils ont fournie soit couplée aux données de dossiers administratifs. Nous décrivons les résultats des analyses effectuées pour comparer les constatations faites lorsque les enregistrements sont produits par des méthodes déterministe et probabiliste de couplage d'enregistrements et nous discutons également des limites des diverses méthodes de couplage, ainsi que de leur incidence sur la puissance statistique et le biais. Nos conclusions sont présentées à la section 4.

¹ Claude Nadeau, Statistique Canada, Division des méthodes d'enquête auprès des ménages, 16-F, immeuble R.-H.-Coats, Ottawa, Canada K1A 0T6 (claude.nadeau@statcan.ca); Marie P. Beaudet, Statistique Canada, Division de la statistique de la santé, 2200-H, immeuble Principal, Ottawa, Canada K1A 0T6 (marie.p.beaudet@statcan.ca); Jocelyne Marion, Statistique Canada, Division des méthodes d'enquête auprès des ménages, 16-Q, immeuble R.-H.-Coats, Ottawa, Canada K1A 0T6 (jocelyne.marion@statcan.ca)

2. Couplage d'enregistrements

Cette section décrit le volet couplage d'enregistrements de l'étude. Les deux sources de données dont les enregistrements ont été couplés sont présentées à la section 2.1. Plusieurs méthodes de couplage sont décrites à la section 2.2, et les résultats sommaires sont présentés à la section 2.3.

2.1 Sources de données et variables d'appariement

Les sources de données qui ont été couplées sont les bases de données de l'Information-santé orientée vers la personne (ISOP) et de l'Enquête sur la santé dans les collectivités canadiennes (ESCC).

Au moment de l'analyse, la base de données de l'ISOP contenait des renseignements sur chaque sortie de l'hôpital survenue au Canada entre l'exercice 1992-1993 et l'exercice 2003-2004 (du 1^{er} avril 1992 au 31 mars 2004) pour laquelle il existait un numéro d'assurance-maladie (NAM). En plus des caractéristiques générales du patient, telles que la date de naissance, le sexe et le NAM, l'ISOP contient des renseignements sur le diagnostic et le traitement pour chaque sortie de l'hôpital.

À la fin de chaque exercice, les provinces transmettent leurs données sur les sorties de l'hôpital à l'Institut canadien d'information sur la santé (ICIS). Ce dernier envoie alors une version épurée des données à Statistique Canada qui procède à un traitement supplémentaire afin de vérifier la cohérence et la qualité des enregistrements. Une partie du traitement supplémentaire a pour but de confirmer, en se servant d'un NAM particulier, que la date de naissance, le sexe et les dispositions concernant la sortie de l'hôpital concordent dans tous les enregistrements appartenant à une même personne. Ce traitement permet de repérer les personnes qui utilisent un même NAM (p. ex. un enfant utilisant le NAM de sa mère). La base de données contient environ 3 millions de sorties de l'hôpital par exercice. Les données correspondant aux deux premiers exercices (1992-1993 et 1993-1994) n'ont pas été comprises dans le couplage, car l'information sur les sorties de l'hôpital n'était pas disponible pour chaque province et territoire. Par conséquent, seules les données recueillies pour l'exercice 1994-1995 et les suivants (du 1^{er} avril 1994 au 31 mars 2004) ont été utilisées. L'information pour cette période est complète, sauf pour le Yukon pour les exercices 1994-1995 à 1996-1997. De surcroît, les données provenant du Québec n'ont pas été incluses dans le couplage pour les raisons exposées plus loin.

L'ESCC est une enquête annuelle conçue pour recueillir des renseignements sur la santé des Canadiens, leur utilisation des services de santé et certains facteurs susceptibles d'avoir une incidence sur leur santé. Elle comporte deux cycles, «.1» et «.2». Les enquêtes régionales («.1») sont réalisées auprès d'environ 130 000 répondants. Elles ont été menées tous les deux ans à compter du cycle 1.1 qui a eu lieu en 2001. Les enquêtes provinciales («.2») sont réalisées auprès d'environ 35 000 répondants et ont été menées tous les deux ans depuis 2002. Des renseignements détaillés sur l'ESCC sont présentés dans Béland (2002).

Lors de chaque interview de l'ESCC, il est demandé aux répondants s'ils autorisent le couplage de l'information qu'ils ont fournie à des données administratives. Seules les personnes qui ont consenti à ce couplage ont été retenues pour l'étude. Comme le montre le tableau 1, environ 90 % de répondants ont donné leur consentement au cycle 1.1. Lors des cycles suivants, la proportion était d'environ 85 %.

Certaines personnes peuvent être interviewées plus d'une fois dans le cadre de l'ESCC. Cela se produit habituellement d'un cycle à un autre, mais dans des cas rares, cela peut arriver au cours d'un même cycle, à cause de l'utilisation d'une base de sondage double, comportant une liste de ménages et une base de sondage téléphonique. La sélection répétée d'un répondant ne pose pas de problème pour le couplage d'enregistrements, à condition d'être prudent et de s'assurer que les résultats du couplage sont cohérents pour les personnes concernées.

Tableau 1. Nombre et pourcentage de répondants à l'Enquête sur la santé dans les collectivités canadiennes qui ont consenti au couplage et nombre et pourcentage ne possédant pas de numéro d'assurance-maladie, selon le cycle de l'enquête

Cycle (année)	Nombre de répondants*	Répondants qui ont consenti au couplage*		Répondants sans NAM*	
		Nombre	Pourcentage	Nombre	Pourcentage
1.1 (2001)	13 1535 (108 868)	119 383 (98 450)	90,8 % (90,4 %)	36 103 (29 767)	30,2 % (30,2 %)
1.2 (2002)	36 984 (31 652)	32 269 (27 370)	87,3 % (86,5 %)	5 114 (4 067)	15,8 % (14,9 %)
2.1 (2003)	135 573 (106 473)	114 287 (89 536)	84,3 % (84,1 %)	32 923 (24 792)	28,8 % (27,7 %)
2.2 (2004)	35 107 (30 327)	30 141 (25 866)	85,9 % (85,3 %)	5 385 (4 386)	17,9 % (17,0 %)
3.1 (2005)	132 947 (103 056)	115 399 (89 108)	86,8 % (86,5 %)	36 239 (26 674)	31,4 % (29,9 %)

* Les chiffres entre parenthèses ne comprennent pas le Québec.

Bien que les bases de données contiennent de nombreuses variables (plus de 100 pour l'ISOP et plus de 1 000 pour l'ESCC), quelques-unes seulement ont été choisies comme zones d'appariement pour les couplages, à savoir : le NAM, la date de naissance, le code postal, la province et le sexe. Les noms ne figurent pas dans la base de données de l'ISOP. Dans cette dernière, les données sur les variables clés pour le couplage manquaient ou étaient incomplètes pour le Québec, parce que les NAM sont brouillés, que les codes postaux sont tronqués aux trois premiers chiffres et que la date de naissance manque. Comme nous l'avons mentionné antérieurement, les résidents du Québec ont été exclus des analyses.

