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

Actes du symposium de Statistique Canada, 2006
Enjeux méthodologiques reliés à la mesure de la santé des populations

Application de méthodes de contrôle de la divulgation statistique à la base de données du Système canadien hospitalier d'information et de recherche en prévention des traumatismes

Ann Brown et Margaret Herbert[1]

Résumé

Nous décrivons ici les méthodes de contrôle de la divulgation statistique (CDS) mises au point pour la diffusion publique du fichier de microdonnées du Système canadien hospitalier d'information et de recherche en prévention des traumatismes (SCHIRPT). Le SCHIRPT est une base de données nationale de surveillance des blessures administrée par l'Agence de santé publique du Canada (ASPC). Après une description du SCHIRPT, nous présentons un bref aperçu des concepts de base du CDS en guise d'introduction à la procédure de sélection et d'élaboration des méthodes de CDS applicables au SCHIRPT, compte tenu des défis et des besoins particuliers qui sont associés à ce système. Nous résumons ensuite quelques-uns des principaux résultats. Le présent article se conclut par une discussion sur les répercussions de ces travaux sur le domaine de l'information en matière de santé et des observations finales sur certaines questions méthodologiques qu'il convient d'examiner.

MOTS-CLÉS : fichier de microdonnées à grande diffusion; surveillance des blessures; suppression de cellules.

1. Introduction

Le présent article fait état de la méthodologie de contrôle de la divulgation statistique mise au point pour la diffusion publique du fichier de microdonnées du Système canadien hospitalier d'information et de recherche en prévention des traumatismes (SCHIRPT). Nous présentons tout d'abord un bref aperçu du SCHIRPT et de ses exigences, et nous décrivons ensuite les principaux concepts de la méthodologie de contrôle de la divulgation statistique (CDS). La section suivante porte sur les grandes étapes de l'élaboration de la méthodologie de CDS dans le cadre du SCHIRPT. L'article présente les principaux résultats des procédures de suppression dans la base de données du SCHIRPT ainsi que des exemples touchant des champs précis. Il se conclut par une discussion sur les répercussions potentielles de ces travaux sur le domaine de l'information en matière de santé et par des observations sur certaines questions méthodologiques qu'il conviendrait d'examiner.

2. Qu'est-ce que le SCHIRPT?

2.1 Programme de surveillance des blessures du Canada

Le SCHIRPT est un programme de surveillance des blessures administré par l'Agence de santé publique du Canada (ASPC). Il recueille des renseignements sur les blessures subies par les personnes traitées dans certaines salles d'urgence de diverses régions du Canada. Ce système sentinelle de surveillance regroupant 14 hôpitaux (dont 10 hôpitaux pédiatriques) a pour mission de contribuer à réduire le nombre et la gravité des blessures au Canada (Mackenzie et Pless). La base de données actuelle du SCHIRPT renferme des renseignements sur plus de 1,6 million de blessures, soit quelque 100 000 cas par année.

[1]Ann Brown, Groupe de consultation statistique, Division des méthodes d'enquêtes sociales, Statistique Canada, 15e étage, Immeuble R.-H.-Coats, pré Tunney, Ottawa (Ontario) Canada K1A 0T6 (ann.brown@statcan.ca, ac.ann.brown@gmail.com); Margaret Herbert, Division de surveillance de la santé et de l'épidémiologie, Agence de santé publique du Canada, Immeuble Jeanne-Mance, 200, promenade Eglantine, Ottawa (Ontario) Canada, K1A 0K9 (Margaret_Herbert@ASPC-aspc.gc.ca)

Institué en 1990, le SCHIRPT regroupait alors les salles d'urgence de 10 hôpitaux pédiatriques (Herbert et Mackenzie). Au cours de la première décennie d'existence du programme, six hôpitaux généraux s'y sont joints, mais deux l'ont abandonné par la suite. Les seuls provinces et territoires qui ne sont pas représentés dans le SCHIRPT sont l'Île-du-Prince-Édouard, le Nouveau-Brunswick, la Saskatchewan et le Yukon.