Pour les autres provinces et territoires, le NAM pourrait manquer dans le fichier de l'ESCC, mais il est toujours disponible dans celui de l'ISOP. La date de naissance figure presque toujours dans l'ISOP. Une date de naissance est toujours disponible dans l'ESCC, mais selon le cycle, de 1 % à 6 % sont des dates partielles, habituellement l'année sans le mois ou le jour de la naissance. Le code postal est rarement omis dans l'ESCC ou l'ISOP, moins de 1 % des enregistrements étant affectés dans chaque fichier. Dans l'ISOP, la province renvoie à celle qui émet le NAM et les données sont toujours disponibles. Dans l'ESCC, deux variables indiquent la province, l'une correspondant à la province qui a émis le NAM et l'autre, à la province de résidence du répondant. La première manque lorsque le NAM manque; la seconde est toujours disponible. Enfin, les données sur le sexe sont toujours disponibles dans les deux fichiers.

2.2 Méthodes déterministe et probabiliste de couplage d'enregistrements

Dans le cas d'un couplage déterministe, deux enregistrements sont appariés si, et seulement si, les zones d'appariement ne manquent pas et concordent parfaitement. Nous avons effectué six couplages déterministes. Le premier a été exécuté en utilisant la province, le NAM et le sexe comme variables d'appariement et est appelé méthode 1. Dans la méthode 2, l'année de naissance a été ajoutée au premier ensemble de variables d'appariement. Dans la troisième méthode, le mois de naissance a été inclus également. Pour la méthode 4, la date de naissance complète (année, mois et jour) a été utilisée en plus du premier ensemble de variables d'appariement, c'est-à-dire la province, le NAM et le sexe. Dans la méthode 5, le code postal a été inclus dans l'ensemble de variables d'appariement utilisé pour la méthode 2. Dans la méthode 6, le NAM a été exclu et la province, le sexe, la date de naissance et le code postal ont été utilisés pour coupler les deux ensembles de données. La province, lorsqu'elle est utilisée comme variable d'appariement, s'entend de la province qui a émis le NAM, sauf pour la méthode 6. Dans cette approche, un appariement sur la province a été accepté lorsque la valeur de l'une des deux variables de l'ESCC indiquant la province correspondait à la valeur de la province dans l'ISOP. Le tableau 2 de la section 2.3 résume les zones d'appariement utilisées dans les couplages déterministes.

Par définition, les méthodes 1 à 4 sont emboîtées. Cela signifie que les appariements faits par la méthode 4 le seront aussi par les méthodes 1, 2 et 3. De même, les appariements obtenus par la méthode 3 seront produits par les méthodes 1 et 2, mais pas nécessairement par la méthode 4. La méthode 5 correspond à la méthode 2 avec l'exigence supplémentaire d'obtenir une concordance pour le code postal. Les méthodes 1, 2 et 5 sont, par conséquent, également emboîtées. La méthode 6 est la seule ne s'appuyant pas sur le NAM comme variable d'appariement. Contrairement aux autres méthodes déterministes, elle est la seule qui permet de coupler deux enregistrements

lorsque le NAM manque dans la base de données de l'ESCC. Puisque le NAM est une zone d'appariement très importante, afin de minimiser la probabilité de faux appariements, il n'a été ignoré que lorsque toutes les autres variables d'appariement étaient utilisées.

Le couplage probabiliste n'exige pas que la concordance soit complète pour les variables d'appariement. Des systèmes de pondération ont été conçus pour toutes les zones d'appariement. La valeur de pondération peut être choisie de diverses façons, y compris à l'aide de probabilités, d'où le nom de couplage probabiliste. Lorsque deux enregistrements sont comparés, des points sont attribués ou soustraits en se fondant sur les similarités ou les différences entre les zones appariées. Par exemple, si les valeurs du code postal sont les mêmes dans les deux enregistrements, une cote positive est attribuée; si les valeurs paraissent semblables selon un algorithme de comparaison de chaînes de caractères, une cote positive plus faible est attribuée, reflétant la concordance partielle; si les valeurs figurant dans les deux enregistrements sont totalement différentes, des points sont soustraits. Le nombre de points doit refléter l'importance de la variable d'appariement, laquelle est habituellement fonction du caractère unique de la variable. Dans notre étude, les NAM sont uniques, si bien que nous leur avons attribué une cote comparativement élevée en valeur absolue.

Les cotes sont additionnées sur l'ensemble des zones d'appariement afin d'obtenir un poids total d'appariement pour chaque paire possible. Puis, des seuils sont choisis en se basant sur la distribution des poids d'appariement. La distribution optimale est bimodale. Les paires d'enregistrements dont le poids est supérieur à un seuil choisi sont considérées comme des appariements vrais, celles dont le poids est inférieur à un seuil choisi sont rejetées comme n'étant pas des appariements, et celles dont le poids est compris entre deux seuils d'exclusion choisis sont considérées comme des appariements possibles et sont habituellement examinées manuellement. Pour les besoins du projet, afin de réduire au minimum l'examen manuel, les deux seuils d'exclusion étaient identiques. Newcombe (1988) décrit en détail la façon d'attribuer les points et de les utiliser. Le Système généralisé de couplage d'enregistrements (SGCE), qui est un logiciel développé par Statistique Canada, a été utilisé pour exécuter le couplage probabiliste. Il s'appuie sur la théorie du couplage probabiliste élaborée par Fellegi et Sunter (1969). Dans la présente étude, l'approche par couplage probabiliste est nommée méthode 0. Une description détaillée de cette approche du couplage d'enregistrements figure dans Nadeau (2007).

2.3 Résultats sommaires

Le tableau 2 présente les taux d'appariement des enregistrements du fichier de l'ESCC et de la base de données de l'ISOP pour les répondants qui ont consenti au couplage. La proportion varie selon le cycle et la méthode de couplage. Les taux ne sont pas pondérés par les poids de sondage et ne comprennent pas les résidents du Québec.

Comme le montre le tableau 1, la proportion de NAM manquant dans le fichier de l'ESCC est près de deux fois plus élevée pour les cycles « .1 » que pour les cycles « .2 », soit 30 % et environ 15 %, respectivement. Par conséquent, il faut s'attendre à un taux d'appariement plus élevé pour les cycles « .2 ». Il en est ainsi pour le cycle 1.2 (tableau 2), sauf pour la méthode 6, qui n'utilise pas le NAM. Toutefois, le taux d'appariement n'est pas plus élevé pour le cycle 2.2, parce que les enfants n'ont pas été suréchantillonnés pour ce cycle et qu'ils sont moins susceptibles d'être hospitalisés.

Tableau 2. Proportion d'enregistrements de l'ESCC couplés à ceux de l'ISOP, selon la méthode et le cycle

Type de couplage	Probabiliste	Déterministe					
	0	1	2	3	4	5	6
Variables d'appariement	Province NAM Sexe Date de naissance Code postal	Province NAM Sexe	Province NAM Sexe Année de naissance	Province NAM Sexe Année de naissance Mois de naissance	Province NAM Sexe Date de naissance	Province NAM Sexe Année de naissance Code postal	Province Sexe Date de naissance Code postal
Cycle (année)	%	%	%	%	%	%	%
1.1 (2001)	34,6	29,2	28,1	27,4	25,3	20,1	24,7
1.2 (2002)	37,2	35,1	34,0	33,4	31,0	23,3	24,5
2.1 (2003)	33,4	29,1	27,6	26,6	24,5	19,0	22,4
2.2 (2004)	27,7	26,4	25,2	24,8	22,6	15,6	16,5
3.1 (2005)	30,8	26,4	26,1	25,7	23,6	15,1	18,9

Nota : Les taux ne comprennent pas les répondants du Québec. Les proportions ne sont pas pondérées par les poids de sondage.

Quelle que soit la méthode, les taux d'appariement pour les cycles « .1 » ont tendance à diminuer avec le temps. Cela est dû, en partie, à la diminution du nombre d'enregistrements dans la base de données de l'ISOP, probablement en raison du recours plus fréquent à des services de soins ambulatoires et à la chirurgie d'un jour. La diminution est plus importante lorsqu'on passe du cycle 2.1 au cycle 3.1 que du cycle 1.1 au cycle 2.1, parce que, pour le cycle 1.1, la proportion de répondants de plus de 65 ans est plus faible. Puisque les personnes âgées sont plus susceptibles d'être hospitalisées, le nombre plus faible de répondants âgés réduit les taux d'appariement pour le cycle 1.1.

La méthode 5 produit systématiquement le taux d'appariement le plus faible, parce qu'elle exige la concordance d'un grand nombre de zones différentes. La seule différence entre les méthodes 2 et 5 est l'exigence d'une concordance des codes postaux dans la seconde. L'effet de cette exigence peut être évalué en comparant les taux d'appariement des deux méthodes.

La diminution progressive des taux d'appariement pour les méthodes 1 à 4 était prévue, étant donné la nature emboîtée de ces approches de couplage. L'écart entre les taux d'appariement pour les méthodes 3 et 4 est plus important, principalement parce que certaines valeurs manquaient pour le jour de la naissance dans les bases de données de l'ESCC et de l'ISOP. Dans l'ISOP, l'information manquante pour le jour de la naissance est imputée comme étant le premier jour du mois et la valeur imputée possède une probabilité plus faible d'appariement.

La méthode 6 a tendance à produire des appariements qui diffèrent de ceux obtenus par les autres méthodes déterministes. Parmi les 24 292 enregistrements du cycle 1.1 de l'ESCC qui ont été appariés à des enregistrements de l'ISOP selon cette méthode, 6 373 ne l'ont pas été dans le cas des autres méthodes déterministes; 91,9 % de ces enregistrements ne contenaient pas de NAM dans la base de données de l'ESCC. La plupart (97,4 %) des 24 292 appariements obtenus par la méthode 6 ont été reproduits par la méthode 0, qui est la méthode probabiliste.

3. Facteurs de risque associés à l'hospitalisation pour une maladie cardiaque

Nous avons créé sept fichiers d'enregistrements couplés de l'ESCC 1.1 et de l'ISOP, un pour chaque méthode de couplage d'enregistrements, qui contenaient les variables nécessaires pour les analyses. Dans chaque fichier, le nombre d'observations était de 72 493 au lieu de 119 383, parce que les répondants du Québec et tous ceux de moins de 30 ans ont été exclus de l'analyse. La prévalence de la maladie cardiaque est faible chez les personnes de moins de 30 ans. Sept régressions logistiques ont été ajustées pour estimer l'association entre le résultat, c'est-à-dire être hospitalisé pour un diagnostic de maladie cardiaque dans les deux ans après l'interview de l'ESCC, et certaines caractéristiques sociodémographiques, les antécédents d'hospitalisation pour une maladie cardiaque et les facteurs de risque modifiables pour lesquels des données sont disponibles dans la base de données de l'ESCC. Les sept

méthodes de couplage d'enregistrements ont produit sept distributions de la variable de résultat et de la variable indépendante Antécédents d'hospitalisation pour une maladie cardiaque au cours des cinq années précédant l'interview de l'ESCC. Afin de tenir compte des effets du plan de sondage, la méthode du bootstrap a été utilisée pour estimer les erreurs types des coefficients (Rao, Wu et Yue, 1992; Rust et Rao, 1996; Yeo, Mantel et Liu, 1999; Kovacevic et Roberts, 2002). Les données et la méthodologie sont décrites à la section 3.1. Les résultats sont présentés à la section 3.2 et sont suivis d'une section consacrée aux limites de l'étude.

3.1 Données et méthodologie

Les données de l'ESCC (cycle 1.1) ont fourni l'information sur les caractéristiques sociodémographiques des répondants et sur certains facteurs modifiables de risque de maladie cardiaque. Le résultat dichotomique, hospitalisation pour une maladie cardiaque dans les deux ans après la date de l'interview de l'ESCC, a été tiré de la base de données de l'ISOP. Les répondants ont été catégorisés comme ayant été hospitalisés pour une maladie cardiaque, événement auquel on a attribué une valeur de 1, si le diagnostic principal (diagnostic responsable de la plus grande partie de l'hospitalisation) correspondait à un code de la CIM-9 commençant par 402, 404, 41 ou 42, ou à un code de la CIM-10 commençant par I11, I13, I2, I3, I4, I50 ou I51. Une valeur de 0, indiquant qu'aucune hospitalisation pour maladie cardiaque n'avait eu lieu, a été attribuée pour tous les autres codes de diagnostic. L'ISOP a également fourni un indicateur dichotomique d'antécédents d'hospitalisation pour maladie cardiaque. Une valeur de 1, reflétant des antécédents d'hospitalisation pour maladie cardiaque, a été attribuée aux répondants pour lesquels l'un des codes de la CIM-9 ou de la CIM-10 énumérés plus haut figuraient dans n'importe laquelle des zones de diagnostic. Une valeur de 0, indiquant l'absence d'antécédents, a été attribuée à tous les autres répondants.