Le SCHIRPT dépend de l'appui des hôpitaux participants. Dans chaque centre, un directeur du SCHIRPT – agissant à titre bénévole –, généralement un urgentologue, est responsable du programme au sein de l'hôpital et de la collectivité. L'ASPC verse des fonds pour couvrir, dans chaque centre, le salaire d'un coordonnateur du SCHIRPT chargé de la distribution des formulaires du SCHIRPT aux patients ou aux parents et de la collecte des formulaires remplis. Le coordonnateur analyse également les données du SCHIRPT de son centre hospitalier et fournit des données et des renseignements sur les blessures à l'hôpital et à la collectivité. Certains hôpitaux sont affiliés à des centres de prévention des blessures ou à des groupes de recherche, et on s'efforce de recruter les codirecteurs du SCHIRPT dans ces organismes.

2.2 Données du SCHIRPT

Les données réunies dans la base de données du SCHIRPT sont tirées d'un questionnaire à réponses libres rempli par les patients ou leurs parents. Une fois les données saisies, chaque enregistrement compte environ 40 variables que l'on peut regrouper en diverses catégories. Les **caractéristiques des patients** comprennent l'âge, le sexe, le code postal, la profession et le secteur d'activité (lorsque la blessure est liée au travail). Les **renseignements cliniques** portent sur la nature de la blessure, la partie du corps touchée, et l'état du dossier (patient ayant reçu son congé, patient admis, etc.). Ces renseignements servent de variable substitutive indiquant la gravité de la blessure.

Les **renseignements sur le traumatisme** se rapportent aux circonstances de la blessure et se distinguent des données cliniques relatives aux blessures que renferment d'autres bases de données sur la santé. Parmi les variables classées sous cette rubrique, notons la ville et le **lieu** où s'est produite la blessure, le **contexte** (activité à laquelle s'adonnait le patient), la **cause** (blessure accidentelle, mauvais traitements, blessure auto-infligée), les **facteurs** ayant pu contribuer à la blessure (p. ex., consommation d'alcool ou de drogues, mauvais usage ou défaut de fonctionnement des produits, sports, conditions environnementales défavorables, etc.), le **type de transfert d'énergie** (mécanique, électrique, chimique, etc.). Une description complète du traumatisme est également consignée.

2.3 Collecte et saisie des données du SCHIRPT

La collecte des données du SCHIRPT se fait selon les étapes suivantes :

- Le patient ou un parent remplit le questionnaire. Dans certains cas, le personnel administratif ou le personnel chargé du triage remplit le formulaire. Dans certains centres, les coordonnateurs tirent les données du dossier lorsque les patients ne sont pas en mesure de remplir le formulaire.
- Les formulaires sont envoyés à l'ASPC.
- Le codage et la saisie électronique des données sont effectués à l'ASPC par une équipe de codeurs très bien formés qui traduisent les réponses aux questions ouvertes en variables codées numériquement et rédigent les descriptions.
- Des vérifications logiques de base et un contrôle de la qualité des données sont effectués lors de la saisie – de manière à rejeter, par exemple, une date de blessure antérieure à la date de naissance du patient.

La base de données nationale est une base relationnelle sur une plate-forme Oracle hébergée dans un serveur de l'ASPC. Au terme de ces étapes, les hôpitaux participant au SCHIRPT reçoivent des mises à jour périodiques de leurs propres données sous la forme d'ensembles de données ACCESS.

2.4 Utilisateurs et utilisations des données du SCHIRPT

Les données du SCHIRPT servent à diverses fins. Les données et les renseignements sur les circonstances du traumatisme complètent les données cliniques d'autres bases de données comme celles de la mortalité et des

hospitalisations. Le SCHIRPT fournit de l'information sur des genres de blessures qui ne sont pas définis dans les codes normalisés de santé de la Classification internationale des maladies (p. ex., blessures associées à des sports précis ou à des produits précis, comme les outils électriques).

Le SCHIRPT donne des renseignements détaillés sur les traumatismes qui permettent d'éclairer les mesures de prévention. Ainsi, l'ASPC utilise les données pour produire des rapports sur les blessures, des publications et l'information affichée sur son site Web. L'ASPC répond aux demandes des chercheurs et d'autres intervenants, et l'information tirée du SCHIRPT alimente également l'application interactive Web de consultation de l'Agence, la Surveillance des blessures en direct (SBD).

On retrouve parmi les utilisateurs des données du SCHIRPT l'administration fédérale, les organisations non gouvernementales s'intéressant à la sécurité et à la prévention des blessures, les bureaux de santé publique, les professionnels de la santé, les chercheurs, les étudiants, le secteur privé, les médias et le grand public.

2.5 Fichier de microdonnées à grande diffusion du SCHIRPT

On a mis au point, dans le cadre du SCHIRPT, un fichier de microdonnées à grande diffusion (FMGD) tiré de la base de données pour accroître et élargir l'accès à cette information. Le FMGD, qui se fonde sur les nouvelles méthodes et technologies, constituera un fichier source mieux adapté à la SBD. Il assurera un accès aux données à un groupe plus vaste d'utilisateurs – des utilisateurs qui n'auront pas à respecter les critères stricts actuellement imposés aux chercheurs présentant une demande d'accès aux données.

L'élaboration du FMGD doit s'appuyer sur deux valeurs fondamentales : la protection de la vie privée des personnes dont les blessures sont décrites dans la base de données et la diffusion de données utiles et pertinentes. Pour mettre au point le FMGD du SCHIRPT, nous avons dû adopter des méthodes :

- systématiques;
- compatibles avec les pratiques exemplaires en matière de contrôle de la divulgation;
- économiques en ce qui a trait au personnel et aux coûts de consultation;
- reproductibles pour chaque nouvelle année de données;
- facilement modifiables pour tenir compte de l'évolution du programme et de la base de données du SCHIRPT.

Pour répondre aux besoins particuliers de certains utilisateurs, l'ASPC a imposé des restrictions qui limitent le choix des méthodes de contrôle de la divulgation utilisées pour créer le FMGD du SCHIRPT. On a décidé d'utiliser les données du plus grand nombre d'enregistrements possible sans modification de l'information consignée dans les enregistrements, de sorte que les méthodes de perturbation des données ont été exclues.

3. Méthodologie de contrôle de la divulgation statistique (CDS) du SCHIRPT

3.1 Types de données et méthodes de CDS

Le choix des méthodes appropriées de CDS dépend du type des données auxquelles ces méthodes seront appliquées. Les types de données consignées dans le SCHIRPT et leur description sont présentés ci-dessous.

- **Données qualitatives et quantitatives**. Les données du SCHIRPT sont qualitatives, c'est-à-dire que tous les champs sont de type nominal et comportent un nombre fini de catégories. Par conséquent, les méthodes, l'arrondissement par exemple, qui s'appliquent aux données quantitatives (numériques), comme le revenu qui est associé à un grand nombre de catégories, ne sont pas requises.
- **Données sous forme de tableaux.** Les méthodes qui s'appliquent aux chiffres sommaires présentés sous forme de tableaux ne sont pas requises puisque les fichiers du SCHIRPT sont constitués de tous les enregistrements individuels des traumatismes, ou microdonnées.

- Deux grands types de méthode de CDS peuvent s'appliquer aux **microdonnées**.

    - En règle générale, les **méthodes de perturbation des données** permettent de modifier les données de manière à préserver les propriétés des statistiques sommaires, comme la moyenne et la variance, et à conserver le plus possible les données individuelles. Parmi ces méthodes figurent l'ajout aléatoire de bruit ou la perturbation aléatoire et la permutation des microdonnées « synthétiques » (avec le souci de limiter le biais). Toutefois, ces méthodes n'ont pas été retenues, en raison de la décision mentionnée précédemment de ne pas modifier les données du fichier du
    - SCHIRPT.

      Seules les méthodes de **réduction ou de restriction des données** étaient donc disponibles. Il s'agit là de méthodes de base du contrôle de la divulgation statistique qui consistent à regrouper ou à réduire l'information. Si la réduction de l'information limite l'utilité des données, les renseignements fournis sont, du moins, aussi exacts et dignes de confiance que les données originales. En général, ces méthodes couvrent :

      . la suppression d'enregistrements visant de très petites sous-populations ou des enregistrements sensibles facilement identifiables;

      . la suppression des champs signalétiques, comme ceux du nom et de l'adresse;

      . la réduction du niveau de détail, ce qui implique le regroupement de catégories, le recodage de certaines variables, l'échantillonnage ou le sous-échantillonnage, le traitement des valeurs aberrantes et des catégories rares, le codage selon des limites inférieures et supérieures déterminées;

      . la suppression de valeurs dans des enregistrements particuliers.

3.2 Concepts du CDS

La présente section donne un bref aperçu des principaux concepts des méthodes de CDS. Ces concepts sont tirés de la version (externe) du guide de Statistique Canada pour la création de fichiers de microdonnées à grande diffusion (2006). Ce document se fonde sur la longue expérience de SC en matière de diffusion des FMGD des enquêtes sociales et des enquêtes auprès des ménages, et sur l'équilibre entre le maintien de la qualité et l'utilité des données, d'une part, et la protection de la confidentialité qu'exige la *Loi sur la statistique*, d'autre part.