Les caractéristiques personnelles comprennent le sexe, l'âge, le niveau de scolarité, le revenu, l'état matrimonial et les modalités de logement. Un ensemble de variables muettes a été créé pour refléter chaque mesure. Pour le sexe, un code de zéro a été attribué aux hommes et un code de 1, aux femmes. Cinq groupes d'âge ont été créés, à savoir : 30 à 39 ans, 40 à 49 ans, 50 à 59 ans, 60 à 69 ans et 70 ans et plus. Le groupe des 30 à 39 ans a été utilisé comme catégorie de référence. Quatre groupes ont été établis pour le niveau de scolarité, à savoir : pas de diplôme d'études secondaires, diplôme d'études secondaires, études postsecondaires partielles et diplôme d'études postsecondaires, qui ont été choisis comme catégorie de référence. La variable de revenu du ménage tient compte des revenus de tous les membres du ménage en provenance de toutes les sources et de la taille du ménage. Quatre groupes ont été créés, à savoir : inférieur, moyen-inférieur, moyen-supérieur et supérieur. Le dernier a été utilisé comme catégorie de référence. Afin de maximiser la taille de l'échantillon, une variable muette a été créée pour indiquer si l'information sur le revenu du ménage était disponible. Les répondants pour lesquels la valeur de cette variable dérivée manquait ont reçu un code de 1. L'état matrimonial a été réparti en trois groupes, à savoir : marié(e) ou vit avec un(e) conjoint(e), célibataire et marié(e) antérieurement, qui comprend les répondants séparés, divorcés ou veufs. Le premier groupe a servi de catégorie de référence. Un code de 1 pour les modalités de logement a été attribué aux répondants qui vivaient seuls. Un code de 0 a été attribué à tous les autres.

Une auto-évaluation du stress a été incluse dans l'analyse. On a demandé aux répondants : « En pensant à la quantité de stress dans votre vie, diriez-vous que la plupart de vos journées sont : pas du tout stressantes, pas tellement stressantes, un peu stressantes, assez stressantes ou extrêmement stressantes? » Trois groupes de réponses à peu près égaux ont été créés, à savoir : pas du tout ou pas tellement stressantes (groupe de référence), un peu stressantes et assez ou extrêmement stressantes.

Les facteurs de risque modifiables pour lesquels l'ESCC fournit des données comprennent l'usage du tabac, l'activité physique durant les loisirs, la consommation d'alcool et l'indice de masse corporelle. La situation d'usage du tabac a été déterminée en demandant aux participants à l'enquête s'ils fumaient des cigarettes tous les jours, à l'occasion ou jamais. Trois groupes ont été créés, à savoir : personnes n'ayant jamais fumé, anciens fumeurs et fumeurs (tous les jours ou à l'occasion). La catégorie « n'ayant jamais fumé » a été utilisée comme catégorie de référence. Deux niveaux d'activité physique pendant les loisirs ont été définis, à savoir : actif ou moyennement actif (1,5 kilocalorie ou plus par kilo de poids corporels par jour), qui a été choisi comme catégorie de référence et inactif (moins de 1,5 kilocalorie par kilo de poids corporels par jour). Un niveau moyen d'activité correspondrait, par exemple, à marcher pendant une heure quatre fois par semaine. En suivant la méthodologie de Wilkins (2002), quatre groupes ont été spécifiés pour préciser le niveau de consommation d'alcool. Les abstinentes tout au long de la vie, qui

représentent la catégorie de référence, regroupaient les répondants ayant déclaré n'avoir jamais consommé d'alcool. Les anciens buveurs regroupaient les répondants ayant déclaré ne pas avoir bu d'alcool l'année précédente, mais avoir bu au moins un verre avant cela. Le niveau de consommation d'alcool des répondants ayant déclaré qu'ils avaient bu au moins un verre d'alcool au cours des douze mois qui ont précédé l'interview a été déterminé d'après le nombre de verres qu'ils ont déclaré avoir bus durant la semaine qui a précédé l'entrevue de l'ESCC. Les buveurs occasionnels étaient ceux ayant déclaré ne pas avoir bu d'alcool au cours de la semaine qui a précédé l'interview. Les petits buveurs étaient ceux ayant déclaré avoir bu un verre la semaine précédente. Comme les hommes et les femmes métabolisent l'alcool différemment, des seuils particuliers au sexe ont été utilisés pour classer les buveurs modérés et les grands buveurs. Les buveurs modérés ont été définis comme ayant bu de 2 à 9 verres l'année qui a précédé l'entrevue pour les femmes et de 2 à 14 pour les hommes. Les grands buveurs ont été définis comme ayant bu 10 verres ou plus la semaine qui a précédé l'entrevue pour les femmes et 15 verres ou plus pour les hommes. L'indice de masse corporelle (IMC) a été calculé en divisant le poids en kilogrammes par le carré de la taille en mètres. Quatre groupes d'IMC ont été créés. Les répondants dont l'IMC était inférieur à 20 ont été considérés comme ayant un poids insuffisant. Ceux dont l'IMC était égal ou supérieur à 20, mais inférieur à 25 ont été considérés comme ayant un poids normal et ont été choisis comme catégorie de référence. Les répondants dont l'IMC était égal ou supérieur à 25, mais inférieur ou égal à 27 ont été considérés comme ayant un certain excès de poids et ceux dont l'IMC était supérieur à 27, comme ayant un surpoids.

3.2 Résultats

Pour chaque méthode de couplage, le tableau 3 donne la distribution de la variable de résultat, c'est-à-dire l'hospitalisation dans les deux ans après l'interview de l'ESCC et de l'une des variables indépendantes, c'est-à-dire des antécédents d'hospitalisation pour maladie cardiaque. En général, les méthodes dont le taux d'appariement est faible (tableau 2) ont produit des nombres plus faibles de personnes ayant une réponse positive pour ces deux variables. Il fallait s'y attendre, puisque la détermination des cas dépend de la méthode de couplage d'enregistrements. Les résultats des régressions logistiques sont présentés au tableau 4 en annexe. Faute d'espace, les résultats des régressions pour les méthodes 2 et 3 ne sont pas présentés. Ces résultats se situaient généralement entre ceux des méthodes 1 et 4 et peuvent être obtenus sur demande auprès des auteurs.

Tableau 3. Distribution de l'hospitalisation pour maladie cardiaque dans les deux ans après l'interview de l'ESCC et des antécédents d'hospitalisation pour maladie cardiaque, selon la méthode de couplage d'enregistrements

Variable	Valeur	Méthode						
		0	1	2	3	4	5	6
Hospitalisation pour maladie cardiaque après l'ESCC	Oui	1 707	1 410	1 349	1 315	1 239	1 158	1 410
	Non	70 786	71 083	71 144	71 178	71 254	71 335	71 083
Hospitalisation antérieure pour maladie cardiaque	Oui	3 937	3 345	3 221	3 142	2 905	2 674	3 145
	Non	68 556	69 148	69 272	69 351	69 588	69 819	69 348

L'examen du tableau 4 révèle un haut niveau de convergence des résultats des régressions logistiques pour toutes les méthodes de couplage et leur similarité à ceux publiés dans la littérature scientifique (Wilkins, 2002). Comme prévu, une hospitalisation antérieure pour maladie cardiaque augmentait la cote exprimant le risque d'une future hospitalisation pour maladie cardiaque. En outre, les femmes étaient moins susceptibles que les hommes d'être hospitalisées pour une maladie cardiaque, même après neutralisation de l'effet des antécédents d'hospitalisation pour maladie cardiaque. Le vieillissement est associé à une plus grande probabilité d'être hospitalisé pour une maladie cardiaque. Ces résultats convergent pour toutes les méthodes de couplage. Dans le cas des méthodes 4 et 5, aucun écart ne se dégage entre les probabilités d'une hospitalisation pour maladie cardiaque calculées pour le groupe des 40 à 49 ans et pour la catégorie de référence, c'est-à-dire le groupe des 30 à 39 ans, probablement en raison de la puissance statistique plus faible.