- La **réidentification** ou la **divulgation de l'identité d'une personne** peut se produire lorsque la diffusion d'un fichier de microdonnées mène à l'identification d'une personne ou la rend possible. Cette situation peut survenir même lorsque les identificateurs directs, comme le nom, le numéro de téléphone ou d'autres numéros d'identification (le numéro d'assurance sociale, par exemple, qui peut servir au couplage avec d'autres fichiers), ont été supprimés du fichier. Aucun enregistrement du fichier ne devrait permettre l'identification d'une personne.
- Les **identificateurs indirects** se trouvent notamment dans les champs de la base de données qui renferment des renseignements sur le lieu géographique et les caractéristiques (démographiques) des personnes. La base de données peut également comporter des champs **de nature délicate** qui renferment des renseignements personnels, par exemple sur la santé, les antécédents criminels ou le revenu. Par conséquent, on retrouve dans les variables clés, dont la combinaison peut permettre la réidentification, des variables d'identificateurs indirects et de nature délicate.
- Les **variables clés** comprennent les champs/variables du fichier de microdonnées diffusé qui peuvent servir à la réidentification des personnes. Il s'agit notamment des champs renfermant des identificateurs indirects et d'autres champs relatifs aux personnes qui, en combinaison avec les identificateurs indirects, peuvent mener à la réidentification. Ces autres champs sont les champs de nature délicate et ils sont propres à la base de données. Les champs de nature délicate du SCHIRPT seront présentés un peu plus loin dans les tableaux des résultats.
- Les méthodes relatives au **risque de divulgation** mesurent le risque de divulgation associé à une base de données. Elles font généralement appel à un ensemble de variables clés et se fondent sur le calcul de la fréquence des enregistrements **uniques**, doubles, triples et quadruples, en pourcentage de l'ensemble des enregistrements de la base de données, un enregistrement unique étant un enregistrement qui figure une seule fois dans la base pour une valeur donnée des variables clés à l'étude.

- La **suppression** ou le traitement du risque se rapporte à des valeurs précises figurant dans des combinaisons de cellules de champs clés. Ces mesures exigent que l'on détermine un nombre précis d'enregistrements ou **seuil** comportant les mêmes valeurs ou les combinaisons de variables clés qui posent un risque inacceptable de réidentification. On peut établir le seuil aux fins du traitement du risque à 3 ou plus. Le traitement de suppression s'applique aux valeurs des variables clés des enregistrements qui n'atteignent pas le seuil défini.

3.3 Principales caractéristiques du SCHIRPT

Globalement, on compte parmi les caractéristiques et les défis propres au fichier du SCHIRPT la quantité de données, le type de données, le niveau de détail de même que les indicateurs indirects et les variables de nature délicate. Ceux-ci sont décrits de façon plus détaillée à la prochaine section. Cependant, la caractéristique la plus frappante du système tient à son asymétrie et à la présence de certains enregistrements rares. Cette caractéristique tient au fait que le SCHIRPT renferme des renseignements sur les blessures traitées dans les hôpitaux participants, lesquels ne couvrent pas l'ensemble des provinces et des territoires. En effet, les hôpitaux participants ne constituent pas un échantillon représentatif d'hôpitaux choisis au hasard. Par conséquent, les groupes suivants sont sous-représentés : les adolescents plus âgés et les adultes traités dans les hôpitaux généraux; les populations des régions rurales (la plupart des hôpitaux participant au SCHIRPT sont situés dans les grands centres urbains); les populations autochtones vivant dans des régions rurales ou éloignées. À titre d'exemple, les distributions de l'âge, du secteur d'activité et de la profession figurant dans le SCHIRPT diffèrent de celles de l'ensemble de la population. Seulement 20 % des enregistrements du SCHIRPT visent des personnes âgées de 15 ans et plus, de sorte que les variables relatives au secteur d'activité, par exemple, ne s'appliquent qu'à une petite proportion du sous-ensemble déjà restreint des personnes âgées de 15 ans et plus.

3.4 Méthodologie de CDS du SCHIRPT

Nous décrivons ici le processus et les principales étapes ayant mené à la sélection des méthodes de CDS adaptées au SCHIRPT, compte tenu des caractéristiques et des exigences du système, ainsi que l'application et l'optimisation de ces méthodes. Celles-ci ont été appliquées à chacune des années de données du SCHIRPT.

Les premières étapes consistent à examiner et à analyser les données du SCHIRPT afin de cerner les petites sous-populations à risque élevé, les champs à risque élevé et tout ensemble de variables clés et de nature délicate qui, seules ou en combinaison, posent un risque de réidentification. Le contenu du fichier de microdonnées du SCHIRPT a donc été analysé selon les paramètres suivants.

- **Géographie du SCHIRPT –** Code postal et code de l'hôpital qui indiquent la province ou le territoire ainsi que la municipalité de l'hôpital.
- **Identificateurs directs** – Renseignements identifiant clairement la personne ayant subi la blessure. Le SCHIRPT comporte un champ indiquant les trois premières lettres du nom de famille de la personne ayant subi la blessure.
- **Identificateurs indirects** – Champs présentant les caractéristiques des personnes. Dans le SCHIRPT, il s'agit de l'âge, du sexe, de la profession, de la branche d'activité et de la langue parlée.
- **Renseignements de nature délicate** – Par exemple, la description des circonstances médicales et autres relatives au traumatisme, considérée de nature délicate en raison de la nature de la blessure ou de sa fréquence peu élevée.
- **Niveau de détail**. Détermination de la fréquence par cellule des catégories de chacun des champs du SCHIRPT.

Par la suite, on a mis au point des méthodes de réduction des données adaptées au SCHIRPT. Il s'agit notamment des suivantes.

- Retrait des enregistrements comportant une caractéristique rare
  Les enregistrements de cas de décès sont retirés du fichier parce qu'on en compte très peu — environ 50 décès par année. Comme ils représentent une très petite sous-population dans le SCHIRPT, ils posent un risque indûment élevé de divulgation involontaire. Les décès sont souvent bien connus dans la collectivité

où ils surviennent et en raison de la couverture qu'en font les médias — p. ex., dans les cas d'accidents ferroviaires graves, de collisions routières, etc. De plus, les renseignements sur les décès sont facilement accessibles dans d'autres sources, notamment les notices nécrologiques, les statistiques sommaires publiques et les bases de données en ligne sur les décès. Le couplage des renseignements publics et des renseignements figurant dans le SCHIRPT pourrait permettre la réidentification d'une personne et la divulgation de nouveaux renseignements personnels sur le patient ou sur le traumatisme. À titre d'exemple, les décès causés par un accident ferroviaire sont rares (souvent moins de 10 par année au Canada) et il serait donc facile de retracer l'identité des personnes à partir de leur âge ou de leur sexe. Un enregistrement du SCHIRPT qui toucherait un décès causé par un accident ferroviaire pourrait aussi révéler des renseignements supplémentaires comme une tentative de suicide, la consommation d'alcool ou l'intervention funeste d'un tiers.

- Retrait de certains champs
  Quinze champs qui posaient un risque indûment élevé de divulgation ont été retirés de la base de données. Notons à cet égard :
  - les identificateurs directs – nom partiel, code de l'hôpital, numéro du dossier;
  - les champs géographiques – code postal, code de l'hôpital;
  - les dates – naissance, traumatisme, visite de la salle d'urgence;
  - les circonstances médicales rares – causes;
  - les champs des textes des descriptions – retrait de la description des traumatismes parce qu'il serait trop laborieux d'examiner manuellement une masse considérable de textes.