Lorsque l'on tient compte de l'effet de certains facteurs figurant dans les ensembles de données et énumérés à la section 3.1, certains facteurs de risque modifiables accroissent systématiquement la probabilité d'une future hospitalisation pour maladie cardiaque. Ces facteurs sont le fait d'être un fumeur ou un ancien fumeur, l'inactivité et le surpoids. L'effet d'un certain excès de poids atteignait la signification statistique pour les données provenant des

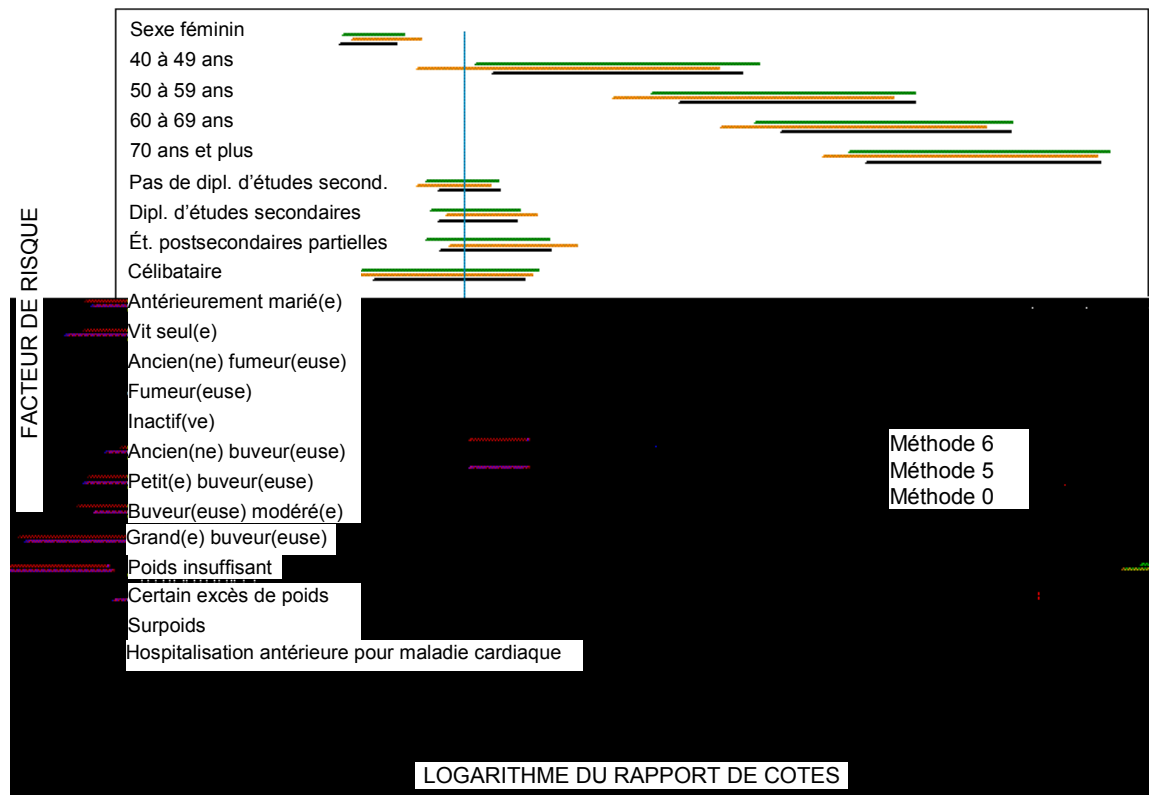
méthodes 0 et 6, tandis qu'une insuffisance pondérale réduisait la cote dans le cas des données provenant des méthodes 5 et 6.

En général, le niveau de scolarité, le revenu, l'état matrimonial, les modalités de logement et l'autoévaluation du stress n'influençaient pas la cote exprimant le risque d'une future hospitalisation pour maladie cardiaque, après neutralisation de l'effet des autres facteurs. Les méthodes 1, 4 et 5 révèlent une plus forte probabilité d'hospitalisation pour les répondants vivant dans un ménage appartenant au groupe de revenu moyen-inférieur et la méthode 1 a permis de déceler une association significative pour le groupe de revenu inférieur.

La figure 1 illustre les intervalles de confiance à 95 % des coefficients de régression logistique, c'est-à-dire le logarithme des rapports de cotes. Seules les méthodes 0, 5 et 6 ont été retenues, car ce sont celles qui ont montré les différences les plus importantes entre les intervalles de confiance. Cela n'est pas étonnant, puisque les méthodes 0 et 5 ont, respectivement, les taux d'appariement le plus élevé et le plus faible, et que la méthode 6 n'utilise pas le NAM. De même, afin de rendre la figure plus compréhensible, le stress et le revenu ne sont pas représentés. Les intervalles de confiance produits par les trois méthodes tendaient à se chevaucher considérablement. Toutefois, l'alignement n'était pas bon pour les antécédents hospitalisation pour maladie cardiaque. Il est possible que la méthode de couplage choisie ait un effet sur la variable de réponse et la variable d'antécédents d'hospitalisation pour maladie cardiaque, puisque toutes deux ont été dérivées du couplage. Par contre, les données sur les autres variables choisies dans la base de données de l'ESCC sont fondées sur une autodéclaration.

Les intervalles de confiance ont tendance à être plus petits pour la méthode 0 et plus grands pour la méthode de 5. La création d'un moins grand nombre d'appariements aboutit moins souvent à une variable de réponse qui est égale à 1. Il est bien connu que la régression logistique produit des estimations des paramètres ayant une plus grande variabilité dans ces circonstances. Les erreurs de couplage pourraient aussi avoir joué un rôle. La méthode 0 est la meilleure des sept méthodes étudiées; elle produit un moins grand nombre de faux appariements et un moins grand nombre d'appariements manqués. Neter, Maynes et Ramanathan (1965) et Krewsky, Wang, Bartlett, Zielinski et Mallick (2001) ont montré que les erreurs d'appariement peuvent entraîner un biais et un accroissement de la variabilité. Bien que leurs articles aient été rédigés, respectivement, dans le contexte des modèles linéaires et de l'analyse de survie, il est raisonnable de penser que leurs résultats soient applicables à la régression logistique. Scheuren et Winkler (1993) ont introduit une technique qui pourrait atténuer le biais causé dans les analyses par les erreurs d'appariement. Elle n'a pas été appliquée ici, car la méthodologie a été mise au point pour les modèles linéaires.