- Diffusion des champs clés
  L'application des méthodes suivantes de contrôle de la divulgation a permis la diffusion des champs clés :
  - **Regroupement –** Agrégation de catégories de faible fréquence. On a regroupé, par exemple, les blessures rares, comme les amputations. On a dû faire appel à des compétences spécialisées pour établir des catégories à la fois utilisables et pertinentes. On a effectué des vérifications à partir des fréquences par cellule avant et après le regroupement.**Catégories normalisées** – Les catégories « autre » et « non déclaré » sont des codes originaux valides pour toutes les variables. Les renseignements relatifs à certaines variables du SCHIRPT ne s'appliquent pas à toutes les blessures et sont alors consignés dans la catégorie « non déclaré », p. ex, la position assise dans le véhicule dans le cas de blessures qui ne sont pas associées à un véhicule. La catégorie « autre » regroupe les situations qui ne sont pas couvertes par les autres catégories de code.
    Suppression de cellules – On a examiné et traité des combinaisons de champs clés de manière à supprimer les cellules renfermant des combinaisons de valeurs peu fréquentes, en fonction des limites minimales par cellule. Cette mesure s'est appliquée aux cellules renfermant des valeurs peu fréquentes pour les cinq combinaisons des identificateurs démographiques indirects, quatre combinaisons de l'âge et du sexe et toutes les combinaisons de sept champs de nature très délicate (partie du corps, concours de circonstances, contexte, lieu, facteur mécanique, mécanisme du traumatisme, nature de la blessure); quatre combinaisons de l'âge et du sexe et toutes les combinaisons de quatre champs de nature moins délicate (état du dossier, type d'incident, équipement protecteur 1 et position assise dans un véhicule). Toutes les combinaisons de dimension inférieure de chacune de celles-ci ont également été traitées. Le contrôle des suppressions de cellules au niveau de l'enregistrement s'est appuyé sur la définition d'un indicateur permettant de signaler qu'au moins un des champs a été supprimé.

Les dernières étapes ont été consacrées à la mise à l'essai du pilote et à l'optimisation de la méthode. On a effectué un nouveau regroupement des données de certaines variables clés parce que, après les routines de suppression des cellules mises au point aux fins du pilote, certaines variables utiles touchant les circonstances de la blessure présentaient un niveau indûment élevé de suppression. Dans l'essai pilote, plusieurs combinaisons de variables clés dans les routines de suppression des cellules étaient de nature ponctuelle et ne visaient aucun champ de données démographiques. Par conséquent, l'optimisation a consisté, tout d'abord, à définir systématiquement les routines de suppression des cellules de manière à englober toutes les combinaisons possibles de variables clés en fonction des renseignements de nature délicate auxquels elles sont associées et, deuxièmement, à rajuster l'ordre des variables

clés de façon plus systématique dans les routines de suppression des cellules. Cette optimisation s'est traduite par une réduction du nombre de combinaisons des variables clés analysées par les routines de suppression des cellules et par l'augmentation de la masse des données diffusées.

3.5 Principaux résultats

Le tableau 1 montre de façon détaillée la réduction du nombre de catégories des champs clés du SCHIRPT. On y compare le nombre de catégories figurant dans le fichier original avant et après la procédure de regroupement. Dans le cas du facteur mécanique, le champ clé le plus détaillé du SCHIRPT, les 629 catégories initiales possibles ont été regroupées en 13 catégories.

TABLEAU 1 : Réduction du nombre de catégories	Nbre DE VALEURS POSSIBLES	
VARIABLE CLÉ	AVANT	APRÈS
Groupe d'âge	7	5
Partie du corps	31	15
Concours de circonstances	29	10
Contexte	49	10
État du dossier	10	5
Langue parlée à la maison	31	4
Sexe	3	3
Secteur d'activité	36	6
Type d'incident	7	4
Lieu	57	7
Mécanisme	33	9
Facteur mécanique	629	13
Nature de la blessure	38	11
Profession	187	4
Équipement protecteur	10	6
Position assise dans un véhicule	22	4

Le tableau 2 compare les résultats issus de l'application de la procédure de suppression des cellules, effectuée dans le cadre de la méthode de CDS du SCHIRPT, aux données originales non regroupées du SCHIRPT et de la méthode d'optimisation pour les données du SCHIRPT de 2002. Il convient de noter que les données non regroupées sont associées à un niveau de suppression nettement supérieur, ce qui signifie que la qualité des données, mesurée par la quantité de données qu'il est possible de diffuser, est considérablement plus élevée lorsque la méthode optimisée s'applique aux données regroupées. Par exemple, la suppression additionnelle s'appliquant au champ du contexte dans la méthode optimisée s'établit à 23,9 % pour les données non regroupées et à 1,8 % pour les données regroupées. De plus, dans le cas de l'ensemble de données de 2002, le niveau global de suppression (fondé sur le nombre d'enregistrements comptant au moins un champ supprimé) est passé de 72 % à 7 %.