Figure 1. Intervalles de confiance à 95 % des logarithmes des rapports de cotes



Nota : Les logarithmes des rapports de cotes sont équivalents aux coefficients de régression logistique.

Les analyses ont été refaites en incluant uniquement les répondants qui n'avaient pas été hospitalisés pour maladie cardiaque dans les cinq années avant l'interview de l'ESCC. Les résultats détaillés ne sont pas présentés ici, mais peuvent être obtenus sur demande auprès des auteurs. Comme pour les ensembles de données précédents, les résultats étaient relativement convergents pour les diverses méthodes de couplage d'enregistrements. En outre, nombre d'associations décelées durant le premier ensemble d'analyses ont été reproduites.

Les femmes étaient moins susceptibles de subir une hospitalisation pour laquelle le diagnostic principal était une maladie cardiaque. La probabilité d'être hospitalisé augmentait avec l'âge. Cependant, l'association entre les groupes d'âge et l'hospitalisation n'a pas été observée pour le groupe des 40 à 49 ans. Être un ancien fumeur augmentait la probabilité d'une hospitalisation pour maladie cardiaque dans les deux ans après l'entrevue de l'ESCC. Être un fumeur et avoir un surpoids augmentaient la cote exprimant le risque d'hospitalisation dans le cas de toutes les méthodes de couplage, sauf la méthode 5, c'est-à-dire celle pour laquelle le taux d'appariement est le plus faible. Avoir un poids insuffisant réduisait la cote exprimant le risque d'hospitalisation dans les ensembles de données obtenus par les méthodes 4, 5 et 6. L'inactivité était associée positivement à une future hospitalisation, sauf dans le cas des méthodes 4 et 5. Pour ces ensembles de données, être un ancien buveur augmentait la cote de risque d'une future hospitalisation pour maladie cardiaque. Le résultat n'est pas inattendu, puisque les anciens buveurs pourraient être en moins bonne santé (Wilkins, 2002). Vivre seul et déclarer être exposé à un niveau moyen de stress étaient deux facteurs associés positivement à l'hospitalisation, dans le cas des méthodes 0 et 6 pour le premier et des méthodes 0, 1, 4 et 6 pour le second.

Comme dans le premier ensemble d'analyses, le niveau de scolarité et l'état matrimonial n'étaient pas associés à une future hospitalisation pour maladie cardiaque. Vivre dans un ménage ayant un revenu moyen-inférieur était un prédicteur d'une future hospitalisation pour maladie cardiaque dans les ensembles de données obtenus par les méthodes 0, 1, 4 et 5.

3.3 Limites

La présente étude a plusieurs limites. L'échantillon de l'ESCC utilisé était constitué des répondants ayant consenti à ce que l'information qu'ils avaient fournie soit couplée à des sources de données administratives. L'importance du biais créé par l'exclusion des répondants qui n'ont pas autorisé ce genre de couplage est inconnue. Toutefois, il devrait être faible, puisque 90 % des répondants ont donné leur consentement au couplage et que les poids de sondage ont été corrigés pour s'assurer que l'échantillon réduit demeure représentatif, de même que les résultats des régressions logistiques.

Les personnes vivant en établissement ne faisaient pas partie du champ d'observation de l'étude. En général, les résidents des établissements de soins ont tendance à être en moins bonne santé que leurs homologues vivant dans la collectivité. Leur exclusion pourrait avoir affaibli la force des associations publiées.

Les données de l'enquête sont recueillies par autodéclaration et leur exactitude est inconnue. Des erreurs pourraient survenir, parce que les répondants ne sont pas disposés à révéler certains renseignements, ont de la difficulté à se souvenir de l'information ou souhaitent plaire à l'intervieweur ou projeter une image positive. La distorsion qui résulte de ces erreurs de déclaration est inconnue, mais devrait être faible, puisqu'elle est en principe aléatoire, c'est-à-dire non corrélée à la variable de résultat.

Deux types d'erreurs peuvent survenir lorsque l'on procède au couplage d'enregistrements. Deux enregistrements peuvent être appariés, mais ne pas appartenir à la même personne ou bien deux enregistrements peuvent appartenir à la même personne, mais ne pas être appariés. Les deux types d'erreurs pourraient s'être produits dans l'étude. Il est possible qu'un enregistrement de l'ESCC ait été apparié à plusieurs enregistrements de l'ISOP et que ces enregistrements n'appartenaient pas à la même personne. Cet événement était le plus susceptible de se produire dans le cas de l'approche probabiliste, c'est-à-dire la méthode 0, ou dans celui de la méthode 6, où le NAM ne faisait pas partie des variables d'appariement. Dans tous les cas où des problèmes ont été relevés, certains appariements, voire tous, ont été rejetés.

De même, étant donné la mobilité à l'intérieur d'une province, l'utilisation du code postal pour les méthodes 5 et 6 pourrait avoir réduit le nombre d'appariements obtenus, surtout dans le cas des tentatives d'appariements d'anciens enregistrements d'hospitalisation à des enregistrements de l'ESCC. La mobilité d'une province à l'autre aggrave le problème, puisque les variables d'appariement changent après un déménagement interprovincial (sauf la date de naissance et le sexe) et que toutes les méthodes de couplage, y compris la méthode probabiliste, sont affectées de la même façon. L'importance de la sous-estimation d'une hospitalisation antérieure ou future pour maladie cardiaque ou des deux, dépend du nombre de déménagements et du moment où ils ont eu lieu. L'accès aux noms des répondants comme variable d'appariement aurait réduit l'importance de la sous-estimation, mais ceux-ci ne sont pas disponibles dans la base de données de l'ISOP.

L'obtention des noms aurait aussi facilité l'évaluation de la qualité des appariements. Néanmoins, l'examen des appariements produits par la méthode 6 en se fondant sur un sous-échantillon pour lequel le NAM était disponible au cycle 1.1 de l'ESCC donne à penser que cette méthode donne lieu à des appariements faux dans 3 % des cas environ. Autrement dit, quand il était disponible, le NAM a corroboré environ 97 % des appariements.

L'effet de l'exclusion des résidents du Québec sur les résultats est inconnu et limite la généralisation des résultats de l'étude. L'effet de cette exclusion est proportionnel à la mesure dans laquelle les caractéristiques personnelles et le profil de risque des résidents du Québec diffèrent de ceux des autres répondants et de la mesure dans laquelle la prestation des soins de santé dans cette province diffère de celle des autres provinces.

Afin de confirmer que les personnes incluses dans l'analyse avaient été hospitalisées pour une maladie cardiaque et de réduire l'effet des différences entre les méthodes de diagnostic et de codage appliquées dans les divers hôpitaux, seul le diagnostic principal a été choisi pour la variable de résultat. Des études par seconde saisie des données indiquent que les pratiques de diagnostic et de codage sont relativement uniformes d'une province à l'autre, surtout si l'on examine le diagnostic principal et que la prévalence du problème de santé est élevée (Juurink, Preyra, Croxford, Chong, Austin, Tu et Laupacis, 2006; CHIMA et ICIS, 2005; ICIS, 2004; ICIS, 2003; ICIS, 2002). La

détermination d'une hospitalisation antérieure pour maladie cardiaque a été fondée pour tous les diagnostics disponibles afin de maximiser les chances de repérer tous les répondants présentant ce genre d'antécédents.

Puisque la maladie cardiaque a été choisie comme résultat, les personnes qui ont été hospitalisées pour un infarctus aigu du myocarde (IAM) et qui ont subi une revascularisation durant leur hospitalisation pourraient avoir été omises si le code du diagnostic principal était celui d'une maladie coronarienne. Cette omission touche un petit nombre de patients et pourrait avoir contribué à une sous-estimation de la relation présentée ici, puisque les caractéristiques de ces patients devraient être semblables à celles des patients pour lesquels le diagnostic principal était une maladie cardiaque.

Les personnes qui ont été hospitalisées pour une maladie cardiaque dans les deux ans après l'entrevue de l'ESCC, mais qui n'avaient pas reçu leur congé de l'hôpital à la fin de la période de deux ans, ont été traitées comme si elles n'avaient pas été hospitalisées. Celles qui sont décédées durant la période de deux ans sans avoir été hospitalisées ont été considérées comme n'ayant pas été hospitalisées. Le biais résultant est inconnu, mais affaiblit vraisemblablement les associations qui se dégagent de la présente étude, en réduisant la puissance statistique de cette dernière, puisqu'il est fort probable que les caractéristiques personnelles et le profil de risque de ces personnes soient semblables à ceux des répondants qui ont été hospitalisés pour une maladie cardiaque.

4. Conclusion

Les résultats des diverses méthodes de couplage d'enregistrements sont convergents, que l'on inclue ou non les sujets ayant des antécédents d'hospitalisation pour maladie cardiaque dans l'analyse. Les résultats concordent aussi avec ceux d'études antérieures portant sur les facteurs de risque de maladie cardiaque. Selon la présente analyse, les femmes sont moins susceptibles d'être hospitalisées pour un diagnostic de maladie cardiaque. Le vieillissement est associé à l'accroissement de la cote exprimant le risque d'hospitalisation pour maladie cardiaque chez le groupe des 50 ans et plus. Les anciens fumeurs et les personnes considérées comme ayant un surpoids ont aussi plus de risques d'être hospitalisés pour une maladie cardiaque.

Les différences entre les deux ensembles d'analyses, c'est-à-dire l'ensemble comportant les sujets ayant des antécédents d'hospitalisation pour maladie cardiaque et celui ne les contenant pas, méritent d'être soulignées. Dans le second ensemble d'analyses, le fait d'être un ancien buveur est associé positivement aux résultats. Par contre, dans le premier ensemble, l'inactivité est un prédicteur. Comme prévu, dans le second ensemble, avoir des antécédents d'hospitalisation pour maladie cardiaque est associé à un accroissement de la cote exprimant le risque d'une hospitalisation future pour maladie cardiaque. En général, le niveau de scolarité, le revenu du ménage, les modalités de logement et l'état matrimonial ne sont pas des prédicteurs de l'hospitalisation pour maladie cardiaque.

Malgré les limites de la présente étude, notamment le pouvoir réduit de généraliser les résultats à d'autres classes de maladies, la méthode probabiliste semble donner des résultats qui ont été reproduits par la plupart des méthodes de couplage déterministes. Il est donc recommandé de la privilégier, puisqu'elle accroît la puissance de détection des associations statistiquement significatives, en produisant peu d'appariement faux et en manquant un moins grand nombre d'appariements que les approches de couplage déterministes.

Remerciements

Nous remercions Helen Johansen, Brad Thomas et Richard Trudeau de leurs commentaires et discussions qui nous ont permis d'améliorer le présent article. Nous remercions aussi Mary Sue Devereaux qui en a amélioré considérablement la lisibilité.

Références

- Béland Y. (2002), "Enquête sur la santé dans les collectivités canadiennes - aperçu de la méthodologie", *Rapports sur la santé*, 13(3), 9-16.
- Canadian Health Information Management Association (CHIMA) & Canadian Institute for Health Information (CIHI) 2005, "Reabstraction Study of the Ontario Case Costing Facilities for Fiscal Years 2002/2003 and 2003/2004", Toronto.
- Canadian Institute for Health Information (CIHI) 2004, "Data Quality of the Discharge Abstract Database Following the First-Year Implementation of ICD-10-CA/CCI – Final report", Toronto.
- Canadian Institute for Health Information (CIHI) 2003, "Discharge Abstract Database (DAD)/ CMG/Plx Data Quality", Toronto.
- Canadian Institute for Health Information (CIHI) 2002. "Discharge Abstract Database Data Quality Re-abstraction Study. Combined Findings for Fiscal Years 1999/2000 and 2000/2001", Toronto.
- Fellegi I.P. and A.B. Sunter (1969), "A Theory for Record Linkage", *Journal of the American Statistical Association*, 64, 1183-1210.
- Juurink D., C. Preyra, R. Croxford, A. Chong, P. Austin, J. Tu and A. Laupacis (2006), "Canadian Institute for Health Information Discharge Abstract Database: A Validation Study", Institute for Clinical Evaluative Sciences, Toronto.
- Kovacevic, M and G. Roberts (2002), "Notes on Estimating Equations Bootstrap", rapport non publié, Statistique Canada.
- Krewski D., Y. Wang, S. Bartlett, J.M. Zielinski et R.Mallick (2001), "D'appariement sur l'inférence statistique impact d'erreurs lors d'études de mortalité par cohortes", *Symposium de Statistique Canada - Recueil*, 1-13.
- Nadeau C. (2006), "Linking HPOI to CCHS", rapport non publié, Ottawa, Canada: Statistique Canada.
- Newcombe H. (1988), *Handbook of Record Linkage: Methods for Health and Statistical Studies, Administration, and Business*. Oxford Medical Publications, Oxford.
- Neter J., E.S. Maynes and R.Ramanathan (1965), "The Effect of Mismatching on the Measurement of Response Errors", *Journal of the American Statistical Association*, 60, 1005-1027.
- Rao J.N.K., C.F.J Wu. et K. Yue (1992), "Quelques travaux récents sur les méthodes de rééchantillonnage applicables aux enquêtes complexes", *Techniques d'Enquête* (Statistique Canada, Catalogue 12-001), 18(2), 225-234.
- Rust K.F. and J.N.K. Rao (1996), "Variance estimation for complex surveys using replication techniques", *Statistical Methods in Medical Research*, 5, 281-310.
- Scheuren, F. et W.E. Winkler (1993), "Analyse de régression de fichiers de données couplés par ordinateur", *Techniques d'Enquête*, 19, 45-66.
- Wilkins K. (2002), "Consommation modérée d'alcool et maladie cardiaque", *Rapports sur la santé*, 14(1), 9-28.
- Yeo D., H. Mantel and T.P. Liu (1999), "Bootstrap variance estimation for the National Population Health Survey", American Statistical Association: Proceedings of the Survey research Methods Section, Baltimore, August 1999.