TABLE 2: % INC. NOT STATED	% RECORDS	
KEY VARIABLE	2002 orig., no coarsening	2002 Optimized
Age group	22.4	7.0
Body Part	37.5	2.6
Breakdown Event	45.8	1.9
Context	23.9	1.8
Disposition	5.1	0.1
Home Language	5.2	5.4
Gender	5.7	0.2
Industry	1.0	0.2
	1,2	0,1
Lieu	44,4	1,1
Mécanisme	13,8	1,5
Facteur mécanique	62,6	2,3
Nature de la blessure	32,7	1,4
Profession	1,7	0,2
Équipement protecteur	2,5	0,1
Position assise dans un véhicule	1,8	0,1
TOTAL - Enregistrements	72,7	7,0

Les tableaux 3A et 3B montrent les résultats de l'application de la procédure de suppression des cellules dans le cadre de la méthode de CDS du SCHIRPT relativement à la distribution des catégories de deux champs du SCHIRPT, en guise d'illustration de la perte d'information attribuable à la suppression. Si le taux initial de non-réponse était très faible pour la « nature de la blessure » et nettement plus élevé pour la « profession », la catégorie « non-déclaré » a enregistré une hausse négligeable, toutes les catégories ayant suivi une distribution comparable avant et après la suppression des cellules.

TABLEAU 3A : INCIDENCE SUR LA VALEUR DES CELLULES

CHAMP DU SCHIRPT : PROFESSION	% AVANT	% APRÈS
Gestionnaires/professionnels/paraprofessionnels	0,6 %	0,5 %
Métiers/commis/ventes/opérateurs de machines/conducteurs	1,3 %	1,3 %
Ouvriers/travailleurs de la construction	0,9 %	0,8 %
Non déclaré	97,2 %	97,4 %
TOTAL	100,0 %	100,0 %

TABLEAU 3B : INCIDENCE SUR LA VALEUR DES CELLULES

CHAMP DU SCHIRPT : NATURE DE LA BLESSURE	% AVANT	% APRÈS
Blessure superficielle	10,3 %	10,2 %
Plaie ouverte	17,8 %	17,7 %
Entorse ou foulure	9,9 %	9,8 %
Blessure à l'oeil	2,4 %	2,2 %
Traumatisme crânien mineur/intracrânien/commotion	9,0 %	8,9 %
Corps étranger	2,8 %	2,6 %
Effet thermal/électrique/d'empoissonnement/toxique	3,5 %	3,3 %
Blessure des tissus mous	12,2 %	12,0 %
Fracture/dislocation/subluxation du coude	24,4 %	24,3 %
Autre	7,6 %	7,6 %
Non déclaré	0,0 %	1,4 %
TOTAL	100,0 %	100,0 %

4. Discussion

À notre connaissance, c'est la première fois qu'un FMGD a été créé à partir des bases de données administrées par l'ASPC. Ces travaux se sont avérés un processus d'apprentissage pour nous, et une grande partie de cet apprentissage est applicable à d'autres ensembles de données du domaine de la santé. Les nouvelles pratiques exemplaires en matière de contrôle de la divulgation guident déjà l'ASPC dans l'élaboration de lignes directrices relatives à la diffusion de l'information tirée des ensembles de données sur la santé. On aurait pu éviter bien des difficultés liées à la production du FMGD du SCHIRPT si la création d'un tel fichier avait été envisagée lors de la conception de la base de données. L'utilisation des FMGD s'intensifie, de sorte qu'on devra tenir compte, dans la conception et la mise au point des bases de données, des avantages associés à la réduction au minimum de la collecte de renseignements personnels et de variables peu utilisées ainsi que du nombre de catégories dans le cas de certaines variables.

L'application de méthodes de CDS au SCHIRPT n'a pas été une procédure simple et directe. Elle a nécessité une série d'essais et d'ajustements qui devaient nous permettre de diffuser une grande quantité de données tout en assurant un niveau élevé de protection des renseignements personnels. Ainsi, nous avons établi à 0,5 % du fichier le seuil des fréquences univariées initialement employé pour regrouper les catégories détaillées en catégories sommaires. Les essais effectués en fonction de ce seuil ont donné lieu à des niveaux indûment élevés de suppression. Compte tenu de l'asymétrie des données du SCHIRPT, nous avons finalement haussé ce seuil de regroupement des catégories détaillées des champs du SCHIRPT à environ 2,5 %.