Annexe

Tableau 4. Rapports de cotes corrigées pour une hospitalisation pour maladie cardiaque dans les deux ans après l'interview de l'ESCC, selon certaines caractéristiques, y compris les antécédents d'hospitalisation pour maladie cardiaque, population à domicile, Canada, y compris les territoires, mais non compris le Québec

Facteur de risque	Méthode 0		Méthode 1		Méthode 4		Méthode 5		Méthode 6	
	Rapport de cotes	IC à 95 %	Rapport de cotes	IC à 95 %	Rapport de cotes	IC à 95 %	Rapport de cotes	IC à 95 %	Rapport de cotes	IC à 95 %
Sexe féminin	0,54***	0,45-0,65	0,57***	0,46-0,69	0,59***	0,48-0,73	0,61***	0,49-0,76	0,56***	0,46-0,68
40 à 49 ans	2,68*	1,20-5,98	2,31*	1,01-5,30	2,26	0,94-5,46	1,95	0,74-5,13	2,68*	1,07-6,69
50 à 59 ans	8,51***	3,97-18,2	6,82***	3,12-14,9	6,41***	2,77-14,8	6,41***	2,60-15,8	7,75***	3,31-18,1
60 à 69 ans	16,1***	7,65-33,8	12,9***	6,12-27,1	11,4***	5,15-25,2	12,2***	5,19-28,6	14,8***	6,46-34,1
70 ans et plus	28,2***	13,3-59,9	24,9***	11,5-53,9	23,7***	10,5-53,9	24,3***	10,0-58,8	27,5***	11,9-63,3
Pas de diplôme d'études secondaires	1,03	0,84-1,26	0,95	0,77-1,18	0,93	0,74-1,18	0,94	0,74-1,19	0,99	0,78-1,25
Diplôme d'études secondaires	1,09	0,85-1,40	1,14	0,87-1,49	1,10	0,82-1,47	1,19	0,89-1,59	1,07	0,81-1,43
Études postsecondaires partielles	1,22	0,85-1,74	1,31	0,90-1,90	1,29	0,87-1,89	1,37	0,91-2,06	1,16	0,78-1,72
Revenu inférieur	1,33	0,95-1,87	1,55*	1,09-2,22	1,43	0,99-2,06	1,38	0,91-2,10	1,13	0,78-1,64
Revenu moyen-inférieur	1,28	0,93-1,77	1,48*	1,08-2,03	1,47*	1,05-2,07	1,49*	1,05-2,12	1,22	0,85-1,76
Revenu moyen-supérieur	0,89	0,66-1,19	1,02	0,75-1,39	1,01	0,73-1,40	1,01	0,72-1,43	0,85	0,62-1,15
Revenu non déclaré	1,26	0,90-1,78	1,31	0,90-1,90	1,10	0,74-1,64	1,26	0,84-1,90	1,09	0,75-1,58
Célibataire	0,91	0,56-1,47	0,76	0,42-1,37	0,71	0,35-1,42	0,74	0,36-1,55	0,91	0,51-1,61
Antérieurement marié(e)	1,18	0,86-1,63	1,17	0,84-1,63	1,08	0,75-1,55	1,17	0,79-1,72	1,11	0,76-1,61
Vit seul(e)	1,06	0,75-1,48	1,01	0,71-1,43	1,07	0,72-1,58	1,02	0,68-1,53	1,12	0,75-1,67
Journées un peu stressantes	1,12	0,92-1,35	1,09	0,89-1,34	1,09	0,88-1,34	1,11	0,89-1,39	1,15	0,92-1,43
Journées assez stressantes, extrêmement stressantes	1,04	0,84-1,30	1,14	0,90-1,45	1,11	0,86-1,44	1,16	0,89-1,53	1,02	0,80-1,32
Ancien(ne) fumeur(euse)	1,42***	1,16-1,75	1,43**	1,14-1,78	1,44**	1,14-1,82	1,39**	1,09-1,78	1,41**	1,12-1,78
Fumeur(euse)	1,54***	1,20-1,97	1,53**	1,16-2,02	1,41*	1,05-1,90	1,37*	1,01-1,86	1,45**	1,10-1,91
Inactif(ve)	1,33**	1,11-1,58	1,31**	1,09-1,59	1,30**	1,06-1,59	1,32*	1,07-1,63	1,33**	1,10-1,62
Ancien(ne) buveur(euse)	1,31	1,00-1,73	1,30	0,96-1,76	1,27	0,91-1,77	1,22	0,87-1,72	1,29	0,95-1,75
Petit(e) buveur(euse)	1,00	0,79-1,25	1,01	0,80-1,29	1,03	0,80-1,33	0,99	0,76-1,29	0,99	0,77-1,27
Buveur(euse) modéré(e)	0,99	0,75-1,31	1,06	0,79-1,43	1,05	0,77-1,44	1,12	0,81-1,57	0,98	0,72-1,34
Grand(e) buveur(euse)	0,74	0,46-1,18	0,87	0,53-1,42	0,92	0,55-1,53	0,90	0,52-1,57	0,92	0,49-1,73
Poids insuffisant	0,74	0,49-1,12	0,76	0,47-1,20	0,68	0,42-1,09	0,57*	0,36-0,90	0,60**	0,41-0,88
Certain excès de poids	1,30*	1,03-1,65	1,20	0,92-1,57	1,23	0,93-1,62	1,24	0,92-1,68	1,35*	1,04-1,75
Surpoids	1,35**	1,13-1,62	1,32**	1,08-1,61	1,35**	1,09-1,67	1,37**	1,09-1,72	1,36**	1,11-1,68
Hospitalisation antérieure pour maladie cardiaque	6,28***	5,20-7,59	7,36***	6,01-9,01	8,61***	6,97-10,6	9,25***	7,37-11,6	8,00***	6,47-9,89

Sources des données : Enquête sur la santé dans les collectivités canadiennes, cycle 1.1, 2001 et base de données de l'Information-santé orientée vers la personne, 1994-1995 à 2003-2004.

* p [0,05; ** p [0,01; *** p [0,001.

Nota : Fondés sur les répondants de l'échantillon qui ont consenti au couplage des renseignements qu'ils ont fournis. Les rapports de cotes pour les catégories de référence sont toujours égaux à 1,00 et ne sont pas présentés. Les résultats sont fondés sur des données pondérées; les erreurs types utilisées dans le calcul des intervalles de confiance ont été estimées par la méthode du bootstrap.