Il importe de reconnaître que l'application de procédures de contrôle de la divulgation au SCHIRPT ou à toute autre base de données a pour effet de réduire le risque de réidentification des personnes – elle ne permet pas toutefois de l'éliminer. Illustrons cette observation par un exemple – un cas hypothétique du SCHIRPT. Une fillette de 9 ans, qui reçoit de l'argent pour distribuer des dépliants publicitaires, marche sur le trottoir lorsqu'elle est attaquée par un chien, mordue à la figure puis admise à l'hôpital. L'enregistrement du SCHIRPT qui la concerne (après regroupement des catégories) comportera les éléments d'information suivants sous diverses variables : 5 à 9 ans / sexe féminin / blessure lors d'un travail rémunéré / coupure, piqûre ou morsure / mettant en cause un animal / blessure survenue sur un trottoir, dans un stationnement ou un accès de voiture / patient admis à l'hôpital.

Tous ces éléments d'information sont couramment consignés dans le SCHIRPT, mais cette combinaison précise se produit rarement, et chaque nouvel élément d'information facilite l'association entre l'enregistrement du SCHIRPT et un traumatisme connu ou rapporté. Si l'incident est rapporté dans les médias, il existe un risque résiduel que l'on puisse rattacher l'information à la victime. Ce risque serait nettement plus élevé dans le cas d'une blessure mortelle – on ne recense, chaque année, que quelques cas de décès d'enfants causés par des morsures de chien, et ces cas sont très médiatisés. C'est pour cette raison que les enregistrements visant des blessures mortelles ont été supprimés par les procédures de contrôle de la divulgation du SCHIRPT.

5. Conclusions

La méthodologie de CDS du SCHIRPT a permis d'évaluer et de traiter le risque inhérent au fichier du SCHIRPT de manière complète, systématique et efficace, conformément aux pratiques exemplaires reconnues de CDS. Ces méthodes de CDS ne devraient pas être appliquées aveuglément. Leur application doit se faire au cas par cas, parce que la science qui sous-tend ces méthodes évolue et que le processus d'application des méthodes pertinentes est un art qui consiste à prendre les meilleures décisions possibles compte tenu des outils et de l'information dont on dispose alors.

On devra contrôler et évaluer l'application des méthodes de CDS ainsi que les résultats issus de cette application en raison de la demande accrue des utilisateurs, des progrès technologiques et de l'évolution des méthodes de CDS. On devra, tout particulièrement, déterminer un niveau acceptable de risque – risque qui pourrait être attribuable à des facteurs extérieurs à la base de données. En bout de ligne, la décision relative à un niveau acceptable de risque revient aux propriétaires/gestionnaires de la base de données qui la prendront après avoir évalué rigoureusement les avis éclairés qu'ils auront reçus.

Références

Boudreau, J. R. (2005), "L'échange de données n'est pas la panacée", *Recueil du Symposium 2005 de Statistique Canada: Défis méthodologiques pour les besoins futurs d'information*, Session 1.

Herbert, M. et Mackenzie, S. G. (2004), "Injury Surveillance in Paediatric Hospitals: the Canadian Experience", Paediatrics and Child Health, 2004, Volume 9, Number 5, pp. 306-308.

De Waal, A. G. et Willenborg L.C.R.J (1998), "Optimal Local Suppression in Micro-Data", *Journal of Official Statistics Sweden*, 14, pp. 421-435.

Mackenzie, S. G. et Pless, I. B. (1999), "SCHIRPT. Canada's Principle Injury Surveillance Program", *Injury Prevention*, 1999, Volume 5, pp.208-213.

Santos, M. J. (2005), "Contrôle de la divulgation statistique: Cadre juridique et aspects méthodologiques", *Recueil du Symposium 2005 de Statistique Canada: Défis méthodologiques pour les besoins futurs d'information*, Session 16.

Skinner, C. J. et Holmes, D.J. (1998), "Estimating the Risk of Re-Identification Risk per Record", *Journal of Official Statistics Sweden*, 14, pp. 361-372.

Statistique Canada (2003), catalogue no. 12-587-XPF, "Contrôle de la confidentialité et de la divulgation" dans *Méthodes et pratiques d'enquête*, chapitre 12 section 5, pp 292-300

Statistics Canada (June 2006), "Handbook for Creating Public Use Micro-data files".

U.S. Office of Management and Budget, Federal Committee on Statistical Methodology, (December 2005), "Chapter II – Statistical Disclosure Limitation Methods: A Primer, Section F. Microdata "*Report on Statistical Disclosure Limitation Methodology*, pp 25-33

Remerciements

Nous aimerions souligner l'apport inestimable de Bev Cleary de l'Agence de santé publique du Canada et de Lori Stratychuk de Statistique Canada.